



Department of Computational Molecular Biology

(Established: 10/2000)



Head

Prof. Dr. Martin Vingron
Phone: +49 (0)30 8413-1150
Fax: +49 (0)30 8413-1152
Email: vingron@molgen.mpg.de

Secretary

Birgit Löhmer
Phone: +49 (0)30 8413-1151
Fax: +49 (0)30 8413-1152
Email: vinoffic@molgen.mpg.de

Scientific assistant

Dr. Patricia Marquardt (since 11/01,
part time)
Phone: +49 (0)30 8413-1716
Fax: +49 (0)30 8413-1671
Email:
patricia.marquardt@molgen.mpg.de

IMPRS coordinator

Dr. Hannes Luz (since 01/07)
Phone: +49 (0)30 8413-1716
Fax: +49 (0)30 8413-1154
Email: luz@molgen.mpg.de

Computer systems administrator

Wilhelm Rüsing

Group leaders of the Department

Dr. Peter Arndt (since 10/03)
Dr. Stefan Haas (since 01/01)
Dr. Sebastiaan Meijnsing (since 09/09)
Dr. Rainer Spang (09/01-12/06)
Dr. Alexander Schliep (05/02-06/09)
Dr. Eike Staub (09/03-02/06)
Dr. Roland Krause (01/05-05/08)

Introduction

Computational biology studies biological questions with mathematical and computational methods. In the area of molecular biology and genomics, the possibility to apply such formal methods, of course, comes from the availability not only of genome sequences, but also of large amount of functional data about biological processes. Computational molecular biology encompasses both development and adaptation of methods in the areas of mathematics, statistics, and computer science, as well as pursuing biological questions applying these tools and close collaborations with experimentalist. In the context of the MPI for Molecular Genetics, computational approaches have become an integral part of most of the research projects pursued.

The research interest of the Computational Molecular Biology Department lies in understanding gene regulatory mechanisms as well as structure and evolution of the eukaryotic genome. To this end, mathematical, computational, and also experimental approaches are being developed and employed. The department is struc-

tured into several research groups, the largest of which is the *Transcriptional Regulation Group* headed by Martin Vingron. The work of this group focuses on theoretical concepts in the prediction of cis-regulatory elements, gene regulatory networks and epigenetic aspects of regulation. Peter Arndt heads the *Evolutionary Genomics Group* which works on developing models how the DNA in primates has evolved. Stefan Haas is heading the *Gene Structure and Array Design Group*. The focus of this group is on transcriptomics and gene structure. As of September 2009, Sebastiaan Meijsing has been building up the *Mechanisms of Transcriptional Regulation Group*, a new, experimental group that will work on transcription factors and protein-DNA interaction.

The current structure and focus of the department differs from what it used to look like a few years ago. In the beginning – the department was founded in 2000 – the groups in the department worked on a broad spectrum of questions ranging from protein evolution to microarray data analysis. The latter was the topic of the *Computational Diagnostics Group* of Rainer Spang, who left end of 2006 for a professorship in Regensburg. His group dealt with microarray data analysis in the context of cancer profiling. The *Algorithmics Group* headed by Alexander Schliep developed machine learning algorithms for pattern recognition in a number of biological problems. Alexander Schliep left in 2009 to take a professorship at Rutgers University, New Jersey. From January 2005 until May 2008 Roland Krause led a group on *Microbial Virulence*, linking the MPI for Molecular Genetics and the MPI for Infection Biology. The *Protein Families and Evolution Group* was headed by Eike Staub who left for industry in February 2006. For the Algorithmics Group this report contains a summary of its activities, while the other groups had been described in earlier research reports.

In contrast to this diversified structure of the department, the last couple of years have brought about an increased focus and concentration. The existing groups share a general interest in mechanisms and evolution of gene regulation. In fact, the groups of Arndt, Haas and Vingron now cooperate very closely and have recently been joined by the experimental lab headed by Meijsing, who also works on transcriptional regulation. The general goal of the department is to elucidate biological processes and mechanisms, albeit based more on theoretical and computational methods than on experimental ones, and, to this end, apply the most adequate and state-of-the-art mathematical and computational techniques.

