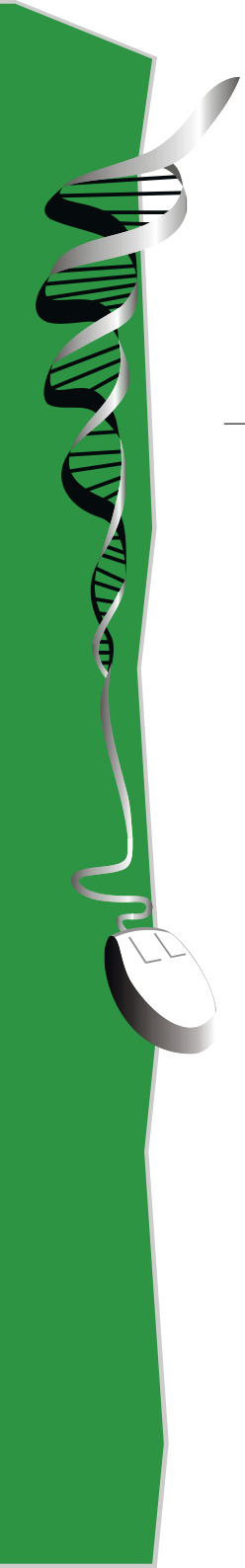

Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber
and Peter Stadler (Editors)

28th to 30th September 2009
Martin Luther University Halle-Wittenberg, Germany

German Conference on Bioinformatics 2009

Short Papers and Poster Abstracts

Preface

This volume contains three of four short papers and the 128 abstracts of posters presented at the German Conference on Bioinformatics, GCB 2009, held in Halle (Saale), Germany, September 28-30, 2009.

The German Conference on Bioinformatics is an annual, international conference, which provides a forum for the presentation of current research in bioinformatics and computational biology. It is organized on behalf of the Special Interest Group on Informatics in Biology of the German Society of Computer Science (GI) and the German Society of Chemical Technique and Biotechnology (Dechema) in cooperation with the German Society for Biochemistry and Molecular Biology (GBM).

We like to thank all local organizers and helpers for the efforts. Thanks are also due to all contributing to and participating in GCB 2009 and the sponsors for their financial support of the conference. Special thanks to Matthias Hübenthal for compiling this abstract book.

August 2009,

Ivo Große, Martin Luther University Halle-Wittenberg
Steffen Neumann, IPB Halle
Stefan Posch, Martin Luther University Halle-Wittenberg
Falk Schreiber, University of Halle-Wittenberg
& IPK Gatersleben, Germany
Peter F. Stadler, University Leipzig

Supporting scientific societies

Gesellschaft für Biochemie und Molekularbiologie e.V. (GBM)
http://www.gbm-online.de/

Gesellschaft für Chemische Technik und Biotechnologie e.V. (DECHEMA)
http://www.dechema.de/en/start_en.html

Gesellschaft für Informatik e.V. (GI), Fachgruppe 4.0.2
http://www.cebitec.uni-bielefeld.de/groups/fg402/

Non-profit sponsors

Leibniz Institute of Plant Biochemistry (IPB) Halle
http://www.ipb-halle.de/

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben
http://www.ipk-gatersleben.de/

Martin Luther University Halle-Wittenberg
http://www.uni-halle.de/

Commercial sponsors

BIOBASE GmbH
http://www.biobase-international.com/

ClusterVision BV
http://www.clustervision.com/

Genomatix Software GmbH
http://www.genomatix.de/

Illumina, Inc.
http://www.illumina.com/

Kapelan GmbH
http://www.kapelan-bioimaging.com/

KWS SAAT AG
http://www.kws.de/

Lehmanns Fachbuchhandlung GmbH
http://www.lob.de/

MEGWARE Computer GmbH
http://www.megware.com/

Microsoft Deutschland GmbH
http://www.microsoft.de

NVIDIA Corporation
http://www.nvidia.de/

SunGene GmbH
http://www.sungene.business.t-online.de/

TraitGenetics GmbH
http://www.traitgenetics.de/

transtec AG
http://www.transtec.de/

Table of Contents

Short Papers

Mutually exclusive spliced exons show non-adjacent and grouped patterns

Martin Pohl[1,4], Dirk Holste[2], Ralf Bortfeldt[3], Konrad Grützmann[1] and Stefan Schuster[1]

[1]Department of Bioinformatics, Friedrich-Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

[2]Austrian Institute of Technology, Donau-City-Straße 1, A-1220 Vienna, Austria

[3]Breeding Biology and Molecular Genetics, Humboldt-University, Invalidenstrasse 42, 10115 Berlin, Germany

[4]To whom correspondence should be addressed. E-mail: `m.pohl@uni-jena.de`

Abstract

The deciphering of the human genome provides the base for many bioinformatic methods researching genome related matters. Splicing of mRNA is one of the processes which depend on the genomic sequence. We present an analysis of the results from a method that detects mutually exclusive spliced exons in the genome. The method is based on transcript data mapped to the genomic sequence. Applying it to human as well as mouse genome we detected more than 1000 mutually exclusive spliced exon pairs per species. The events we analysed broaden the view on mutually exclusive regulated splicing and reveal some unexpected characteristics.

1 Introduction

Alternative splicing (AS) of pre-mRNAs plays many different roles and in higher eukaryotes it is one regulatory mechanism for tissue-specific protein variability [Bla00; Gra01]. Estimations of the scope of AS are often different. Most recent ones (based on short cDNA reads/mRNA-Seq data) show almost every (95%) human multi-exon gene to undergo AS [Wan+08]. Traditionally, four main AS patterns are considered: exon skipping, alternative acceptor or donor site, intron retention [Fer+07; Kim+08]. Among the different types of AS, mutually exclusive exons (MXEs) constitute a comparatively rare, but very intriguing type [Sam+08]. In its simple form, two internal exons are spliced in such a way that one exon is excised, while the other one is kept

within the mature mRNA (and vice versa), thereby linking the regulation of different exons. So far, MXEs have been usually assumed to be genomic adjacent [Hol+06; Nag+06; Zhe+05]. One hypothesis about its functional role is the selection of context specific sites within protein domains. Splicing of MXEs has not been analysed in as much detail as more frequent types of AS, e.g., exon-skipping or alternative exon boundaries [Bor+08; Hil+04; ML03]. Nonetheless, such AS events contribute to our understanding of this still largely unclear phenomenon. Several regulatory mechanisms for MXEs have been proposed, including spliceosome incompatibility, steric hindrance, docker-selector mechanism, or the regulation by splicing factors coupled with nonsense-mediated decay [Smi05]. In this study we used a computational approach to infer about two thousand MXEs across more than 20,000 of the mapped genes for human and mouse genome, respectively. We then compared and contrasted our data for their compatibility with known AS patterns as well as genomic features predicted by existing models for the regulation of MXEs.

2 Results

All analyses are based on annotations of protein coding genomic regions retrieved from Ensembl (`http://www.ensembl.org/`). Onto these regions, we mapped transcripts from the UCSC database (`http://genome.ucsc.edu/`; 4.8 million ESTs and 150,000 Fl-mRNAs). For the mapping we relied on the transcript to genome alignments provided by UCSC [Kuh+09]. We considered a pair of exons to be mutually exclusive if all transcripts, mapped such that they span the genomic region of both exons, exactly one of the exons is included (for further details on mapping, filtering and constraints please contact authors). For about 20,000 human genes, we inferred 1,300 MXE pairs (3.4% of all genes). In a similar study, we inferred 1,200 MXE pairs (3.3% of all genes) for over 21,000 mouse genes. Comparing the detected pattern and their features (genomic, transcriptional) with existing regulatory models, we found for both studies that nearly all MXEs were not adjacent with respect to their genomic location (99%, cf. Fig. 1.1), which is in contrast to the predicted patterns of existing models. Furthermore, while we could not infer complex patterns, such as docker-selector sites in the fruitfly gene Dscam [Ana+06], we found MXEs to be involved in mutually exclusive splicing of more than one exon simultaneously (cluster-spliced exons, cf. Fig. 1.1). None of the

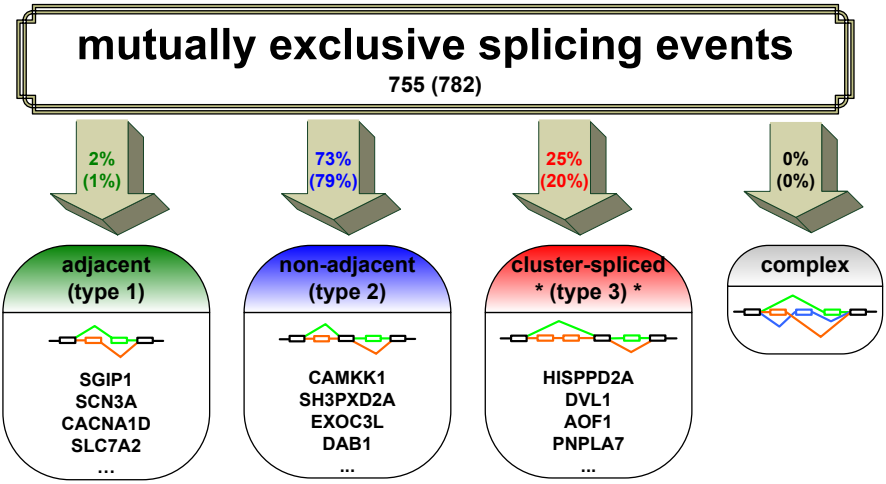

Figure 1.1: Representation of human (mouse) MXEs inferred within this study. We divided MXE events into four distinct groups based on known regulatory mechanisms and detected splicing pattern. Examples of splicing pattern are shown at the bottoms of each type. Left hand side: three quarter of all inferred mutually exclusive splicing events comprised single alternative exons, split into adjacent (type 1) and non-adjacent (type 2) patterns. Middle: one quarter of inferred events featured the newly introduced mutually exclusive splicing of exon clusters (type 3). Right hand side: we did not observe more complex patterns, e.g., like the selection of an exon, singled-out from a cluster of possible MXEs in the Dscam gene of the fruitfly.

inferred MXEs obsessed the /AT and AC\splice site motifs for the minor spliceosome. At first glance, checking for orthologous genes we could infer for 21% (11%) of human genes with adjacent events (non-adjacent) an event for mouse as well. But a closer look at exon specific sequence conservation remains for a more substantive reasoning on conservation of events. While we could find some previously reported events as for voltage-gated sodium and calcium channel protein alpha subunits or human glycine receptor a2 gene (SCNA, CACNA1d, GLRA2) there is as well a number of reported events not among our results like KCNMA1, TCL6 [Cop04; Sam+08]. This is in particu-

lar due to that mutually exclusive properties are frequently defined on tissue specific or transcript pair context while our approach requires global mutual exclusivity. Frame shift analysis revealed higher frame conservation for adjacent than for non-adjacent MXEs. For the latter ones we also found a variable number of enclosed constitutive exons reaching up to 67 with a strong bias towards two.

3 Conclusions

One central outcome of this study is that non-adjacent exons constitute the most frequent MXE event. Consequently, the inferred AS patterns, determined by transcript alignments, do not further substantiate existing models and hence challenge their suitability as widely functioning mechanisms for mutually exclusive splicing in higher eukaryotes, e.g., mammals. The model incorporating spliceosome incompatibility found no evidence in this study, the model incorporating steric hindrance is most conceivable for small introns intervening MXEs, while the model incorporating docker-selector sites is not conceivable for non-adjacent MXEs. The regulation by splicing-factors accompanied by the NMD mechanism remains a candidate for adjacent MXEs and ought to be validated in future studies. Regulation of the new subtype of cluster-spliced MXEs cannot be explained well by current models and new hypotheses for AS regulation are necessary. Evolutionary origin, effects on evolvability [Che+06], splice site recognition as well as open questions raised by our results will be discussed.

References

[Ana+06] D. Anastassiou et al. “Variable window binding for mutually exclusive alternative splicing”. In: *Genome Biology* 7.1 (2006), R2. DOI: 10.1186/gb-2006-7-1-r2.

[Bla00] D. L. Black. “Protein diversity from alternative splicing: a challenge for bioinformatics and postgenome biology”. In: *Cell* 103.3 (2000), pp. 367–370. DOI: 10.1016/S0092-8674(00)00128-8.

[Bor+08] R. Bortfeldt et al. “Comparative analysis of sequence features involved in the recognition of tandem splice sites”. In: *BMC Genomics* 9.1 (2008), p. 202. DOI: 10.1186/1471-2164-9-202.

[Che+06] F.-C. Chen et al. “Alternatively and Constitutively Spliced Exons Are Subject to Different Evolutionary Forces”. In: *Molecular Biology and Evolution* 23.3 (2006), pp. 675–682. DOI: 10.1093/molbev/msj081.

[Cop04] R. R. Copley. “Evolutionary convergence of alternative splicing in ion channels”. In: *Trends in Genetics* 20.4 (2004), pp. 171–176. DOI: 10.1016/j.tig.2004.02.001.

[Fer+07] E. N. Ferreira et al. “Alternative splicing: a bioinformatics perspective”. In: *Molecular BioSystems* 3.7 (2007), pp. 473–477. DOI: 10.1039/b702485c.

[Gra01] B. R. Graveley. “Alternative splicing: increasing diversity in the proteomic world”. In: *Trends in Genetics* 17.2 (2001), pp. 100–107. DOI: 10.1016/S0168-9525(00)02176-4.

[Hil+04] M. Hiller et al. “Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity”. In: *Nature Genetics* 36.12 (2004), pp. 1255–1257. DOI: 10.1038/ng1469.

[Hol+06] D. Holste et al. “HOLLYWOOD: a comparative relational database of alternative splicing”. In: *Nucleic Acids Research* 34.suppl_1 (2006), pp. D56–62. DOI: 10.1093/nar/gkj048.

[Kim+08] E. Kim et al. “Alternative splicing: current perspectives”. In: *BioEssays* 30.1 (2008), pp. 38–47. DOI: 10.1002/bies.20692.

[Kuh+09] R. M. Kuhn et al. “The UCSC Genome Browser Database: update 2009”. In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D755–761. DOI: 10.1093/nar/gkn875.

[ML03] B. Modrek and C. J. Lee. “Alternative Splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss”. In: *Nature Genetics* 34.2 (2003), pp. 177–80. DOI: 10.1038/ng1159.

[Nag+06] H. Nagasaki et al. “Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns”. In: *Bioinformatics* 22.10 (2006), pp. 1211–1216. DOI: 10.1093/bioinformatics/btl067.

[Sam+08] M. Sammeth et al. “A General Definition and Nomenclature for Alternative Splicing Events”. In: *PLoS Computational Biology* 4.8 (2008), e1000147. DOI: 10.1371/journal.pcbi.1000147.

[Smi05] C. W. J. Smith. “Alternative Splicing — When Two’s a Crowd”. In: *Cell* 123.1 (2005), pp. 1–3. DOI: 10.1016/j.cell.2005.09.010.

[Wan+08] E. T. Wang et al. “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221 (2008), pp. 470–476. DOI: 10.1038/nature07509.

[Zhe+05] C. L. Zheng et al. “MAASE: An alternative splicing database designed for supporting splicing microarray applications”. In: *RNA* 11.12 (2005), pp. 1767–1776. DOI: 10.1261/rna.2650905.

Novel intron positions in Drosophila are mostly caused by intron sliding and tandem duplications

Jörg Lehmann[1], Carina Eisenhardt[2], Peter F. Stadler[1] and Veiko Krauss[1,3]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany

[2]Genetics Group, Department of Biology II, University of Leipzig, Leipzig, Germany

[3]To whom correspondence should be addressed. E-mail: `krauss@rz.uni-leipzig.de`

Abstract

Introns that reside in non-conserved positions are either novel or remnants of frequent losses of introns in some evolutionary lineages. Here we identified 36 unambiguous cases of novel intron positions in 35 Drosophila genes. A near intron pair (*NIP*) in Drosophila consists of an ancient and a novel intron position that are separated by less than 32 nt. Within a specific gene, only one of the two intron positions can exist at the same time. In general, the ancient intron position has disappeared in favor of the novel one. A survey for *NIPs* among 12 Drosophila genomes identifies intron sliding (migration) as the most frequent cause of novel intron positions. Seven other novel introns seem to have been gained by regional tandem duplications of coding sequences containing a proto-splice site.

1 Introduction

Comparative studies of spliceosomal intron densities suggested relatively high rates of intron gain during eukaryote evolution. In conserved coding sequences at least, recent intron gain events at *novel* intron positions appear to be rare [RI09]. The mechanisms are still poorly understood. Exon sizes smaller than 50 nt are scarce [Sae+07] and in general functionally detrimental [Wei+06]. Introns less than 50 nt away from each other therefore typically exclude each other. Among introns observed at such nearby positions in orthologous genes, one has to be evolutionarily younger and should define a monophyletic group [Kra+08]. Here we use the relatively recently diverged genomes of 12 Drosophila species [Dro07] to identify recent intron gain events in a comparative analysis of gene structures and evaluate possible mechanisms of their origin.

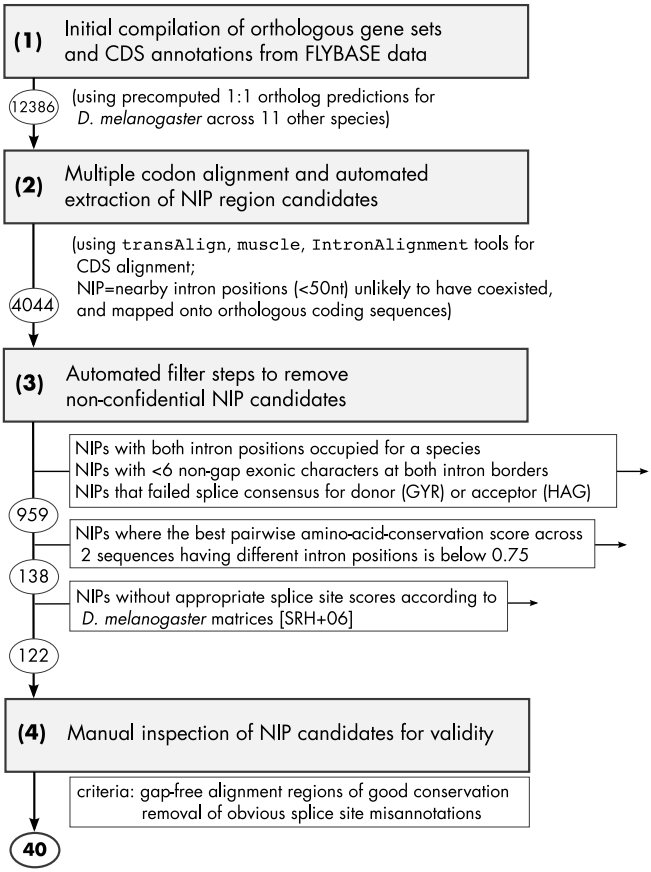

Figure 1.2: Pipeline for extracting *NIP* candidate regions from `FLYBASE` and subsequent filtering. Ovals show the number of candidates retained after each step.

2 Methods

We started with 12386 sets of 1:1 orthologous protein coding genes extracted from FLYBASE data, Figure 1.2. Coding sequences were aligned at the protein level. Their intron positions mapped onto the alignments resulted in 4044 *NIP* candidate regions. We excluded *NIPs* that contained sequences with both introns of a pair and those that failed to show the splice site consensus for both donor (GYR) and acceptor (HAG). In addition, 6 non-gap characters were required at both exon borders. A minimal conservation score of 0.75 further reduced the candidate list. Drosophila splice site matrices [She+06] were used detect likely misannotations of introns. The computational filtering concluded with 122 *NIP* regions that were then inspected manually, finally resulting in 40 regions comprising 41 *NIPs*. ESTs support both introns in 6 and only one intron in 23 of these cases. To obtain a validated set of small exons, we used the same orthologous dataset and considered all internal exons of *D. melanogaster* that were fully experimentally supported (transcript evidence score 15) according to FLYBASE.

Although our *NIP* pipeline was tuned to be very restrictive in order to effectively remove misannotated gene structures, the experimental evidence for our *NIPs* is not strong. Only 35 of the 72 introns of the 36 *NIPs* are supported by ESTs. RT-PCR experiments to improve the direct evidence are therefore under way.

3 Results

To see whether an exon length less than 50 nt is typically excluded also in Drosophila, we derived (1) 41 *NIPs* with an intron-distance less than 50 nt, and (2) 106 internal exons shorter than 50 nt. Figure 1.3 compares the two length distributions and shows that short internal exons are more abundant than *NIPs* above a cut-off value of 31 nt. *NIPs* of larger distance were therefore excluded from further analysis.

NIPs were introduced as reliable phylogenetic markers for insect evolution [Kra+08]. Out of the 36 validated Drosophila *NIPs*, only one character distribution contradicts the established tree of the 12 species [Dro07] and renders the position of *D. ananassae* uncertain. We conclude that *NIPs* are reliable phylogenetic markers also for recent radiations. Short branch lengths, as observed between Drosophila species, could critically limit the number of

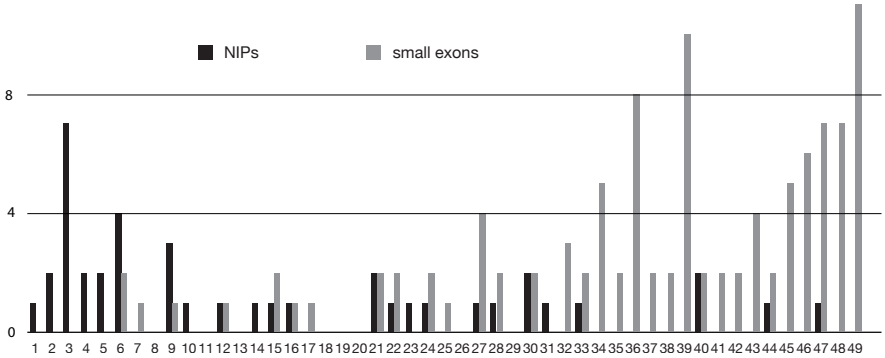

Figure 1.3: Comparison of intron-intron distances in *NIPs* with the length distribution of internal exons in Drosophila. Only *NIPs* derived from genes structures supported by splice site quality and coding sequence conservation and small exons supported by experimental evidence are included in the data.

available characters, however.

Our data show a dominant role of intron sliding (migration) as cause of novel intron positions. Seven of the analyzed *NIPs* clearly emerged by sliding because they show significant sequence similarities between introns of both positions. In another six *NIPs*, three of the following four properties expected for positionally migrated introns were observed: (1) there is no intermediate, intron-less state in the tree; (2) *NIP* distance is a multiple of 3, allowing a stepwise intron shift; (3) cryptic splice sites occur in *NIP* distances and are supported by amino acid conservation; and (4) one-sided shifts or `GYRGYR/NAGNAG` splice sites in some species. While the first two conditions are also consistent with mechanisms of intron gain, cryptic splice sites within the corresponding distance and one-sided shifts of intron borders in some species are specific requirements of intron sliding.

The second most frequent cause of novel introns appears to be local duplications of coding sequences. During such hypothetical events, a proto-splice site (AG/GY) within a duplicated region turns one copy of the exonic sequence into an intron. Seven novel introns from our data set probably have been gained in this way. This is supported by potentially functional

proto-splice sites at both exonic borders of the introns, as measured by the Drosophila 5' and 3' splice site scores [She+06]. These sites are conserved also in Drosophila species which have no introns at these positions. The patterns thus may have been present already at the time of intron origin. In two of these cases, regional repetitivity enhanced the probability of duplications in the CDS.

Other mechanisms of intron gain in conserved coding sequences could not be identified in our Drosophila dataset. In particular, we did not find any relevant sequence conservation to other introns, to transposons, or to exons.

4 Conclusions

During our study of novel intron positions in evolutionarily conserved genes of Drosophila we confirmed that near intron pairs (*NIPs*) are reliable phylogenetic markers. Furthermore, our results support that intron sliding (migration) is the most frequent cause of recently emerged intron positions within Drosophila genomes, at least if the novelty of intron positions is based on the *NIP* definition. We also found evidence for the rise of novel introns by tandem duplication of exonic DNA. The origin of 18 remaining *NIPs* stayed unknown. Contrary to expectations, the gain of novel introns by other mechanisms could not be proved, for example, by insertion of a spliceosomal intron via reverse splicing into a new position, by insertion of a transposable element or by gene conversion. Recent origin of spliceosomal introns in eukaryotic genes might be, therefore, mainly due to local mutations and not due to insertions of alien sequences.

References

[Dro07] Drosophila 12 Genomes Consortium. "Evolution of genes and genomes on the Drosophila phylogeny". In: *Nature* 450.7167 (7167 2007), pp. 203–18. ISSN: 1476-4687. DOI: 10.1038/nature06341.

[Kra+08] V. Krauss et al. "Near intron positions are reliable phylogenetic markers: an application to holometabolous insects". In: *Molecular Biology and Evolution* 25.5 (5 2008), pp. 821–830. ISSN: 1537-1719. DOI: 10.1093/molbev/msn013.

[RI09] S. W. Roy and M. Irimia. "Mystery of intron gain: new data and new models". In: *Trends in Genetics* 25.2 (2 2009), pp. 67–73. ISSN: 0168-9525. DOI: 10.1016/j.tig.2008.11.004.

[Sae+07] Y. Saeys et al. "In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists". In: *Bioinformatics* 23.4 (4 2007), pp. 414–420. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btl639.

[She+06] N. Sheth et al. "Comprehensive splice-site analysis using comparative genomics". In: *Nucleic Acids Research* 34.14 (14 2006), pp. 3955–3967. ISSN: 1362-4962. DOI: 10.1093/nar/gkl556.

[Wei+06] M. Weir et al. "Challenging the spliceosome machine". In: *Genome Biology* 7.1 (1 2006), R3. ISSN: 1465-6914. DOI: 10.1186/gb-2006-7-1-r3.

A cis-regulatory signature for chordate anterior neurectodermal genes

Maximilian Haeussler[1,2], Yan Jaszczyszyn[1,2], Lionel Christiaen[1] and Jean-Stéphane Joly[1,3]

[1]Institute of Neurosciences Alfred Fessard, CNRS/INRA Gif-sur-Yvette, France

[2]These authors contributed equally to this work

[3]To whom correspondence should be addressed. E-mail: joly@iaf.cnrs-gif.fr

Abstract

Cis-regulatory sequences from different genes with similar expression patterns show very subtle, non-alignable similarities which are difficult to quantify by a score. We analysed one single anterior neurectoderm enhancer by successive mutations and found that it consists of a duplicated structure. Thus, we searched the genome for all conserved non-coding elements with a duplicated pentamer. Motifs were ranked by group specificity score, a simple binomial p-Value to obtain a certain number of anterior nervous system genes, versus a background set of genes, expressed in other annotated tissues. One of the most specific motifs corresponds to the binding site of the transcription factor OTX, known to play a pivotal role in the anterior neurectoderm development of all bilaterian species. We could validate some of the predicted cis-regulatory sequences *in-vivo*. Our results highlight the importance of *in-situ* annotation databases for regulatory analyses and show that it is possible to find similar enhancers starting from just a single well-characterized sequence.

1 Introduction

We have previously described an enhancer (called "D1", 323bp) that controls expression of the Ciona intestinalis Pitx gene in a region overlapping neural and non-neural cells (anterior neural boundary) [Chr+05] which we consider a part of the anterior neurectoderm in the following. As the sequence contains many putative transcription factor binding sites, we wanted to find the most important ones for the anterior neurectodermal expression.

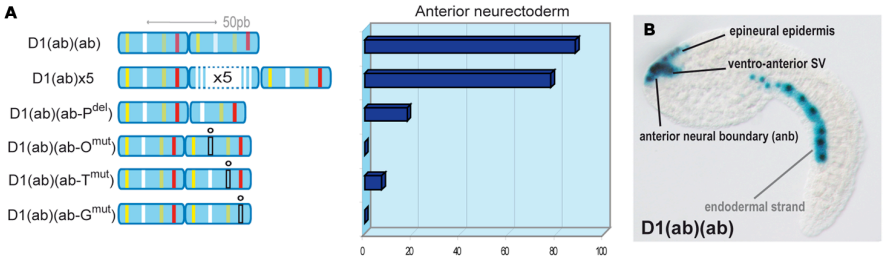

Figure 1.4: Artificial enhancer constructs reveal a tandem-like structure. (A) Two or five copies of the 54bp D1(ab) enhancer drive strong expression (88%, n=167 and 77%, n=72). The P site and G sites do not seem to be completely essential, as some embryos are stained after their deletion (17%, nPde=90 and 7%, nGmut=118). Duplicates of O, T, G are require as mutations in the second copy of (ab) completely abolish lacZ expression (0%, nOmut=137 and 0%, nTmut=84). (B) An embryo electroporated with two repeated copies of D1(ab): Expression of this sequence which is much shorter than the complete D1 enhancer is present in the ANB but extends slighly into neighboring tissues

2 Results

To characterize essential cis-regulatory sequences, we split the D1-enhancer linearily into five parts of similar length, D1 a,b,c,d,e of which a and c were found to be dispensable for expression. We identified four classes of putative binding sites in the remaining sequence. All of these were represented at least twice in the D1 enhancer. We therefore created artificial enhancers containing multiple copies of the subfragment D1(ab) and found that as few as two copies were sufficient to drive strong lacZ expression in the anterior neurectoderm (Fig. 1.4A, B)). .Point mutations in the second D1(ab) copy strongly reduced enhancer activity (Fig. 1.4A). The results demonstrate that duplications of some critical binding sites are essential for D1 enhancer activity and do not merely determine the quantitive readout or are due to redundancy. This observation led us to investigate whether a duplicated motif is overrepresented around neurectodermal genes at this developmental stage.

We downloaded *in-situ* images from the database Aniseed [Ani] and manually selected all 100 genes that are specifically expressed in the anterior and

not the posterior parts of the central nervous system and vice versa . We could use the already present annotations from this database for other structures, for the tissues muscle, epidermis and notochord. To obtain a set of conserved non-coding elements (CNEs), we aligned the genome of *C. intestinalis* with *C. savignyi* and removed transcribed sequences. This resulted in 168306 CNEs with an average length of 143 bp.

We then aimed at studying the distribution of duplicated short DNA motifs around the 904 genes for which we had obtained annotations, to find those motifs that show a bias towards genes expressed in the anterior or posterior nervous system, muscle, epidermis or notochord. We chose to search the CNEs for all possible 512 pentamer consensus sequences. The rationale for using consensus and not matrix based searches was that we wanted to focus on homeodomain proteins, a class of transcription factors with the most best characterized binding sites that have been shown to resemble pentamer motifs with few degenerate positions [Ber+08; Noy+08]. We observed from our case study that the essential site had to occur twice. As for the maximum distance between the two copies, we noted that the unknown motif had to be located in the parts D1ab and also in D1de. Both fragments span around 120 bp in the original enhancer.

The positive genes (foreground) are the ones expressed in a given tissue and the background are all 904 annotated genes. The group specificity score [Hug+00] is the negative logarithm of the binomial probability to obtain a certain number of positives among the predicted genes by chance. The better predicted genes match target genes, the higher the score. Based on the experimental results, we required two matches of a pentamer within 125bp, conserved in the alignment of a CNE; we used only the gene closest for each match. For each tissue, we calculated the group specificity score on all pentamers. The program is also available as an interactive website at

`http://www.ciona.cnrs-gif.fr/scripts/cionator2/wordSearchForm.cgi`.

3 Discussion

In the anterior nervous system, the three highest ranking pentamers are found in the D1 enhancer, but only the third one, GATTA, is located in the minimal fragment and its mutation leads to complete disruption of the expression. It corresponds to a binding site for an OTX/Bicoid-like transcription factor

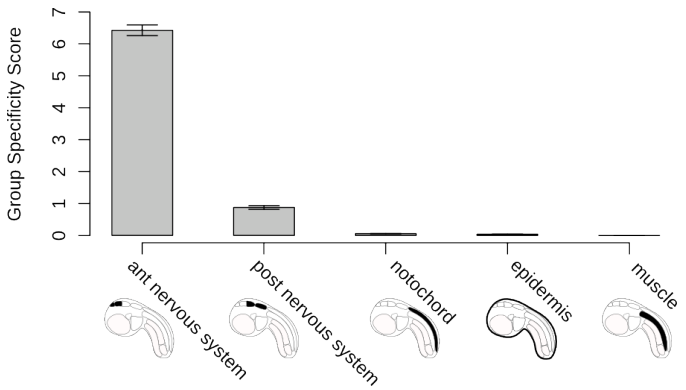

Figure 1.5: Group specificity scores of two GATTA motif within 125bp against foreground gene sets expressed in different tissues. Error bars correspond to standard deviations obtained from shuffling 10% of the annotations to show the possible impact of errors in tissue assignments

which is part of the anterior nervous system foreground gene set itself. It is known as a head marker gene: OTX knock-out embryos lack various anterior nervous system structures [Aca+95]. If we search the motif model across the different tissue gene sets, scores show an anterior-posterior gradient for this motif (Figure 1.5). We selected 23 of these matches that flank a gene expressed in the anterior nervous system and cloned them into an expression vector; ten of them drove expression in the anterior neurectoderm when tested *in-vivo*.

Our analysis is currently limited to non-degenerate consensus sequences and could be extended to combinations of transcription factor weight matrix matches that exceed a certain cutoff. Similar approaches have been applied to gene sets usually derived from microarray data (e.g. [Pen+07]). *In-situ* images hold the promise to be virtually noise-free and the group specificity score can take into account any combination of motifs to find a link between motif content and flanking tissue-specific gene sets.