Some significant research results of the last years are:

- Development of biophysically motivated prediction methods for transcription factor binding sites and determination of tissue-specific transcription factors;
- Development of an analysis pipeline for next generation sequencing data;
- Comprehensive description of the statistics of transcription factor binding site prediction;
- Identification of strand-specific patterns of mutagenesis due to the process of transcription;
- Evolutionary model explaining long-range correlations in the genome.

The study of gene regulation and of structure and evolution of the genome is currently under rapid development. In particular, it has become increasingly clear that in addition to transcription factors, chromatin structure plays an important role for regulation and for understanding the genome in general. We are currently in the process of extending in the direction of theoretical studies of this so-called epigenetic regulation. At the same time, evolution is shaping the genome and the



regulatory networks, just as it is shaping the protein world. Thus, we are also working on extending our earlier studies on evolution of gene families to the study of evolution of gene regulation.

Members of the department come from various backgrounds, ranging from mathematics, statistics and computer science *via* physics to biology and genetics. The largely theoretical type of work of course relies heavily on powerful computers. The *computer equipment* of the department comprises PCs under Linux, a cluster of 32-bit processors and a cluster of 64-bit Opteron processors, largely for sequence analysis and typical bioinformatics applications. Recently, two new 32-processor, 64-bit computers with large memory were acquired for numerical computations and for processing the new sequencing data. Several RAID arrays with together more than 95 TB of disk space are available. The computer set-up is maintained by the department system administrator, Willi Rüsing, in close cooperation with the institute computing unit.

Department members contribute substantially to the *bioinformatics curriculum* at Free University of Berlin. We teach a number of courses and offer students to do internships, practical courses, and thesis work with us. This brings many bright, young students to the department and at the same time allows the university to show the students a much larger spectrum of bioinformatics than would normally be possible in the university framework. In cooperation with the university we have established an *International Max Planck Research School on Computational Biology and Scientific Computing* (IMPRS-CBSC, <http://www.imprs-cbse.mpg.de/>). In 2009, this IMPRS was reviewed and recommended for an extension. With funding from the Max Planck Research Award we have initiated the *International Otto Warburg Summer Schools*. The schools bring together international lecturers with a select group of national and international students and combine lecture-style teaching and research seminars to give an overview of new important areas in computational biology. This initiative is being continued in spite of the expiration of the funds from the award.

During the reporting period we have been involved in a number of *national and international projects* and collaborations. On a national level, we are part of BMBF funded project in the context of NGFN, the National German Network on Genome Research, of an SFB and a graduate school funded by DFG, and of a project funded by Volkswagenstiftung. On an international level, we are participating in several EU projects. A number of research visitors are financing their stay at the department from fellowships from Alexander von Humboldt Foundation or DAAD (German Academic Exchange Service). From 2001 to 2006, significant funding to the department came from the BMBF-funded Berlin Center for Genome Based Bioinformatics (BCB), a large network made up of several Berlin bioinformatics groups and coordinated by Martin Vingron. Since 2006 Martin Vingron has also acted as one of the directors at the newly founded CAS-MPG Partner Institute for Computational Biology in Shanghai, China. He spends around six weeks a year there, spread over six individual visits roughly every other month. In terms of service to Max Planck Society, Vingron has also been a member of the BAR commission on purchase of large computing equipment in MPG and of several other ad-hoc committees. He has acted as managing director of MPIMG from 2003-2008. In 2004, Vingron was elected to the National German Academy Leopoldina and, in the same year, also received the Max Planck Research Award. In 2009 he became chair of the steering committee of the RECOMB conference series, a renowned international conference on computational biology.