References

[Aca+95] D. Acampora et al. "Forebrain and midbrain regions are deleted in Otx2-/- mutants due to a defective anterior neuroectoderm specification during gastrulation". In: *Development* 121.10 (1995), pp. 3279–3290. URL: http://dev.biologists.org/cgi/content/abstract/121/10/3279.

[Ani] *Ascidian Network for InSitu Expression and Embryological Data (ANISEED)*. 2007. URL: http://aniseed-ibdm.univ-mrs.fr/.

[Ber+08] M. F. Berger et al. "Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences". In: *Cell* 133.7 (2008), pp. 1266–1276. DOI: 10.1016/j.cell.2008.05.024.

[Chr+05] L. Christiaen et al. "A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodæum". In: *Developmental Biology* 277.2 (2005), pp. 557–566. DOI: 10.1016/j.ydbio.2004.10.008.

[Hug+00] J. D. Hughes et al. "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae". In: *Journal of Molecular Biology* 296.5 (2000), pp. 1205–1214. DOI: 10.1006/jmbi.2000.3519.

[Noy+08] M. B. Noyes et al. "Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites". In: *Cell* 133.7 (2008), pp. 1277–1289. DOI: 10.1016/j.cell.2008.05.023.

[Pen+07] L. A. Pennacchio et al. "Predicting tissue-specific enhancers in the human genome". In: *Genome Research* 17.2 (2007), pp. 201–211. DOI: 10.1101/gr.5972507.

Accepted Posters

Modelling changes in amino acid metabolism during day-night cycles

Sascha Schäuble[1], Ines Heiland[1], Olga Voytsekh[2], Maria Mittag[2] and Stefan Schuster[1]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany

[2]Institute of General Botany, Friedrich-Schiller-University Jena, Germany

First, we analysed nitrogen uptake and amino acid synthesis during day-night cycles based on different availability of energy and carbon sources. Second, we searched for UG-repeat sequences in the mRNA of enzymes involved in nitrogen metabolism. The insertion of those sequences into the 3'-UTR of reporter constructs triggers circadian expression [Kia+07]. Therefore, we additionally analysed possible changes in nitrogen uptake and amino acid synthesis due to circadian regulation of enzyme expression.

As kinetic data are not available, we applied elementary mode analysis [Sch+00]. Using this method we are able to simulate day-night changes and the implication of circadian regulation on the nitrogen metabolism of *Chlamydomonas reinhardtii*.

References

[Kia+07] S. Kiaulehn et al. "The Presence of UG-Repeat Sequences in the 3'-UTRs of Reporter Luciferase mRNAs Mediates Circadian Expression and Can Determine Acrophase in Chlamydomonas reinhardtii". In: *Journal of Biological Rhythms* 22.3 (2007), pp. 275–277. DOI: 10.1177/0748730407301053.

[Sch+00] S. Schuster et al. "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks". In: *Nature Biotechnology* 18 (2000), pp. 326–332. DOI: 10.1038/73786.

Temperature regulation of circadian rhythms

Ines Heiland[1], Thomas Hinze[1], Olga Voytsekh[2], Maria Mittag[2] and Stefan Schuster[1]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany

[2]Institute of General Botany, Friedrich-Schiller-University Jena, Germany

Circadian clocks are endogenous rhythms that can be entrained by external cues, like light-dark and temperature cycles. They persist under constant conditions with a period of about 24 hours and allow organisms to anticipate daily environmental variations.

Although the underlying biochemical processes are temperature dependent, the period of the circadian clock is relatively insensitive to temperature [Joh+08]. This phenomenon is called temperature compensation. However, single pulses of temperature or light can shift clock phase.

Temperature regulation of circadian clocks is poorly understood both experimentally and theoretically. To gain a better understanding of the molecular mechanisms behind temperature compensation and entrainment, we built a mathematical model that is based on the experimentally observed temperature regulation of the circadian RNA-binding protein CHLAMY1 from *Chlamydomonas reinhardtii* [Voy+08].

Using this model we analysed the response of the model oscillator to temperature cycles and pulses, and compared the temporal changes of model species to experimental data as far as available, or otherwise made predictions for the behaviour of the circadian system.

References

[Joh+08] C. H. Johnson et al. “A Cyanobacterial Circadian Clockwork”. In: *Current Biology* 18.17 (2008), R816–R825. DOI: 10.1016/j.cub.2008.07.012.

[Voy+08] O. Voytsekh et al. “Both Subunits of the Circadian RNA-Binding Protein CHLAMY1 Can Integrate Temperature Information”. In: *Plant Physiology* 147.4 (2008), pp. 2179–2193. DOI: 10.1104/pp.108.118570.

Structure and functions of TRPA1

Vasilina Zayats[1], Abdul Samad[2] and Rudiger Ettrich[2]

[1]Institute of Physical Biology, Academy of Sciences of the Czech Republic

[2]Institute of Systems Biology and Ecology, Academy of Sciences of the Czech Republic

Transient receptor potential (TRP) channels are a large superfamily of nonselective cation channels. TRPA1 is a candidate for mechanically gated transduction channels potentially mediating the sensations of hearing, touch, and some forms of pain. Human TRPA1 is a 127.4 kDa protein comprised of 1119 amino acids. Like other TRPs, TRPA1 has six predicted membrane-spanning domains (S1 to S6) and the pore between S5 and S6. In this work we focus on homology modeling of its for TRPs unusually long N-terminal intracellular region containing 18 predicted ankyrin repeats. Ankyrin repeats have been implicated in protein-protein interactions, provide elasticity and make molecular springs. Also a calcium-binding domain, EF-hand, was indicated at the N-terminus, consisting of 12 residues involved in Ca-dependent activation. Simulations of dynamic behavior of three-dimensional all-atom models indicate stability and equilibration, and let us describe structural and functional properties to understand the system. Structural models are build using Modeller, visual analyzing and energy minimization is done in Yasara, and molecular dynamics simulations are carried out in GROMACS (molecular dynamics simulation package). The general aim is to embed the results of this work later into an all-atom model of the channel in the membrane to get a stable tetrameric structure of complete TRPA1 in a close-to-natural environment.

References

[Ben+09] J. Benedikt et al. “Essential role for the putative S6 inner pore region in the activation gating of the human TRPA1 channel”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1793.7 (2009), pp. 1279–1288. DOI: 10.1016/j.bbamcr.2009.04.014.

[Doe+07] J. F. Doerner et al. “Transient Receptor Potential Channel A1 Is Directly Gated by Calcium Ions”. In: *Journal of Biological Chemistry* 282.18 (2007), pp. 13180–13189. DOI: 10.1074/jbc.M607849200.

[Mic+02] P. Michaely et al. “Crystal structure of a 12 ANK repeat stack from human ankyrinR”. In: *The EMBO Journal* 21 (2002), pp. 6387–6396. DOI: 10.1093/emboj/cdf651.

Standards and infrastructure for managing experimental metadata

Susanna-Assunta Sansone[1] and Philippe Rocca-Serra[1]

[1]European Bioinformatics Institute (EMBL-EBI), United Kingdom

Today's researchers can perform biological and biomedical studies where the same material is run through a wide range of assays, comprising several technologies such as genomics, transcriptomics, proteomics and metabol/nomics. To enable others to correctly interpret the complex data sets that result, it is necessary to provide contextualizing experimental metadata (i.e., sample characteristics, study design and execution) at an appropriate level of granularity. Several standards initiatives work to develop reporting standard (minimum information checklists, ontologies and file formats) for describing, formatting and exchanging both data and metadata. They normally cater to particular domains, however in parallel several synergistic standards activities foster cross-domain harmonization. Our ISA infrastructure (`http://isatab.sf.net`) is the first to leverage on these synergistic activities to provide a common structured representation and storage mechanism for a variety of studies. The infrastructure's main (open source) components are described here.

- ISAconfigurator enables power users (e.g. curators) to regulate the minimal requirement fields, according to the relevant MIBBI checklist(s) (`http://www.mibbi.org`), and set their allowed values, e.g. to ontology terms. The configuration then used by the ISAcreator editor tool.

- ISAcreator drives the experimentalists to report the metadata following the configured requirements and search and select terms from OBO Foundry ontologies (`http://www.obofoundry.org`), accessed using web services provided by OLS (`http://www.ebi.ac.uk/ontology-lookup`) and BioPortal. (`http://bioportal.bioontology.org`).

- ISAconverter transforms ISA-Tab formatted metadata into several other related tabular and XML-based formats for submission to public repositories (e.g. ArrayExpress, ENA-Reads and PRIDE).

- BioInvestigation Index database enables storing and querying functionalities. An instance of the BioInvestigation Index database has been installed as prototype at EBI (`http://www.ebi.ac.uk/bioinvindex`).

An R package, supporting the ISA-TAB format, is under development. The infrastructure' components can work independently, or as unified system for local use.

Definition of Quality Control check points critical to obtain optimal results in Microarray Data Analysis

Fatima Kreusch[1], Andrea Gaarz[1], Svenja Debey-Pascher[1] and Andrea Staratschek-Jox

[1]LIMES Institute Genomics and Immunoregulation, University of Bonn, Germany

Microarrays are a broadly applied technique to analyse gene expression data. One major application lies in the development of clinical therapeutics. Therefore, a standard analysis approach is the classification of individuals into healthy and diseased, based on biomarkers or gene signatures derived in former experiments. To reduce the error rate and the risk of possible wrong classifications, data quality needs to meet certain standards. However, no gold standard for Microarray analysis and preprocessing exists [All+06]. Setting decent quality control checkpoints could help improve the results of Microarray analysis and set new standards for future analyses. We therefore have developed a clinically oriented processing scheme for quality control. Our criteria encompass a check on present call rate, the amount of background noise, the range of expression values and the intragroup-correlation of samples belonging to the same group of replicates. We show that, using these control criteria, the results of gene expression data analysis show higher quality and contain lower numbers of false positives. We propose a number of quality control points and show the benefit of applying these standards on expression data analysis on a murine data set of immune cells, containing 30 samples. From our analyses we have clear indication that the establishment of quality control check points result in a better reproducibility of microarray experiments, which, so far, remains a challenge [Ioa+09].

References

[All+06] D. B. Allison et al. “Microarray data analysis: from disarray to consolidation and consensus”. In: *Nature Reviews Genetics* 7 (2006), pp. 55–65. DOI: 10.1038/nrg1749.

[Ioa+09] J. P. A. Ioannidis et al. “Repeatability of published microarray gene expression analyses”. In: *Nature Genetics* 41 (2009), pp. 149–155. DOI: 10.1038/ng.295.

The Pools-and-Proteins Approach: Dynamic Networks from Single-Molecule-Reactions

Tihamer Geyer[1], Florian Lauck[2] and Volkhard Helms[1]

[1]Center for Bioinformatics, Saarland University, Saarbrücken, Germany

[2]Department of Biopharmaceutical Sciences, University of California, San Francisco, USA

One of the key issues for the understanding of how a cell works is to reconstruct the metabolic and regulatory networks. The usual—and quite successful—approach is based on the notion of pathways. There, one tries to connect and complete the experimentally determined conversion chains to form a map-like network with the metabolites as nodes and the enzymes hidden in the links.

However, as many pathways were identified under steady state conditions, it is often unclear, how well these top-down models are suited to describe dynamic scenarios.

Here, we propose a bottom-up approach, where the building blocks are the enzymatically active proteins. For every copy of the protein in the cellular context there is one independent copy in the simulation model. The proteins in turn are built up from their elementary reactions like the association of a metabolite or the transfer of an electron from one residue to another. In the simulation, these one-metabolite-at-a-time reactions are propagated stochastically and independently. The metabolites may diffuse freely within their respective compartments. With these well-mixed pools of metabolites and the independent active proteins, it was named the "pools-and-proteins" model. The major difference compared to the standard descriptions is that no *a priori* network is given but that the fluxes may evolve freely within the assigments of which metabolite is connected to which of the proteins.

To develop and evaluate this modelling approach, we used the well known and simple photosynthetic apparatus of the purple bacterium *Rhodobacter sphaeroides*. We also demonstrate that all model parameters can be determined from a few dynamic experiments. The parametrized proteins will be re-used in the modelling of other metabolic systems, iteratively building an expanding library of parametrized protein models.

Graph Measures Reveal Fine Structure of Complexes Forming in Multiparticle Simulations

Florian Lauck[1], Volkhard Helms[2] and Tihamer Geyer[2]

[1]Department of Biopharmaceutical Sciences, University of California, San Francisco, USA

[2]Center for Bioinformatics, Saarland University, Saarbrücken, Germany

Modern simulation techniques are beginning to study the dynamic assembly and disassembly of multi-protein systems. In these many-particle simulations it can be very tedious to monitor the formation of specific structures such as fully assembled protein complexes or virus capsids above a background of monomers and partial complexes. However, such analyses can be performed conveniently when the spatial configuration is mapped onto a dynamically updated interaction graph. On the example of Monte Carlo simulations of spherical particles with either isotropic or directed mutual attractions we demonstrate that this combined strategy allows for an efficient and also detailed analysis of complex formation in many-particle systems [Lau+09].

References

[Lau+09] F. Lauck et al. "Graph Measures Reveal Fine Structure of Complexes Forming in Multiparticle Simulations". In: *Journal of Chemical Theory and Computation* 5.3 (2009), pp. 641–648. DOI: 10.1021/ct800396v.

Distinguishing the exhaled breath condensates of two patient groups with an electronic nose

Akira Hattesohl[1], Sarah Noeske[1], Robert Bals[1], Claus Vogelmeier[1] and Rembert Koczulla[1]

[1]Universitätsklinikum Marburg, Germany

Background. Exhaled respiratory air contains countless volatile organic compounds (VOCs). The composition of those VOCs varies dependent of different pulmonary diseases. With the customizable electronic nose *Cyranose 320* (C-320) [Smi] it is possible to detect VOCs and distinguish between different VOC patterns. It was the aim of this study to analyze and distinguish the VOCs in exhaled breath condensate (EBC) of patients with α_1-antitrypsin deficiency (AATD) and healthy controls (HC).

Materials and methods. The EBC was collected from ten HCs and ten AATD patients with the RTube [Res] by ten minutes of tidal breathing. 250 μl of EBC were heated up to $37°C$ and argon gas was passed through with a constant flow of 350 ml/min for two minutes. Subsequently, the samples were measured by holding the snout of the C-320 above the surface of the samples and drawing gas for ten seconds. Ten samples of ambient lab air were taken as loading controls. A principal component analysis (PCA) (to reduce the dimensionality) and a linear discriminant analysis (LDA) (to separate the different groups) were performed on the data.

Results. The analyses showed that AATD patients, HCs and lab air are clearly separable. The Mahalanobis distances [Mae+00] and the cross-validation values (CVV) were calculated with a 100-fold-cross-validation using ten percent of the training data as test data.

Conclusion. Using the C-320, we could distinguish between EBCs from AATD patients and HCs. This approach also shows EBC contains enough VOCs to separate two different patient groups. An advantage of using the EBC is the lower contamination risk by not measuring for example infectious patients directly with the C-320 and easiness to sample and store EBC.

References

[Mae+00] R. d. Maesschalck et al. “The Mahalanobis distance”. In: *Chemometrics and Intelligent Laboratory Systems* 50.1 (2000), pp. 1–18. DOI: 10.1016/S0169-7439(99)00047-7.

[Res] Respiratory Research, Inc., Charlottesville, Virginia, USA. *RTube*.

[Smi] Smiths Detection - Pasadena, Inc., Pasadena, California, USA. *Cyranose R320*.

Mapping of GABI-Kat Flanking Sequence Tags from A. thaliana to the TAIR9 Annotation Data Set

Nils Kleinbölting[1], Gunnar Huep[1], Yong Li[2] and Bernd Weisshaar[1]

[1]Department of Biology, Bielefeld University, Germany

[2]University of Freiburg, Germany

Knockout mutants of the model plant *Arabidopsis thaliana* are important tools for reverse genetics. The T-DNA based GABI-Kat population contains about 65.000 sequence-indexed insertion alleles and is the second largest T-DNA mutant collection worldwide.

T-DNA insertion into the genome is a random process. Transformed plants can be selected after transformation in the greenhouse, but the position of the T-DNA insertions in the individual plants remains unknown [Ros+03]. Flanking Sequence Tags (FSTs) can be generated by PCR amplification of genome fragments neighbouring the T-DNA insertion by using primers specific to the T-DNA border [Str+03]. Subsequently, the FSTs can be used to predict the position of the insertion in the genome sequence. The annotation of the GABI-Kat insertion alleles is based on the TIGR v5 annotation release of 2004 [Gab]. In June 2009, the TAIR9 *A. thaliana* genome annotation was made available. The release incorporates next-generation re-sequencing data, new genetic features like microRNA genes, and is based on pseudochromosome coordinates [Gab]. In June 2009, the TAIR9 *A. thaliana* genome annotation was made available. The release incorporates next-generation resequencing data, new genetic features like microRNA genes, and is based on pseudo-chromosome coordinates.

A re-mapping of the FSTs to the new sequence and annotation data set is highly desirable to make use of the comprehensive biological annotation data included in TAIR9. This poster presents a strategy to calculate exact positions of the insertion alleles and their classification with respect to transcription units and coding sequences. Therefore, BLAST searches on the nucleotide level will be done and subsequently evaluated with regard to T-DNA fragments and multiple FST-sequences for the same insertion.

References

[Gab] *GABI-Kat*. 2008. URL: http://www.gabi-kat.de/.

[Ros+03] M. G. Rosso et al. "An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics". In: *Plant Molecular Biology* 53.1-2 (2003), pp. 247–259. DOI: 10.1023/B:PLAN.0000009297.37235.4a.

[Str+03] N. Strizhov et al. "High-throughput generation of sequence indexes from T-DNA mutagenized Arabidopsis thaliana lines". In: *BioTechniques* 35.6 (2003), pp. 1164–8.

Systematic evaluation of centroid algorithms for de-novo motif detection

Hendrik Mehlhorn[1] and Ivo Grosse[2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]Martin-Luther-University Halle-Wittenberg, Germany

Understanding the transcriptional regulation of gene expression is of central importance for many branches of molecular biology. One of the fundamental prerequisites is the identification of *cis*-elements and *cis*-regulatory modules (*CRMs*). Modern wet-lab techniques such as *EMSA*, *DNAse footprinting*, *ELISA*, *ChIP-chip*, or *ChIP-seq* allow a systematic identification of putative *cis*-elements and *CRM*s. However, these techniques are time-consuming and expensive, so computer algorithms such as *MEME* or the *Gibbs Sampler* have been developed as cheap alternatives.

One of the most accurate algorithms for the prediction of *CRM*s is the *Centroid Gibbs Sampler* (*CGS*) [New+07]. This algorithm combines three noteworthy features: (i) it is based on a realistic model for *CRM*s, (ii) it uses *phylogenetic footprinting*, and (iii) it applies a new prediction approach based on *centroids*. While it could be clearly demonstrated that the *CGS* outperforms almost any other *CRM* prediction algorithm, it has not been investigated to which degree each of the improvements (i) – (iii) contribute to the observed accuracy.

Here, we present a systematic comparison of the *CGS* with four existing and nine novel *CRM* prediction algorithms, which we apply to 24 different data sets of varying complexity. Interestingly, we find that the *centroid* approach leads to consistently improved predictions in almost all studied cases even when applied to other iterative sampling algorithms. These findings suggest that similar *centroid* algorithms might be useful in other areas of modern genomics and epigenomics, leading to possibly improved predictions of splice sites, translation initiation sites, transcription start sites, nucleosome binding sites, or miRNA binding sites.

References

[New+07] L. A. Newberg et al. “A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction”. In: *Bioinformatics* 23.14 (2007), pp. 1718–1727. DOI: 10.1093/bioinformatics/btm241.

Mayday and RLink

Florian Battke[1], Stephan Symons[2] and Kay Nieselt[2]

[1]Center for Bioinformatics Tübingen, Department of Information and Cognitive Sciences, University of Tübingen, Germany

[2]ZBIT/PAS, University of Tübingen, Germany

DNA Microarrays are the standard method for large scale analyses of gene expression and epigenomics. MAYDAY [Die+06] is a free graphical workbench for visual analysis of microarray data written in Java. New challenges can swiftly be met due to MAYDAY's plugin interface. MAYDAY includes a large variety of plugins for visual data exploration, clustering, machine learning and classification, as well as Gene Set Enrichment Analysis. Interactive visualization tools, the use of metadata to enhance plots as well as the possibility to create publication quality images make MAYDAY a power analysis tool for microarray data.

We present MAYDAY RLink, an interactive shell harnessing the power of R [R D09] within the framework provided by MAYDAY. Within this shell, users can directly work on MAYDAY's core data structures, and apply methods provided by R or R packages such as those from Bioconductor [Gen+04]. Results can either be passed back to MAYDAY in the form of new datasets or attached to the original data as meta-information. Plugins offered by MAYDAY can also be called from the R shell allowing sophisticated analyses by combining the methods both programs offer, manual as well as scripted. Furthermore, any number of R processes can connect to a MAYDAY instance over the network. This allows to move expensive computations to dedicated machines and have them return their results to MAYDAY when they are finished.

Mayday and RLink are available at `http://www.zbit.uni-tuebingen.de/pas/mayday/`

References

[Die+06] J. Dietzsch et al. "Mayday – a microarray data analysis workbench". In: *Bioinformatics* 22.8 (2006), pp. 1010–1012. DOI: `10.1093/bioinformatics/btl070`.

[Gen+04] R. Gentleman et al. “Bioconductor: open software development for computational biology and bioinformatics”. In: *Genome Biology* 5.10 (2004), R80. DOI: 10.1186/gb-2004-5-10-r80.

[R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria 2009. ISBN: 3-900051-07-0. URL: http://www.R-project.org.

Acceleration of hidden Markov model based database searches with PSSM family models

Michael Beckstette[1], Robert Homann[2], Robert Giegerich[3] and Stefan Kurtz[4]

[1]Application-Oriented Bioinformatics Group; Center for Bioinformatics, University of Hamburg, Germany

[2]International NRW Graduate School in Bioinformatics and Genome Research, Center for Biotechnology (CeBiTec), Bielefeld University, Germany

[3]Faculty of Technology, Bielefeld University, Germany

[4]University of Hamburg, Germany

Profile Hidden Markov models (pHMMs) are probably the most popular modeling concept for protein families. They provide sensitive family descriptors, and sequence database searching with models from major pHMM collections has become a standard task in today's sequence analyses and genome annotation pipelines. On the downside, searching for pHMMs with programs like *hmmsearch* or *hmmpfam* is computationally expensive. The application of the programs to complete proteomes, or even whole protein databases like UniProtKB/Swiss-Prot or UniProtKB/TrEMBL, would require vast computational resources.

We propose a new method to speed up *hmmsearch* in runtime critical large scale analysis scenarios. We employ simpler models of protein families called PSSM family models. For fast database search with these models, we combine full text indexing, efficient exact P-value computation of PSSM match scores, and fast fragment chaining. The resulting method is very fast due to its sublinear expected runtime and is well suited to pre-filter the set of sequences to be searched for subsequent database searches with *hmmsearch* using pHMMs.

In experiments on the SCOP database, adressing the methods' sensitivity and specificity, we achieved a classification performance only marginally inferior to *hmmsearch*. Yet, results were obtained in a fraction of runtime and we observed a speedup factor of more than 170. In experiments addressing the methods' ability to pre-filter the sequence space for subsequent database searches with pHMMs, our method reduces the number of sequences to be searched with *hmmsearch* to only 1.32% of all sequences. The filter is very

fast and leads to a total speedup of factor 72 over the unfiltered search, while retaining 99.7% of the original results, including E-values and scores.

A reliable pipeline for in silico identification of sequence polymorphisms by combining 454 and Sanger reads of sugar beet

Daniela Holtgraewe[1], Thomas Rosleff Soerensen[1], Paul Dirksen[1], Prisca Viehoever[1], Cornelia Lange[2], Heinz Himmelbauer[2], Britta Schulz[3] and Bernd Weisshaar[4]

[1]Bielefeld University, Germany

[2]Max Planck Institute of Molecular Genetics Berlin, Germany

[3]KWS Saatzucht AG, Germany

[4]Bielefeld University, Germany

Genetic markers are a basic and highly valuable tool for creating genetic maps as well as for studying genetic linkage between yield, disease resistance or other traits and genetic inheritance in model and crop plants. Sugar beet (*Beta vulgaris*) is an important crop for sucrose production in the temperate climate zone, and its breeding is increasingly supported by genetic markers. SNP- (single or simple nucleotide polymorphism) based markers have a high potential for automatisation and multiparallel analysis, allowing high throughput genotyping of many individuals. Good SNP markers should be locus specific, of intermediate frequency (to be useful in many populations), and easy to genotype. Reliable detection of polymorphic sites is a crucial step for the development of large SNP based marker sets. We have developed a bioinformatic pipeline for the detection, evaluation and verification of transcript-associated SNPs from sugar beet. ESTs derived from 454 sequencing of one parent (P2) of a mapping population where compared with existing Sanger ESTs derived from the other parent (P1). In addition to several sequence (pre-)processing steps like trimming and clustering, the polymorphic site detection was integrated with tracking of quality values. From one 454 GS-FLX run we obtained 266,000 reads with an average read length of 428 bp from P2. These reads assembeled into about 13,000 contigs and 76,000 singlets. In addition, we used 11,119 Sanger ESTs from P1 for the analysis (Ref. 1). By scoring multiple alignments, we obtained 22,847 SNPs from 454 singlets and 3,268 SNPs in 454 contigs respectively. According to empirically optimised parameters for SNP evaluation, we have classified these SNPs in

"good", "useful" and "bad" and came up with 2,481 "good" SNPs. A subset of these "good" *in silico* detected SNPs was used for mismatch verification by amplicon sequencing and subsequent marker generation. All "good" candidate SNPs for which amplicons have been generated were confirmed, and all known SNPs which overlap the *in silico* detected set fall into the "good" group.

References

[Her+02] R. Herwig et al. "Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes". In: *The Plant Journal* 32.5 (2002), pp. 845–857. DOI: 10.1046/j.1365-313X.2002.01457.x.

An implementation of index-based bidirectional pattern search with an application to RNA structural motifs

Fernando Meyer[1], Sebastian Will[2] and Michael Beckstette[1]

[1]Application-Oriented Bioinformatics Group, Center for Bioinformatics, University of Hamburg, Germany

[2]Albert-Ludwig-University of Freiburg, Germany

The dispersal of next-generation sequencing technologies and the resulting exponential growth rate of biological sequence databases ask for tools that allow for fast search of biologically relevant motifs. While several sequence-pattern matching problems can be vastly accelerated by well-known index data structures [Bec+06], like suffix trees and suffix arrays, they are not well suited for efficient search with RNA structural motifs. Since they only support traditional left-to-right pattern search, they cannot handle the constraints introduced by the secondary structures of RNA molecules.

We present an implementation for efficient RNA structural motifs matching that builds an index for a given sequence database using a data structure called affix array [Str07]. The affix array allows for ultra-fast bidirectional pattern search, by performing it from the middle to the left and right boundaries of the pattern. Due to their symmetric layout, the analysis of RNA secondary structures like hairpins is extremely accelerated with this data structure. In an extended application, we use sets of structural motifs for RNA family classification. This is accomplished by extracting patterns from multiple alignments of families and by searching for chains of matches where the order of the patterns is preserved.

Experiments show that our implementation of affix arrays speeds up search for structural motifs by a factor of thousands of times compared to a simple online search implementation. We also show that multiple motifs with preserved order can be used as descriptors for RNA families and allow, in combination with the pattern matching implementation, for efficient and accurate RNA family classification.

References

[Bec+06] M. Beckstette et al. “Fast index based algorithms and software for matching position specific scoring matrices”. In: *BMC Bioinformatics* 7.1 (2006), p. 389. DOI: 10.1186/1471-2105-7-389.

[Str07] D. Strothmann. “The affix array data structure and its applications to RNA secondary structure analysis”. In: *Theoretical Computer Science* 389.1-2 (2007), pp. 278–294. DOI: 10.1016/j.tcs.2007.09.029.

An automatic method to create function-specific sequence patterns for enzymes

Constantin Bannert[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

Models for the simulation of metabolic networks require the accurate prediction of enzyme function. Based on a genomic sequence, enzyme function is today mainly predicted by sequence comparison and operon analysis. Other methods in sequence analysis can be used to support these techniques: We have developed an automatic method that creates function-specific sequence patterns for enzymes.

First, the enzymes in the UniprotKB are identified and compared against each other with BLAST. The pairwise BLAST E-Values are then used as distances in the clustering of the enzymes. The clustering results in a number of trees. Each tree node contains a set of enzymes that we align with ClustalW, if certain constraints are met. The conserved columns in the multiple alignment are used to construct a pattern, representing the conserved evolutionary information in the given node. Finally, we investigate the quality of each pattern. Among other data, we evaluate the number of true and false positives that we find when searching the proteins in the source database. Patterns with sufficient quality can then be used to match microbial genes without functional annotation.

We ran our protocol on a recent SwissProt release and discuss the results. The main advantages of our method are that it is unsupervised, and the search procedure is quite fast once the patterns are ready.

Predicting PDZ domain — motif interaction specificities from large-scale experimental data: a comparative study

Katja Luck[1], Sadek Fournane[1], Sebastian Charbonnier[1], Murielle Masson[1], Bruno Kieffer[1] and Gilles Travé[1]

[1]Ecole Supérieure de Biotechnologie de Strasbourg, Université Strasbourg

PDZ domains constitute a large family of globular domains (250 copies in human proteins), which generally participate in the control of cell polarity and intercellular communication. PDZs bind small motifs at the extreme C-terminus of their target proteins according to selectivity rules which are not fully understood. The E6 oncoprotein of Human Papillomaviruses (HPV) contains a C-terminal PDZ-binding motif. E6 provokes proteasome-driven destruction of some of its PDZ-containing targets, altering cell adhesion and communication, which may contribute to cancer development. Cervical cancer in women is caused by HPV and it is proven that the PDZ-binding motif of E6 proteins is critical for this development. Our aim is to understand the complex network that is perturbed upon binding of E6 to PDZ-containing proteins. This includes the prediction of the PDZ domain-containing proteins targeted by E6 as well as the cellular interaction partners of these proteins. To this end, we compared strategies for predicting PDZ-motif specificities based on two published large-scale studies of PDZ-peptide interactions which have used very different experimental approaches. Tonikian et al. [Ton+08] applied phage display to analyse the binding of PDZ domains to a random library of artificial C-terminal sequences; whereas Chen et al. [Che+08] measured the interaction of PDZ domains to putative PDZ-binding C-terminal sequences derived from cellular proteins. A careful analysis of this data and performed tests show a strong bias related to the experimental approach applied to establish the training data. We discuss the nature and possible causes of the bias, the weaknesses of the inferred predictions, and their possible improvements. We now plan to carry out experimental tests to validate our conclusions.

References

[Che+08] J. R. Chen et al. "Predicting PDZ domain-peptide interactions from primary sequences". In: *Nature Biotechnology* 26 (2008), pp. 1041–1045. DOI: 10.1038/nbt.1489.

[Ton+08] R. Tonikian et al. “A Specificity Map for the PDZ Domain Family”. In: *PLoS Biology* 6.9 (2008). DOI: 10.1371/journal.pbio.0060239.

Expert knowledge enhanced structure based profile HMMs for protein sequence families

Daniel Bindreither[1], Stefan Wegenkittl[2], Felix Auer[2] and Peter Lackner[1]

[1]University of Salzburg, Austria

[2]University of Applied Sciences Salzburg, Austria

Hidden Markov Models (HMMs) are frequently used to encode protein sequence information into stochastic models which subsequently are applied to identify related proteins in sequence databases. Profile HMMs are derived from multiple sequence alignments and thus heavily depend on the quality and information content of the latter. It has been shown that multiple structure comparison improves the performance of HMMs when the sequence similarity in a certain protein family is low [Ber+07]. Further improvement is expected when supplementary biological information such as strictly conserved sites can be considered more explicitly in the HMM. Working along these lines we developed a methodology and the appropriate software tool which provides the expert user a facility to fine tune the representation of the biological properties in the HMMs derived from a multiple structure alignment (MStA). The resulting tool HMModeler is implemented as an extension for the UCSF Chimera [Pet+04] molecular modelling system. HMModeler [Weg+09] provides for a GUI for editing the alignments and supports the user in adding knowledge to the model. This is done by column-wise fine tuning of predetermined breaking points (acceptance of insertions or deletions), degree of conservation, and specified amino acid sets. It allows for a precise definition of the skeleton of the HMM without the need for understanding the stochastic nature and numerous transition and emission probabilities of the model. The approach was tested with SCOP [Mur+95] families and superfamilies using a Viterbi algorithm. Improvements compared to Clustal [Lar+07] and several MStA based alignments used as input for HMMER [Edd98] and SAM [HK96] are shown.

References

[Ber+07] J. S. Bernardes et al. "Improving model construction of profile HMMs for remote homology detection through structural alignment". In: *BMC Bioinformatics* 8 (2007), p. 435. DOI: 10.1186/1471-2105-8-435.

[Edd98] S. R. Eddy. "Profile hidden Markov models". In: *Bioinformatics* 14.9 (1998), pp. 755–763. DOI: 10.1093/bioinformatics/14.9.755.

[HK96] R. Hughey and A. Krogh. "Hidden Markov models for sequence analysis: extension and analysis of the basic method". In: *Computer Applications In The Biosciences* 12.2 (1996), pp. 95–107. DOI: 10.1093/bioinformatics/12.2.95.

[Lar+07] M. A. Larkin et al. "Clustal W and Clustal X version 2.0". In: *Bioinformatics* 23.21 (2007), pp. 2947–2948. DOI: 10.1093/bioinformatics/btm404.

[Mur+95] A. G. Murzin et al. "SCOP: A structural classification of proteins database for the investigation of sequences and structures". In: *Journal of Molecular Biology* 247.4 (1995), pp. 536–540. DOI: 10.1016/S0022-2836(05)80134-2.

[Pet+04] E. F. Pettersen et al. "UCSF Chimera - A visualization system for exploratory research and analysis". In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612. DOI: 10.1002/jcc.20084.

[Weg+09] S. Wegenkittl et al. *Optimierte Modelle zur Beschreibung von Proteinfamilien*. 2009.

Construction of an integrated biochemical reaction database

Michael Stelzer[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

For the construction of cellular models, it is essential to develop organism- specific reaction networks in order to be as close as possible to reality. A number of sources for biochemical reactions exist like the databases *BRENDA* [Cha+09], *KEGG* [Kan+08], and *MetaCyc* [Cas+08]. Due to the fact that the completeness of reaction data differs between the databases, it becomes important to combine the available reaction information of the used source databases in form of an integrated reaction database. One crucial point hereby is that the same pathway occurring in compared databases often varies in size or length, i.e. it consists of different number and type of reactions. So a combination of same pathways will lead to more complete metabolic networks. Therefore it is necessary to find identical reactions between the recognised databases. In this work information of the biological databases *BRENDA*, *KEGG*, and *MetaCyc* was used. Identification of equal reactions between two databases was achieved by an *in silico* approach in which a compound name comparison (incl. synonyms) and, in a second step, a comparison by *InChIs* (linearised chemical structure), are combined. Data download, storage and comparison is realised by *Python* scripts and therein embedded *MySQL* statements. The combined database will contain a unique list of reactions that occur in any of these databases and the associations between equivalent reactions.

References

[Cas+08] R. Caspi et al. "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases". In: *Nucleic Acids Research* 36.suppl_1 (2008), pp. D623–631. DOI: 10.1093/nar/gkm900.

[Cha+09] A. Chang et al. "BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009". In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D588–592. DOI: 10.1093/nar/gkn820.

[Kan+08] M. Kanehisa et al. "KEGG for linking genomes to life and the environment". In: *Nucleic Acids Research* 36.suppl_1 (2008), pp. D480–484. DOI: 10.1093/nar/gkm882.

Improving protein structure prediction using sequence-derived structure profiles

Katrin Wolff[1], Michele Vendruscolo[2] and Markus Porto[1]

[1]Darmstadt University of Technology, Germany

[2]Cambridge University, United Kingdom

A crucial step in the prediction of protein structures is the transition from low-resolution candidates to high-resolution models. There exist various tools that generate candidate sets that contain high-quality, yet coarse-grained, structures. In a subsequent refinement step these structures are improved to all-atom representations and minimized using a high-resolution energy functional. Due to limited computer time it is vital to restrict this refinement step to promising candidates. In particular, it is very important to reliably identify the very best few structures. The energy functional used in the structure generation step, however, is only of limited use for the problem of recognizing and selecting these 'good' structures. We discuss the use of structure profiles for this filtering step. As a proof of principle we show that the exact profile (derived from the native structure) is very secure in choosing candidates with low cRMSD to the native structure and clearly outperforms other filtering methods like filtering by energy or clustering the decoy set. Such structure profiles can also be predicted to good accuracy from sequence. We therefore test the use of profiles as predicted from sequence and show that our approach reliably identifies candidates with low cRMSD to the native structure [Wol+09].

References

[Wol+09] K. Wolff et al. *Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles.* Proteins: Structure, Function, and Bioinformatics, in press. 2009.

New features for the BRENDA enzyme information system

Andreas Grote[1], Maurice Scheer[1], Antje Chang[1], Ida Schomburg[1], Carola Söhngen[2], Michael Stelzer[1], Michael Rother[1], Juliane Thiele[2], Cornelia Munaretto[2] and Dietmar Schomburg[2]

[1]Department of Bioinformatics and Biochemistry, Braunschweig University of Technology, Germany

[2]Braunschweig University of Technology, Germany

BRENDA (BRaunschweig ENzyme DAtabase) is the largest information system on enzymes and enzyme-related data [Cha+09]. It comprises molecular, biochemical and structural data on all classified enzymes. Detailed information on enzyme nomenclature, catalyzed biochemical reactions, substrate specificity, stability, mutants, isolations and preparation, 3D structure, physicochemical parameters and the tissue-specific localization of enzymes was manually extracted from over 91,000 primary literature references. For each enzyme links to the relevant literature references, the source organism and, if available, the protein sequences are provided. The data can be accessed by a web-based query engine (`http:/www.brenda-enzymes.org`) which comprises sophisticated software tools e.g the presentation of 3D structures of enzymes, the search for enzyme-ligands according to a specific substructure or a tool for browsing the genomic-context of the enzyme-encoding genes.

Newly added features of the query engine allow to view protein-specific enzyme data and visualize the active sites (e.g. substrate binding sites) in the 3D structure view of an enzyme. Furthermore, a ligand-centric summary site and SBML export functionalities were introduced. In addition to the manually curated data, BRENDA is supplemented with data from other public databases (e.g. UniProt, PDB) and with information that is automatically extracted from Pubmed abstracts by a text-mining procedure. This text-mining routine generates two additional repositories FRENDA (Full Reference ENzyme DAta) and AMENDA (Automatic Mining of ENzyme DAta) and was recently improved by reducing the number of false-positives hits. Likewise, the routine for the automatic mining of diseases was strongly enhanced increasing the overall number of hits by a factor of nine. Finally, the SOAP interface was recently re-programmed to allow a simplified computational access to the data of BRENDA.

References

[Cha+09] A. Chang et al. “BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009”. In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D588–592. DOI: 10.1093/nar/gkn820.

Next-Generation sequencing and Metagenomics

Suparna Mitra[1], Max Schubach[1] and Daniel Huson[1]

[1]University of Tübingen, Germany

Background. Metagenomics is the study of environmental samples using sequencing. Rapid advances in sequencing technology are fueling a vast increase in the number and scope of metagenomics projects. Most metagenome sequencing projects so far have been based on Sanger or Roche-454 sequencing, as only these technologies provide long enough reads, while Illumina sequencing has not usually been considered suitable for metagenomic studies due to a short read length of only 35bp. However, now that reads of length 75bp can be sequenced in pairs, Illumina sequencing has become a viable option for metagenome studies. We performed a simulation study of three different metagenomes: low complex (LC), moderately complex (MC) and highly complex (HC) metagenome. We simulated 6,000 reads with 250bp for Roche-454 and 200,000 reads with 75bp for the Illumina technology using MetaSim. For Illumina we used paired-reads with both short and long clone libraries.

Results. Our study addresses the problem of taxonomical analysis of paired-reads. We describe a new feature of our metagenome analysis software MEGAN that allows one to process sequencing reads in pairs and makes assignments of such reads based on the combined bit-scores of their matches to reference sequences. Using this new software in a simulation study, we investigate the use of Illumina paired-sequencing in taxonomical analysis and compare the performance of single reads, short clones and long clones. In addition, we also compare against simulated Roche-454 sequencing runs. This work shows that paired-reads perform better than single reads, as expected, but also, less obviously, that long clones allow more specific assignments than short ones.

Availability. A new version of MEGAN that explicitly takes paired-reads into account will be available from: `www-ab.informatik.uni-tuebingen.de/software/megan`

References

[Hus+07] D. H. Huson et al. "MEGAN analysis of metagenomic data". In: *Genome Research* 17.3 (2007), pp. 377–386. DOI: 10.1101/gr.5969107.

[Mit+10] S. Mitra et al. *Short clones or long clones? A simulation study on the use of paired-reads in metagenomics*. Submitted to APBC. 2010.

[Ric+08] D. C. Richter et al. "MetaSim – A Sequencing Simulator for Genomics and Metagenomics". In: *PLoS ONE* 3.10 (2008), e3373. DOI: 10.1371/journal.pone.0003373.

High Quality Protein Sequence Alignment by combining Structural Profile Prediction and Profile Alignment using SABERTOOTH

Florian Teichert[1], Jonas Minning[1], Ugo Bastolla[2] and Markus Porto[1]

[1]Darmstadt University of Technology, Germany

[2]Centro de Biologia Molecular "Severo Ochoa", Spain

While sequence alignments give satisfactory results in the area of high sequence identity, distantly related proteins are better compared using structure alignments. This is especially true when sequence similarity reaches the so-called 'twilight zone' where sequence similarity measures get indistinguishable from random. Unfortunately, structural information is frequently not available which demands for the development of techniques that extend the applicability of accurate sequence alignments to more distantly related proteins.

We present here a new sequence alignment tool which combines structural profile prediction from the protein's sequence with subsequent alignment of these profiles using our recently developed alignment tool SABERTOOTH [Tei+09]. In particular, we predict the residue contact vector of protein structure, utilising position specific scoring matrices output by PSI-BLAST as input to an artificial neural network. We have shown before that the exact contact vector derived from coordinates contains sufficient information to perform state-of-the-art structure alignments. The sequence alignments produced by SABERTOOTH are of superior quality in comparison to the established sequence aligners Clustal W, T-Coffee, MUSCLE, and PSI-BLAST itself. We assess the alignment accuracy by evaluating the quality of the corresponding spatial structure superimposition for a comprehensive set of proteins with known structure, with relationships ranging from very close (SCOP family level) to very distant (SCOP fold level, at which simple sequence identity is unable to detect significant relationships). We utilise the structural PSI to measure global alignment accuracy, as it establishes an objective reference for recognising similarities. Furthermore, we assess the quality of the significance scores output by SABERTOOTH, by quantifying their compliance with the SCOP classification.

References

[Tei+09] F. Teichert et al. *High Quality Protein Sequence Alignment by combining Structural Profile Prediction and Profile Alignment using SABERTOOTH*. Submitted. 2009.

Better Guide Trees from Non-Metric Structural Similarity/Distance Measures

Thomas Margraf[1] and Andrew E. Torda[1]

[1]Center for Bioinformatics, University of Hamburg, Germany

We have developed a new fast and scalable method for the multiple alignment of large numbers of protein structures. One of the biggest problems in the construction of structure based guide trees and phylogenies is the fact that structural distance measures such as RMSD values do not satisfy the conditions of a metric (particularly the triangle inequality). However, metric distances are a necessary precondition for distance based methods to be able to reconstruct the correct tree. In order to overcome this limitation, HANSWURST extends the progressive alignment approach used by other tools such as ClustalW [Tho+94] in several ways: Firstly, we provide a way to compute probabilistic representations for centroids of clusters of protein structures. HANSWURST can then align these centroids in order to more accurately estimate the distance between internal nodes of the guide tree. Secondly, we present a method to project a non-metric distance matrix into a metric space. In order to do this, Crippen's metric matrix method [CH78] first generates high dimensional euclidean coordinates for the points (proteins) represented by the original distance matrix. Subsequently, we reduce the dimensionality of these coordinates by iteratively removing the least significant dimension through a projection that best preserves the original distances.

Our poster presents these method and compares their resulting trees to those obtained from straightforward UPGMA and Neighbor-Joining.

References

[CH78] G. M. Crippen and T. F. Havel. "Stable calculation of coordinates from distance information". In: *Acta Crystallographica* A34.2 (1978), pp. 282–284. DOI: 10.1107/S0567739478000522.

[Tho+94] J. D. Thompson et al. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic Acids Research* 22.22 (1994), pp. 4673–4680. DOI: 10.1093/nar/22.22.4673.

New insights on non-specific protein- DNA interactions: the DNase I model

Josephine Abi Ghanem[1], Marc Guéroult[2], Brahim Heddi[2, 3], Marc Baaden[2] and Brigitte Hartmann[2]

[1]University Pierre et Marie Curie, France

[2]Centre national de la recherche scientifique (CNRS) - UPR 9080

[3]Institut de Biologie Physico-Chimique (IBPC), France

In the cell, DNA interacts almost continuously with proteins in order to ensure its biological functions. Specific and non-specific protein-DNA interactions imply the formation of intermolecular interfaces requiring electrostatic and structural complementarity of the related partners. Nevertheless, the mechanisms underlying the formation of non-specific protein-DNA complexes remain particularly obscure.

In this context, we chose to study the DNase I/DNA system as a representative and rather simple model of non-specific complex. DNase I is a glycoprotein which hydrolyzes the DNA phosphodiester linkages in presence of divalent cations, Ca^{2+} and Mg^{2+}. However, two questions remain open: i) the location and the role of the divalent cations and ii) the molecular basis of the variation in DNA cleavage intensities observed when DNase I binding is the limiting step for hydrolysis.

Combining various experimental and theoretical techniques, we studied DNA oligomers and DNase I, free and bound. First, four strong cation binding sites were identified on free DNase I. We showed then that Ca^{2+} and Mg^{2+}, close or within the DNA/protein interface, are crucial for optimizing the electrostatic fit between DNA and enzyme. Second, DNA cleavage intensity was found to be correlated with the adjacent 3' phosphate linkage flexibility. Indeed, flexible phosphates were associated with large minor groove variations that promote the affinity of DNase I, minimizing the cost of DNA deformation upon binding. In sum, this work demonstrates that DNA-DNase I interaction is mediated by both the presence of cations on DNase I (electrostatic fit) and the DNA sequence dependant flexibility (structural fit).

The Bioscanners project at sourceforge

Detlef Groth[1], Stefan Müller[2] and Joachim Selbig[1]

[1]University of Potsdam, Germany

[2]Free University of Berlin, Germany

Although biologists and bioinformaticians are permanently challenged with parsing and analysing output of various biological tools, often utilising different formats, a standard method for solving such tasks has not been found, yet. To overcome the problems of commonly used standalone applications (difficult data integration), and Bio-frameworks (complex programming interface, slow data parsing), we generated the Bioscanners project. The usage of scanner generators in application coding ensures easy programming, little influence of personal programming styles, a small code base, easy maintenance, and very high processing speed.

Besides easy and fast data scanning, the storage of the scanning result in a format, that can be efficiently used in downstream processing and to selectively extract information, is important for the user. For that reason the output of our scanners is standard database code, suitable to be used for SQL compliant databases like PostgreSQL, MySQL or SQLite. That data integration platform ensure the best possible performance in data exploration, especially in a relational context, as well as in maintaining data integrity.

As an example, our BLASTScanner, available from the sourceforge project page, is a single C source file which can be compiled on any modern computer platform to a small 20Kb executable not depending on any external library or runtime. The data processing time required of the scanner is shorter than the time required by the database to import the data. Even input files of several hundred MB can be translated into database code within several seconds.

The scanner applications can serve as a platform for the development of standardized bioinformatics related tools. The source code is freely available at the sourceforge project page (`http://bioscanners.sf.net`) under an opensource BSD-license. Programmers, who would like to participate in the development are highly welcome.

Conserved introns reveal novel transcripts in eukaryotic genomes

Michael Hiller[1], Sven Findeiß[2], Sandro Lein[3], Manja Marz[2], Claudia Nickel[3], Dominic Rose[2], Christine Schulz[4], Rolf Backofen[5], Sonja J. Prohaska[2], Gunter Reuter[3] and Peter F. Stadler[2]

[1]Department of Developmental Biology, Stanford University, USA

[2]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

[3]Institute of Genetics, Biologicum, Martin-Luther-University Halle-Wittenberg, Germany

[4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

[5]Bioinformatics Group, Albert-Ludwigs-University of Freiburg, Germany

In many eukaryotes most of the genome is transcribed producing large numbers of long non-coding RNAs (ncRNAs) with intrinsic functions [Mer+09]. A subclass of them, similar to mRNAs, gets spliced, capped, and polyadenylated and are therefore called mRNA-like non-coding RNAs (mlncRNAs). However, their inventory is incomplete, even in well studied organisms, and so far no computational methods exist to predict them from genomic sequences only. We introduce two different genome-wide comparative genomics approaches identifying conserved transcripts with high specificity by searching for introns in (1) insect and in (2) mammalian genomes. We predict 369 novel *D. melanogaster* introns, among them EST-confirmed ones, introns refining the repertoires of protein-coding genes, and 129 novel unstructured mlncRNAs. Using RT-PCR, we verified 7 of 12 tested introns in novel mlncRNAs and 11 of 17 introns in novel coding genes. The expression level of all verified mlncRNA transcripts is low but varies during development, which suggests regulation [Hil+09]. Similarly, we computationally identified at least 328 novel *C. elegans* introns of which 255 most presumably are novel ncRNAs. Since insect genomes are usually short and compact, we can identify complete introns in one go, whereas *de novo* intron prediction in mammals, dealing with longer poorly conserved introns, must undergo challenging intermediate steps. Based on sequence data alone and using modern machine learning techniques, we are able to reliably recognize mammalian donor and acceptor splice-sites. Eval-

uating our prediction performance on an independent test set (AUC ∼0.90) yields valuable 60% true positives and only 2% false positives. At the expense of sensitivity, we reach promising specificity and we currently investigate, if it is possible to correctly obtain true mammalian donor/acceptor pairs which define novel introns and mark novel spliced transcripts.

References

[Hil+09] M. Hiller et al. "Conserved introns reveal novel transcripts in Drosophila melanogaster". In: *Genome Research* 19.7 (2009), pp. 1289–1300. DOI: 10.1101/gr.090050.108.

[Mer+09] T. R. Mercer et al. "Long non-coding RNAs: insights into functions". In: *Nature Reviews Genetics* 10 (2009), pp. 155–159. DOI: 10.1038/nrg2521.

SEGA — A Semi-Global Approach to Graph Alignment

Marco Mernberger[1], Florian Finkernagel[1], Eyke Huellermeier[1] and Gerhard Klebe[1]

[1]University of Marburg, Germany

The comparative analysis of biomolecules is a central topic in structural bioinformatics. An especially interesting approach to the comparison of protein structures, in particular protein binding sites, has recently been introduced as a counterpart to the concept of sequence alignment. Using approximate graph matching techniques, this method, called *graph alignment*, enables the robust identi cation of approximately conserved patterns in biologically related structures. Since the calculation of optimal graph alignments is computationally intractable, different heuristic methods have been proposed to solve this problem [Fob+09; Wes+07]. These methods tackle the alignment problem globally in the sense of maximizing a scoring function that evaluates alignments of complete structures. Here, we present an alternative method for graph alignment that seeks to combine advantages from global and local approaches to graph comparison.

Like global methods, our algorithm constructs a complete alignment of the structures to be compared. Yet, like local methods, it performs comparisons only on the level of local substructures, resorting to global structure information only when necessary. Compared with existing approaches, it is therefore more robust toward mutational variation, conformational changes, and noisy data. Another advantage of restricting comparisons to local substructures is an increased computational effciency.

References

[Fob+09] T. Fober et al. “Evolutionary Construction of Multiple Graph Alignments for the Structural Analysis of Biomolecules”. In: *Bioinformatics* (2009), btp144. DOI: 10.1093/bioinformatics/btp144.

[Wes+07] N. Weskamp et al. “Multiple Graph Alignment for the Structural Analysis of Protein Active Sites”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 4.2 (2007), pp. 310–320. DOI: 10.1109/TCBB.2007.358301.

Dual-Specificity Phosphatases Specifically Recognise their STAT Partner

Christophe Jardin[1] and Heinrich Sticht[1]

[1]University of Erlangen-Nuremberg, Germany

Protein phosphorylation presents a mechanism by which the activity of numerous proteins is switched on or off across a wide array of biological processes like signal transduction. The extent of tyrosine phosphorylation in signalling pathways is regulated by the coordinated activity of the partner proteins tyrosine kinase and tyrosine phosphatase (PTP). The dual-specificity protein-tyrosine phosphatases (DSPs) are a subfamily of the PTPs that can remove phosphate groups of both phosphothreonine and phosphotyrosine residues from mitogen-activated protein (MAP) kinases in their pT-X-pY motif (X represents any residue). More recently it has been found that VH1-like DSPs can also *specifically* dephosphorylate activated Signal Transducer and Activator of Transcription (STAT) proteins at a phosphorylated tyrosine: VH1 dephosphorylates STAT1 but neither STAT3 nor STAT5 [Man+08] whereas VHR dephosphorylates specifically STAT5. Tyrosine dephosphorylation of STAT5 by VHR requires both an intact Src homology 2 (SH2) domain of the STAT protein and a phosphorylated tyrosine residue located away from the active site at the phosphatases [Hoy+07].

Molecular dynamics simulations, however, didn't reveal any significant preference of VHR for the tyrosine phosphorylated tail of STAT5 over STAT1. We therefore investigated whether the VH1-like phosphatases recognize their STAT-substrate via specific interactions at a second interface formed between the SH2 domain of the STAT protein and the phosphatase. This hypothesis was tested using a docking approach and revealed interactions at the protein-protein interface that offer a plausible explanation for the observed binding specificity.

References

[Hoy+07] R. Hoyt et al. "Cutting Edge: Selective Tyrosine Dephosphorylation of Interferon-Activated Nuclear STAT5 by the VHR Phosphatase". In: *Journal of Immunology* 179.6 (2007), pp. 3402–3406. URL: `http://www.jimmunol.org/cgi/content/abstract/179/6/3402`.

[Man+08] B. A. Mann et al. “Vaccinia Virus Blocks Stat1-Dependent and Stat1-Independent Gene Expression Induced by Type I and Type II Interferons”. In: *Journal of Interferon & Cytokine Research* 28.6 (2008), pp. 367–380. DOI: 10.1089/jir.2007.0113.

Development of Software System for Serological Analysis to Control Porcine Respiratory Disease Complex on a Herd level

Baek-Sop Kim[1], Ji-Hee Yoon[1], Hoo-Don Joo[2] and Byung-Sik Change[2]

[1]Hallym University, Republic of Korea

[2]Jeno Biotech Inc., Republic of Korea

Porcine respiratory disease complex (PRDC) is a serious health problem in growing and finishing pigs typically around 16–22 weeks of age. PRDC is now recognized as being a disease of major economic importance to the global pig industry. Pneumonia in pigs with PRDC is due to a combination of both viral and bacterial agents, such as porcine reproductive and respiratory syndrome virus (PRRSV), porcine circovirus type 2 (PCV2), swine influenza virus (SIV), mycoplasma hyopneumoniae (MH), actinobacillus pleuropneumoniae (APP) and pasteurella multocida (PM). To control multiple pathogen infection, the antibody profile and infection pattern per herd unit have been required through statistical program. We developed software system for analysis and management of serological dynamics about complex PRDC infection on herd level. The system consists of five parts: data acquisition from ELISA scanner for reading optical density (OD), user interface, data base management, validation check, and data processing by the assay types. The data was expressed by average of sample OD to positive control OD ratio, coefficient of variation (CV%) and standard variation (STD) on herd level. When the animals were vaccinated, the average of antibody titer was increased moderately and STD was measured as low. If the herd was infected with pathogens, the mean of antibody level and CV% were developed 2 times higher than normal herd after one month. Taken together, the developed software gives information of herd antibody profile and infection pattern, and thus it will provide insight into the establishment of an effective control strategy of differentiation infected and vaccinated farm and vaccine schedules for PRDC on the field.

References

[Par+08] C.-K. Park et al. “Infection patterns of porcine reproductive and respiratory syndrome virus by serological analysis on a farm level”. In: *Korean Journal of Veterinary Research* 48.1 (2008), pp. 67–73.

Molecular Dynamics of Viral Glycoproteins

Pia Dirauf[1] and Heinrich Sticht[1]

[1]Institute of Biochemistry, Erlangen University, Germany

Viral envelope fusion glycoproteins, such as vesicular stomatitis virus glycoprotein G (VSV gG), perform large scale motions involving structural rearrangements during fusion of viral and host cell membranes in the process of virus entry. VSV gG forms trimeric complexes and mediates membrane fusion under low pH conditions. Two structures have been solved for VSV gG: One that represents the state before membrane fusion (prefusion) and one that reflects a putative postfusion state [Roc+06; Roc+07]. The aim of this study is to elucidate the dynamic rearrangement process of VSV gG by means of computational methods. Molecular dynamics simulations of VSV gG are performed to study general and correlated motions of gG in both conformations. Analysis of the protein's lowest frequency normal modes is used to explore directed motions between the two conformations. The role of conserved histidine residues as pH-triggered molecular switches of the rearrangement is also examined. Based on structural analogies, herpesviral glycoproteins (e. g. herpes simplex virus glycoprotein B (HSV gB)), are believed to belong to the same class of fusion proteins as VSV gG. However, for HSV gB only a putative postfusion structure has been solved so far [Hel+06]. Insights from VSV gG as a model system will therefore be used to extend our knowledge about herpesviral gB and its mechanism of rearrangement.

References

[Hel+06] E. E. Heldwein et al. "Crystal Structure of Glycoprotein B from Herpes Simplex Virus 1". In: *Science* 313.5784 (2006), pp. 217–220. DOI: 10.1126/science.1126548.

[Roc+06] S. Roche et al. "Crystal Structure of the Low-pH Form of the Vesicular Stomatitis Virus Glycoprotein G". In: *Science* 313.5784 (2006), pp. 187–191. DOI: 10.1126/science.1127683.

[Roc+07] S. Roche et al. "Structure of the Prefusion Form of the Vesicular Stomatitis Virus Glycoprotein G". In: *Science* 315.5813 (2007), pp. 843–848. DOI: 10.1126/science.1135710.

MetaSAMS — Sequence Analysis and Management System for Metagenomics Datasets

Thomas Bekel[1], Christina Ander[1], Martha Zakrzewski[1], Jens Stoye[1] and Alexander Goesmann[1]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, Germany

The Sequence Analysis and Management System — or SAMS for short — has been developed at CeBiTec, the Center for Biotechnology at Bielefeld University [Bek+09]. It has been successfully applied in several sequencing projects. The spectrum of application cases covers bacterial whole genome shotgun sequencing as well as EST clustering, assembly, and functional annotation of eukaryotic cDNA data. In these classical genomics approaches sequence data is obtained from a single cultured target organism or mRNA is isolated from tissues of interest. In contrast to this, metagenomics deals with sequence data obtained directly from environmental samples. Therefore, the taxonomic composition of the analyzed community is investigated by researchers as well as their repertoire of gene functions. Both topics, the taxonomic composition and the functional analysis of a given metagenome, can be analyzed using MetaSAMS.

Due to the modular design of MetaSAMS, bioinformatics tools, e.g. for taxonomic or functional analyses can be integrated very easily. Currently, two taxonomic classification modules are included: The RDP-classifier and CARMA. In addition to BLAST-based analyses, e.g. using the SWISS-PROT database, functional analysis modules are included for different concepts like KEGG or COG.

Within this poster presentation, we will introduce you to the newly developed MetaSAMS platform. Technical background information will be given as well as impressions of MetaSAMS applied to current metagenomics projects run at CeBiTec.

References

[Bek+09] T. Bekela et al. “The Sequence Analysis and Management System – SAMS-2.0: Data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies”. In: *Journal of Biotechnology* 140.1-2 (2009), pp. 3–12. DOI: 10.1016/j.jbiotec.2009.01.006.

EnzymeDetector: a comparison of genome annotations from common databases and an up-to-date BLAST-Search to help the user find the right annotation

Susanne Quester[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

For the construction of models for the simulation of metabolic networks an accurate prediction of enzyme function, based on a genome sequence, is an essential requirement. We found that in 50% of all cases the main annotation databases (e.g. NCBI, KEGG, PEDANT) have contradictory annotations. Therefore we developed the program EnzymeDetector, which automatically compares the annotations of these databases and performs an up-to-date Blast-Search [Joh+08] for every annotated gene of the organism. The BLAST-Search is necessary because the public genome annotations might be based on non-up-to-date protein sequences. The BLAST-Output is afterwards evaluated and, if possible, a decision for one gene function is made. Otherwise more than one function is suggested. The program output is stored in a SQL-database, which contains all BLAST-hits for every gene of the organism, that have an e-value smaller than e^{-10}. Additionally the database contains information about the quality of the annotation. A score is assigned to every gene enzyme combination, if the input databases have this enzyme annotated, if the evaluation program suggests this annotation or if this enzyme is known to be present in the organism based on BRENDA or AMENDA [Sch+02]. The sum of all scores builds the relevance, which gives an evidence for the quality of the annotation. Subsequently the program gapFiller can be used to fill the existing gaps in the metabolic network.

References

[Joh+08] M. Johnson et al. "NCBI BLAST: a better web interface". In: *Nucleic Acids Research* 36.suppl_2 (2008), W5–9. DOI: 10.1093/nar/gkn201.

[Sch+02] I. Schomburg et al. "BRENDA, enzyme data and metabolic information". In: *Nucleic Acids Research* 30.1 (2002), pp. 47–49. DOI: 10.1093/nar/30.1.47.

GapFiller — a tool for the identification of missing enzymes in biochemical pathways

Maren Lang[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

Correct predictions of enzyme functions are crucial for the reconstruction of metabolic networks. Incorrect annotations may lead to gaps in reconstructed networks, which occur when the respective enzyme functions are not predicted by bioinformatic methods. Since these predictions are not always reliable and always incomplete, we have developed an automated method for identifying missing enzymes and the genes encoding them. A complete list of enzymes for the organism of interest is obtained by the EnzymeDetector tool which is also developed in our group. This data is combined with metabolic information of the databases BRENDA, KEGG and MetayCyc in order to detect the enzyme-catalyzed reactions. In the next step the reachability of central metabolites to components that belong to the biomass is used for identification of those paths that are fundamental to be accessible in the organism. If a biomass component is not reachable, possible paths are searched by using known reactions of other organism. Enzymes that are not yet annotated for the evaluated organism are further examined. In these cases a "backwards" BLAST search and a sequence-based motif search is performed to get appropriate enzyme predictions including the genomic location of the encoding genes. For the acceptance of predicted sequence candidates, the probability to be in an operon, the existence of gene annotations and the presence in related organisms is taken into account.

References

[Rah+05] S. A. Rahman et al. "Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC)". In: *Bioinformatics* 21.7 (2005), pp. 1189–1193. DOI: 10.1093/bioinformatics/bti116.

BRAVEMAP: a tool for displaying a metabolic network and its different aspects

Susanne Quester[1], Maurice Scheer[1], Michael Rother[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

During the construction and simulation of metabolic networks an appropriate tool for the visualisation of the respective networks is required that provides an overview of the underlying biochemical reactions. We introduce a tool that, supplies a bipartite reference network of pathways, which contains all known reactions. The networks is created with the network editor Cytoscape [Sha+03]. The network can be displayed at different levels of detail. On the first level of detail only the basic metabolites of all pathways are shown, for example solely alanine for the alanine metabolism. On the second level all main metabolites are displayed. On the third level of detail the secondary metabolites, like NADH and ATP, are added to the picture and on the most detailed level all main and secondary metabolites as well as all enzymes are shown. By default the different pathways are presented in different colours. Additionally, the user has the possibility to highlight all irreversible reactions, all enzymes that e.g. consume ATP or produce CO_2, etc.. Another display option is to focus on the reactions that occur in specific organism. The underlying information is obtained by the programs EnzymeDetector and gapFiller also developed in our group. In another setting flux information, for example from a Flux Balance Analysis, can be visualised. Futile cycles and constrained reactions may be highlighted in this setting. These visualisation possibilities allow to identify discrepancies and errors in a metabolic model.

References

[Sha+03] P. Shannon et al. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks”. In: *Genome Research* 13.11 (2003), pp. 2498–2504. DOI: 10.1101/gr.1239303.

MetaboliteExplorer: A GUI based program for the managment of metabolome data, their visualization and database assisted storage of metabolite libraries

Timo Luehr[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

A flexible administration of metabolome data and the creation of corresponding libraries is essential for laboratories routinely performing metabolome experiments. For this purpose MetaboliteExplorer is designed. It stores metabolome data in an SQL database and creates a library of the metabolites. It can be used without prior knowledge of SQL. All manipulations of the established library are monitored and the user can export a change-log file. The program offers an easy adaptation system to expand the existing library. For this purpose it supports several file-formats and is compatible to the NIST format. The adaptation can be performed in different ways such as using spectrum similarity, string equality with and without synonyms and the comparison of numeric values to a user-given threshold. The program supports several export possibilities. With the integrated SQL Query Manager the user can create individual datasets in a dialogue-based manner like filtering for metabolites only present in a certain pathway. The visualisation of stored data can be done in a customized way and the corresponding settings can be saved. A sophisticated query system allows the convenient presentation of diverse metabolite properties such as filtering for experimental determined Rt values. Furthermore the stored metabolites can be processed for displaying in viewers for structural information in 2D or 3D using the OpenBabel library [Guh+06]. Additional information is derived from InChIs, SMILES or other chemical file formats. In addition MetaboliteExplorer provides a spectrum viewer, an assistant for performing regression calculations and a web assistant for KEGG pathway queries.

References

[Guh+06] R. Guha et al. “The Blue ObeliskInteroperability in Chemical Informatics”. In: *Journal of Chemical Information and Modeling* 46.3 (2006), pp. 991–998. DOI: 10.1021/ci050400b.