Evolutionary Genomics Group

(Established: 10/2003)



Head

Dr. Peter Arndt (since 10/03)
Phone: +49 (0)30 8413-1162
Fax: +49 (0)30 8413-1152
Email: arndt@molgen.mpg.de

Scientist

Dr. Brian Cusack (since 06/08)

PhD students

Federico Squartini (since 10/05,IMPRS)
Paz Polak (since 10/06, IMPRS)
Yves Clement (since 10/08, IMPRS)
Philipp Messer (02/05 – 03/08)

Scientific overview

Unraveling the evolutionary forces responsible for variations of neutral substitution patterns among taxa or along genomes is a major issue for detecting natural selection within sequences. The genomes of many species (and of individuals within a species) have been sequenced today. This gives us the unprecedented opportunity for a quantitative analysis of this data with respect to evolutionary aspects. Due to advances in next generation sequencing technologies this is possible with more power and precision than before.

We use both comparative genomics and the study of genomic fossils (e.g. retroviral sequences, pseudo genes) to learn more about the processes that shape the human genome and the genomes of other species. We investigate processes on short length scales, e.g. nucleotide substitutions, insertions and deletions and long length scales, e.g. insertions of repetitive elements and duplications. Our analysis is complemented by studies of the mathematical underpinnings of models for nucleotide substitutions and phylogeny as well as experimental approaches to study selection *in vitro*.



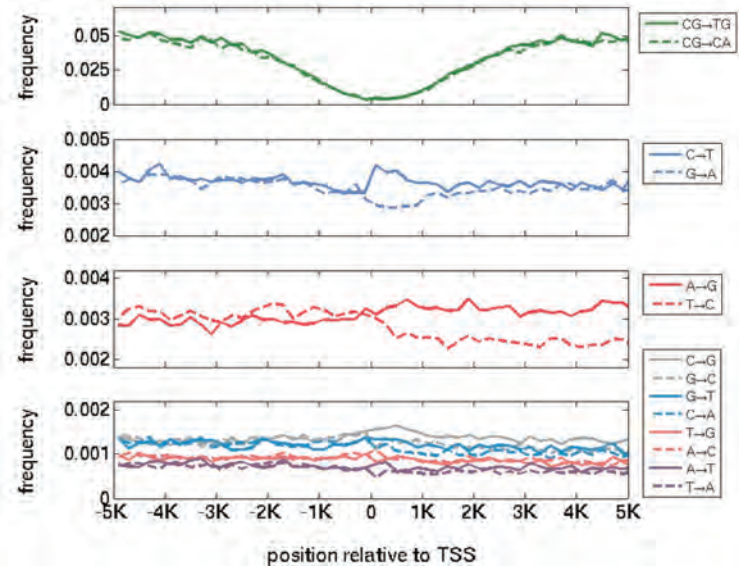
Comparative analysis of nucleotide substitutions

[Clement, Polak, Arndt]

Mammalian genomes show large-scale regional variations of GC-content (the isochors), but the substitution processes at the origin of this structure are poorly understood. It has been shown that meiotic recombination has a major impact on substitution patterns in human, driving the evolution of GC-content.

But also other cellular processes have an influence on nucleotide substitutions. A regional analysis of nucleotide substitution rates along human genes and their flanking regions allowed us to quantify the effect of mutational mechanisms associated with transcription in germ line cells. Our analysis revealed three distinct patterns of substitution rates. First, a sharp decline in the deamination rate of methylated CpG dinucleotides, which is observed in the vicinity of the 5' end of genes. Second, a strand asymmetry in complementary substitution rates, which extends from the 5' end to 1 kbp downstream from the 3' end, associated with transcription-coupled repair. Finally, a localized strand asymmetry, an excess of C->T over G->A substitution in the non-template strand confined to the first 1-2 kbp downstream of the 5' end of genes. We hypothesize that higher exposure of the non-template strand near the 5' end of genes leads to a higher cytosine deamination rate.