Reconstructing metabolic networks in silico using CARMEN

Jessica Schneider[1], Eva Trost[1], Andreas Tauch[1] and Alexander Goesmann[1]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, Germany

Newly developed ultrafast sequencing strategies provide efficient access to huge amounts of microbial genome sequences. One challenge after assembly and annotation is the establishment of an automated and fast functional analysis pipeline to gain insights into metabolic features that can be related to organism-specific lifestyle and pathogenicity of microbes.

Therefore, the tool CARMEN has been developed, which supports the fast *de novo* reconstruction of metabolic pathways as well as the creation of template based SBML models for comparative genomics. The *de novo* pathway reconstruction is based on KEGG [Oga+99] pathway maps in combination with genome annotation data, which can be obtained from NCBI GenBank files. This strategy facilitates a rapidly generated overview of the metabolic properties. Nevertheless, a reliable *de novo* reconstruction depends on a high-quality genome annotation and may require manual curation for species-specific networks. In addition, annotation data of a newly sequenced species can be mapped onto a reference SBML template using reciprocal best BLAST hits with the reference species. This poster presents the workflow of CARMEN combined with an application focused on the central carbohydrate metabolism of various Corynebacteria. Extensive manual verification of the reconstructed networks of *C. kroppenstedtii* and *C. aurimucosum* showed reasonable results with regard to their habitat and lifestyle.

CARMEN stores the reconstructed pathways in standardized SBML- format. This XML-based exchange format can be imported by the CellDesigner [Fun+03] software to visualize and manually curate the data. Moreover, the generated pathways can be used as input for further modeling and simulation tools that are compliant to the SBML standard.

References

[Fun+03] A. Funahashi et al. “CellDesigner: a process diagram editor for gene-regulatory and biochemical networks”. In: *BIOSILICO* 1.5 (2003), pp. 159–162. DOI: 10.1016/S1478-5382(03)02370-9.

[Oga+99] H. Ogata et al. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 27.1 (1999), pp. 29–34. DOI: 10.1093/nar/27.1.29.

Computational tools for the integration and analyses of metagenome data

Martha Zakrzewski[1], Andreas Schlüter[1], Felix Tille[1], Wolfgang Gerlach[1], Jens Stoye[1] and Alexander Goesmann[1]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, Germany

Metagenomics gains insights into the understanding of the diversity, function and interaction of microbial communities. Due to the development of high-throughput sequence technologies, metagenomes can be sequenced rapidly at low cost and without cloning bias. However, they also produce a high amount of sequencing data. Therefore, computational tools are necessary for the analyses of metagenome data.

Herein, tools will be presented, which are helpful for the sequence analysis within metagenome projects. The objective of the projects is for example to identify novel enzymes with biotechnological, industrial, or pharmaceutical applications within metagenomes delivered from soil or water habitats. Applicable tools in the focus of metagenome projects can be CARMA, Provides, GenDB and MetaSAMS.

The software CARMA was developed for predicting the taxonomic origins and functional assignments of short environmental 454 DNA reads using the PFAM database. CARMA produces its results in a text output. Therefore, the framework Provides can be applied, which offers interactive visualization of CARMA results in taxonomic trees, bars and tables. The GenDB genome annotation system can be employed for the structural and functional annotation of long assembled contigs from a metagenome. At first, the Reganor Pipeline for prediction of coding sequences can be employed. Afterwards, the automatic Metanor pipeline can be used to assign gene names, gene products, EC numbers, GO terms and COG functional categories based on BLAST searches against nucleotide and protein databases.

For the analysis of short reads the MetaSAMS system can be employed. MetaSAMS provides insights into the taxonomic as well as the functional composition of a metagenome. The reads are taxonomically classified using CARMA and RDP classifier. For the functional analysis Metanor can be used to compute consistent functional assignments for each read.

Classification and identification of non-coding RNAs using High Throughput Sequencing Data

David Langenberger[1], Steve Hoffmann[1], Clara Bermudez-Santana[1] and Peter F. Stadler[1]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

High throughput sequencing (HTS) of the whole transcriptome for the first time offers the opportunity to sequence small RNA molecules on a large scale. Recently, several strategies have been devised to identify and classify small RNA directly from HT sequences [Fri+08; Hac+09].

After mapping HT sequences to the reference genome using *segemehl* [Hof+09] and merging them to blocks with *blockbuster* [Lan+09] we observe specific block patterns characteristic for different classes of non-coding RNA (ncRNA).

Based on this observation, we have implemented a machine learning algorithm based on the random forest method [Bre01] that accurately classifies three types of ncRNA: microRNAs, tRNAs and snoRNAs. To distinguish the classes we introduced features that reflect the characteristics of block patterns as well as secondary structure. Using this approach, we achieve recall rates and positive predictive values above the 0.9 level. Moreover, we predict several new loci of ncRNA transcription with high probabilities.

References

[Bre01] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[Fri+08] M. R. Friedlander et al. "Discovering microRNAs from deep sequencing data using miRDeep". In: *Nature Biotechnology* 26.4 (2008), pp. 407–415. DOI: 10.1038/nbt1394.

[Hac+09] M. Hackenberg et al. "miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments". In: *Nucleic Acids Research* 37.suppl_2 (2009), W68–76. DOI: 10.1093/nar/gkp347.

[Hof+09] S. Hoffmann et al. *Fast mapping of short sequences with mismatches, insertions and deletions using index structures*. PLoS Computational Biology, in press. 2009.

[Lan+09] D. Langenberger et al. "Evidence for Human microRNA-Offset RNAs in Small RNA Sequencing Data". In: *Bioinformatics* (2009), btp419. DOI: 10.1093/bioinformatics/btp419.

A flowering-time gene network model for association analysis in Arabidopsis thaliana

Karin Klotzbücher[1], Yasushi Kobayashi[2], Nino Shervashidze[2,3], Karsten M. Borgwardt[3] and Detlef Weigel[2]

[1]Center for Bioinformatics Tübingen (ZBIT), Germany

[2]Max Planck Institute for Developmental Biology, Tübingen, Germany

[3]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

In our project we want to determine a set of single nucleotide polymorphisms (SNPs), which have a major effect on the flowering time of *Arabidopsis thaliana*. Instead of performing a genome-wide association study on all SNPs in the genome of *Arabidopsis thaliana*, we examine the subset of SNPs from the flowering-time gene network model. We are interested in how the results of the association study vary when using only the ascertained subset of SNPs from the flowering network model, and when additionally using the information encoded by the structure of the network model. The network model is compiled from the literature by manual analysis and contains genes which have been found to affect the flowering time of *Arabidopsis thaliana* [Far+08; KW07]. The genes in this model are annotated with the SNPs that are located in these genes, or in near proximity to them. In a baseline comparison between the subset of SNPs from the graph and the set of all SNPs, we omit the structural information and calculate the correlation between the individual SNPs and the flowering time phenotype by use of statistical methods. Through this we can determine the subset of SNPs with the highest correlation to the flowering time. In order to further refine this subset, we include the additional information provided by the network structure by conducting a graph-based feature pre-selection. In the further course of this project we want to validate and examine the resulting set of SNPs and their corresponding genes with experimental methods.

References

[Far+08] S. Farrona et al. "The impact of chromatin regulation on the floral transition". In: *Seminars in Cell & Developmental Biology* 19.6 (2008), pp. 560–573.

[KW07] Y. Kobayashi and D. Weigel. “Move on up, it’s time for change – mobile signals controlling photoperiod-dependent flowering”. In: *Genes & Development* 21.19 (2007), pp. 2371–2384. DOI: 10.1101/gad.1589007.

Systematically Screening for Tissue- Specific Alternative Splicing: heart as an example

Pascal Gellert[1], Shizuka Uchida[1] and Thomas Braun[1]

[1]Max Planck Institute Bad Nauheim, Germany

Alternative splicing is a post-translational mechanism to diversify the number of proteins to be produced from pre-mRNA sequences. With the rise of proteomics methods (e.g. 2-D gel and mass spectrometry), alternative splicing is considered to be a reason for the discrepancies between transcriptomics and proteomics data. It is predicted that more than 70% of human genes undergo alternative splicing, and the most recent reports estimated this number to be more than 90%. Due to very frequent splicing events, mutations in alternative splicing machineries result in various diseases.

It is now widely accepted that most of alternative splicing events are controlled in the tissue-specific manner. With the rise of high throughput technologies (e.g. microarray, deep sequencing), it is now able to monitor the alternative splicing events at a particular tissue and physiological condition.

In this study, we applied our algorithm (DGSA = Database-dependent Gene Selection and Analysis) to find tissue-enriched genes from various publicly available data and databases. Then, we developed a user friendly web interface to process Affymetrix's exon array data. Through this web interface, we analyzed exon array data from 11 different tissues and identified tissue-specific alternative splicing events (either exon inclusion or exclusion) by merging to previously derived tissue-enriched genes.

Our in silico results were validated by RT-PCR experiments using cDNA from 15 mouse adult organs. In conclusion, our systematic approach to identify tissue-enriched genes combined with exon arrays is an effective way to screen for tissue-specific alternative splicing events.

Alternative splicing is a post-translational mechanism to diversify the number of proteins to be produced from pre-mRNA sequences. It is predicted that more than 70% of human genes undergo alternative splicing, and the most recent reports estimated this number to be more than 90%. Due to very frequent splicing events, mutations in alternative splicing machineries result in various diseases.

It is now widely accepted that most of alternative splicing events are con-

trolled in the tissue-specific manner. In this study, we applied our algorithm (DGSA = Database-dependent Gene Selection and Analysis) [Uch+09] to find tissue-enriched genes from various publicly available data and databases (microarray, SAGE and EST). Then, we developed a user friendly web interface to process Affymetrixs exon array data. Through this web interface, we analyzed exon array data from 11 different tissues and identified tissue-specific alternative splicing events (either exon inclusion or exclusion) by merging to previously derived tissue-enriched genes. Our *in silico* results were validated by RT-PCR experiments using cDNA from 15 mouse adult organs.

In conclusion, our systematic approach to identify tissue-enriched genes combined with exon arrays is an effective way to screen for tissue-specific alternative splicing events.

References

[Uch+09] S. Uchida et al. "An integrated approach for the systematic identification and characterization of heart-enriched genes with unknown functions". In: *BMC Genomics* 10.1 (2009), p. 100. DOI: 10.1186/1471-2164-10-100.

A Gene Ontology based analysis of high-throughput data via a multivariate approach

Marc Bruckskotten[1], Mario Looso[1], Anne Konzer[1], Franz Cemic[2] and Thomas Braun[1]

[1]Max Planck Institute Bad Nauheim, Germany

[2]Fachhochschule Gießen-Friedberg, Germany

Several tools have been developed to permit exploration and searching of Gene Ontology (GO) databases allowing efficient GO enrichment analysis and visualization of the GO graph. Nevertheless, identification of highly specific GO-terms in complex high-throughput data sets is relatively complicated and the display of GO term assignments and GO enrichment analysis by simple tables or pie charts is not optimal. Valuable information such as the hierarchical position of a single GO term and related terms within the GO tree (topological ordering), or enrichment within a complex set of biological experiments is not displayed. Especially pie charts based on GO tree levels do not present a measurement of hierarchical specificity for the biological system under study.

The new method allows GO analysis of multidimensional experimental settings. We employed principal component analysis (PCA) and developed a new enrichment score, which is defined by taking the relative enrichment of a certain GO term and the specificity (hierarchical position) within the GO graph into account. Our score enables us to identify more specific GO-terms positioned further down the GO-graph. PCA2GO does not depend on specialized experimental settings and allows visualization and detection of multidimensional dependencies both within the acyclic graph (GO tree) and the experimental setting.

RNAz 2.0: improved noncoding RNA detection

Andreas R. Gruber[1], Sven Findeiß[1], Stefan Washietl[2], Ivo L. Hofacker[3] and Peter F. Stadler[1]

[1]University of Leipzig, Germany

[2]European Bioinformatics Institute (EMBL-EBI), United Kingdom

[3]University of Vienna, Austria

With the mounting evidence that noncoding RNAs (ncRNAs) are key players in the regulatory network of cells and the rapidly increasing availability of genomic sequence data the *de novo* prediction of ncRNAs has become of particular interest. While protein gene prediction is a classical problem in computational biology and has been studied for more than 15 years, RNA gene prediction is still in its infancy. Nevertheless, significant progress has been made regarding the prediction of "structured ncRNAs". One of the leading software tools in this field is the program package RNAz [Was+05]. The RNAz algorithm uses the evolutionary conservation and the thermodynamic stability of RNA secondary structures as signals to detect ncRNAs. The wide-spread use of RNAz also helped to identify some of its limitations and to point directions for improvements.

RNAz 2.0 provides significant improvements in several respects: (i) Accuracy is increased by the systematic use of dinucleotide models. (ii) The limitation of the previous version that cannot handle alignments with more than six sequences is overcome by increased training data and the usage of an entropy measure to represent sequence similarities. (iii) The use of structural alignments and a specially trained classification model help to increase the classification power at levels of low sequence conservation.

Evaluation of the predictive power of RNAz 2.0 on human ENCODE regions and a dinucleotide background model [GW08] showed a significant reduction of the estimated false discovery rate compared to the former version (approx. 83% to 55%). RNAz 2.0 is available free of charge and can be obtained at: `http://www.tbi.univie.ac.at/~wash/RNAz/`

References

[GW08] T. Gesell and S. Washietl. "Dinucleotide controlled null models for comparative RNA gene prediction". In: *BMC Bioinformatics* 9 (2008), p. 248. DOI: 10.1186/1471-2105-9-248.

[Was+05] S. Washietl et al. "Fast and reliable prediction of noncoding RNAs". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2454–2459. DOI: 10.1073/pnas.0409169102.

The 'GGG'-bias of Microarray Data — Analysis and Correction

Mario Fasold[1], Peter F. Stadler[2], Stephan Preibisch[3] and Hans Binder[1]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

[2]University of Leipzig, Germany

[3]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Effects due to probe sequence composition represent the major factor prohibiting accurate transformation of probe intensities into expression estimates of microarrays [Bin+03; NM03]. For example, runs of guanines potentially increase the measured intensity by up to two orders of magnitude [Upt+08]. This 'GGG'-bias causes considerable systematic errors of expression estimates and downstream analyses. Presently there is no method available to correct these effects.

We applied the positional-dependent sensitivity model to chip data of different array types. The rank of the model was increased from single-base to next nearest neighbor motifs. We found that non-specific binding is affected by the poly-G bias and that higher order models are required to properly correct sequence effects. We show that a hybrid-rank positional dependent sensitivity model efficiently removes systematic sequence biases from microarray intensity data. The correction applies to different chip-types: gene expression, tiling- and SNP-arrays.

References

[Bin+03] H. Binder et al. "Sequence specific sensitivity of oligonucleotide probes". In: *Proceedings of the German Conference on Bioinformatics*. Vol. 2. Neuherberg and Garching, Germany: belleville Verlag München, 2003, pp. 145–147.

[NM03] F. Naef and M. O. Magnasco. "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays". In: *Physical Review E* 68.1 (2003), p. 011906. DOI: 10.1103/PhysRevE.68.011906.

[Upt+08] G. J. G. Upton et al. "G-spots cause incorrect expression measurement in Affymetrix microarrays". In: *BMC Genomics* 9 (2008), p. 613. DOI: 10.1186/1471-2164-9-613.

The 3'-bias of expression values and RNA degradation — quality control and correction in microarray analysis

Mario Fasold[1] and Hans Binder[1]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

Expression arrays aim at determining the abundance of thousands of mRNA transcripts in parallel. Affymetrix designs their microarrays such that 10-16 probes interrogate a single mRNA with the respective sequences aligning towards the 3' end of the transcript. This specific design allows to analyze a probe positional bias due to RNA degradation that negatively affects microarray measurements and leads to systematic errors in downstream data analysis. In fact, laboratory treatment often leads to differently degraded RNA in a sample-dependent fashion, giving rise to biases of the respective expression profiles and misinterpretation of differential expression [Cop+07].

We present a calibration method that uses probe position and target hybridization modes to quantify the probe positional effect and to correct for the resulting expression bias. The method optionally considers relative or absolute probe positions with consequences for the expression values in a gene-specific fashion. The method allows to identify critical samples in terms of quality control and to calibrate the respective expression measures. We show that our normalization approach improves the inter-sample comparability and illustrate the partial failure of established methods of expression analysis.

References

[Cop+07] V. Copois et al. "Impact of RNA degradation on gene expression profiles: Assessment of different methods to reliably determine RNA quality". In: *Journal of Biotechnology* 127.4 (2007), pp. 549–559. DOI: 10.1016/j.jbiotec.2006.07.032.

Reconstruction, modeling and analysis of Sulfolobus solfataricus metabolism

Thomas Ulas[1] and Dietmar Schomburg[1]

[1]Braunschweig University of Technology, Germany

We reconstructed a model of *Sulfolobus solfataricus*, a bacterium which grows in terrestrial volcanic hot springs with optimum growth occurring at pH 2-3 and a temperature of 75-80℃. *S. solfataricus* is of high interest for environmental and biotechnological applications because of its thermophilic and acidophilic properties. The established network comprised a total of about 300 reactions and about 250 metabolites representing the overall metabolism of this Archaeon, based on the annotated genome and available biochemical information. Experimental data reported in the literature was used to fit the model to phenotypic observations. Using the model in conjunction with constraint-based methods, we simulate the metabolic fluxes induced by different environmental and genetic conditions. The predictions are compared to experimental growth measurements and phenotypes of *S. solfataricus*. *S. solfataricus* produces lipids with ether links between the head group and side chains, making the lipids more resistant to heat and acidity than bacterial and eukaryotic ester-linked lipids was [KM07]. This metabolic network can be used for a better understanding of the metabolite fluxes required for the membrane lipide production in *S. solfataricus*.

References

[KM07] Y. Koga and H. Morii. "Biosynthesis of Ether-Type Polar Lipids in Archaea and Evolutionary Considerations". In: *Microbiology and Molecular Biology Reviews* 71.1 (2007), pp. 97–120. DOI: 10.1128/MMBR.00033-06.

Calcium Oscillations and Jensen's Inequality

Christian Bodenstein[1], Beate Knoke[1], Marko Marhl[2], Matjaz Perc[2] and Stefan Schuster[1]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany

[2]Department of Physics, University of Maribor, Slovenia

The question which physiological advantages calcium (Ca^{2+}) oscillations as compared to an adjustable stationary Ca^{2+} signal may have in non-excitable cells has often been posed [Kno+08]. One of the proposed answers is that oscillations lower the average Ca^{2+} level, which would be advantageous because Ca^{2+} is harmful to the cell in high concentrations. It turns out that Jensen's inequality is useful to check this hypothesis. Jensen's inequality states that for a (strictly) convex function, the function value of the average of a set of argument values is lower than the average of the function values of the arguments from that set. By analytical calculations, we show that due to the special stoichiometric structure of the Ca^{2+} models the kinetics of the Ca^{2+} efflux out of the cell is crucial in this context. If the Ca^{2+} efflux is a convex function of the cytosolic Ca^{2+} level, then oscillations lower the average Ca^{2+} concentration in comparison to the unstable steady state. This is reversed if the efflux is a concave function. Both levels equal one another if that function is linear [Kno+08]. We also analyse the case where the efflux obeys a Hill kinetics, which involves both a convex and a concave part. Numerical simulations of a simple model illustrate our results. As an example for the comparison of the theoretical predictions with experimental data we use Ca^{2+} time series from hepatocytes [SS91]. The obtained results are in good agreement with our theory, however more experimental work is needed.

References

[Kno+08] B. Knoke et al. "Equality of average and steady-state levels in some nonlinear models of biological oscillations". In: *Theory in Biosciences* 127.1 (2008), pp. 1–14. DOI: 10.1007/s12064-007-0018-4.

[SS91] R. Somogyi and J. W. Stucki. "Hormone-induced calcium oscillations in liver cells can be explained by a simple one pool model". In: *Journal of Biological Chemistry* 266.17 (1991), pp. 11068–11077. URL: http://www.jbc.org/cgi/content/abstract/266/17/11068.

Combining the "OMICS": Integrative analysis of proteomics and targeted metabolomics data improves the resolution of classification markers of obesity

Henry Wirth[1], Hans Binder[1], Martin von Bergen[2] and Andreas Oberbach[3]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

[2]Department of Proteomics, Umweltforschungszentrum Leipzig, Germany

[3]Devision of Endocrinolgy, University of Leipzig, Germany

According to the WHO, more than one billion adults are overweight and 300 million are clinically obese. Obese subjects have a decreased life quality and expectancy as well as an increased risk of developing comorbidities such as insulin resistance and type 2 diabetes. Thus, early detection tools for a precise identification of individuals at risk are needed. Since biomarkers for this phenotype are often low-molecular-weight proteins and metabolites secreted into the bloodstream as a result of the disease process, according profiles are topics of current research.

Based on a study aiming on impacts of obesity, a statistical pipeline is introduced and applied to blood serum proteins and metabolites. Feature selection, component analysis and interaction analysis are illustrated and assessed on 'Single-Omic' data. To improve detection of biomarkers and to enable investigation of protein-metabolite interactions, the subsets are integrated and processed in TransOMIC analysis. We show that the simultaneous application of complementary statistical tools is necessary to translate purely descriptive proteomic and metabolomic profiles into a functional context and provide important insights into pathophysiological mechanisms of obesity.

Uphill Unfolding of Native Protein Conformations

Leonidas Kapsokalivas[1], Andreas Albrecht[2] and Kathleen Steinhofel[1]

[1]King's College London, United Kingdom

[2]Queen's University Belfast, United Kingdom

We present results from simulations of unfolding in cubic lattices with two types of simplified energy functions, namely the Miyazawa-Jernigan (MJ) energy function (see [FP06] for details) and the hydrophobic-polar(HP) model [BD96]. The simulations are executed on six benchmark problems for the MJ model proposed by Faisca and Plaxco [FP06] and ten well-known benchmark problems for the HP model devised by Beutler and Dill [BD96]. The unfolding procedure utilizes the pull-move set as a neighbourhood relation and a new population-based search method. For all sixteen benchmark problems we establish the existence of short pathways with monotonically increasing energy functions from ground states to contact-free unfolded states, which includes the three sequences with a high contact order number studied in the MJ model. The number of pull-move transitions (length of unfolding pathways) differs only slightly for the sixteen benchmark problems (all of length 48) and ranges from 27 to 31 for both types of benchmarks. For a given protein sequence S, the search for unfolding paths with monotonically increasing energy values starts with a population of $m = 100$ copies of the same minimum energy conformation $\widetilde{\alpha}_{\mathrm{S}}$. Each individual is treated independently from the remaining elements of the population. The actual transition $\widetilde{\alpha}_{\mathrm{S}} \Rightarrow \widetilde{\alpha}'_{\mathrm{S}}$ to a neighbouring conformation $\widetilde{\alpha}'_{\mathrm{S}}$ is executed if and only if the following conditions are satisfied simultaneously (a) $0 \leq \mathrm{Z}\,(\widetilde{\alpha}'_{\mathrm{S}}) - \mathrm{Z}\,(\widetilde{\alpha}_{\mathrm{S}}) \leq \vartheta$; (b) $0 < \kappa \leq \delta\,(\widetilde{\alpha}_{\mathrm{S}}, \widetilde{\alpha}'_{\mathrm{S}})$, where ϑ is an upper bound for the maximum increase of the energy functions, and $\delta\,(\widetilde{\alpha}_{\mathrm{S}}, \widetilde{\alpha}'_{\mathrm{S}})$ denotes the number of contacts erased during the transition. Here, we have chosen $\vartheta = 2$ and $\kappa = 1$. We observed that without the upper bound ϑ and the lower bound κ the search for unfolding paths tends to be trapped in compact local maxima.

References

[BD96] T. C. Beutler and K. A. Dill. "A fast conformational search strategy for finding low energy structures of model proteins". In: *Protein Science* 5.10 (1996), pp. 2037–2043. DOI: 10.1002/pro.5560051010.

[FP06] P. F. N. Faísca and K. W. Plaxco. "Cooperativity and the origins of rapid, single-exponential kinetics in protein folding". In: *Protein Science* 15.7 (2006), pp. 1608–1618. DOI: 10.1110/ps.062180806.

RepFish: A program for guided repeat-finishing of genome-scale shotgun sequencing projects

Patrick Schwientek[1]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, Germany

In contrast to the ever increasing number of genome sequencing projects, caused by continuous advances in high-throughput technologies, finishing of these projects becomes decreasingly common [Eic+04]. One of the main reasons for this trend is the time consuming manual correction of repetitive elements which most of todays assembly programs can not resolve correctly. Despite these limitations, repetitive elements can be efficiently identified and resolved post assembly by a four step process:

1. Identification of repetitive elements by read-coverage.
2. Splitting of the identified elements into a number of copies according to the ratio of observed read-coverage in the repetitive region to the average read-coverage of the genome.
3. Searching for suitable joining positions within the complete genome sequence by using a highly optimized search algorithm [Ras+06].
4. Joining of each split sequence with a retrieved joining position in an iterative manner.

Using this algorithm, the quality of the assembly can at least be greatly improved if not completely finished at all. RepFish is being developed to incorporate these steps in a single, purely java-based application. Its graphical user interface offers a convenient way for visualization and handling of all necessary actions and, if desired, can also automatically execute the full finishing procedure.

References

[Eic+04] E. E. Eichler et al. “An assessment of the sequence gaps: Unfinished business in a finished human genome”. In: *Nature Reviews Genetics* 5 (2004), pp. 345–354. DOI: 10.1038/nrg1322.

[Ras+06] K. R. Rasmussen et al. “Efficient q-Gram Filters for Finding All ϵ-Matches over a Given Length”. In: *Journal of Computational Biology* 13.2 (2006), pp. 296–308. DOI: 10.1089/cmb.2006.13.296.

Modelling of ADP-induced Platelet Activation

Heinrich Brinck[1], Frank Domurath[1] and Verena Tischler[1]

[1]University of Applied Sciences of Gelsenkirchen, Germany

Platelets are best suited for computational simulations because they lack a nucleus. ADP binds to the GPCR P2Y1, which results in an activation of PLC-b. This enzyme cleaves PIP3 into DAG and IP3, which induces the efflux of calcium. In the cytoplasm it activates PKC in cooperation with DAG, which finally results in an increased affinity of membrane protein IIb/IIIa for fibrin. However PKC also inhibits PLC-b and thus limits the Ca^{2+} signal. These cell-signalling pathways were modelled deterministically using SBML and the SBToolbox2 for MATLAB [Pur+08]. We tested the calcium response to different ADP concentrations and showed that ADP provokes calcium bursts only from 100nM-10μM. We tried to excite a second calcium signal in vain. Analysing the signal cascade we concluded that P2Y1 retains all ADP independently from the surrounding ADP concentrations. A second hint, that the initial value for P2Y1 was too high, was the fact that unlike IP3 receptor, a 100-fold increase in the number of P2Y1 influences the calcium signal negatively. Lewandrowski et al. [Lew+09] used mass spectrometric data of platelets to calculate the exponentially modified Protein Abundance Index (emPAI). It is directly proportional to the amount of proteins. We deduced a regression line from emPAI data and measured protein amounts and compared simulation results with the estimated absolute protein contents. This work demonstrates poor agreement and further research is needed to identify the reasons. Potentially emPAI is not well-suited for protein abundance estimation. Other reasons may be that the model includes only a small part of the physiological relevant reactions and that the initial values were constructed to fit calcium efflux and do not resemble measured data.

References

[Lew+09] U. Lewandrowski et al. “Platelet membrane proteomics: a novel repository for functional research”. In: *Blood* 114.1 (2009), e10–19. DOI: 10.1182/blood-2009-02-203828.

[Pur+08] J. E. Purvis et al. “A molecular signaling model of platelet phosphoinositide and calcium regulation during homeostasis and P2Y1 activation”. In: *Blood* 112.10 (2008), pp. 4069–4079. DOI: 10.1182/blood-2008-05-157883.

Harnessing the phylogenetic signal in mobile elements to understand the evolution of fin-footed marine mammals

Stefanie Grosser[1], Christiane Schroeder[1], Christoph Bleidorn[1], Detlef Groth[1], Ralph Tiedemann[1], Joachim Selbig[1] and Stefanie Hartmann[1]

[1]Institute for Biochemistry and Biology, University of Potsdam, Germany

Traditional molecular phylogenetic analyses of single or low copy-number genes have not been able to resolve the relationships of pinnipeds, which are fin-footed, semi-aquatic marine mammals that include seals, walruses, and sea lions. New approaches are therefore needed to resolve their evolutionary relationships. Retrotransposons are mobile elements that have the ability to integrate into the genome at a new site within their cell of origin. These elements have successfully been used as phylogenetic markers, but their identification in genomes of non-model organisms is generally very time-consuming. To address this problem, we have established an approach for the large-scale identification of retrotransposons that can be used as alternative phylogenetic markers for pinnipeds and their terrestrial relatives.

Our pipeline first identifies conserved exons between the two major lineages of carnivores, using as representatives the genomes of domestic dog (*Canis familiaris*) and cat (*Felis catus*). Introns that contain a retrotransposon in the dog but not in cat are then selected. This step takes advantage of the recently developed BlastScanners to parse the large amounts of generated BLAST output. Exons that flank introns containing the dog retrotransposons are subsequently aligned in an automated fashion, and the software Primer3 is used to design PCR primers that are exon-priming and intron-spanning. These primers can then be used in the lab for amplification and sequencing of the corresponding regions of other taxa.

We were recently able to successfully use sequences of a small number of short interspersed elements (SINEs) in introns as phylogenetic markers. The pipeline introduced here will allow to greatly expand this analysis and to obtain a robust phylogeny of pinnipeds and their terrestrial relatives.

Two novel tools for promoter analysis

Vladimir Shelest[1], Eugen Fazius[1], Daniela Albrecht[1] and Ekaterina Shelest[1]

[1]Hans Knoell Institute (HKI), Jena, Germany

We suggest two novel promoter analysis tools developed in our department. SiteTracker is based on a novel approach to the prediction of transcription factor binding sites (TFBS). The approach is alternative to widely used PWMs and HMMs. The query sequence is scanned for motifs with a sliding window; each motif is assigned a weight that reflects its similarity to the motifs of the training set and simultaneously the non-randomness of its occurrence. Important is, that the motifs of the training set are not aligned. Thus, we avoid the uncertainties introduced by the procedure of alignment, which is especially problematic for very degenerate motifs. Instead, the calculated weight takes into account the number and "quality" of the matches between the training set and the query. The method is compared with the MatchTM (utilizing TRANSFAC matrices). We demonstrate the better performance of our tool.

DistanceScan utilizes the previously developed method of distance distributions of TFBS pairs. It is based on modeling the distribution of distances between the constituents of TFBS pairs in random sequences. We describe the randomness of the occurrence of TFBS pairs on certain distances, comparing the random distance distribution with the distributions in query sequences. This approach is used as a filtering technique for the selection of TFBS pairs for promoter modeling. We demonstrate the applicability of the tool to the construction of promoter models on the example of iron-dependent pathways in fungi.

References

[JL02] U. S. Jung and D. E. Levin. "Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway". In: *Molecular Microbiology* 34.5 (2002), pp. 1049–1057. DOI: 10.1046/j.1365-2958.1999.01667.

[She06] E. Shelest. "Genetic networks of antibacterial responses of eukaryotic cells. Bioinformatics analysis and modeling". PhD thesis. Technical University of Braunschweig, 2006. URL: http://www.digibib.tu-bs.de/?docid=00000045.

Workflow-based Integration of Metabolite Identification

Michael Gerlich[1] and Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry (IPB), Halle, Germany

Mass spectrometry (MS) is a valuable analysis method in metabolomics. Today, it's major drawback is the tedious (and mostly manual) process of metabolite identification. Tandem-MS spectra contain evidence on a metabolite's structure. We developed workflows and a browser-based application to provide scientists with a simple user interface to metabolite identification.

Our simple workflows are targeted at identification of metabolites using the spectral library MassBank [Hor+08] or analysis by prediction algorithms [Wol09]. Unfortunately, spectral libraries only cover a small number of metabolites, whereas the results of prediction algorithms are difficult to assess without further evidence.

Therefore the integration workflow creates an "alignment" of hypotheses between MassBank and MetFrag for a given spectrum, showing the unique and common hits and their scores and ranks. Consistently, high scores confirm individual results, and good results for a related compound in the library confirm predictions of metabolites not included in MassBank. Together, the two identification approaches provide a comprehensive view and help to find the correct metabolite identification by its spectrum.

References

[Hor+08] H. Horai et al. "Comparison of ESI-MS Spectra in MassBank Database". In: *International Conference on BioMedical Engineering and Informatics*. Vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, 2008, pp. 853–857.

[Wol09] S. Wolf. *MetFrag*. 2009. URL: `http://msbi.ipb-halle.de/MetFrag/`.

Evolution across Scales: enhancing evolutionary thinking and knowledge in the biosciences

Nadine Bernhardt[1], Stefanie Grosser[1], Florian Hollandt[1], Kai Kruse[1] and Henriette Seichter[1]

[1]Institute for Biochemistry and Biology, University of Potsdam, Germany

An interdisciplinary initiative "Evolution across Scales" was established at the University of Potsdam in 2008. The overall goal of this initiative is to enhance evolutionary thinking and knowledge in biosciences as well as in other natural sciences. The initiative integrates (a) genomics, metabolomics, and bioinformatics to allow the genome-wide analysis of molecular events in model organisms, and (b) geosciences and climate impact research, which provide a superior opportunity to study the interaction between the biosphere and the geosphere.

Bioinformatics, which occupies a position at the intersection of the biosciences, computer sciences, mathematics, and statistics, plays a central role in transforming raw data into biological knowledge. In addition, bioinformatics approaches are often comparative in nature and thus contain a strong evolutionary component. Research that is carried out by Masters students enrolled in the "Evolution across Scales" program therefore frequently relies on using bioinformatics approaches. Projects of these students include the comparative analysis of metabolic networks, the large-scale identification of markers in non-model systems that can be used for phylogenetic studies, a mathematical model of endocrine disruption in humans, analysis of genetic variation in the context of conservation genetics efforts, and the modeling of dormancy in algae.

One of the specific projects that will be presented is the identification of elements of the central circadian oscillator in the green algae *Chlamydomonas reinhardtii*, and the prediction of targets for the C3 subunit of the clock protein CHLAMY1, which is a homologue of the mammalian protein CUGBP1. Relationships between these homologues can be used to establish a basic mathematical model of the circadian oscillations in *C. reinhardtii*.

Transposons Shape the Architecture of Plant Genomes

Heidrun Gundlach[1], Georg Haberer[1], Manuel Spannagl[1] and Klaus F. X. Mayer[1]

[1]Institute for Bioinformatics and System Biology, Helmholtz Center Munich, German Research Center for Environmental Health

Transposable elements constitute a major part of plant genomic DNA and are often responsible for drastic increases in genome size. Their overall amount ranges from 15 to over 90 percent. The large genome size variations even between related taxonomic groups pose many intriguing questions regarding the concerted evolution of transposons and host plant. Within the family of grasses diploid genome sizes vary between 0.3 Gb up to ∼8 Gb. All grasses have evolved within the last 60 mio years from a common ancestor and have retained large stretches of syntenic regions, with still conserved gene order. However the intergenic space has evolved much more rapidly under the influence of different transposon families. We have established workflows and analyses protocols for an exhaustive structural annotation not only of the genic, but also of the repetitive space of plant genomes. An in-depth repeat annotation has to deal with element identification, defragmentation and the reconstruction of insertion events. Our repeat annotation concept is based on mips-REcat, a generic repeat element classification catalog, and mips-REdat, an exhaustive database of plant repeat elements. The detection layer of our ANGELA pipeline (Automated Nested Genetic ELement Analysis) combines intrinsic repeat detection approaches with homology based methods. The processing layer integrates genes or other additional data. It handles element overlaps, followed by the identification of nested structures and the timing of LTR retrotransposon insertion events. Implemented standard evaluations cover composition tables, copy numbers, target-insert pair preferences or insertion age distributions. Heat maps are used for the visualisation of chromosome organization. The composition and differential distribution of transposon and gene space will be shown for different plant species.

References

[Pat+09] A. H. Paterson et al. “The Sorghum bicolor genome and the diversification of grasses”. In: *Nature* 457.7229 (2009), pp. 551–556. DOI: 10.1038/nature07723.

Phylogenetic networks from multilabeled trees

Thomas Bonfert[1], Regula Rupp[1] and Daniel Huson[1]

[1]University of Tübingen, Germany

The evolutionary history of a set of species is usually described using rooted phylogenetic trees. But such trees can't be used to display reticulate events, such as horizontal gene transfer, hybridization, recombination or reassortment. There is a need for methods which can construct phylogenetic networks from different types of data.

We present three methods to construct phylogenetic networks from multilabeled trees (MUL-trees). Such trees can be obtained when the evolution of polyploids is analysed (e.g. [For+08]). We have implemented these methods in Dendroscope [Hus+07]. Thus, a future version of this program will be able to construct and visualise networks from MUL-trees.

References

[For+08] P. M. Fortune et al. "Molecular phylogeny and reticulate origins of the polyploid Bromus species from section Genea (Poaceae)". In: *American Journal of Botany* 95.4 (2008), pp. 454–464. DOI: 10.3732/ajb.95.4.454.

[Hus+07] D. Huson et al. "Dendroscope: An interactive viewer for large phylogenetic trees". In: *BMC Bioinformatics* 1 (2007), p. 460. DOI: 10.1186/1471-2105-8-460.

Automatic annotation of metabolomic ESI-LC/MS Data with CAMERA

Carsten Kuhl[1] and Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry (IPB), Halle, Germany

Metabolomics deals with the characterisation and identification of metabolites in biological samples. For profiling experiments LC/MS serves as state of the art method, providing a high-throughput analysis. But the following identification procedure is a time consuming step, with much manual work from an experimentalist. The typical ionsiation method ESI provides highly redundant signals for a single compound, like adduct ions (e.g. $[M + Na]^+$), fragment ions ("insource" CID of the precursor ion) and cluster ions $[2M + H]^+$). These increase the time for identification and complicate database queries. As first step of an automatic procedure for MS data, we need to assign which peaks originate from one substance and what kind of ion the peaks are. This process is called annotation. We developed the recently released Bioconductor-Package CAMERA [Kuh+09], as the successor of ESI [Tau+07]. It contains methods based on retention time and peak correlation comparison for the grouping and a ruleset-based algorithm for the annotation. In comparison to ESI, we developed new procedures, which reduce the number of false positive annotations. On a sample with known substances, we show that CAMERA is able to separate most of the substances, calculate the correct substance mass and finally annotate many adduct and isotope ions.

References

[Kuh+09] C. Kuhl et al. *CAMERA: Collection of annotation related methods for mass spectrometry data*. 2009. URL: `http://bioconductor.org/packages/bioc/html/CAMERA.html`.

[Tau+07] R. Tautenhahn et al. "Bioinformatics Research and Development". In: vol. 4414. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007. Chap. Annotation of LC/ESI-MS Mass Signals, pp. 371–380. ISBN: 978-3-540-71232-9. DOI: `10.1007/978-3-540-71233-6_29`.

Unraveling the Barley Genome from Short Sequence Reads

Mihaela Martis[1], Heidrun Gundlach[1], Klaus F. X. Mayer[1], Stefan Taudien[2], Hana Simkova[3], Pavla Suchankova[3], Jaroslav Dolezel[3], Nils Stein[4], Uwe Scholz[4], Burkhard Steuernagel[4], Andreas Graner[4], Thomas Wicker[5], Andreas Petzold[2], Marius Felder[2] and Matthias Platzer[2]

[1]Institute for Bioinformatics and System Biology, Helmholtz Center Munich, German Research Center for Environmental Health

[2]Friedrich-Loeffler-Institute Jena, Germany

[3]Laboratory of Molecular Cytogenetics and Cytometry, Institute of Experimental Botany, Czech Republic

[4]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[5]Institute of Plant Biology, University of Zuerich, Switzerland

Barley is the number four cereal crop in the world. It is a major resource for animal feed and for the brewing and distilling industry. Whole genome sequencing of barley is complicated by the huge genome size of ∼5.1 Gbp and the inherent genome complexity caused by a content of 80 – 90% repetitive elements. In the present study, we combined chromosome sorting and next generation sequencing (NGS) to gain insight at unprecedented density into the gene content of an entire Triticeae chromosome. Integration of marker data with high resolution synteny data from grass model genome sequences of rice, sorghum and brachypodium allowed us to propose a virtually ordered gene inventory of anchored genes from all seven barley chromosomes. The syntenic integration provided specifically added value for regions with limited genetic resolution of barley chromosomes, i.e. centromeric and sub-centromeric regions. Thus, integration of genomic, transcriptomic and synteny-derived information represents a major step towards developing reference sequences of chromosomes and complete genomes of the most important plant tribe for mankind with genome sizes exceeding mammalian genomes by far.

Protein phosphorylation patterns affected by nuclear DNA polymorphisms in a genome-wide scale in Arabidopsis

Sabrina Kleessen[1], Diego Mauricio Riaño-Pachón[1], Pawel Durek[1], Jost Neigenfind[1], Wolfgang Engelsberger[1], Dirk Walther[1], Joachim Selbig[1], Waltraud Schulze[1] and Birgit Kersten[1]

[1]Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany

Protein phosphorylation is an important post-translational modification which influences virtually every aspect of dynamic cellular behavior by controlling, e.g., cellular signaling. Site-specific phosphorylation of the amino acid residues serine (S), threonine (T) and tyrosine (Y) can have profound effects on protein structure, activity, stability and the interaction with other biomolecules. Hence, the systematic analysis of protein phosphorylation by phosphoproteomic and bioinformatic approaches is of great importance in understanding cellular functions [Ker+09]. In Arabidopsis thaliana, recent progress in the identification of phosphosites has prompted the creation of dedicated web-resources (e.g., PhosPhAt, `http://phosphat.mpimp-golm.mpg.de/`) and has made it possible to develop Arabidopsis-specific predictors of phosphosites [Hea+08].

Based on sets of experimentally identified phosphosites and predicted pS-sites (from PhosPhAT) as well as newly predicted pT- and pY-sites we computed (i) over- and under-represented GO terms in the proteins harboring phosphosites, (ii) over- and under-represented domains co-occurring with phosphosites and (iii) hot spots of phosphorylation. Moreover, we analysed how the phosphorylation patterns are influenced by non-synonymous single nucleotide polymorphisms (nsSNPs). For that purpose we made use of the polymorphisms between several inbreed Arabidopsis accessions, that have been extensively studied by re-sequencing arrays [Cla+07] and by ultra-deep sequencing technologies such as Illumina/Solexa [Oss+08]. Loss and gain of phosphorylation sites by nsSNPs were identified and will be publicly available via the GABI Primary Database (GabiPD; `http://www.gabipd.org/`) [RP+09].

This work is supported by the BMBF (GABI-Future grant 0315046).

References

[Cla+07] R. M. Clark et al. “Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana”. In: *Science* 317.5836 (2007), pp. 338–342. DOI: 10.1126/science.1138632.

[Hea+08] J. L. Heazlewood et al. “PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor”. In: *Nucleic Acids Research* 36.suppl_1 (2008), pp. D1015–1021. DOI: 10.1093/nar/gkm812.

[Ker+09] B. Kersten et al. “Plant phosphoproteomics: An update”. In: *Proteomics* 9.4 (2009), pp. 964–988. DOI: 10.1002/pmic.200800548.

[Oss+08] S. Ossowski et al. “Sequencing of natural strains of Arabidopsis thaliana with short reads”. In: *Genome Research* 18.12 (2008), pp. 2024–2033. DOI: 10.1101/gr.080200.108.

[RP+09] D. M. Riano-Pachon et al. “GabiPD: the GABI primary database–a plant integrative ’omics’ database”. In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D954–959. DOI: 10.1093/nar/gkn611.

MetFrag — Match Predicted Fragments with Mass Spectra

Sebastian Wolf[1] and Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry (IPB), Halle, Germany

Tandem- and multistage mass spectrometry has become a valuable tool for identification and elucidation of small molecules in metabolomics. Unfortunately, libraries of reference spectra have a rather small coverage, and manual interpretation is a time consuming, non-trivial task.

We are developing MetFrag — an open-source combinatorial fragmenter for identifying small organic compounds. The application takes tandem-MS peak lists as input, and searches generic compound libraries such as PubChem (`http://pubchem.ncbi.nlm.nih.gov/`) or KEGG (`http://www.genome.jp/kegg/`) for molecules which might explain the measured spectrum, applying in-silico fragmentation to each candidate.

We compare MetFrag against the results from [Hil+08], who used MassFrontier 4, a commercial fragmentation software. [Hil+08] used a PubChem snapshot from February 2006 with an average of 272 candidate structures where the correct compound was ranked in the top 44 hits. An exact mass lookup in a current PubChem snapshot returns 1120 compounds and MetFrag ranks the correct solution among the first 87 of the hits.

The software is available through its web interface

`http://msbi.ipb-halle.de/MetFrag/`,

as a BioMoby web service and Java command line tool. The web interface features a ranked tabular display of candidate compounds and shows matched fragments for each entry.

References

[Hil+08] D. W. Hill et al. “Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra”. In: *Analytical Chemistry* 80.14 (2008), pp. 5574–5582. DOI: 10.1021/ac800548g.

A Data Analysis Workflow for the Identification of Expression Signatures Functioning as Molecular Biodosimeters

Sonja Boldt[1], Katja Knops[2], Ralf Kriehuber[2] and Olaf Wolkenhauer[1]

[1]Chair in Systems Biology and Bioinformatics, University of Rostock, Germany

[2]Radiation Biology Unit, Department of Safety and Radiation Protection, Forschungszentrum Jülich, Germany

Purpose. Radiation biodosimetry aims at predicting exposure doses after ionizing radiation in a fast and highly accurate manner. Current experimental techniques for dose determination are time-consuming, which is why we here utilize high-throughput gene expression data to overcome limitations of existing approaches.

Materials and Methods. Human peripheral blood from 6 healthy donors was irradiated *ex vivo* with 0.5 Gy, 1 Gy, 2 Gy and 4 Gy (?-ray; Cs-137). Afterwards the gene expression of the isolated peripheral blood lymphocytes was measured at 6 h, 24 h and 48 h after radiation exposure with whole human genome DNA arrays from Agilent [Kno+09]. Based on these data, we established a partially automated data analysis workflow, with which the exposure dose can be classified.

Results. Gene expression signatures, consisting of significantly up- and down-regulated radiation responsive genes, were used to build k-nearest neighbour classifier. Considering each time point separately, the classifier correctly predicted the radiation dose of all samples. After pooling the data from the given time points, the same procedure resulted in a prediction accuracy of 82%. An information-gain-based method allowed us to reduce the original sets of characteristic signatures while maintaining high prediction accuracy.

Conclusion. We identified small sets of expression signatures that allow an accurate classification of blood samples for different radiation doses. This is a first step towards a fast and reliable tool to support medical decision making after accidental radiation exposure.

References

[Kno+09] K. Knops et al. “Gene expression pattern analyis as a tool for radiation biodosimetry”. In: *37th Annual Meeting of the European Radiation Research Society, 26–29th August*. Prague, Czech Republic 2009.

GABI GAIN: Data Integration in Genomics-based Plant Breeding

Robert Wagner[1], Pawel Durek[1] and Birgit Kersten[1]

[1]Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany

Modern plant breeding programs produce a huge amount of genomic and phenotypic data. Efficient data mining is therefore crucial regarding exploiting the generated information and optimising the design process of the experiments. In the scope of the GABI GAIN project tools are being developed for the access, visualisation, standardisation and mining of data produced in genomics-based plant breeding. A data schema has been developed in order to store integrated breeding data in an efficient and flexible way. The model has been designed platform independently using the Unified Modeling Language and can be realised in all common database management systems. An exchange format, Gain-Tab, has been defined to ease the integration of new data into the database as well as into other analysis tools. Gain-Tab is a table-based format, so that Spreadsheet tools like Excel or OpenOffice can be utilised to easily create files for data exchange. Finally, a graphical user interface has been developed that supports performing CRUD operations (create, read, update, delete) on the data and provides the possibility of exporting into different formats like Excel or CSV. Different views on the data exist in terms of tables or plots; e.g., a pedigree view visualises the inheritance relationships of a genotype in a tree-like manner together with the mean values of different traits and years of the individual generations. Also an interface to the PLABSTAT program allows for further analysing the integrated data. The tool-package is provided as a locally installable application for partners within the GABI GAIN project to optimise genomics-based plant breeding.

This work is supported by the German Federal Ministry of Education and Research BMBF (GABI-FUTURE grant 0315072C).

Characterisation of non-coding RNAs and RNA-RNA interactions in S. coelicolor

Alexander Herbig[1] and Kay Nieselt[1]

[1]Center for Bioinformatics Tübingen, Department of Information and Cognitive Sciences, University of Tübingen, Germany

Several studies about non-coding RNAs (ncRNAs) have shown that they are involved in a wide spectrum of different processes, far beyond those of tRNAs or rRNAs. Nevertheless, the functions of most ncRNA transcripts are still unknown. A number of tools for the prediction of ncRNA regions in prokaryotes and also in eukaryotes has been published. *RNAz* [Was+05], for example, is a program for the detection of such loci. As most such programs, however, *RNAz* produces an unknown number of false positives. Furthermore, the locus prediction does not include information about the functional RNAs that might be contained in these regions. In addition to the comparison with ncRNA families that are already known, we present complementary approaches to refine the annotation of predicted ncRNA regions. In particular we try to annotate ncRNA loci with respect to features related to the transcription process, such as promoter regions and transcription terminator signals. This enables us to specify the coordinates and the strand of putative ncRNA transcripts. In addition, we utilize *IntaRNA* [Bus+08] to predict interactions between predicted ncRNA transcripts and mRNAs. These methods are applied to the genome of the antibiotics producing soil bacterium *S. coelicolor*. First results suggest that several key proteins of *S. coelicolor*, which are involved in regulatory or catalytic functionalities related to important metabolic processes, are regulated by ncRNA transcripts. These results indicate that ncRNAs may act as regulators in gene regulatory networks. This is supported by genome wide high resolution timeseries expression data.

References

[Bus+08] A. Busch et al. "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions". In: *Bioinformatics* 24.24 (2008), pp. 2849–2856. DOI: 10.1093/bioinformatics/btn544.

[Was+05] S. Washietl et al. “Fast and reliable prediction of noncoding RNAs”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2454–2459. DOI: 10.1073/pnas.0409169102.

Control System-Based Reverse Engineering of Circadian Oscillators

Benedict Schau[1], Thomas Hinze[1], Thorsten Lenser[1], Ines Heiland[1] and Stefan Schuster[1]

[1]Friedrich-Schiller-University Jena, Germany

Control systems are a concept from engineering to achieve a desired dynamical behaviour like adjusting temperature. Later, they came into the scope of life sciences as part of a cybernetic approach to understand biological systems. The control system-based description of the circadian clock found in *New Zealand Weta* can be seen as pioneering example [CL83]. Control systems benefit from a strict modularisation that allows a clear decomposition of a complex system into functional units interconnected by signalling channels. Signal processing is commonly represented by block diagrams that map input or memorised signals into output signals. Its correspondence to modular functional units and dedicated reaction network motifs was shown in [Wol+05]. Within an ongoing study, we combine the specification of block diagrams with the ability to an *artificial evolution* of reaction network candidates exhibiting a desired input/output interdependency. Here, dynamical behaviour analysis enables selection of the fittest candidates. In this way, each component of a control system (e.g. controller, actuator, plant, sensor) can be independently reconstructed by providing numerous, topologically different network candidates. Finally, the arrangement of these candidates leads to valid models of the entire system. By means of this modular network evolution, the search space is significantly reduced while keeping a high probability of heuristical success. With the SBMLevolver [Len+07], a suitable software to this task is available. We obtained building blocks with non-linear transmission behaviour to be composed towards distinct circadian control systems at various levels of description.

References

[CL83] N. D. Christensen and R. D. Lewis. "The circadian locomotor rhythm of hemideina thoracica (Orthoptera; Stenopelmatidae): A population of weakly coupled feedback oscillators as a model of the underlying pacemaker". In: *Biological Cybernetics* 47.3 (1983), pp. 165–172. DOI: 10.1007/BF00337006.

[Len+07] T. Lenser et al. “Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics”. In: vol. 4447. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007. Chap. Towards Evolutionary Network Reconstruction Tools for Systems Biology, pp. 132–142. ISBN: 978-3-540-71782-9. DOI: 10.1007/978-3-540-71783-6_13.

[Wol+05] O. Wolkenhauer et al. “A systems- and signal-oriented approach to intracellular dynamics”. In: *Biochemical Society Transactions* 33 (2005), pp. 507–515. URL: http://www.biochemsoctrans.org/bst/033/bst0330507.htm.

Assemblies of Uncommon DNA Patterns at the Transcription Start Sites of Genes in Human and Mouse Genomes

Miklos Cserzo[1], Gabor Turu[1], Peter Varnai[1] and Laszlo Hunyady[1]

[1]Semmelweis University, Hungary

We identified subsets of DNA motifs in Human and Mouse genomes based on their statistical properties. These motifs are less frequent than one would expect by random chance and they occur in clusters in close proximity of distinguished genetic positions. The length of these segment types is well defined – between 18 and 26 bases in length. The actual list of motifs is 75% common between Human and Mouse while random motif lists expected to show only ∼0.6% identity. In this way this feature is linked to the transcription initiation point of genes and it is highly conserved between Human and Mouse. The clusters significantly overlap with experimentally determined Polymerase II binding sites in the Human genome and small ncRNA hits in the ENCODE region.

These findings suggest that this type of segments is serving a particular biological function therefore their number of occurrence is limited and restricted to certain regions of DNA. Currently this functionality is not known however, very likely linked to gene expression.

Exploiting the Characteristics of Metabolomics Data

Diana Boronczyk[1], Yvonne Pöschl[2], Ivo Grosse[2] and Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry (IPB), Halle, Germany

[2]Martin-Luther-University Halle-Wittenberg, Germany

Mass spectrometry (MS) is an analytical method to measure metabolites. A major application is the detection of differentially abundant metabolites. The statistical questions posed are similar to transcriptomics, but MS data has their own characteristics and quirks. To date there are no satisfying methods to discover metabolites produced differentially under various conditions. One fundamental shortcoming of existing methods is that they treat *all* peaks to be statistically independent, which they are not. We investigate how commonly used statistics can adapt to or even exploit these characteristics. Several methods are based on variants of the classical t-test. [ORS07] shrink the variances of each gene towards an overall variance, still treating signals statistically independent. Another approach [Smy+05] uses replicated spots of each gene on the microarray for a more stable variance estimation. We incorporate the dependencies between signals from the same metabolite using them as replicates to improve the prediction of differential signals. In addition, we verify the annotation of signals originating from a single metabolite using Hypergeometric Testing, in analogy to Gene Set enrichment analysis. These tests unveil, for each metabolite, whether related features are also enriched in (non-)differential features. We evaluate the methods on a dataset with a series of different concentrations of *Arabidopsis Thaliana* leaf extracts, and show that the use of related signals as replicates in modified Students-tests improves identification of differential signals.

References

[ORS07] R. Opgen-Rhein and K. Strimmer. "Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007), p. 9. DOI: 10.2202/1544-6115.1252.

[Smy+05] G. K. Smyth et al. "Use of within-array replicate spots for assessing differential expression in microarray experiments". In: *Bioinformatics* 21.9 (2005), pp. 2067–2075. DOI: 10.1093/bioinformatics/bti270.

Interaction prediction and classification of PDZ domains

Sibel Kalyoncu[1], Ozlem Keskin[1] and Atilla Gursoy[1]

[1]Chemical and Biological Engineering, Koc University, Istanbul, Turkey

Protein interaction domains play key roles in cellular signalling and protein localization. Clarification of binding specificities of these domains is very crucial in order to understand the complexity of most cellular pathways. PDZ domains, one of the largest families of protein interaction domains, bind the C termini of their specific binding partners. PDZ domains generally fall into two classes according to the C termini of target proteins; Class I PDZ domains bind to the C terminal motif with the sequence [S/T-X-Φ-COOH] and Class II PDZ domains are specific for [Φ-X-Φ-COOH] target sequence. Although PDZ domains are very specific, they are also promiscuous, binding to more than one protein, so it is a hard problem to have insight about their binding specificity. In this study, PDZ domains were classified and their binding specificity was predicted by using only their primary sequences. Our model is based on Support Vector Machine (SVM) analyses of triplet frequencies of the primary sequences of PDZ domains and their interaction peptides. Triplet frequencies were obtained by frequency calculation of overlapping triplet sequences whose amino acids were clustered into seven classes according to their dipoles and volumes of the side chains [She+07]. The model was trained by using interaction data of 85 PDZ domains and 217 peptides encoded in the mouse genome [Sti+07]. The interaction prediction score of our method (AUC=0.98) is better than other sequence-based studies and the classification ability of our method (AUC=0.86) is supported by motif extraction. The model also predicts the effects of point mutations in peptide ligands on the binding selectivity of PDZ domains. These results indicate that our approach can be a powerful model to predict binding selectivity of PDZ domains.

References

[She+07] J. Shen et al. "Predicting protein-protein interactions based only on sequences information". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.11 (2007), pp. 4337–4341. DOI: 10.1073/pnas.0607879104.

[Sti+07] M. A. Stiffler et al. "PDZ Domain Binding Selectivity Is Optimized Across the Mouse Proteome". In: *Science* 317.5836 (2007), pp. 364–369. DOI: 10.1126/science.1144592.

A pipeline to identify SNPs causing variation in the splicing pattern

Kirsten Faber[1], Karl-Heinz Glatting[1], Angela Risch[1] and Agnes Hotz-Wagenblatt[1]

[1]German Cancer Research Center, Heidelberg, Germany

Single nucleotide polymorphisms (SNPs) become more and more important in disease research. The aim of most existing SNP tools is the annotation of SNPs mainly by analyzing variances in the coding region of the affected gene and possible effects on the transcriptome and proteome levels [Che+09]. In contrast, our AASsites pipeline (Automatic Analysis of SNP sites) indicates if a change in the splice pattern due to a SNP is likely to occur. Implemented in the W3H-pipeline system [Ern+03], AASsites uses a combination of different gene prediction methods (GenScan, GeneID, HMMgene, GlimmerHMM and GrailExp), and an ESE (Exonic Splice Enhancer) detection to predict changes in the elements which are relevant for splicing. Additionally, Genewise analysis determines changes in the ORF (open reading frame). The rule based classification system ranks the SNPs and their effects into different categories. Modifications in the exon-intron structure and variances in ESE and ORF are shown in the final result of the pipeline including new, disappeared and modified exons and introns. Individual SNPs in the same sequence are analyzed independently.

We extracted a set of 60 SNPs from the literature that were described as changing the splicing behavior and used them as a benchmark set. AASsites was able to predict 80 percent of these changes correctly.

This pipeline will be used for finding candidate human genes whose splicing pattern might be affected by SNPs.

References

[Che+09] C. Chelala et al. “SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms”. In: *Bioinformatics* 25.5 (2009), pp. 655–661. DOI: 10.1093/bioinformatics/btn653.

[Ern+03] P. Ernst et al. “A task framework for the web interface W2H”. In: *Bioinformatics* 19.2 (2003), pp. 278–282. DOI: 10.1093/bioinformatics/19.2.278.

Structural Modelling of Signal Transduction in Hepatocytes exemplified by the Insulin Network

Jörn Behre[1], Regina Samaga[2], Steffen Klamt[2] and Stefan Schuster[1]

[1]Friedrich-Schiller-University Jena, Germany

[2]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

The concept of Elementary flux modes (EFMs) is important for detecting nondecomposable pathways in metabolic networks. The method is based on the steady state condition meaning that metabolic networks must obey a mass balance condition. In signal transduction networks that condition is of less importance because here the flow of information matters. Nevertheless, it is interesting to implement pathway detection methods also for signalling systems. Here we present a formalism that affords the application of elementary flux modes in the case of enzyme cascades operating, for example, by phosphorylation and dephosphorylation. Our approach [BS09] is based on the ideas that the signal must not attenuate along each cascade and that the system has to return to its initial state after each signalling event. We illustrate our method by several prototypic single-phosphorylation and doublephosphorylation cascades including convergent and divergent branching. Moreover we employ our formalism to a part of the insulin signalling network. Beyond that we present the insulin signalling network also as a Boolean model for the program CellNetAnalyzer [Kla+06]. In this form it will be merged with a Boolean model of the EGF signalling network.

References

[BS09] J. Behre and S. Schuster. "Modeling Signal Transduction in Enzyme Cascades with the Concept of Elementary Flux Modes". In: *Journal of Computational Biology* 16.6 (2009), pp. 829–844. DOI: 10.1089/cmb.2008.0177.

[Kla+06] S. Klamt et al. "A methodology for the structural and functional analysis of signaling and regulatory networks". In: *BMC Bioinformatics* 7.1 (2006), p. 56. DOI: 10.1186/1471-2105-7-56.

Exploring the Genomic Diversity in Rye Using a SNP Detection Pipeline

Thomas Schmutzer[1], Uwe Scholz[1], Nils Stein[1], Chris-Carolin Schön[2], Grit Haseneyer[2], Eva Bauer[2], Klaus F. X. Mayer[3] and Michael Seidel[4]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]Technical University of Munich, Germany

[3]Institute for Bioinformatics and System Biology, Helmholtz Center Munich, German Research Center for Environmental Health

[4]Helmholtz Center Munich, Germany

Rye (*Secale cereale L.*) reveals high tolerance to abiotic stresses. The identification of genes involved in stress tolerance represents targets which might be transferred to other important crop plants. Our approach aims to unlock the genetic potential of rye by providing a comprehensive sequence resource. Five highly diverse rye inbred lines were selected for cDNA sequencing on a Roche/454 GS FLX platform. Further, our goal is to detect a robust set of polymorphic differences, namely single nucleotide polymorphism (SNP), to establish a high-density rye transcript map.

As input we used a data set of 2.5 Mio sequences. The developed pipeline comprises a set of essential working modules. It involves preprocessing steps like clipping reads regarding adaptors and quality, a computational high complex assembly process and in subsequence a SNP detection. Four assembly strategies are established using the Mira assembler to identify and proof a reliable set of contigs that form a comprehensive basis of unigenes for in silico polymorphism mining. A number of 84,451 contigs with average length of 496 bp refer as our prime data proofed by other assembly strategies with diverse combination of genotype reads.

SNP candidates are identi ed using the tool GigaBayes. Our results are quality-checked by scanning for reliable patterns in haplotype distribution. Defined parameters avoid candidates in paralogs, false positives caused by repeats or homopolymer uncertainties. A manually curated set of SNPs guides the pipeline to high sensitivity and selectivity. Over 10,000 preliminary SNPs have been detected and will be evaluated concerning our criteria. Results will lead into a detailed annotation of SNP positions or motifs like detected mi-

crosatellites, exon splice-sites or known repeats. The outcome will generate a high-density rye transcript map and facilitate studies on association mapping.

The GABI RYE EXPRESS project is funded by the BMBF (FKZ 0315063B).

Automated data acquisition and image processing for high-throughput phenotyping of barley

Anja Hartmann[1,2], Tobias Czauderna[2] and Falk Schreiber[1,2]

[1]Martin-Luther-University Halle-Wittenberg, Germany

[2]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

In the last few years image-guided high-throughput analysis methods became state-of-the-art in the life sciences. By screening plants non-destructivly over a period of time by means of image acquisition techniques a large amount of images are recorded and have to be analysed. This poster describes the establishment of automatic data acquisition and the development of an image-based analysis of barly plants. The images are generated by an automatic greenhouse system, the high-throughput phenotyping platform developed by LemnaTec. We present the development of an image analysis pipeline implemented as a plugin for ImageJ (an open source image processing software). This plugin improves barley plants color image analysis taken from the high-throughput phenotyping platform in top view and side view. Different parameters for plants such as height, width and leaf area are calculated. The results of the analysis can be used to derive new biological insights.

ConquestExplorer — a genomic and phenotypic Data Repository and Processing Tool towards Omics Data

Rico Basekow[1], Jost Neigenfind[1], Axel Nagel[1], Christiane Gebhardt[2] and Birgit Kersten[1]

[1]Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany

[2]Max Planck Institute for Plant Breeding Research, Koeln, Germany

Efficient storing, managing, retrieval and processing of data are important tasks for any project which handles huge amount of data. The ConquestExplorer tries to combine these tasks in one application. Therefor it uses a database system for storing, provides a graphical user interface for managing and retrieving and has tools integrated to visualize and analyze the data. The coupling of the data schema and processing tools within the ConquestExplorer is very "loosely", so it is easy to adapt the data schema for new requirements and to add new tools. The advantage of such a combined approach is that it provides an easy and consistent handling of huge datasets. The ConquestExplorer was developed in the GABI-II Conquest2 project and it was used for handling and analyzing genomic (e.g. sequence trace files, SNP-, PCR-markers, haplotypes) and phenotypic (e.g. starch content, resistance ratings, maturity) data of 192 potato plants with overall more than 100.000 genomic and phenotypic data points. The underlying database schema is an extended version of PoMaMo [Mey+05]. The sequence trace file viewer Jtrev was adapted to show SNPs within the trace. The statistic software R was integrated to compute associations between genomic (e.g. SNPs, haplotypes) and phenotypic features (e.g. resistance ratings). Other integrated tools are SATlotyper [Nei+08], which computes haplotypes from unphased SNP data of heterozygous polyploids and YAMB [RP+09] to visualize genetic maps. In the current GABI-FUTURE Papatomics project ConquestExplorer will be extended to handle omics data like transcriptomic, proteomic and metabolomic data in addition. Concerning these extensions, tools like MapMan [Usa+05] will be integrated to process these new data. This work is/was funded by the German Federal Ministry of Education and Research BMBF (GABI-FUTURE grant 0315065B, GABI-II grant 0313112).

References

[Mey+05] S. Meyer et al. “PoMaMo–a comprehensive database for potato genome data”. In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D666–670. DOI: 10.1093/nar/gki018.

[Nei+08] J. Neigenfind et al. “Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT”. In: *BMC Genomics* 9.1 (2008), p. 356. DOI: 10.1186/1471-2164-9-356.

[RP+09] D. M. Riano-Pachon et al. “GabiPD: the GABI primary database–a plant integrative ’omics’ database”. In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D954–959. DOI: 10.1093/nar/gkn611.