Recently we also established that the presence of CpG Island has an asymmetric influence on nucleotide substitution up and downstream indicative of a cellular process that starts at a CpG Islands and moves outwards beyond its 5' and 3' end.



Substitution rates in introns and in intergenic regions in the vicinity of 5' end of human genes. The panels show the estimated 12 single-nucleotide substitution rates and the CpG deamination rates in non-overlapping 200-bp-long windows along the nontemplate strand.

Models of genome evolution

[Arndt, Bielow, Engleitner, de la Chaux, Messer]

In the recent past it became clear that besides nucleotide substitutions also the insertion and deletion of short pieces of DNA as well as the insertion of repetitive elements have a substantial influence on the evolution of GC isochors in mammals.

We have shown that simple expansion randomization systems (ERS) are able to generate long-range correlation of the GC content, which is one of the hallmarks of isochors. A wide range of such ERSs fall within one universality class and the characteristic decay exponent of the correlation function can easily be calculated from the rates of the underlying processes. This result gives us also a simple method to simulate long-range correlated sequences and recently we were able to quantify the influence of such correlations on the alignment statistics of sequence, which turned out to be quite substantial. Corresponding corrections should be taken into account when calculating p-values for the alignment of genomic sequences. At the basis of expansion randomization systems are processes that duplicate, insert, or delete segments of a sequence.

Comparing the genome of humans to the ones of its closest relatives, the chimpanzee and rhesus monkeys, gave us the opportunity to investigate instances of nucleotide insertions and deletions on small scales and quantify their rates in the genomic context. In the future we also want to extend our analysis and also include

insertion of repetitive elements into the vertebrate genomes. In particular we want to understand the quantitative differences in variations of the GC-content between hominoids and rodents. At the end we will generate a much richer null model of genomic evolution, especially for the evolution of promoter regions, which often include multiple binding sites for the same transcription factors.

In vitro selection

[Arndt]

The advancements of next generation sequencing technologies give us a novel tool for the quantitative analysis of Systematic Evolution of Ligands by Exponential Enrichment (SELEX) experiments. Such experiments are conducted in close collaboration by the Glökler group (Dept. Lehrach). Starting from a highly diverse pool of DNA sequences ligands to particular molecules, e.g. transcription factors or other cellular relevant molecules are enriched through subsequent rounds of selection. In-house sequencing capabilities give us the opportunity to sequence the DNA pools after each round of selection. This way we are going to study the dynamics of selection for strong binding ligands in lieu of a highly diverse background of unspecific ligands. Since very high diversities can be charted using Illumina sequencing we will also be able to study non-dominant secondary clones and follow the dynamics of their frequency in the population during rounds of selections. New approaches to cluster and analyze the clonal structure of synthetic sequence pools have to be developed.

Mathematics of evolutionary models

[Arndt, Squartini]

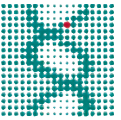
Markov models describing the evolution of the nucleotide substitution process, widely used in phylogeny reconstruction, usually assume the hypotheses of stationarity and time reversibility. Although these models give meaningful results when applied to biological data, it is not clear if the two assumptions mentioned above hold and, if not, how much sequence evolution processes deviate from them. To this aim, we introduced two sets of indices that can be calculated from the nucleotide distribution and the substitution rates. The stationarity indices (STIs) can be used to test the validity of the equilibrium assumption. The irreversibility indices (IRIs) are derived from the Kolmogorov cycle conditions for time reversibility and quantify the degree of non-time-reversibility of a process. Computations of these indices for genomic nucleotide substitutions in *Drosophila simulans* and *Homo sapiens* reveal statistically significant deviations from the ideal case of a process that has reached stationarity and is time reversible.

Phenotypic mutations

[Arndt, Cusack]

Recent studies have hinted at the importance of “phenotypic mutations” (errors made in transcription and translation) in molecular evolution. These are thought to facilitate positive selection for adaptations that require multiple-substitutions but the generality of this phenomenon has yet to be explored.

Our research in this area focuses on the importance of phenotypic mutations to negative selection and to the maintenance of genomic robustness by selective constraint. We initially approached this in the context of Nonsense Mediated Decay (NMD)-based surveillance of human gene transcription. We have discovered a pattern of codon usage in human genes that compensates for the variable NMD efficiency by minimizing nonsense errors during transcription. Our future work will focus on whether phenotypic mutations due to other types of mis-transcription constitute a similar selective force.



Selected information

Selected publications

Polak P. and Arndt P.F. (2008): *Transcription induces strand-specific mutations at the 5' end of human genes*. *Genome Research* 18 (8), 1216-23

Duret L. and Arndt P.F. (2008): *The impact of recombination on nucleotide substitutions in the human genome*. *PLoS Genet* 4(5), e1000071

Squartini F. and Arndt P.F. (2008): *Quantifying the stationarity and time reversibility of the nucleotide substitution process*. *Molecular Biology and Evolution* 25(12), 2525-35

Messer P.W. and PF Arndt (2007): *The majority of recent short DNA insertions in the human genome are tandem duplications*. *Molecular Biology and Evolution* 24(5), 1190-7

Messer P.W., Bundschuh R., Vingron M. and Arndt P.F. (2006): *Alignment Statistics for Long-Range Correlated Genomic Sequences*. *Lecture Notes in Computer Science* 3909, 426-440

Selected invited talks

Arndt P.F. *Parity Rules and their Violation in Molecular Biology*, BMS Days 2009, Berlin, 16. – 17.2.2009

Arndt P.F. *The Evolutionary Processes that shape the human genome*, Kavli Institute for Theoretical Physics, Santa Barbara, 18.3. 2007

Arndt P.F. *The Evolutionary Processes that shape the human genome*, Center for Theoretical Biological Physics, San Diego, 9.3. 2007

Arndt P.F. *Substitution pattern of mammalian transposable elements*, 1st International Conference on the Genomic Impact of Eukaryotic Transposable Elements, Asilomar, 31.3. – 4.4.2006

Work as scientific editor

- Journal of Statistical Mechanics – Theory and Experiment

Work as scientific referee

Peter Arndt serves as scientific referee for the following journals and conference series: *Nature*, *Proceedings of the National Acad-*

emy of Sciences, *Molecular Biology and Evolution*, *Journal of Molecular Evolution*, *Biophysical Journal*, *Nucleic Acids Research*, *Europhysics Letters*, *BioSystems*, *Physical Review Letters*, *Research in Computational Molecular Biology (RECOMB)*, *Gene*, *Genetics*.

In addition, Peter Arndt serves as scientific referee for the following institutions: *National Science Foundation*, *National Institute of Health*, *German Israeli Foundation*.

Appointments of former members of the Group

Philipp Messer: HFSP fellow at the Biology Department, Stanford University

Nicole de la Chaux: Doctoral Student, University of Zurich

Teaching activities

Winter 05/06; Summer 08: Seminar on Population Genetics

Winter 06/07: Lecture and Tutorials on Theoretical Genetics

Summer 07: Seminar on the Evolution of Sex

Winter 07/08; 08/09: Lecture and Tutorials on Population Genetics

Organization of scientific events

Otto Warburg International Summer School and Workshop on Regulatory (Epi-)Genomics, Harnack House Berlin, August 29 - September 6, 2009

Otto Warburg International Summer School and Workshop on Computational Systems Biology, Harnack House Berlin, August 27 - September 5, 2007

Symposium on the Evolution of Sex, XI Congress of The European Society for Evolutionary Biology, Uppsala, August 20-25, 2007

Otto Warburg International Summer School and Workshop on Evolutionary Genomics, Berlin, August 29 - September 8, 2006

Gene Structure & Array Design Group

(Established: 01/2001)



Head

Dr. Stefan Haas (since 01/01)
Phone: +49 (0)30 8413-1164
Fax: +49 (0)30 8413-1152
Email: haas@molgen.mpg.de