[Usa+05] B. Usadel et al. “Extension of the Visualization Tool MapMan to Allow Statistical Analysis of Arrays, Display of Coresponding Genes, and Comparison with Known Responses”. In: *Plant Physiology* 138.3 (2005), pp. 1195–1204. DOI: 10.1104/pp.105.060459.

Mutually exclusive spliced exons show non-adjacent and grouped patterns

Martin Pohl[1], Dirk Holste[2], Konrad Grützmann[1], Ralf Bortfeldt[3] and Stefan Schuster[1]

[1]Friedrich-Schiller-University Jena, Germany

[2]Austrian Research Centers, Vienna, Austria

[3]Humboldt-University of Berlin, Germany

Among the different types of alternative splicing (AS), mutually exclusive exons (MXEs) constitute a comparatively rare, but very intriguing and interesting type [Sam+08]. In its simple form, two internal exons are spliced in such a way that one exon is excised, while the other is transcribed (and vice versa), thereby linking the regulation of different exons. Splicing of MXEs has not been analysed in as much detail as more frequent types of AS, e.g., exon-skipping, though nonetheless such AS events contribute to our understanding of this still largely unclear phenomenon. This study used a computational approach to infer about one thousand MXEs across tens of thousands of mapped genes. Comparing the detected pattern and their features (genomic, transcriptional) with existing regulatory models[Smi05], we found for both studies that nearly all MXEs were non-adjacent with respect to their genomic location. Furthermore, while we could not infer complex patterns as in DSCAM [Ana+06] we surprisingly found MXEs to be involved in mutually exclusive splicing of more than one exon simultaneously. We carried out a number of additional analyses for more detailed characterization of our findings. One central outcome of this study is that gapped, rather than adjacent exons are typical MXE events. Consequently, currently inferred AS patterns do not further substantiate existing model predictions.

References

[Ana+06] D. Anastassiou et al. “Variable window binding for mutually exclusive alternative splicing”. In: *Genome Biology* 7.1 (2006), R2. DOI: 10.1186/gb-2006-7-1-r2.

[Sam+08] M. Sammeth et al. “A General Definition and Nomenclature for Alternative Splicing Events”. In: *PLoS Computational Biology* 4.8 (2008), e1000147. DOI: 10.1371/journal.pcbi.1000147.

[Smi05] C. W. J. Smith. “Alternative Splicing — When Two’s a Crowd”. In: *Cell* 123.1 (2005), pp. 1–3. DOI: 10.1016/j.cell.2005.09.010.

Towards a workflow of cross-species analysis of alternative splicing - A case study of the fungal domain

Konrad Grützmann[1], Martin Pohl[1], Karol Szafranski[2] and Stefan Schuster[1]

[1]Friedrich-Schiller-University Jena, Germany

[2]Fritz Lipmann Institute Jena, Germany

Background. In the past decades, the importance of alternative splicing (AS) of pre-mRNAs has been recognized increasingly. AS now is considered a frequent and complex process in eukaryote cells. AS is supposed to contribute to protein diversity, can act as a further layer of gene expression regulation, and may contribute to species evolution. But how has AS developed during eukaryote evolution? How are AS subclasses distributed through eukaryote taxa? Is the distribution correlated to gene structure evolution? To address these questions, comparative analysis of AS is a promising approach. Although computational capacities allow for such an effort, so far, no gold standard exists.

Methods. Here, we present a universal workflow for the investigation of AS from the genome and transcript sequences of a species list. The first step consists in stringent filtering and cleaning of the transcript data (e.g. ESTs) as sequencing techniques are often error-prone. Secondly, transcripts are mapped to genome sequences mainly using EXALIN [ZG06]. We obtain so-called spliced alignments that contain long gaps that correspond to introns. The crucial part of the method is to read AS events from the alignments in a reliable way. Thereby, in our procedure, a predicted alternative isoform always relies on multiple transcripts.

Results. Alternative splicing has not been studied as extensively in fungi as in other eukaryotes. Applying our approach to EST and genome sequences from 25 fungal species, we find that fungal lines markedly differ in the extent of AS as well as in the distribution of AS subclasses.

References

[ZG06] M. Zhang and W. Gish. "Improved spliced alignment from an information theoretic approach". In: *Bioinformatics* 22.1 (2006), pp. 13–20. DOI: 10.1093/bioinformatics/bti748.

The LAILAPS Search Engine: A Text Index Infrastructure for Relevance Ranking over Life Science Database Entries

Matthias Lange[1], Mandy Weißbach[1], Matthias Klapperstück[1], Steffen Flemming[1] and Uwe Scholz[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Searching scientific databases effectively necessitates the use of contemporary software to locate desired and meaningful information in the context of the users scientific or project priorities. However, the combination of relevance ranking with life science data retrieval is missing in nowadays life science information systems. Here, we present the **LAILAPS** search engine, an open Web-application to search over life science databases and provide a user specific relevance ranking of found records. Well known ranking criteria are:

- *database entry properties* like authors, publishing journal,the described organism, evidence scores, surrounding keywords etc.
- *Search Term Frequency — Inverse Document Frequency* (TF-IDF),
- *probabilistic relevancy ranking* using pre-trained models,
- *user feedback systems*,
- *semantic search engines* in combination with methods from natural language processing and dictionaries

We combined aspects of the above criteria with further relevance criteria to a general feature model. This feature model is implemented in a neuronal network and was trained by a training set of about 1000 manual ranked database entries, resulting from 19 querys.

The **LAILAPS** implementation concept is to combine an intuitive and slim Web user interface on top of

1. a machine learning (ML) ranking system,
2. the LUCENE text indexing JAVA-API,
3. an automatic synonym expansion of search terms,
4. a novel feature model and
5. a user feedback system.

LAILAPS is public available for SWISSPROT data at `http://LAILAPS.ipk-gatersleben.de`.

Automatic Detection of Fluorescence Labeled Neurites in Microscope Images

Danny Misiak[1], Stefan Posch[1], Nadine Stöhr[2], Stefan Hüttelmaier[2] and Birgit Möller[1]

[1]Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Germany

[2]Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Germany

Neurons are important constituents of higher organisms which have the ability to establish complex networks via their neurites. The analysis of neurite morphology is essential to understand brain development as well as learning and memory. Furthermore it reveals fundamental information of neuronal diseases which result in severe morphological defects. Within the research group of Stefan Hüttelmaier, the transport of β-actin mRNA from the cell nucleus into neurites up to the growth cones is analyzed. Due to this the β-actin protein concentration in the growth cone is increased and promotes the protrusion of the neurite [H¨05]. During the last years, fluorescence microscopy techniques have been improved to analyze such mechanisms in detail. A large amount of high resolution images can be obtained, requiring fully automatic techniques for processing and evaluation. Our main goal is to combine automatic neurite segmentation and morphology characterization with the spatial and quantitative assessment of protein ratios on the single cell level in given non-homogeneously textured neurites. For that purpose we developed a novel and fully automatic approach for neuron cell localization and neurite segmentation. The approach combines two phases, where the first one yields an initial pre-segmentation of all neuronal cells in the analyzed optical field applying thresholding techniques. Subsequently, one of the localized neurons is selected to be further analyzed. In the second stage active contour models [Kas+88] based on hierarchical Gradient Vector Flow fields [XP98] are employed to improve the initial neurite segmentation. Neuron localization results as well as segmented neurite contours from a set of test images demonstrate the appropriateness of our approach for practical biomedical research.

References

[H¨05] S. Hüttelmaier et al. “Spatial regulation of [beta]-actin translation by Src-dependent phosphorylation of ZBP1”. In: *Nature* 438.7067 (2005), pp. 512–515. DOI: 10.1038/nature04115.

[Kas+88] M. Kass et al. “Snakes: Active contour models”. In: *International Journal of Computer Vision* 1.4 (1988), pp. 321–331. DOI: 10.1007/BF00133570.

[XP98] C. Xu and J. L. Prince. “Snakes, shapes, and gradient vector flow”. In: *IEEE Transactions on Image Processing* 7.3 (1998), pp. 359–369. DOI: 10.1109/83.661186.

MicroRNA target prediction improved by RNA secondary structure calculations

Ray Marin[1] and Jiri Vanicek[1]

[1]Laboratory of Theoretical Physical Chemistry, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Among the small RNAs that regulate gene expression, miRNAs in animals are characterized by an incomplete complementarity to their 3'UTR targets, making it difficult to predict functional interactions by standard bioinformatic tools. Recently, a new algorithm was proposed to predict miRNA targets, based on the overrepresentation of conserved sites complementary to the miRNA seed, and it was successfully used to predict the immediate early gene IE1 in the human cytomegalovirus[Mur+08]. Nevertheless, it is well known that the accessibility of the binding site also plays a key role in the biological function of miRNAs. Here we introduce an extension of the original algorithm from Ref. [Mur+08], in which we consider the accessibility of complementary sites, obtained from the Boltzmann ensemble of all possible secondary structures for the corresponding 3'UTR [Muc+06]. In order to test the merits of this extended algorithm, we applied it to the genome of *Drosophila melanogaster*. Comparison with the previous version of the method shows that including accessibility of the complementary sites increases the specificity of the algorithm. Moreover, comparison with more elaborate standard methods shows that our approach is more specific for a given sensitivity or, on the other hand, that higher sensitivity is achieved, preserving the same level of specificity.

References

[Muc+06] U. Muckstein et al. "Thermodynamics of RNA-RNA binding". In: *Bioinformatics* 22.10 (2006), pp. 1177–1182. DOI: 10.1093/bioinformatics/btl024.

[Mur+08] E. Murphy et al. "Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: Implications for latency". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.14 (2008), pp. 5453–5458. DOI: 10.1073/pnas.0711910105.

Interactive real time ray tracing in Molecular Visualization

Anna Katharina Dehof[1], Lukas Marsalek[2], Iliyan Georgiev[2], Daniel Stoeckel[1], Stefan Nickels[1], Hans-Peter Lenhof[1], Philipp Slusallek[1,3] and Andreas Hildebrandt[1]

[1]Center for Bioinformatics, Saarland University, Saarbrücken, Germany

[2]Department of Computer Science, Saarland University, Saarbrücken, Germany

[3]German Research Center for Artificial Intelligence, Germany

Molecular visualization is one of the most valuable tools in structural bioinformatics, computational chemistry, and related fields. Data in this field can be very diverse in nature and, optimally, the visualization tools should cover three-dimensional scalar quantities - cryo-em data or electrostatic potentials, for instance - as well as large different representations and coloring schemes for molecular structures.

For this visualization, a particularly important goal is to provide researchers with an accurate, yet intuitive three-dimensional spatial representation of molecular structural arrangements and the effective extraction of relevant information in the volume data. A second aim of molecular graphics is the creation of publication-quality images. For this task, ray tracers are usually the method of choice; although these traditionally take minutes to hours to produce high-quality results.

Recently, we have been able to integrate real-time, CPU-based polygonal ray tracing techniques, provided by the RTfact library, and custom-made GPU-based volume ray tracing methods into the versatile molecular viewer and modeller BALLView.

This combination allows the user to employ advanced visualizations, such as realistic shadows and reflections on (polygon-based) structural representations of molecular systems, and to create real-time, high-quality visualization of three-dimensional volume data, offering fine-grained control over the selection of scalar values and the mapping to colors. Combined with an intuitive user interface, this allows even non-expert users to gain new insights into their data, and to easily create exciting publication-quality images or movies.

SoupViewer — efficient analysis of large cluster trees

Jan Engelhardt[1], Peter F. Stadler[1] and Kristin Reiche[2]

[1]Department of Computer Science, University of Leipzig, Germany

[2]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

Genome-wide computational and experimental screens for structured non-coding RNA (ncRNA) genes in different taxa resulted in an extensive list of putative structured ncRNAs. Only a minority of them could be assigned to already known ncRNAs by sequence similarity. We presented a clustering method which detects homologous secondary structure motifs in a set of sequences by pairwise sequence-structure alignments [Wil+07]. The clustering procedure results in a hierarchical cluster-tree which reflects structural similarities. NcRNAs sharing similar secondary structures are found in the same group, divergent ncRNAs are separated. We provided a method which extracts reasonable subtrees of high-similar structure motifs automatically [Rei08].

However, such full-automatic clustering is not free of misclassifications. This is the reason why the cluster tree may still to be examined by a human expert to check if the partitions are reasonable. The graphical cluster-tree viewer `SoupViewer` which reports for each subtree sequence and structural conservation as well as a significance measure that the subtree consists of ncRNA candidates sharing the same structural motif, provides a user-friendly framework for semi-automatic detection of high-interesting groups. The most important advantage compared to already existing tree viewers is that for each internal node the secondary consensus structure and the alignment of the sequences belonging to this node can be displayed easily.

The structure-based clustering pipeline is a very useful approach for annotating ncRNAs. The new developed `SoupViewer` makes the analyses of large cluster trees much easier and increases the efficiency of the pipeline.

References

[Rei08] K. Reiche. *RNAsoup Documentation*. Fraunhofer Institute for Cell Therapy and Immunology. 2008. URL: `http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAsoup/manual.pdf`.

[Wil+07] S. Will et al. “Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering”. In: *PLoS Computational Biology* 3.4 (2007), e65. DOI: `10.1371/journal.pcbi.0030065`.

Reverse Engineering of Signaling Pathways from RNAi Data

Lars Kaderali[1], Eva Dazert[1], Betina Knapp[1], Narsis Aftab Kiani[1], Michael Frese[1], Ulf Zeuge[1] and Ralf Bartenschlager[1]

[1]BIOQUANT, University of Hiedelberg, Germany

The inference of signal transduction and genetic regulatory networks is a major goal in systems biology. Systematic screenings of RNA interference (RNAi) offer the possibility to identify genes related with a particular phenotype or cellular pathway of interest (Fire, 1998). The temporal and spatial placement of these genes in the respective cellular pathway remains a challenging problem (Moffat and Sabatini, 2006). While Sacher et al. (2008) cluster phenotypes, Markowetz et al. (2007) use the nested structure of effects of different knockdowns to solve this problem. We propose a stochastic model with Boolean networks for pathway inference where the activation probabilities for each gene are described by sigmoid functions. A Markov chain Monte Carlo approach is used to infer model topology and model parameters simultaneously, by sampling from the posterior distribution over model parameters given the knockdown data in a Bayesian setting. We compute the exact transition probabilities between different network states using the effect of single- or combinatorial knockdowns. Incomplete observations are integrated out via marginalization over unobserved nodes. To address the problem of under determined model parameters we use a prior distribution on the model parameters. We then approximate the likelihood allowing to sample from the posterior distribution without explicit evaluation of the likelihood whereas the sampling from the posterior for networks with larger number of nodes is permitted. We evaluated our method on a small artificial network with five nodes and we present results from the inference of the Jak/Stat signal transduction pathway in a hepatoma cell line given the knockdown data of 11 genes of the core Jak/Stat pathway.

Maltcms — an Application Framework for Processing of Metabolomics-Data

Nils Hoffmann[1], Mathias Wilhelm[1] and Jens Stoye[1]

[1]Genome Informatics Group, Faculty of Technology, Bielefeld University, Germany

We present the current state of our modular application toolkit for chromatography-mass spectrometry (Maltcms), and two derived applications: ChromA [HS09], which is applicable to gas-chromatography (GC) and liquid-chromatography (LC) data with single-dimension detectors (FID, FL) or multi-dimension detectors (MS), and ChromA4D, which is applicable to data from GCxGC-MS experiments. The framework Maltcms allows to setup and configure individual processing components with few effort. All processing steps center around the pipeline paradigm, where each step can define dependencies on previous steps.

ChromA is a configuration of Maltcms, which includes preprocessing, in the form of time-scale alignment, mass binning, and annotation of signal peaks found within the data, as well as visualizations of unaligned and aligned data. The multiple alignment is based on a star-wise or tree-based application of an enhanced variant of pairwise dynamic time warping (DTW). To reduce both runtime and space requirements, we identify conserved signals throughout the data, constraining the search space of DTW. These alignment anchors can be augmented or overriden by user-defined anchors, such as previously identified compounds, characteristic mass or MS/MS identifications. They are then paired by means of a bidirectional best-hits criterion, which compares the candidates for similarity. Paired anchors are then extended to k-cliques with configurable k, which help to determine the conservation of signals across measurements.

ChromA4D also applies DTW for alignment, but in this case to data with one additional chromatographic dimension from GCxGC-MS experiments. It includes peak finding, integration by seeded region growing and bidirectional best-hits peak matching between chromatograms. Its visualizations represent aligned chromatograms as color overlay images. This allows a direct visual comparison of signals present in one sample, but not present in another sample.

Maltcms is freely available at `http://maltcms.sourceforge.net` under the L-GPL v3 license.

References

[HS09] N. Hoffmann and J. Stoye. "ChromA: signal-based retention time alignment for chromatography-mass spectrometry data". In: *Bioinformatics* 25.16 (2009), pp. 2080–2081. DOI: 10.1093/bioinformatics/btp343.

IBIS — Improved base calling for the Illumina Genome Analyzer

Martin Kircher[1], Udo Stenzel[1] and Janet Kelso[1]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The Illumina Genome Analyzer (GA) is a high-throughput sequencing platform that generates millions of short (18-100nt) reads using parallel fluorescence-based readout of immobilized DNA templates. Technical and chemical limitations make base calling for this platform a non-trivial, multi-class classification problem. Different approaches to base calling [Ben+08; Erl+08; Rou+08] have been proposed. These are generally limited in accuracy due to an incomplete modeling of the underlying sequencing process or too time consuming. We show that given a training data set, fast and accurate base calling can be performed using a statistical learner without the need to specifically model the sequencing process.

Using a set of Support Vector Machines (SVMs) together with training data obtained from the standard Illumina base caller (Bustard), we show that the interconnected chemical and technical parameters can be efficiently captured without any prior model. The SVM parameters can then be used to base call a run more accurately. We present IBIS (Improved Base Identification System), an accurate and easy-to-use base caller for the Illumina sequencer. On test data sets, IBIS significantly reduces the error rate associated with Bustard base calling by 26-73% and thereby increases the output of usable reads on all available chemistries (FC-104-100x, FC-204-20xx, FC-103-300x, FC-103-300x+RDP) and all technical versions (GAI, GAII/IIx). IBIS outperforms other publicly available base calling packages like AltaCyclic [Erl+08] and Rolexa [Rou+08] in accuracy while requiring less than 5% of the computational time.

References

[Ben+08] D. R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456 (2008), pp. 53–59. DOI: 10.1038/nature07517.

[Erl+08] Y. Erlich et al. “Alta-Cyclic: a self-optimizing base caller for next-generation sequencing”. In: *Nature Methods* 5.8 (2008), pp. 679–682. DOI: 10.1038/nmeth.1230.

[Rou+08] J. Rougemont et al. “Probabilistic base calling of Solexa sequencing data”. In: *BMC Bioinformatics* 9.1 (2008), p. 431. DOI: 10.1186/1471-2105-9-431.

Reverse Engineering of Gene Regulatory Networks with a Nonlinear ODE-Model Embedded into a Bayesian Framework

Johanna Mazur[1], Daniel Ritter[1], Gerhard Reinelt[2] and Lars Kaderali[1]

[1]Viroquant Research Group Modeling, Bioquant, University of Heidelberg, Germany

[2]Research Group Combinatorial Optimization, Institute for Scientific Computing, University of Heidelberg, Germany

One of the most challenging and interesting problems in the field of systems biology is the reconstruction of gene regulatory networks from gene-expression time-series data. We present a new conceptual approach for this purpose using nonlinear differential equations, which we embed into a Bayesian learning framework. We thus address the desire for quantitative dynamic models, and take into account the stochasticity of the underlying data. Additionally, the Bayesian framework provides an easy way to incorporate prior biological knowledge, such as sparsity of the network, into the learning process. Knowledge from literature or other data sources can also be integrated easily in this framework.

By using stochastic sampling via Markov chain Monte Carlo methods, we consider the full Bayes' posterior distribution and can nicely address the problem of nonidentifiability of network structures or parameters. Furthermore, the ability to calculate the full distribution gives a good starting point for experimental design to resolve ambiguities in reconstructed networks, e.g., discriminate between two alternative highly-probable network structures. Altogether, with the presented conceptual approach we are able to tackle relevant questions in systems biology from a quantitative dynamic perspective.

On simulated data, robustness is shown by varying the dataset sizes and the levels of noise. Furthermore, we evaluated our method on the challenge 3 data from the DREAM 2 initiative, real experimental data from a synthetic, engineered 5-gene network. We are able to reconstruct correctly the dynamics and main regulations of the underlying biological system.

OpenMS — An Open Source Framework for Mass Spectrometry Data

Chris Bielow[1], Stephan Aiche[1], Sandro Andreotti[1], Andreas Bertsch[2], Clemens Gröpl[3], Rene Hussong[4], Nico Pfeifer[2], Marc Sturm[2], Alexandra Zerck[1], Andreas Hildebrandt[4], Knut Reinert[1] and Oliver Kohlbacher[2]

[1]Algorithmic Bioinformatics, Free University Berlin, Germany

[2]Center for Bioinformatics, Eberhard-Karls-University Tübingen, Germany

[3]Institute for Mathematics and Informatics, Ernst-Moritz-Arndt-University of Greifswald, Germany

[4]Center for Bioinformatics, Saarland University, Saarbrücken, Germany

The development of algorithms for mass spectrometry can be greatly facilitated by using efficient modules in a C++ library. Here we present our C++ based open source framework OpenMS [Stu+08], which provides algorithms for analyzing and manipulating mass spectrometry data. The algorithms range from routines for file I/O of complex XML formats (mzML, mzData, ...), file conversion and data visualization to elaborated algorithms for preprocessing (BaselineFilter, NoiseFilter, ...), quantitation (FeatureFinder, FeatureLinker, ...), identification (MascotAdapter, ConsensusID, ...) and more. OpenMS can be used as a library and also offers precompiled, stand-alone tools that can be readily deployed to users.

The user benefits from highly configurable tools (TOPP [Koh+07] — The OpenMS Proteomics Pipeline), which are build on top of OpenMS and are chainable into custom analysis pipelines. TOPP offers more than 40 tools exposing algorithms from the library along with a GUI based assistant to build custom pipelines and a versatile viewer to inspect the resulting data. Installation on Mac OS X, Linux and Windows is automated via Packages and Installers.

References

[Koh+07] O. Kohlbacher et al. “TOPP–the OpenMS proteomics pipeline”. In: *Bioinformatics* 23.2 (2007), e191–197. DOI: 10.1093/bioinformatics/btl299.

[Stu+08] M. Sturm et al. “OpenMS - An open-source software framework for mass spectrometry”. In: *BMC Bioinformatics* 9.1 (2008), p. 163. DOI: 10.1186/1471-2105-9-163.

Adequate Usage of Affymetrix' background probes on Exon and Gene 1.0 ST arrays

Jan Brücker[1], Peter F. Stadler[1,2,3] and Hans Binder[1]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

[2]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

[3]Santa Fe Institute, Santa Fe, USA

Affymetrix oligonucleotide probes are affected by parasitic effects that distort the linear relation of measured intensity to transcript concentration. This fact requires to develop suitable algorithms that correct (calibrate) the intensity measures.

A key step of these correction methods requires to estimate the non-specific background of each probes intensity that is caused by polynucleotides not targeted by the probes. An instrument to estimate background content of a probe was supplied for generations of Affymetrix arrays with the mismatch probes. However, with Exon 1.0 ST and Gene 1.0 ST arrays Affymetrix' last array generations are designed as PM-only chips without mismatched MM-probes to double the number of genomic targets. For these arrays they introduced instead sets of background control probes. One set is build from bacterial genome (16943 probes), another one from intronic regions of the human genome (10990 probes).

We analyzed the sequence specific intensity bias of these probes and propose appropriate sequence models of the sets of background probes. We found that the specially designed sets of background probes are suboptimal for predicting background corrections for the studied chips. Instead we suggest to use subensembles of the "normal" probes printed on the chip, which were selected by applying a modified version of the recently developed "Hook"-method [Bin+08; BP08]. By extending our hook algorithm to PM-only chips we show how the sets of of background probes can be used to deduce a suitable set of background probes.

References

[Bin+08] H. Binder et al. ""Hook"-calibration of GeneChip-microarrays: Chip characteristics and expression measures". In: *Algorithms for Molecular Biology* 3.1 (2008), p. 11. DOI: 10.1186/1748-7188-3-11.

[BP08] H. Binder and S. Preibisch. ""Hook"-calibration of GeneChip-microarrays: Theory and algorithm". In: *Algorithms for Molecular Biology* 3.1 (2008), p. 12. DOI: 10.1186/1748-7188-3-12.

TInA (T-Invariant Analysis) A Tool Box for Exploring Pathways in Biochemical Systems at Steady State

Anja Thormann[1,2], Konrad Rudolph[2] and Ina Koch[3,4]

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany

[2]Free University of Berlin, Germany

[3]Beuth University for Technology, Berlin, Germany

[4]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Background. Combining experimental and theoretical approaches to enlighten the functioning of molecular processes in the cell is a main goal in systems biology. Theoretical modelling of biological systems in particular will become more and more important to reduce wet lab experiments for target finding. Based on qualitative Petri net models, a minimal set of basic pathways at steady state, the transition invariants (T-invariants) or elementary modes, respectively, can be computed [KH08; Vos+03]. In the case of complex and large networks, the number of T-invariants increases exponentially. Thus, new concepts for further network decomposition into biologically meaningful pathways have been developed, such as MCT-sets [Sac+06], T-clusters [GB+08], and Mauritius maps [Gru+08].

We developed a software tool which compiles all these concepts providing a user-friendly interface and interfaces to Petri nets and systems biology standards.

Methods. Petri nets can be used to model and analyze biochemical systems [KH08]. We focus on the computation of T-invariants and their further exploration using MCT-sets, T-clusters, and Mauritius maps. MCT-sets represent disjunctive sets of transitions, which occur always together and exhibit the same expression behaviour. T-clusters summarise transitions, which are important for special biological functions. Mauritius map is a data structure in terms of a binary tree to reflect dependencies of T-invariants. It can be represented graphically and is helpful in knockout analyses.

Results. We provide different techniques for the exploration of Petri nets applicable in the context of biological pathway analysis and bring them together in a Java based graphical user interface. Our tool exhibits high potential to be used for analyzing complex Petri nets. To demonstrate this, we present as case study the modelling of the gene regulation of the Duchenne muscular dystrophy comprising 63 positions and 87 transitions [Gru+08].

References

[GB+08] E. Grafahrend-Belau et al. "Modularization of biochemical networks based on classification of Petri net t-invariants". In: *BMC Bioinformatics* 9.1 (2008), p. 90. DOI: 10.1186/1471-2105-9-90.

[Gru+08] S. Grunwald et al. "Petri net modelling of gene regulation of the Duchenne muscular dystrophy". In: *Biosystems* 92.2 (2008), pp. 189–205. DOI: 10.1016/j.biosystems.2008.02.005.

[KH08] I. Koch and M. Heiner. "Analysis of Biological Networks". In: ed. by B. H. Junker and F. Schreiber. Wiley Book Series in Bioinformatics. Wiley-Interscience, 2008. Chap. 7, pp. 139–180. ISBN: 0470041447.

[Sac+06] A. Sackmann et al. "Application of Petri net based analysis techniques to signal transduction pathways". In: *BMC Bioinformatics* 7.1 (2006), p. 482. DOI: 10.1186/1471-2105-7-482.

[Vos+03] K. Voss et al. "Steady state analysis of metabolic pathways using Petri nets". In: *In Silico Biology* 3.3 (2003), pp. 367–387. URL: http://www.bioinfo.de/isb/2003030031/.

Evidence for a functional role of CAG/glutamine repeats

Martin Schaefer[1] and Miguel A. Andrade-Navarro[1]

[1]Computational Biology and Data Mining, Max Delbrück Center for Molecular Medicine, Berlin, Germany

Several neurodegenerative conditions such as Huntington's disease are caused by the expansion of CAG trinucleotide repeats or polyglutamine (polyQ) stretches over specific length thresholds. During disease progression the affected proteins form intracellular aggregates and recruit additional proteins into these inclusions. Nevertheless, the pathomechanism as well as the wild type function of the repeats remain largely unknown. It is even under debate if function and toxicity of the repeats are properties at the RNA or at the amino acid level.

To address these questions we first provide indications for a functional role on RNA level by showing that the number of CAG repeats in Untranslated Regions exceed the number that is expected by chance in several species. Additionally, motivated by the observation that polyQ containing proteins have a significantly higher amount of interaction partners we investigated a role for polyQ as a protein interaction domain by assessing the deviation of the polyQ protein-protein interaction network from random topology. To this end, we set up a statistical testing framework which allows for directly modeling functional and degree distribution biases in the set of polyQ containing proteins. We observed several enriched domains in the interaction network surrounding the polyQ protein set. These findings suggest a dual role for CAG/glutamine repeats both at the RNA and at the amino acid level.

Nutrilyzer — a Tool for Deciphering Atomic Compositon of Differentially Expressed Orthologous Proteins

Katrin Lotz[1,2], Falk Schreiber[2,3] and Röbbe Wünschiers[4]

[1]SunGene GmbH, Gatersleben, Germany

[2]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[3]Martin-Luther-University Halle-Wittenberg, Germany

[4]Hochschule Mittweida, Germany

Organisms try to maintain homeostasis by balanced uptake of nutrients from their environment. From an atomic perspective this means that e.g. carbon:nitrogen:sulfur ratios are kept within given limits.

We are particular interested whether the carbon, nitrogen, sulfur or oxygen content of paralogous proteins correlates with their expression level under carbon, nitrogen, oxygen or sulfur limitations, respectively. In order to facilitate such analyses for a broad range of organisms we created a generic, web-based software pipeline, doped Nutrilyzer. It extracts paralogous protein coding sequences from an annotated genome sequence and evaluates their atomic composition. When fed with gene-expression data from nutrient limited and normal conditions, Nutrilyzer provides a list of genes that are signi cantly differently expressed and contain signi cantly different amounts of the limited nutrient in their atomic compostion. The software is distributed under the GPL licence and can be obtained from `http://nutrilyzer.sourceforge.net` as a linux live image.

The software architecture is based on the Java Spring Web MVC and Web Flow frameworks. It runs under an Apache Tomcat webserver on an openSuse operating system. Additionally, a MySQL database connection is used for storing genome and cluster information. For finding all paralogs from the whole genome, the program blastclust from the NCBI toolbox is used. The signi cance tests are done by the statistical language R from the R-project and Bioconductor.

Improved Automatic Annotation of Metazoan Mitochondrial Genomes

Alexander Donath[1], Fabian Externbrink[1], Frank Jühling[1], Matthias Bernt[2], Martin Middendorf[2] and Peter F. Stadler[1]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany
[2]Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Germany

The NCBI RefSeq [Pru+06] is the most up-to-date source for mitochondrial genomes and their annotation. Despite efforts on the part of the RefSeq project, the annotation is still highly inconsistent and contains numerous errors [Boo+05; Per+08], such as i) annotations on the reverse complement strand, ii) wrong names (most error-prone are the pairs of tRNAs L1, L2 and S1, S2), iii) missing genes, iv) genes that are annotated too often, and v) wrong start and end positions. Earlier attempts [Wym+04] to challenge this task needed user-interaction and subsequent manual improvement of the results to perform reasonably well.

Here we present a tool that uses state-of the art methods for the automatic annotation of metazoan mitochondrial genomes. tRNAs and rRNAs are detected through structure-aware covariance models, whereas proteins are annotated via homology search, incorporating BLAST and subsequent start and stop codon prediction. The method simplifies and standardizes novel genome annotations, significantly corrects existing annotations, and allows, e.g., subsequent analyses such as genome rearrangement and improved phylogenetic studies. We also provide an improved re-annotation of all metazoan RefSeq mitogenomes to the community.

References

[Boo+05] J. L. Boore et al. "Sequencing and Comparing Whole Mitochondrial Genomes of Animals". In: *Methods in Enzymology* 395 (2005), pp. 311–348. DOI: 10.1016/S0076-6879(05)95019-2.

[Per+08] M. Perseke et al. "Evolution of mitochondrial gene orders in echinoderms". In: *Molecular Phylogenetics and Evolution* 47.2 (2008), pp. 855–864. DOI: 10.1016/j.ympev.2007.11.034.

[Pru+06] K. D. Pruitt et al. “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. In: *Nucleic Acids Research* 35.Database issue (2006), pp. D61–D65. DOI: 10.1093/nar/gkl842.

[Wym+04] S. K. Wyman et al. “Automatic annotation of organellar genomes with DOGMA”. In: *Bioinformatics* 20.17 (2004), pp. 3252–3255. DOI: 10.1093/bioinformatics/bth352.