Scientist

Dr. Hughes Richard (since 09/06)

PhD students

Marcel Schulz (since 08/06, IMPRS)
Anne-Katrin Emde (since 10/08, IMPRS)
Shen Lin (since 01/09)
Helge Roeder (10/05 – 07/09)

Scientific programmers

Sean O'Keeffe (since 11/2006)
Ramu Chenna (12/2006 – 11/2009)

Scientific overview

The group focuses mainly on the analysis of genome-wide expression data to decipher mechanisms driving tissue-specific activation/repression of genes or distinct transcripts. As a complement we provide bioinformatics support related to the design of experiments and develop tools to facilitate processing/analysis of expression studies based on high-throughput technologies like microarrays or next-generation sequencing.

Prediction of alternative transcripts

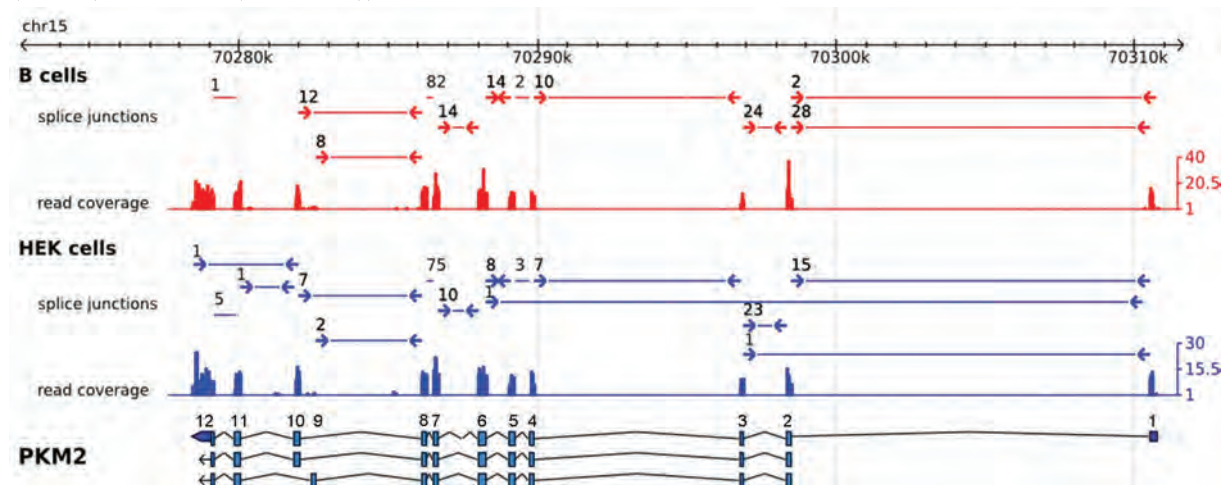
[Gupta, Haas]

The huge variety of gene products expressed in an organism is to a major extent caused by differential usage of exons leading to alternative transcripts encoded by a single gene. In general, different mechanisms contribute to the formation of alternative transcript isoforms. While alternative splicing *via* the spliceosome causes variations in the usage of internal exons, alternative promoters and alternative polyadenylation will change the usage of alternative first or last exons, respectively.

By integrating public data sets targeted to determine transcriptional start sites (TSS) like CAGE and full-length clone data (DBTSS) together with Ensembl/RefSeq transcript annotations and our EST-based transcript predictions, we generated a comprehensive set of potential alternative transcriptional start sites. Our analysis revealed that most genes usually express a single, dominant transcript, whereas alternative transcripts are often expressed on lower levels. Although methods like



CAGE are tailored to detect TSSs, still the traditional data sets uniquely contribute a significant number of TSSs to the overall set. Such a set of TSSs can be used to select a representative transcript per gene based on expression rather than on location, which might be more appropriate when studying gene regulation. Furthermore, the integration of TSS predictions allows fine-tuning of TSS localization e.g. important in cases where only small proximal promoter regions have to be analyzed. The predicted TSSs are visualized in our Promotion Genome Browser (promotion.molgen.mpg.de) together with additional genomic features dedicated to facilitate the interpretation and integration of important aspects associated with gene regulation (e.g. sequence conservation, transcription factor binding affinities (TRAP), EST data (GeneNest)).