Identifying metabolic markers of Verticillium longisporum infection by means of one-dimensional self-organizing maps

Cornelia Göbel[1], Kirstin Feussner[2], Alexander Kaever[3], Peter Meinicke[3], Ivo Feussner[3], Manuel Landesfeind[3] and Burkhard Morgenstern[1]

[1]Department of Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-University Goettingen, Germany

[2]Department of Developmental Biochemistry, Institute for Biochemistry and Molecular Cell Biology, University Medical Center Goettingen, Germany

[3]Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Goettingen, Germany

Infection of rapeseed (*Brassica napus*) by the soil-borne, vascular fungal pathogen *Verticillium longisporum* results in severe economic damage worldwide. Till today knowledge on either involvement of phytohormones or on other metabolic processes altered by this plant-pathogen interaction is rather scarce. Using the model plant *Arabidopsis thaliana*, an unbiased high-throughput UPLC-TOF-MS approach was established to compare metabolite profiles in aqueous and non-polar extracts with the aim to identify metabolome changes induced by infection with *V. longisporum*. A characteristic sample set consists of eight different sample groups (infected vs. non-infected plants using two different *A. thaliana* ecotypes and two time points after infection). Sample-based principal component analysis (PCA) shows a clear sample separation in correlation to the eight conditions. In contrast to that, the corresponding PCA loadings plot does not allow clear identification of interesting groups of metabolic markers. To overcome this problem we applied one-dimensional self-organizing maps (1D-SOMs) as implemented in the tool "MarVis" for metabolite-based clustering and visualization[Kae+09]. The 1D-SOMs allow to overview datasets of high complexity and to identify relevant groups of marker candidates. The corresponding visualization makes it possible to discriminate between general, early or late occurring as well as ecotype-specific infection markers. Searching databases (KEGG, LIPID MAPS) with the accurate molecular masses of markers with infection correlated intensity pattern, metabolites of the phenylpropanoid and suberin biosynthetic pathway could be identified as infection markers.

References

[Kae+09] A. Kaever et al. "MarVis: a tool for clustering and visualization of metabolic biomarkers". In: *BMC Bioinformatics* 10.1 (2009), p. 92. DOI: 10.1186/1471-2105-10-92.

Prediction of reversibly oxidized proteins in human tissue and organelle

Hang-Mao Lee[1], Karl-Josef Dietz[1] and Ralf Hofestaedt[1]

[1]Bielefeld University, Germany

Cells constantly suffer from endogenous and exogenous oxidative stress. Once the balance between oxidative stress generation and elimination is broken, the exceed free radicals may oxidize bioactive molecules and lead to their malfunction. The thioredoxin and glutardoxin systems are the rescuers for the oxidized proteins in disulfide bond form. We are interested to predict proteins which's oxidized form could be reduced by thioredoxin and glutaredoxin. These interacting proteins of thioredoxin and glutaredoxin have intermolecular reversible oxidative cysteines which are in thiol form in reduced environment and form disulfide bond under oxidation.

We implement the algorithm from [San+08] and test on thioredoxin target proteins described in [Bal+04] and apply this tool on human protein dataset. Since PDB structure is essential for the prediction, we use Swiss Model to model the structure of proteins which's PDB structure is undetermined. This tool could scan the human protein set of user-specified tissue and organelle then output the proteins containing reversibly oxidized proteins.

References

[Bal+04] Y. Balmer et al. "Thioredoxin links redox to the regulation of fundamental processes of plant mitochondria". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.8 (2004), pp. 2642–2647. DOI: 10.1073/pnas.0308583101.

[San+08] R. Sanchez et al. "Prediction of reversibly oxidized protein cysteine thiols using protein structure properties". In: *Protein Science* 17.3 (2008), pp. 473–481. DOI: 10.1110/ps.073252408.

CELLmicrocosmos 2.2: Advancements in Modeling of three-dimensional PDB Membranes

Björn Sommer[1], Tim Dingersen[1], Sebastian Schneider[2], Sebastian Rubert[1] and Christian Gamroth[1]

[1]Bielefeld University, Germany

[2]Technical University of Munich, Germany

Background. The CELLmicrocosmos 2 project develops a tool providing a simplified workflow to create membrane (bi-)layers by using PDB based lipid and protein files [Ber+00]: The MembraneEditor (CmME). In comparison to the former version the capabilities have been enhanced to meet the needs of the simulation community.

Results. To cover the demand of Molecular Dynamics Simulations like Gromacs [Lin+01], advanced packing algorithms has been developed, achieving high lipid densities and enabling the definition of exact lipid quantities. The placement of proteins has been advanced in two ways: The precision of the manual alignment has been improved and is now possible while looking at the concrete membrane structure. On the other hand, automatic placement of many membrane proteins has been implemented, by using data from the PDB_TM [Tus+04] and OPM [Lom+06] databases. In addition, microdomains and multilayers are now supported. The Plug-In-Interface for the development of user-defined algorithms can now also be used to access the atomic level. A reengineering function enables the editing of PDB membranes changed or simulated by external programs.

Conclusions. The CmME has been advanced to meet the requirements of visualization as well as simulation environments. The documentation and the Java Webstart application, needing only an internet connection and Java 6, is accessible at: `http://Cm2.CELLmicrocosmos.org`

References

[Ber+00] H. M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[Lin+01] E. Lindahl et al. “GROMACS 3.0: a package for molecular simulation and trajectory analysis”. In: *Journal of Molecular Modeling* 7.8 (2001), pp. 306–317. DOI: 10.1007/s008940100045.

[Lom+06] M. A. Lomize et al. “OPM: Orientations of Proteins in Membranes database”. In: *Bioinformatics* 22.5 (2006), pp. 623–625. DOI: 10.1093/bioinformatics/btk023.

[Tus+04] G. E. Tusnady et al. “Transmembrane proteins in the Protein Data Bank: identification and classification”. In: *Bioinformatics* 20.17 (2004), pp. 2964–2972. DOI: 10.1093/bioinformatics/bth340.

Modeling RNA loops based on sequence homology and geometric constraints

Christian Schudoma[1], Patrick May[1] and Dirk Walther[1]

[1]Max Planck Institute of Molecular Plant Physiology, Golm, Germany

Understanding the three-dimensional structure of RNA plays an important role in gaining insight into its function. Of special interest are those regions of an RNA molecule that appear unstructured when considering secondary structure alone – the loop regions. Previous studies have shown that, for example, RNA hairpin loops can be grouped into different clusters according to their sequence composition and non-canonical basepair patterns. In this poster, we present RLooM – RNA LoopModeling, a web database reflecting structures in the PDB with additional functionality for the modeling of RNA loop structures.

Identification and classification of ncRNA molecules using graph properties

Liam Childs[1], Zoran Nikoloski[1], Patrick May[1] and Dirk Walther[1]

[1]Max Planck Institute of Molecular Plant Physiology, Golm, Germany

The study of non-coding RNA genes has received increased attention in recent years fuelled by accumulating evidence that larger portions of genomes than previously acknowledged are transcribed into RNA molecules of mostly unknown function, as well as the discovery of novel non-coding RNA types and functional RNA elements. Here, we demonstrate that specific properties of graphs that represent the predicted RNA secondary structure reflect functional information. We introduce a computational algorithm and an associated web-based tool (GraPPLE) for classifying non-coding RNA molecules as functional and, furthermore, into Rfam families based on their graph properties. Unlike sequence-similarity-based methods and covariance models, GraPPLE is demonstrated to be more robust with regard to increasing sequence divergence, and when combined with existing methods, leads to a significant improvement of prediction accuracy. Furthermore, graph properties identified as most informative are shown to provide an understanding as to what particular structural features render RNA molecules functional. Thus, GraPPLE may offer a valuable computational filtering tool to identify potentially interesting RNA molecules among large candidate datasets.

Inference of Synechocystis sp. strain PCC 6803's carbon metabolism using elementary modes

Jost Neigenfind[1], Jan Huege[1], Joachim Kopka[1], Oliver Ebenhoeh[1], Martin Hagemann[2], Joachim Selbig[3] and Birgit Kersten[1]

[1]Max Planck Institute of Molecular Plant Physiology, Golm, Germany

[2]University of Rostock, Germany

[3]University of Potsdam, Germany

The amount of inorganic carbon represents one of the main environmental factors determining productivity of photoautotrophic organisms like cyanobacteria. The CO_2 fixating enzyme RuBisCO also shows oxigenase activity. Thus, O_2 is also fixated in Calvin Cycle resulting in photosynthetic useless intermediates [Eis+08a]. Therefore, higher plants gained the energy consumptive photorespiration, a pathway which is able to recycle such intermediates and feed the products back to the Calvin Cycle.

To compensate for the shortage of inorganic carbon, cyanobacteria developed an efficient CO_2 concentrating mechanism [Bad+06]. However, it was recently demonstrated that the efficiency of the CO_2 concentrating mechanism is not sufficient to completely suppress RuBisCO's oxygenase activity which resulted in addition of the photorespiratory pathway to the *Synechocystis* carbon matabolism [Eis+08b].

Here we propose several new knockout mutants based on elementary mode [Sch+00] and minimal cut set [KG04] analysis of the *Synechocystis* carbon metabolism. The calculated knockout mutants make predictions which can be tested experimentally. Verification of the computed predictions in experiments are necessary conditions for accepting the proposed model of photorespiration in *Synechocystis*.

Furthermore, by means of a parsimony approach the available metabolite levels from different knockout mutants of *Synechocystis* [Eis+08a] will be used to develop a method for the calculation of possible flux distributions which represent the best explanation of the observed changes in metabolite levels in respect to the number of assumptions.

References

[Bad+06] M. R. Badger et al. “The environmental plasticity and ecological genomics of the cyanobacterial CO2 concentrating mechanism”. In: *Journal of Experimental Botany* 57.2 (2006), pp. 249–265. DOI: 10.1093/jxb/eri286.

[Eis+08a] M. Eisenhut et al. “Metabolome Phenotyping of Inorganic Carbon Limitation in Cells of the Wild Type and Photorespiratory Mutants of the Cyanobacterium Synechocystis sp. Strain PCC 6803”. In: *Plant Physiology* 148.4 (2008), pp. 2109–2120. DOI: 10.1104/pp.108.129403.

[Eis+08b] M. Eisenhut et al. “The photorespiratory glycolate metabolism is essential for cyanobacteria and might have been conveyed endosymbiontically to plants”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.44 (2008), pp. 17199–17204. DOI: 10.1073/pnas.0807043105.

[KG04] S. Klamt and E. D. Gilles. “Minimal cut sets in biochemical reaction networks”. In: *Bioinformatics* 20.2 (2004), pp. 226–234. DOI: 10.1093/bioinformatics/btg395.

[Sch+00] S. Schuster et al. “A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks”. In: *Nature Biotechnology* 18 (2000), pp. 326–332. DOI: 10.1038/73786.

Application of Granger causality testing to the detection of cause-effect relationships between metabolites and transcripts in yeast adapting to temperature stress

Katrin Strassburg[1], Pawel Durek[2], Joachim Kopka[2] and Dirk Walther[2]

[1]Leiden University, Netherlands

[2]Max Planck Institute of Molecular Plant Physiology, Golm, Germany

We characterized the molecular response of yeast to increased and lowered temperatures relative to optimal reference conditions across two levels of molecular organization: the transcriptome and the metabolome. When analyzed with regard to known metabolic pathways, pairwise metabolite correlation levels were found to carry more pathway-related information and to extend to farther distances within the metabolic pathway network than associated transcript level correlations. Metabolite - transcript correlations were stronger for metabolites correlated with transcript levels encoding their cognate enzyme than for more distant metabolite-transcript pairs. Furthermore, we observed a temporal hierarchy between the two levels of molecular organization both under heat and cold stress. Changes of metabolite pools were most significantly correlated to the transcript levels encoding metabolic enzymes, when metabolites were defined leading in time-shifted correlation analyses. By applying the concept of Granger-causality, we detected directed metabolite-transcript relationships and identified the metabolite serine as a major causative metabolite, while glutamic acid levels changed predominantly in response to preceding changes of enzyme-transcript levels. Selected examples are presented, illustrating the intertwined relationships between metabolites and transcripts.

The AnnotationSketch genome annotation drawing library

Sascha Steinbiss[1], Gordon Gremme[1], Christin Schaerfer[1], Malte Mader[1] and Stefan Kurtz[1]

[1]University of Hamburg, Germany

For intuitive orientation in the vast amount of genome annotation data available today, a visual representation of genomic features in a given sequence range is often desirable. We present *AnnotationSketch* [Ste+09], a C library for drawing linear maps of genomic features. It accepts annotation data given in a graph format as specified by the Sequence Ontology [Eil+05] which can be read from several common input formats (such as GTF, GFF3 and BED) but can also be constructed on the fly. *AnnotationSketch* supports multiple output formats (PNG, PDF, PostScript and SVG) as well as direct drawing to GUI widgets. It is configurable to allow user-defined color and layout customizations. Furthermore, it is extensible via a custom track mechanism, allowing drawing of arbitrary user-defined content (such as curves etc.). The software can easily be integrated into custom C applications using an object-oriented C API. This object-oriented approach makes it straightforward to create bindings to scripting languages. To exemplify such use of *AnnotationSketch*, bindings to the scripting languages Ruby, Python and Lua are bundled with the library.

References

[Eil+05] K. Eilbeck et al. "The Sequence Ontology: a tool for the unification of genome annotations". In: *Genome Biology* 6.5 (2005), R44. DOI: 10.1186/gb-2005-6-5-r44.

[Ste+09] S. Steinbiss et al. "AnnotationSketch: a genome annotation drawing library". In: *Bioinformatics* 25.4 (2009), pp. 533–534. DOI: 10.1093/bioinformatics/btn657.

Flux Coupling Analysis meets Gene Expression Analysis

Frank Wessely[1], Christoph Kaleta[1] and Stefan Schuster[1]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany

Flux Coupling Analysis allows for the detection of reactions within metabolic models that are always operating together at steady state [Bur+04]. Reactions working in unison suggest that the corresponding enzyme genes are transcriptionally co-regulated. Previous studies comparing coupled reactions with gene expression data used either an approach based on a small microarray data set [RP04] or investigated only pairs of coupled reactions [Not+08]. Our new approach takes into account a much larger compendium of gene expression profiles and coupled sets of any size. We measure the co-regulation of genes using a mutual information based approach. Clusters of co-regulated genes are computed and compared to the coupled reaction sets of the most recent genome-scale metabolic model of *Escherichia coli*. About one sixth of the coupled sets are significantly transcriptionally co-regulated. The majority of this fraction is reproduced by obtaining the coupled sets that significantly correlate with operons. Our results suggest that coupled reaction sets are only weakly co-regulated if they do not belong to the same operon. Thus, they are either expressed in a sequential fashion or regulated by post-transcriptional mechanisms.

References

[Bur+04] A. P. Burgard et al. "Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions". In: *Genome Research* 14.2 (2004), pp. 301–312. DOI: 10.1101/gr.1926504.

[Not+08] R. A. Notebaart et al. "Co-Regulation of Metabolic Genes Is Better Explained by Flux Coupling Than by Network Distance". In: *PLoS Comput Biology* 4.1 (2008), e26. DOI: 10.1371/journal.pcbi.0040026.

[RP04] J. L. Reed and B. O. Palsson. "Genome-Scale In Silico Models of E. coli Have Multiple Equivalent Phenotypic States: Assessment of Correlated Reaction Subsets That Comprise Network States". In: *Genome Research* 14.9 (2004), pp. 1797–1805. DOI: 10.1101/gr.2546004.

A systems biological approach to heterosis: Analysis of molecular network structures in Arabidopsis thaliana

Sandra Andorf[1], Joachim Selbig[2], Thomas Altmann[3], Hanna Witucka-Wall[2] and Dirk Repsilber[1]

[1]Research Institute for the Biology of Farm Animals, Dummerstorf, Germany

[2]Institute for Biochemistry and Biology, University of Potsdam, Germany

[3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

The phenomenon of heterosis is well-known but so far little understood. We propose a systems biological approach to contribute to the understanding of heterosis on the molecular level. We use partial correlations to characterize the global interaction structure of regulatory networks from observational data [Wer+06]. Our hypothesis is that heterosis comes with increased partial correlations which we interpret as increased numbers of regulatory interactions which are leading to an enlarged adaptability of the hybrids [And+09]. We test our hypothesis on two homozygous parental lines of *A. thaliana* and the reciprocal crosses that show heterosis in their phenotypes. Our hypothesis holds for the gene-expression data as well as for the metabolite profile data of 7 time points of early development of *A. thaliana*. The heterozygous crossings that show heterosis show also a higher connectivity in their partial correlation networks.

References

[And+09] S. Andorf et al. "Towards Systems Biology of Heterosis: A Hypothesis about Molecular Network Structure Applied for the Arabidopsis Metabolome". In: *EURASIP Journal on Bioinformatics and Systems Biology* 2009.147157 (2009). DOI: 10.1155/2009/147157.

[Wer+06] A. V. Werhli et al. "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks". In: *Bioinformatics* 22.20 (2006), pp. 2523–2531. DOI: 10.1093/bioinformatics/btl391.

A simulation approach to analyse genotype-phenotype-mapping via the metabolome level

Nina Melzer[1], Dörte Wittenburg[1] and Dirk Repsilber[1]

[1]Research Institute for the Biology of Farm Animals, Dummerstorf, Germany

Currently, we compare two different simulation approaches to predict milk phenotypes using SNP data. The conventional genotype-phenotype-mapping approach (e.g. Meuwissen et al [Meu+01]) is compared to our genotype-phenotype-mapping that incorporates the metabolome level. The aim is to simulate more realistic data by integrating the metabolome. To address this issue we use a curated SBML model.

The bovine genotype is simulated using annotated SNP positions. The allele frequencies are randomly drawn from the allele frequencies of bovine chromosome 6. Further, QTL positions are fixed, which code for the enzymes occuring in the SBML model. Afterwards, the SNP closest to the QTL position is treated as QTL. For the conventional approach, the allele substition effects are obtained from a gamma distribution [HG01]. In contrast, our integrated approach determines those effect by varying enzyme parameters in an allele-dependent manner similar to Mendes et al. [Men+03]. This approach results in different outputs of the SBML model. To receive the phenotype, a random error component is added to the simulated genotype in both approaches. The different simulated data sets are evaluated using different Bayesian methods.

References

[HG01] B. Hayes and M. E. Goddard. "The distribution of the effects of genes affecting quantitative traits in livestock". In: *Genetics Selection Evolution* 33.3 (2001), pp. 209–229. DOI: 10.1051/gse:2001117.

[Men+03] P. Mendes et al. "Artificial gene networks for objective comparison of analysis algorithms". In: *Bioinformatics* 19.suppl_2 (2003), pp. ii122–129. DOI: 10.1093/bioinformatics/btg1069.

[Meu+01] T. H. E. Meuwissen et al. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps". In: *Genetics* 157.4 (2001), pp. 1819–1829. URL: http://www.genetics.org/cgi/content/abstract/157/4/1819.

Composition and Characterization of a Type 2 Diabetes Mellitus Topological Model

Anja Thormann[1], Axel Rasche[1] and Ralf Herwig[1]

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany

Background. Based on a meta-analysis approach that combines high-throughput data of heterogeneous origin in the domain of type 2 diabetes mellitus (T2DM), in particular in connection with obesity as a risk factor we computed a set of marker genes most relevant for obesity-induced T2DM. Different data sources such as DNA microarrays, ChIP on chip and qualitative data from multiple tissues from human and mouse are integrated and validated by a scoring system in order to assign disease relevance to the genes.

Results. Using these marker genes we derived a T2DM topological model containing metabolic reactions and protein-protein interactions. As resources we used CPDB. We evaluated graph properties of the PPI network like the degree distribution, the average cluster coefficients and all-pairs shortest paths. In order to compare these attributes we computed them also for random networks and for a complete PPI network. Random networks contain the same number of nodes and interactions as the T2DM model. The complete network comprises the union of all protein-protein interactions contained CPDB. In-depth investigations of the PPI network reveal gene regulatory interactions and overlaps with signalling pathways annotated by KEGG.

Conclusion. We built a topological model that relates novel disease genes to the functional context of metabolic pathways and protein-protein interactions. Thereby the model exhibits high potential for revealing new connections in diabetic pathways and pathway cross-talk.

fragrep3: Combining fragmented sequence homology with secondary constraints for annotating non-coding RNA

Liang Zhu[1], Peter F. Stadler[2] and Axel Mosig[1]

[1]CAS-MPG Partner Institute for Computational Biology, Shanghai, China

[2]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

Annotating non-coding RNAs (ncRNAs) in genome-wide surveys is still a ma jor challenge, in particular when dealing with rapidly evolving families of ncRNA or searching across large evolutionary time scales. We present the homology search tool `fragrep3`, which combines fragmented homology search on sequence level with a flexible model of including secondary structure constraints.

Despite ma jor recent improvements to hidden Markov models and covariance models [Naw+09], these tools still fail to cover larger evolutionary gaps for many families of ncRNA. As its preceeding versions [Mos+07], `fragrep3` adresses this issue through utilizing homology models that can be easily modified by hand, and thus allowing to integrate further understanding of conservation patterns into the query patterns. The major extension implemented in `fragrep3` is the integration of secondary structure constraints using an approach derived from rnabob [Edd96]. Compared to covariance models, this approach also allows to formulate pseudoknotted secondary structure constraints. Combined with fragrep2's fractional programming approach for fragmented sequence homology [Mos+07], this allows to specify more specific search patterns while keeping the flexibility of manually editing search patterns. Thus, `fragrep3` provides a promising approach to complete the evolutionary picture of several families of ncRNA.

References

[Edd96] S. R. Eddy. *RNABOB: a program to search for RNA secondary structure motifs in sequence databases*. 1996. URL: `http://selab.janelia.org/software.html`.

[Mos+07] A. Mosig et al. “Algorithms in Bioinformatics”. In: vol. 4645. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007. Chap. Homology Search with Fragmented Nucleic Acid Sequence Patterns, pp. 335–345. ISBN: 978-3-540-74125-1. DOI: 10.1007/978-3-540-74126-8_31.

[Naw+09] E. P. Nawrocki et al. “Infernal 1.0: inference of RNA alignments”. In: *Bioinformatics* 25.10 (2009), pp. 1335–1337. DOI: 10.1093/bioinformatics/btp157.

Dynamic simulation; exploring a fit between immunity and hormones in plants

Muhammad Naseem[1], Nicole Philippi[1] and Thomas Dandekar[1]

[1]Department of Bioinformatics, Biocenter, University of Wuerzburg, Germany

Plant hormones play a pivotal role in regulating signaling networks involved in a given host pathogen interaction. Implications of hormones in plant immunity were relatively neglected until recently. Besides their active role in developmental processes, phytohormones such as auxins (Aux), cytokinins (CKs), gibberellic acid (GA) and abscisic acid (ABA) are as important as salicylic acid (SA) and jasmonic acid/ethylene (JA/ET) in plant diseases [RS+07]. Hormones act in concert to sustain physiological harmony. However, pathogenic attack causes hormonal imbalances which add to complexity in cellular networks as compared to steady state of the plant. Much progress has been made in understanding individual hormone signaling in plant disease resistance. Nevertheless, most of these endeavors are independent of each other[GJ09; Pie+09].

To consolidate these vital proceedings, we came up with a novel approach where a dynamic simulation of plant immunity in complex hormonal network has been investigated. Biological information in terms of key pathway enzymes leading to the synthesis of phytohormones, their hormonal end-points and immunologically important regulatory proteins were fed to CellDesigner being a systems biology modeling tool. Celldesigner based data was subjected to SQUAD analysis which provides a dynamic simulation of signaling networks. Upon challenging the system with Flagellin being a PAMP and AvrPto being an Effector, the expected activation and inhibition of plant resistance machinery is quite up to the expectations. This dynamic modeling[Car+07] approach not only pave way to better understanding of intricate hormonal networks but also very much instrumental in immunological aspects of plants.

References

[Car+07] A. di Cara et al. “Dynamic simulation of regulatory networks using SQUAD”. In: *BMC Bioinformatics* 8.1 (2007), p. 462. DOI: 10.1186/1471-2105-8-462.

[GJ09] M. R. Grant and J. D. G. Jones. "Hormone (Dis)harmony Moulds Plant Health and Disease". In: *Science* 324.5928 (2009), pp. 750–752. DOI: 10.1126/science.1173771.

[Pie+09] C. M. J. Pieterse et al. "Networking by small-molecule hormones in plant immunity". In: *Nature Chemical Biology* 5.5 (2009), pp. 308–316. DOI: 10.1038/nchembio.164.

[RS+07] A. Robert-Seilaniantz et al. "Pathological hormone imbalances". In: *Current Opinion in Plant Biology* 10.4 (2007), pp. 372–379. DOI: 10.1016/j.pbi.2007.06.003.

Deterministic Effects Propagation Networks for Reconstructing Protein Signalling Networks from Multiple Interventions

Holger Froehlich[1] and Tim Beissbarth[1]

[1]Division Molecular Genome Analysis, German Cancer Research Center, Heidelberg, Germany

Measuring the expression of genes after interventions performed via siRNA knockdowns provides a rich pool of information on biological systems. While the Nested Effects Models of Fröhlich et al. [Fro+08] give the means to analyse intervention effects in gene expression microarray data, a similar analysis tool for protein expression data is needed. Here we propose the framework of Deterministic Effects Propagation Networks (DEPN) to investigate high throughput data from reverse phase protein arrays and model the interventional effects of silencing specific components in the system via siRNA-Knockdown. Signalling networks are reconstructed via monitoring the intervention effects in the protein expression data in few timepoints and combining the signal propagation through the network via definition of boolean functions at every node. We applied the DEPN framework to reverse phase protein array measurements of several components in the ERBB signalling network measured in de novo trastuzumab resistant human breast cancer cells [Sah+09] and reconstructed regulation patterns occurring after the knockdowns in form of a protein signalling network. While retrieving good concordance with literature knowledge, we also could identify new interesting feed back regulations involving the ERBB1 receptor and pERK1/2 with strong support from the data, that are not yet described in the current literature.

References

[Fro+08] H. Frohlich et al. “Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data”. In: *Bioinformatics* 24.22 (2008), pp. 2650–2656. DOI: 10.1093/bioinformatics/btm634.

[Sah+09] O. Sahin et al. “Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance”. In: *BMC Systems Biology* 3.1 (2009), p. 1. DOI: 10.1186/1752-0509-3-1.

Transcription Factor Binding Site Detection Using Nucleotide Covariance

Erola Pairo[1], Santiago Marco[1] and Alexandre Perera[2]

[1]Institute for BioEngineering of Catalonia, Universitat de Barcelona, Spain

[2]Universitat Politècnica de Catalunya, Spain

In order to regulate transcription, transcription factors bind to short DNA sequences located at the promoter of a gene. Unraveling the mechanisms of gene expression implies the localtization of transcription factors binding sites (TFBS) and many detectors have been proposed. While it is experimentally proven that TFBS have position interdependencies, most detectors use Position Specific Scoring Matrices (PSSM) [Sto00], which assume that each position in a binding site is independent. Symbolical DNA can be mapped to numerical sequences. Using a conversion where each nucleotide is mapped to the vertex of a regular tetrahedron [SL86] we propose a detector based on the Q-residuals of a Principal Component Analyisis of numerical DNA sequences. The purpose is show that a second order statistics can capture information about position interdependencies and improve the detection based on PSSM algorithms. Two different treatments of missing values have been proposed, improving the detectors accuracy. The detector has been compared to existing PSSM methods, using transcription factor binding sites from S. cerevisiae extracted from TRANSFAC database. To show the performance of the detectors ROC curves and its AUC are computed, our detector ouperforms MATCH algorithm [Kel+03]. Comparison to MAST [BG98] using pseudo-random data and the TRANSFAC Drosophila TFBS, has also been carried out.

References

[BG98] T. L. Bailey and M. Gribskov. "Combining evidence using p-values: application to sequence homology searches". In: *Bioinformatics* 14.1 (1998), pp. 48–54. DOI: 10.1093/bioinformatics/14.1.48.

[Kel+03] A. E. Kel et al. "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences". In: *Nucleic Acids Research* 31.13 (2003), pp. 3576–3579. DOI: 10.1093/nar/gkg585.

[SL86] B. D. Silverman and R. Linsker. "A measure of DNA periodicity". In: *Journal of Theoretical Biology* 118.3 (1986), pp. 295–300. DOI: 10.1016/S0022-5193(86)80060-1.

[Sto00] G. D. Stormo. "DNA binding sites: representation and discovery". In: *Bioinformatics* 16.1 (2000), pp. 16–23. DOI: 10.1093/bioinformatics/16.1.16.

Analysis of A Cell Cycle Model Using the Wildau Interaction Knowledgebase WINTER

Gabriele Petznick[1], Sarah Strunk[1], Stephanie Tscherneck[1], Paul Hammer[1], Ronny Amberg[1] and Peter Beyerlein[1]

[1]Technical University of Applied Sciences Wildau, Germany

In the WINTER (Wildau INTERaction) knowledge base public and proprietary interaction data is integrated. In addition it provides three additional features: mutual information based selection of important proteins [Tsc+09], semi-automatic ODE system generation (using GMA-kinetics) and a control theory analysis of the obtained ODE system.

A key process in growth and development is the cell cycle. The related ODE model [Mar08; Wod06] which we optimized shows increasing differences between consecutive cycles, the oscillation fails after 20 cell divisions (indicating 'in-silico' cell death). Control theory [SJ04] is used to analyse the oscillating behavior and its failure. This analysis shows that the ODE system with tens of differential equations and tens of kinetic parameters has multiple steady states, some of them with stable and some of them with unstable characteristics. The control theory analysis aims at identifying the origin of the oscillation failure.

References

[Mar08] H. Marquardt. "Numerical analysis of nonlinear, kinetic ordinary differential equation systems for in-silico modeling of biochemical processes using the cell cycle". MA thesis. Technical University of Applied Sciences Wildau, 2008.

[SJ04] H. Schmidt and E. W. Jacobsen. "Linear systems approach to analysis of complex dynamic behaviours in biochemical networks". In: *Systems Biology* 1.1 (2004), pp. 149–158. DOI: 10.1049/sb:20045015.

[Tsc+09] S. Tscherneck et al. "On the Mutual Information Between Interaction Perplexity and Function of Proteins". In: *IEEE International Workshop on Genomic Signal Processing and Statistics, 17-–19th May.* Minneapolis, MN, USA 2009.

[Wod06] J. Wodke. "Qualitative Modelling of the Human Cell Cycle". MA thesis. Humboldt University Berlin, 2006.

Efficient ncRNA gene finding: Scanning whole genomes using a fast variant of the Sankoff algorithm

Michael Siebauer[1], Kristin Reiche[2], Sebastian Will[3], Peter F. Stadler[1,2,4,5,6] and Rolf Backofen[3]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

[2]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

[3]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University of Freiburg, Germany

[4]Institute for Theoretical Chemistry, University of Vienna, Austria

[5]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[6]Santa Fe Institute, Santa Fe, USA

Genome-wide transcriptomic studies demonstrated that mammalian genomes are almost completely transcribed into transcripts that are not translated into protein [The07]. Evidences emerged that the majority of such non-coding RNA (ncRNA) genes are likely to be functional and participate in a wide range of molecular processes [Mer+09]. Most known functions of ncRNAs require specific structural motifs inducing conserved secondary structures; while less evolutionary pressure is observed for their primary sequences. While efficient methods exist to scan whole genomes for homologous genes, that are conserved on sequence level, nothing comparable is currently available for functional transcripts that are just conserved in secondary structure.

We present here a true scanning variant of `LocARNA` [Wil+07], an efficient variant of the Sankoff algorithm [San85]. It is suitable for genome-wide ncRNA gene finding, in case the ncRNA exhibits a conserved secondary structure. It computes semi-global alignments based on the assumption that for small ncRNAs the entire transcript should be conserved. For the query ncRNA the full ensemble of secondary structures is required as input and provided as a probability distribution (McCaskills algorithm [McC90]). A map of secondary structure predictions of a certain length must be provided for the target genome which is time-consuming but needs to be computed just once for different queries of similar length.

References

[McC90] J. S. McCaskill. "The equilibrium partition function and base pair binding probabilities for RNA secondary structure". In: *Biopolymers* 29.6-7 (1990), pp. 1105–1119. DOI: 10.1002/bip.360290621.

[Mer+09] T. R. Mercer et al. "Long non-coding RNAs: insights into functions". In: *Nature Reviews Genetics* 10 (2009), pp. 155–159. DOI: 10.1038/nrg2521.

[San85] D. Sankoff. "Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems". In: *SIAM Journal on Applied Mathematics* 45.5 (1985), pp. 810–825. URL: http://www.jstor.org/stable/2101630.

[The07] The ENCODE Project Consortium. "Identification and analysis of functional elements in 1the human genome by the ENCODE pilot project". In: *Nature* 447.7146 (2007), pp. 799–816. DOI: 10.1038/nature05874.

[Wil+07] S. Will et al. "Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering". In: *PLoS Computational Biology* 3.4 (2007), e65. DOI: 10.1371/journal.pcbi.0030065.

Conditional profile hidden Markov models for microRNA target prediction

Jan Grau[1], Claus Weinholdt[1], Daniel Arend[1] and Stefan Posch[1]

[1]Martin-Luther-University Halle-Wittenberg, Germany

MicroRNAs are short (∼22 nt) RNAs that bind specifically to mRNA targets and repress the translation of the encoded protein. Experimentally verified microRNAs exhibit nearly perfect base complementarity at nucleotides 2 to 7 at the 5'-end of the microRNA, which is called the seed-region, whereas the 3'-end shows less complementarity. G:U wobble pairs in the seed-region reduce repression to a greater extent than expected from thermodynamics, and strong binding in the 3' region may partially compensate mismatches in the seed-region.

Many existing algorithms for predicting microRNA targets combine alignments and prediction of thermodynamic stability or employ RNA secondary structure prediction. Here, we propose conditional profile hidden Markov models (CoProHMMs) for the prediction of microRNA targets. CoProHMMs are based on an extension of the PLAN9 architecture of original profile HMMs [Kro+94] in which the emission probabilities at the main states are replaced by *conditional* probabilities of the target nucleotides given the nucleotide at a certain position of the microRNA. By this means, we can jointly learn a CoProHMM from a number of different microRNAs and the corresponding targets and automatically fit general properties of microRNA-target binding.