Alternative splicing of the PKM2 gene. Three transcripts annotated in ENSEMBL are shown next to the gene name, exons numbered. The read coverage is shown for each exon (blue for HEK and red for B cells). Splice-junction reads are shown as arrows; the numbers above the arrows represent the number of reads at junctions. Two different sequenced junctions connecting to either exon 9 or exon 10 reflect alternative transcripts with mutually exclusive exons in HEK and in B-cells.

Tissue-specific gene regulation

[Roeder, O’Keeffe, Haas]

Regulation of gene expression is mainly controlled by chromatin modifications and the activity of specific transcription factors acting either on promoters close to the transcriptional start, or on distant enhancer elements. Since many genes are involved in a discrete biological context, the use of such contextual information is crucial to successfully unravel regulatory relationships between genes and transcription factors (TF). We therefore categorize genes according to their significance of expression in a certain tissue based on the statistical evaluation of EST data (T-STAG) or DNA-microarray/RNA-Seq expression measurements in a number of organisms (human, mouse, sheep etc.). Given the TSSs of such co-expressed genes we rank their potential proximal promoter regions by DNA-binding affinity of known transcription factors using our method TRAP. In order to detect candidate regulatory TFs we developed the method PASTAA, which iteratively tests for genes significantly ranked by tissue-specific expression as well as by high TF binding affinity. This way we were able to computationally predict a large number of functional TF-tissue associations well supported by literature. Intriguingly, we found that TFs predicted to regulate genes expressed in a certain tissue are frequently themselves significantly expressed in the respective tissue. In line with results revealed in another computational study performed in our department we observed potential auto-regulatory loops for many of the TFs involved in tissue-specific gene regulation.

In a subsequent study we could show that the success in predicting functional associations is strongly related to the CpG content of the promoters investigated. Functional predictions for transcription factors with respect to tissue-specificity are far more successful for promoters with low CpG content. However, even in high CpG promoters we could find functionally plausible TF-tissue association, e.g. NRSF-brain, which could be hardly detected when analyzing the full set of promoters. Thus, our computational analysis highlights the importance of categorizing promoters into high and low CpG groups since those promoters may be regulated by alternative biological mechanisms.

Next generation sequencing

[Richard, Emde, Schulz, O’Keeffe, Chenna, Haas]

The recent advances in next-generation sequencing technologies (NGS) provide the opportunity to tackle biological questions on genome scale in an unprecedented quality. However, the huge amount of data generated using these technologies requires the development of new algorithms to handle the data efficiently but also to analyse this new type of sequence information.

In this context, we were among the first studying the performance of next-generation sequencing in transcriptome sequencing (RNA-Seq) of two human cell lines. In collaboration with the group of Marie-Laure Yaspo we could show that NGS clearly outperforms state-of-the-art microarray technology in terms of sensitivity and noise thus revealing a more complete picture of the transcriptome. Although the transcriptomes in B- and HEK-cells are well studied we were able to discover extensions of exonic regions and a limited number of potential, so far unknown exons.

While the basic mapping of sequencing reads mainly provides information about genomic regions that are transcribed, details about differential splicing has to be determined by those sequencing reads mapping on exon junctions. We therefore generated an artificial set of all exon-junctions of a gene based on our comprehensive set of gene structures derived from EST data and the Ensembl database. The mapping of sequence reads to these junctions revealed a large number of alternative splicing events even with a relatively low sequencing depth.