As a proof of concept, we learn CoProHMMs on a number of microRNAs and corresponding targets predicted by miRanda, TargetScan, and RNAhybrid, and we demonstrate how CoProHMMs adapt to the characteristics of these three approaches.

References

[Kro+94] A. Krogh et al. "Hidden Markov Models in Computational Biology : Applications to Protein Modeling". In: *Journal of Molecular Biology* 235.5 (1994), pp. 1501–1531.

Predicting nucleosome positioning from DNA sequence

Jan Grau[1], Martin Porsch[1], Ioana Lemnian[1], Jens Keilwagen[2], Ivo Grosse[1] and Stefan Posch[1]

[1]Martin-Luther-University Halle-Wittenberg, Germany

[2]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Nucleosomes are the fundamental building-blocks of eukaryotic chromatin. In each nucleosome, a stretch of DNA of approximately 147 bp is wound around a histone octamer. Since the DNA bound in nucleosomes is virtually inaccessible to other DNA-binding proteins like transcription factors, the prediction of nucleosome positions is inevitable for predicting functional transcription factor binding sites.

Recent studies indicate that nucleosome positioning can be predicted from DNA sequence [Fie+08]. Here, we propose a two-stage process for predicting nucleome positioning: First, we learn three *component classifiers* that discriminate nucleosome-bound sequences from linkers: one for nucleosome-bound sequences vs. linkers in coding DNA, one for those in non-coding DNA, and one for those at the borders between coding and non-coding DNA. Second, we learn a model that assigns each input sequence probabilities of being coding, non-coding, or at a border region. These probabilities are used as weights on the votings of the component classifiers to obtain the final classification.

We apply this two-stage process to the genome of *S. cerevisiae* using the annotation of nucleosome positions of [Fie+08] obtained by parallel sequencing. Using simple component classifiers based on homogeneous Markov models, we achieve a prediction performance comparable to the approach of [Fie+08], and the prediction performance can be further improved by including additional features such as the length of poly(A/T)-tracts, the number of CTG trinucleotides, and helical properties like persistence length.

References

[Fie+08] Y. Field et al. "Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals". In: *PLoS Comput Biology* 4.11 (2008), e1000216. DOI: 10.1371/journal.pcbi.1000216.

Predicting related traits from SNP markers by multi-task learning

Christoph Lippert[1], Oliver Stegle[1], Stefanie Jegelka[1], Yasemin Altun[1] and Karsten M. Borgwardt[1]

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Advances in sequencing technology have triggered large-scale genotyping initiatives [100a; 100b]. The availability of the genetic background of large sample groups motivates genome-wide association studies which map the phenome of the sample to various genetic loci. A multitude of complex, high-dimensional phenotypes, like physiological measurements, cannot be treated as independent entities. Instead, pleiotropy, i.e. single genetic loci influencing large sets of phenotypes, calls for methods that account for interdependencies of the phenotypic traits.

We address this important challenge using multi-task learning, providing a systematic framework to exploit the correlation structure of the phenome. Our method is scalable, allows large numbers of putative regulators and phenotypes to be modeled in a joint fashion:

$$\arg\min_{\theta,\beta} \sum_{t=1}^{K} \left(\sum_{i=1}^{N} \left(\mathbf{Y}[i,t] - \sum_{j=1}^{K} \mathbf{X}[i,:]\theta[:,j]\beta[j,t] \right)^2 \right) + \lambda_1 \left(\sum_{i=1}^{M} \|\theta[i,:]\|_2 \right) + \lambda_2 \|\beta\|_{\mathrm{F}} \quad (2.1)$$

Here, $\mathbf{Y}$ is a *phenotype matrix* of N individuals $\times$ K phenotypes, $\mathbf{X}$ is a *genotype matrix* of N individuals $\times$ M SNPs, θ is a $M \times K$ matrix that assigns phenotype-specific weights to SNPs and β is a $K \times K$ matrix combining the predictions of correlated phenotypes.

To ensure interpretability of the solutions, we employ a transparent sparsity prior, regularizing the weights θ on a per SNP level. This choice of regularization encourages sparse relations to SNPs while sharing information across correlated phenotypes.

Methodologically, our approach advances the state-of-the-art [Kim+09] by

jointly modeling of the correlation structure of the phenome and the relationships between SNPs and phenotypes. This leads to an optimization problem that is not jointly convex; however, can be efficiently tackled using an alternating procedure.

References

[100a] *1000 Genomes – A Deep Catalog of Human Genetic Variation*. 2009. URL: http://www.1000genomes.org/.

[100b] *1001 Genomes — A Catalog of Arabidopsis thaliana Genetic Variation*. 2009. URL: http://www.1001genomes.org/.

[Kim+09] S. Kim et al. "A multivariate regression approach to association analysis of a quantitative trait network". In: *Bioinformatics* 25.12 (2009), pp. i204–212. DOI: 10.1093/bioinformatics/btp218.

GenDisMix: combining generative and discriminative learning approaches for the recognition of sequence motifs

Jens Keilwagen[1], Jan Grau[2], Stefan Posch[2], Marc Strickert[1] and Ivo Grosse[2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]Martin-Luther-University Halle-Wittenberg, Germany

The recognition of functional sites such as donor or acceptor splice sites, transcription start sites, transcription factor binding sites, or insulator binding sites remains one of the important challenges of genome research. During the last decades, a plethora of different and well-adapted models has been developed, but only little attention has be payed to the development of different and similarly well-adapted learning principles. Only recently it was noticed that discriminative learning principles can be superior over generative ones in diverse bioinformatics applications, too. Here, we propose a generalization of both learning principles, containing each of them as limiting case, which we call generative-discriminative-mixture learning principle (GenDisMix). We investigate theoretical aspects of this learning principle, and we illustrate its efficacy for the recognition of human donor splice sites.

Agnostic RNA-seq Transcriptome Analysis for Quantification of Gene and Transcript Expression

Ronny Amberg[1], Alexander Auner[1], Paul Hammer[1], Gabriele Petznick[1], Chong Wang[1], Andreas Beutler[2] and Peter Beyerlein[1]

[1]Technical University of Applied Sciences Wildau, Germany

[2]Mayo Clinic, Rochester, Minnesota, USA

Next Generation Sequencing machines (e.g. Illumina Genome Analyser) are capable of producing millions of reads (short DNA sequences) in less than a week. More efficient but at the same time less expensive than traditional Sanger sequencingq, these new technologies have motivated many authors to analyse hundreds of genomic samples [Mar+08; Sri+08; WM08]. While the power of the technology has been readily recognized, it has become increasingly clear, that the computational analyses required to arrive at relevant results are highly complex, shifting the frontier of research further from data acquisition to high throughput data analysis. The present study responded to this challenge with the development of an analysis pipeline that can be adopted for a variety of RNA-seq investigations. We have based our pipeline on aligners like Illuminas ELAND as well as our in-house aligner YANALA. We employed the newly developed pipeline in a differential expression study with mRNA obtained from CNS tissue of control rats and of chronic pain rats. Chronic pain is among the most common and most costly medical problems and a leading cause of disability. Currently available analgesic treatments frequently fail or lead to unacceptable side effects [Sto+08]. We analysed 320 million mRNA reads of the length of 50 base pairs (Illumina GA). We extracted exon, gene and transcript expression levels and identifyed a set of yet nonannotated exons and genes.

References

[Mar+08] J. C. Marioni et al. “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays”. In: *Genome Research* 18.9 (2008), pp. 1509–1517. DOI: 10.1101/gr.079558.108.

[Sri+08] A. Srivatsan et al. “High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies”. In: *PLoS Genetics* 4.8 (2008), e1000139. DOI: 10.1371/journal.pgen.1000139.

[Sto+08] B. Storek et al. “Sensory neuron targeting by self-complementary AAV8 via lumbar puncture for chronic pain”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.3 (2008), pp. 1055–1060. DOI: 10.1073/pnas.0708003105.

[WM08] B. Wold and R. M. Myers. “Sequence census methods for functional genomics”. In: *Nature Methods* 5.1 (2008), pp. 19–21. DOI: 10.1038/nmeth1157.

HMM based Identification of Polyglutamine-Islands

Anika Tillich[1], Steffen Pallartz[1], Ronny Amberg[1], Paul Hammer[1], Gabriele Petznick[1], Chong Wang[1], Diego Walther[2] and Peter Beyerlein[1]

[1]Technical University of Applied Sciences Wildau, Germany

[2]Max Planck Institute for Molecular Genetics, Berlin, Germany

The Hidden Markov Model (HMM) is commonly used in bioinformatics to represent sequence motifs. It is the model of choice if variability in terms of insertions, deletions and substitutions is to be represented in a well-defined statistical framework to reflect the uncertain nature of biological structure [Gal+95; Won+04]. We developed a method that automatically labels polyglutamine stretches and polyglutamine islands in nonannotated amino acid sequences. Although polyglutamine stretches are related to a series of neurodegenerative diseases [Bar+07; Ger+94], they have not yet been unambiguously (and quantitatively) defined. To attack this problem a brute-force statistical analysis was performed to generate seed data to be used in the training process of the polyglutamine HMM. With help of the trained HMM we relabeled the seed database and could identify in addition a set of approximately 2000 polyglutamine candidate islands. The analysis of the candidate islands shows their biological relevance as some of them are found to be annotated as part of disease relevant genes or as domain in transcription factors.

References

[Bar+07] S. Barton et al. "The Length Dependence of the PolyQ-mediated Protein Aggregation". In: *Journal of Biological Chemistry* 282.35 (2007), pp. 25487–25492. DOI: 10.1074/jbc.M701600200.

[Gal+95] W. Gales et al. *The Htk Large Vocabulary Recognition System For The 1995 Arpa H3 Task*. 1995.

[Ger+94] H. P. Gerber et al. "Transcriptional activation modulated by homopolymeric glutamine and proline stretches". In: *Science* 263.5148 (1994), pp. 808–811. DOI: 10.1126/science.8303297.

[Won+04] K.-J. Won et al. "Training HMM structure with genetic algorithm for biological sequence analysis". In: *Bioinformatics* 20.18 (2004), pp. 3613–3619. DOI: 10.1093/bioinformatics/bth454.

[You+06] S. J. Young et al. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK. 2006. URL: http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml.

Using micro arrays to detect natural variation in hormone induced expression changes

Yvonne Pöschl[1], Carolin Delker[2], Marcel Quint[2] and Ivo Grosse[1]

[1]Martin-Luther-University Halle-Wittenberg, Germany

[2]Leibniz Institute of Plant Biochemistry (IPB), Halle, Germany

Natural variation is the phenomenon that not all members of a species respond equally to environmental stimuli. Even though the individuals are highly conserved on sequence level they can show differences in numerous physiological traits such as flowering time, growth and pathogen responses. Phytohormones like auxin are critical for physiological responses to a multitude of internal and external stimuli. To assess the differences in response to auxin on a molecular level a transcriptome analysis of seven *Arabidopsis* ecotypes was performed using ATH1 micro array chips.

We investigate which of the genes coding for known components of the auxin signaling network show different expression patterns among the seven ecotypes using the modified t-test for small sample sizes from [ORS07], and we find that a surprisingly high fraction of these genes is regulated differently despite their coding sequences are highly conserved. A co-expression analysis by using the Local Context Finder (LCF) [KG03] shows that the regulation of the auxin signaling network differs considerably among the ecotypes, suggesting that also the downstream response might be different. To test this hypothesis, we perform a combined cluster and LCF-network analysis based on the whole-genome expression data, and we find differences in the transcriptional response of the seven *Arabidopsis* ecotypes to auxin. These findings indicate that, despite the high degree of sequence conservation among the seven ecotypes studied, the transcriptional regulatory networks are different, demonstrating naturally occurring variation on the molecular level.

References

[KG03] F. Katagiri and J. Glazebrook. "Local Context Finder (LCF) reveals multidimensional relationships among mRNA expression profiles of Arabidopsis responding to pathogen infection". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.19 (2003), pp. 10842–10847. DOI: 10.1073/pnas.1934349100.

[ORS07] R. Opgen-Rhein and K. Strimmer. "Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007), p. 9. DOI: 10.2202/1544-6115.1252.

MotifAdjuster: a tool for computational reassessment of transcription factor binding site annotations

Jens Keilwagen[1], Jan Baumbach[2], Thomas Kohl[3,4] and Ivo Grosse[5]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]International Computer Science Institute, Berkeley, USA

[3]International NRW Graduate School in Bioinformatics and Genome Research, Center for Biotechnology (CeBiTec), Bielefeld University, Germany

[4]Institute for Genome Research and Systems Biology (IGS), Center for Biotechnology (CeBiTec), Bielefeld University, Germany

[5]Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Germany

Valuable binding site annotation data are stored in databases. However, several types of errors can, and do, occur in the process of manually incorporating annotation data from the scientific literature into these databases. Here, we introduce the open source software MotifAdjuster (`http://dig.ipk-gatersleben.de/MotifAdjuster.html`), a tool that helps to detect these errors, and we demonstrate its efficacy on public data sets.

How to assess de-novo motif discovery approaches?

Jens Keilwagen[1], Jan Grau[2], Stefan Posch[2] and Ivo Grosse[2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
[2]Martin-Luther-University Halle-Wittenberg, Germany

De-novo motif discovery has been in the focus of genome research for more than a decade, but none of the existing approaches are fully satisfactory, and the recognition of short motifs such as donor or acceptor splice sites, splicing enhancers or silencers, translation initiation sites, transcription start sites, transcription factor binding sites, nucleosome binding sites, miRNA binding sites, or insulator binding sites remains a challenging problem of bioinformatics today. Unbiased assessments and systematic comparisons of all of these approaches on different data sets are crucial for further progress, but interestingly this task is highly nontrivial and has been neglected in the past. Here, we discuss different alternatives for the evaluation of motif discovery algorithms including different ways of choosing appropriate test data sets, of defining correct locations of binding sites, and of selecting adequate performance measures.

tRNA Cluster

Clara Bermudez-Santana[1], Camille Stephan-Otto Attolini[2], Toralf Kirsten[3], Sonja J. Prohaska[1], Stephan Steigele[4], Jan Engelhardt[1] and Peter F. Stadler[1]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

[2]Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, USA

[3]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

[4]Genedata AG, Basel, Switzerland

Surprisingly little is known about the organization and distribution of tRNAs and tRNA-related sequences in whole genome data [GP06]. Occasionally, tandem arrangements of tRNAs have been observed, e.g. in *Entamoeba histolytica* [Taw+08]. We therefore set out to compare the genomic locations tRNA of a wide range of taxonomic groups to identify common patterns and taxon-specific peculiarities. Based on the efficient and accurate detection of tRNA by `tRNAscan-SE` [LE97] we developed a pipeline to extract and analyze tRNAs locations to screen systematically eukaryotic genomes.

The distributions of tRNA numbers are very broad, with standard deviations on the order of the mean. There are, however, several significant outliers, such as the *sea-anemone* with its 88.73% clustered tRNAs, or *Trichoplax adhaerans* without any tRNA clusters. In teleosts, only in *zebrafish* 69% of tRNAs are clustered and in *tiddler* its 87%. In Mammalian species, where tRNA numbers can reach 170.000, only a 23% of the loci are organized in clusters, while this percentage increases to 40% in the armadillo. In higher primates, the average is 616,120(SD) tRNAs of which 17% to 36% clustered, the exception being *Otolemur* where only 5.6% of its 45225 tRNAs are clustered. Not surprisingly, species with high copy numbers of tRNAs often tend to have extensive genomic clusters.

Both copy numbers and peculiarities of tRNA arrays appear to vary dramatically at very short evolutionary time-scales suggesting that eukaryotic tRNA distribution is more complex and diversified than previously considered.

References

[GP06] J. M. Goodenbour and T. Pan. "Diversity of tRNA genes in eukaryotes". In: *Nucleic Acids Research* 34.21 (2006), pp. 6137–6146. DOI: 10.1093/nar/gkl725.

[LE97] T. M. Lowe and S. R. Eddy. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence". In: *Nucleic Acids Research* 25.5 (1997), pp. 955–964. DOI: 10.1093/nar/25.5.955.

[Taw+08] B. Tawari et al. "Patterns of Evolution in the Unique tRNA Gene Arrays of the Genus Entamoeba". In: *Molecular Biology and Evolution* 25.1 (2008), pp. 187–198. DOI: 10.1093/molbev/msm238.

Rapid and Accurate Semi-Global Alignment of Diverged Sequencing Reads

Udo Stenzel[1]

[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Applications of modern high-throughput sequencing techniques typically require the alignment of millions of reads with non-trivial error profiles to a suitable reference genome. Presently this is usually done using either local alignments from a heuristic similarity search program such as *megablast* [Zha+00], or a fast mapping program such as bwa [LD09] specialized to the matching of short sequences with low divergence, as encountered in resequencing projects. In our experience, both kinds of tools are unsatisfactory when applied to rather short DNA sequences with higher divergence from the reference and possible chemical modifications.

Here we present Anfo, a program and associated post-processing tool set, that is designed to quickly and accurately find the best alignment of short sequences to genome-sized databases, even in the presence of miscoding lesions and substantial divergence. Unlike fast mappers, Anfo produces genuine alignments including affinely scored gaps and a variable number of mismatches, and is guaranteed to find the best such alignment. Anfo implements a best-first search strategy, which optimizes search times in the common case, since it is fastest when alignments for most sequences are expected to have only few differences. In our experience, Anfo is at least an order of magnitude faster than *megablast* for the mapping of shotgun sequencing reads. Full support for IUPAC ambiguity codes and a flexible alignment algorithm adaptable to complex models of both errors and evolution make Anfo uniquely suited to the mapping of chemically modified DNA, such as in studies of ancient DNA or sequencing of bisulphite treated samples in studies of DNA methylation.

References

[LD09] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[Zha+00] Z. Zhang et al. "A Greedy Algorithm for Aligning DNA Sequences". In: *Journal of Computational Biology* 7.1-2 (2000), pp. 203–214. DOI: 10.1089/10665270050081478.

Apples and oranges: avoiding different priors in Bayesian DNA sequence analysis

Jens Keilwagen[1], Jan Grau[2], Stefan Posch[2] and Ivo Grosse[2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

[2]Martin-Luther-University Halle-Wittenberg, Germany

One of the challenges of bioinformatics remains the recognition of functional sites in genomic DNA as for instance donor or acceptor splice sites, splicing enhancers or silencers, transcription factor binding sites, miRNA binding sites, or insulator binding sites. During the last decade, a wealth of algorithms for the recognition of such sites has been developed and compared. While in many studies different statistical models or different learning principles have been analyzed, the influence of choosing different prior distributions for the model parameters has been overlooked, leading to questionable conclusions. With the goal of allowing direct comparisons of different models from the family of Markov Random Fields and different learning principles based on the same prior, we derive a generalization of the commonly-used product-Dirichlet prior. We compare the derived prior with other existing priors and illustrate its utility for a direct comparison of different models and different learning principles for the recognition of human donor splice sites and binding sites of the transcription factor Sp1.

Combining Phylogenetic Footprinting with Realistic Motif Models

Martin Gleditzsch[1], Stefan Posch[1] and Ivo Grosse[1]

[1]Martin-Luther-University Halle-Wittenberg, Germany

The identification of cis-regulatory modules is one of the prerequisites for improving our understanding of gene regulation. Several different algorithmic approaches exist, but none of them is fully satisfactory. For example, traditional algorithms for de-novo discovery of cis-regulatory modules ignore information stemming from promotors of orthologous genes of evolutionary related species, whereas phylogenetic footprinting and phylogenetic shadowing algorithms, which use this information, are based on motif models that ignore the presence of statistical dependencies among different motif positions. Here, we present an approach that combines both ideas, and we apply the resulting phylogenetic footprinting algorithm based on Bayesian trees, PhyloBT, to target promoters of Cat8 in *Saccharomyces cerevisiae* and its relatives. Cat8 activates glucose inhibited genes and is therefore one of the key players in the evolutionarily conserved regulatory network of Snf1. In contrast to the wealth of knowledge about Cat8, only little is known about its binding sites, and existing algorithms could not provide acceptable predictions. Guided by preliminary data suggesting that Cat8 binding sites are evolutionarily conserved *and* exhibit strong statistical dependencies among non-neighboring motif positions, we choose a set of approximately 100 experimentally determined target promoters of Cat8 as example for testing the efficacy of PhyloBT.

Functional Characterization of Hepatitis C Virus Host Factors identified by RNA Interference

Sarah Diehl[1], Hagen Blankenburg[1], Fidel Ramírez[1], Ilka Wörz[2], Thomas Lengauer[1], Ralf Bartenschlager[2] and Mario Albrecht[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany

[2]University of Heidelberg, Germany

The hepatitis C virus (HCV) is the major causative agent of chronic hepatitis with 180 million infected patients, 3% of the world's population. Over 70% of whom are chronic HCV carriers at high risk of developing liver cirrhosis and hepatocellular carcinoma. The current standard therapy takes 48 weeks and consists of a combination of peginterferon and ribavirin. This treatment, however, shows significant side effects and a limited success rate of 50%. New antiviral drugs directly targeting the virus have been applied successfully in clinical trials, but HCV rapidly develops drug resistance due to the error-prone nature of the viral RNA polymerase.

A promising alternative approach to controlling HCV infection is the search for host cell factors required by the viral life cycle in human liver cells. The functional characterization of these factors will aid in elucidating the cellular processes involved with HCV infection and may point to new drug targets. Importantly, targeting host factors may pose a major obstacle for the virus to overcome by resistance development.

In experiment, specific host genes can be effectively silenced by RNA interference (RNAi). Therefore, we conduct a comprehensive bioinformatics analysis of experimental data obtained by large-scale RNAi screens to discover human genes that are particularly vital for HCV entry and replication. To support the biological interpretation of the RNAi results, we implement novel computational methods and integrate functional annotations, expression profiles, protein interaction data, molecular pathways, and further knowledge about viral host factors.

Acknowledgment: We would like to acknowledge the work of further colleagues involved with the RNAi screens in Heidelberg: Holger Erfle, Petr Matula, Lars Kaderali, Marion Pönisch, Karl Rohr, Roland Eils, Artur Kaul, Sandra Bühler, Rainer Pepperkok, Volker Lohmann.

Crows Nest, a synteny driven plant comparative genome framework component

Stephan Karl Rößner[1] and Klaus F. X. Mayer[1]

[1]Helmholtz Center Munich, Germany

We are now entering a new era in plant genomics. Besides the already sequenced plant genome there is a drastically increasing number of genomes in the pipeline. Revolutionary advances in sequencing technology and methods have caused a rapid accumulation in plant genetic and genomic data for a multitude of different species. The availability of intuitive comparative genome mapping and visualization tools have become an invaluable resource to investigate genetic and genomic data effectively. Although there are several comparative mapping tools available, such as the ensemble genome browser (`http://www.ensembl.org`) or CMap (`http://www.gmod.org/cmap`), they operate on specific data types and visualize the results according to the different genome types. Here we describe CrowsNest, a whole genome comparative mapping and visualization tool between genetic, genomic and FPC data of plants. The comparative map viewer is a web-based community resource and is integrated into a Plant comparative genome framework. CrowsNest is specifically designed to visualize synteny at macro and micro levels. It allows to intuitively explore macro- and microsynteny and to transfer knowledge between several plant species. CrowsNest is an integral part of PlantsDB, a database which was specifically designed to build a platform for integrative and comparative plant genome research [Spa+07]. AVAILABILITY of CrowsNest: `http://mips.helmholtz-muenchen.de/plant/crowsNest/index.jsp`.

References

[Spa+07] M. Spannagl et al. “MIPSPlantsDB–plant database resource for integrative and comparative plant genome research”. In: *Nucleic Acids Research* 35.suppl_1 (2007), pp. D834–840. DOI: 10.1093/nar/gkl945.

LipidX: a truly platform independent lipid analysis suite

Ronny Herzog[1], Dominik Schwudke[2], Michael Schroeder[3] and Andrej Shevchenko[1]

[1]Max Planck Institute of Molecular and Cell Biology and Genetics (MPI-CBG), Dresden, Germany

[2]National Center for Biological Sciences, Bangalore/Bengaluru, India

[3]Biotechnology Center (BIOTEC), Dresden University of Technology, Dresden, Germany

Lipidomics aims at large-scale study of the cellular lipid complement. Today, lipids are recognized to play an important role in many metabolic diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes. The field is highly driven by recent advances in the technology of mass spectrometry. A direct infusion of total lipids mixtures in the mass spectrometer, called shotgun lipidomics, dramatically enhances the speed of the acquisition, especially for a large number of samples. Despite becoming a popular approach, shotgun lipidomics still suffers from a lack of available software for data analysis and interpretation.

We present a free software framework for shotgun lipidomics, called LipidX. It features mass spectrometric alignment of multiple samples and import into a flat-file database. In addition, LipidX accounts for machine-dependent variables such as accuracy, resolution, resolution change over m/z, etc providing users with a true platform-independent data analysis suite.

LipidX's lipid interpretation is based on the in-house developed Molecular Fragmentation Query Language (MFQL). To our knowledge, this is the first time that a query language is applied to interpret mass spectra and identify lipids. MFQL enables the emulation of the major types of shotgun lipidomics applications: high-throughput screening, data-dependent experiments and targeted quantitative analysis. Given that common database-based approaches hardly contain every possible mass spectrometer- and acquisition setting-dependent lipid fragmentation pattern, LipidX overcomes this issue using its inherently platform-independent, MFQL-based, de-novo lipid interpretation.

LipidX was applied to several biological studies, such as the lipidomic manifestations of metabolic syndrome-related disorders [Gra+09]; to functional annotation of lipid metabolism-related genes in *C. elegans*; and to the quantitation of rare tetradecasphing-4,6-dienine containing sphingolipids of *D.*

melanogaster. We also compared our approach with the softwares LipidQA and LipidProfiler to verify its correctness.

References

[Gra+09] J. Graessler et al. "Top-Down Lipidomics Reveals Ether Lipid Deficiency in Blood Plasma of Hypertensive Patients". In: *PLoS ONE* 4.7 (2009), e6261. DOI: 10.1371/journal.pone.0006261.

[Sch+05] D. Schwudke et al. "Lipid Profiling by Multiple Precursor and Neutral Loss Scanning Driven by the Data-Dependent Acquisition". In: *Analytical Chemistry* 78.2 (2005), pp. 585–595. DOI: 10.1021/ac051605m.

[Sch+07] D. Schwudke et al. "Top-Down Lipidomic Screens by Multivariate Analysis of High-Resolution Survey Mass Spectra". In: *Analytical Chemistry* 79.11 (2007), pp. 4083–4093. DOI: 10.1021/ac062455y.

Scoring geometries of protein-protein complexes

Florian Krull[1], Myong-Ho Chae[1] and Ernst-Walter Knapp[1]

[1]Free University of Berlin, Germany

Interactions between proteins are known to play a central role in many biological processes. However, it is often easier to obtain structural information of the individual proteins instead of the whole protein complex. Hence, there is need for theoretical procedures that can predict protein complex structures given the structures of the individual proteins. Many algorithms try to solve this problem. The currently used methods typically generate many possible protein complex geometries (decoys) considering shape complementarity only. To discriminate near-native geometries from false positives these decoys are ranked by a scoring function. To train learning-based scoring functions we generate decoys primarily for the purpose of training. We demonstrate the impact of carefully chosen sets of training decoys on the performance of two different learning-based scoring methods.

Modelling of Biotechnological Processes

Karl-Michael Meiß[1], Wolfgang Eisenberg[1] and Stefan Fiedler[1]

[1] Arnold-Sommerfeld-Gesellschaft e.V. Leipzig, Germany

Our planned project at controlled modelling of biotechnological processes is using the production of Poly-3-hydroxybutyrate (PHB) in an exemplary way. It can be divided into the tree steps: i fermentation, ii extraction, iii modification (FEM). Especially in this project a virtual microorganism (VMO) is used as an integrated model representing the microbial attitude during growth phase and accumulation phase of the storage material PHB. The characteristics of the VMO are compatible adapted to the empirical data of the bioproduction of PHB and describe the properties in relevance of the tree steps. The "model-handling" of the VMO concerns all process steps. With our fermentation model we are able to estimate usage of substrate and production of PHA in time and space precisely.

Further, we use a process-adapted ensemble of simulation models and methods: network methods, event driven simulation, cross catalytic networks and cellular automata — generic simulated in a computing grid. So we are able to describe the network structure of a complex biotechnological system, the effects of emergence and the predictability of the systems behaviour and impacts of extreme events to the complex system. Usually, biochemical processes are simulated by use of molecular dynamics and grid computing techniques. In contrast to that we will use process-adapted methods. We expect that our ensemble of methods is generalizable and upgradable and can be applied for varios biosystems in future.

References

[Mei+03] K.-M. Meiß et al. *Implementationsstudie zur biotechnologischen Produktion von Biopolymeren unter Einsatz digitaler Modelle auf der Basis nachwachsender Rohstoffe und organischer Abfälle*. Forschungsbericht 200 66 302, UBA-FB 000455. 2003. URL: `http://www.umweltbundesamt.de/uba-info-medien/dateien/2303.htm`.

[Mei05] K.-M. Meiß. *Der virtuelle Mikroorganismus*. 2005. URL: `http://www.dr-meiss.de`.

RNA Sequence Design, Newtonian Dynamics and Mean Fields

Marco Matthies[1], Stefan Bienert[1] and Andrew E. Torda[1]

[1]Center for Bioinformatics, University of Hamburg, Germany

Quantum chemists often think they have a monopoly on self consistent mean eld methods. Classical simulators believe they own Newtonian dynamics. Both of these methods can be put to use or perverted in other contexts.

Imagine you have a favourite RNA molecule. You are not allowed to change its shape (base-pairing) but you would like to re-design its sequence, hoping to find a molecular biologist who can synthesise the molecule for you. This is a discrete problem, but there are methods to find ones way through the $4n$ possible sequences as if it were a continuous space. Each site can be a vector of probabilities or coordinates in a 4-dimensional space [A, C, G, U]. Given a literature energy model, the bases interact with their base-pair and sequence neighbours. With a mean-field approach, one cycles between probability and energy updates. With a dynamics approach, one gives each site a mass, initial velocity and lets the particles roam as they please while being cooled.

Following literature tradition, we treat RNA as a flat molecule and only treat base-pairing. Within this framework, the problem is much easier than protein design and the methods work remarkably well. One can even build "negative design" directly into the energy function and the methods are not upset when they are fed pseudo-knots. By all computational measures, the resulting sequences appear to be better than those in nature. The reader can decide whether to blame nature or the ugly quasi-energy functions.

Expression QTL Infrastructure

S. Möller[1], H. N. Krabbenhöft[2], A.-K. Grimm[2], B. Bauer[2] and S. Ibrahim[2]

[1]Debian Linux Community

[2]Institute for Neuro- and Bioinformatics, University of Luebeck, Germany

The analysis of genotype-phenotype relationships in complex diseases follows the biologically motivated trend towards molecular phenotypes. This led to the development of quantitative trait loci (eQTL) for gene expression levels. The here presented "Expression QTL Infrastructure" is a template for the seeding dynamic web sites for single eQTL projects. The platform organises the computational analysis, presents the findings to biologists, and may possibly be added as functional supplement to publications.

This setup allows for the submission of jobs to computational grids or local queueing systems. Results are collected in a MySQL database, presented as web pages, and via web-services. The latter allow for the direct access by workflow tools like Taverna. This caters for extra feedback loops for biological users, who may desire to optimise parameters of the computation to further investigate the role of covariates or interacting effects between loci or genes.

Availability: `http://eqtl.berlios.de` (AGPL-3)

Modular Modeling with ProMoT in Systems Biology

Katrin Kolczyk[1], Sebastian Mirschel[1], Michael Rempel[1] and Ernst Dieter Gilles[1]

[1]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

In recent years Systems Biology has become a massive information science, driven by data that is produced by powerful high throughput technologies. Based on these data, models are growing fast in terms of size, number and complexity. Thus, modular modeling concepts become more and more attractive.

Our open-source modeling tool ProMoT [Mir+09] facilitates efficient and comprehensible setup and editing of modular models coupled with customizable visual representations. ProMoT has its origins in chemical process engineering but in recent years more and more functionality has been incorporated to handle models in system biology. The key features of ProMoT comprise (i) the modular modeling which can be done in an object-oriented fashion based on extensible modeling libraries, (ii) the exchange of models with other tools via the international standard format SBML and (iii) advanced visual capabilities.

ProMoT supports different modeling approaches. Quantitative models can be based on differential-algebraic equations, qualitative models on a recently introduced boolean logic modeling formalism. Both types of models can be structured and modular. This results in several advantages such as reusable flexible modules. By decomposing into modules huge and complex models can be handled. The basic concepts, main GUI components of ProMoT and ongoing projects will be presented. Up-to-date versions of the software, installation instructions and several tutorials can be obtained on ProMoT's website (`http://www.mpi-magdeburg.mpg.de/projects/promot`).

References

[Mir+09] S. Mirschel et al. "ProMoT: modular modeling for systems biology". In: *Bioinformatics* 25.5 (2009), pp. 687–689. DOI: `10.1093/bioinformatics/btp029`.

Authors of Accepted Short Papers

Authors of Accepted Posters

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S