In light of the steadily increasing sequence output of NGS we are currently implementing a generalised computational pipeline comprising statistical measures for quality control of sequencing runs as well as improved methods for efficient mapping of reads. In this context we are involved in the development of RazerS, a tool to perform read mapping allowing for short insertion/deletions, which is an important feature with respect to future applications aiming at the discovery of disease causing mutations by deep genomic sequencing.



Selected information

Selected publications

Roider, H.G., Lenhard, B., Kanhere, A., Haas, S.A., Vingron, M. (2009). *CpG-depleted promoters harbor tissue-specific transcription factor binding signals - implications for motif overrepresentation analyses*. Nucleic Acids Res., in press

Roider, H.G., Manke, T., O’Keeffe, S., Vingron, M., Haas, S.A. (2009). *PASTAA: identifying transcription factors associated with sets of co-regulated genes*. Bioinformatics 25(4):435-442

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O’Keeffe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, M.L. (2008). *A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome*. Science 321 (5891): 956-960

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Koenig, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006). *Gene-expression based classification of neuroblastoma patients using a customized oligonucleotide-microarray outperforms current clinical risk stratification*. J. Clin. Oncol. 24:5070-5078

Hecht, H., Kuhl, H., Haas, S.A., Bauer, S., Poustka, A.J., Lienau, J., Schell, H., Stiege, V., Seitz, V., Reinhardt, R., Duda, G.N., Mundlos, S. and Robinson, P.N. (2006). *Gene Identification and Analysis of Transcripts Differentially Regulated in Fracture Healing by EST Sequencing in the Domestic Sheep*. BMC Genomics, 7:172.

Selected invited talks

Haas S.A. *Lessons learned from 2nd Generation Sequencing of entire Transcriptomes*, Sixth Annual Meeting of the Italian Bioinformatics Society, Genoa, Italy, 18.-20.03.2009

Haas S.A. *Deep Sequencing of the Transcriptome of two Human Cell Lines*, ESF Symposium: Computational Challenges of the Next-Generation DNA Sequencing, Uppsala, Sweden, 15.-17.01.2009

Haas S.A. *Deep Sequencing of the Transcriptome of two Human Cell Lines*, Workshop: Parallel Sequencing, Basel, Switzerland, 02.10.2008

Richard H. *Estimating genetic richness in EST libraries*, JOBIM conference, Marseilles, France, 10.-12.07.2007

Haas S.A. *Integrating ESTs into a comprehensive Picture of Alternative Transcriptional Start Sites*, Symposium: Alternative Transcript Diversity, Heidelberg, Germany, 21.03.2006

Work as scientific referee

Stefan Haas serves as scientific referee for the following journals: BMC Genomics, BMC Bioinformatics, Bioinformatics, Genome Research, Genome Biology, Nucleic Acids Research.

Teaching activities

EMBO Course: *Next Generation Sequencing: ChIP-seq and RNA-seq*, 02/09

Public relations

Lange Nacht der Wissenschaft (06 – 09)

Tag der Talente, BMBF (09/08)

Presentations for High-School classes (07 – 09)

Mechanisms of Transcriptional Regulation Group

(Established: 09/2009)



Head:

Sebastiaan H. Meijnsing (since 09/09)
Phone: +49 (0)30 8413-1176
Fax: +49 (0)30 8413-1152
Email: meijnsing@molgen.mpg.de

Technician

Edda Einfeldt (since 09/09)

Scientific overview

Recent technological advances facilitate the rapid genome-wide identification of transcription factor binding sites. *The long-term goal of our research is to understand the mechanisms by which these binding sites specify where and when genes are expressed.* Metazoan nuclear hormone receptors regulate the expression level of target genes to orchestrate development and to respond to changes in their environment. Target genes are not simply turned on or off, but instead their expression is fine-tuned to meet the needs of a cell. Hence mechanisms exist in metazoans that specify not only where and when genes are expressed but also at which level. To study gene regulation, we investigate transcriptional regulation by the glucocorticoid receptor (GR), a member of the steroid hormone receptor family. Upon

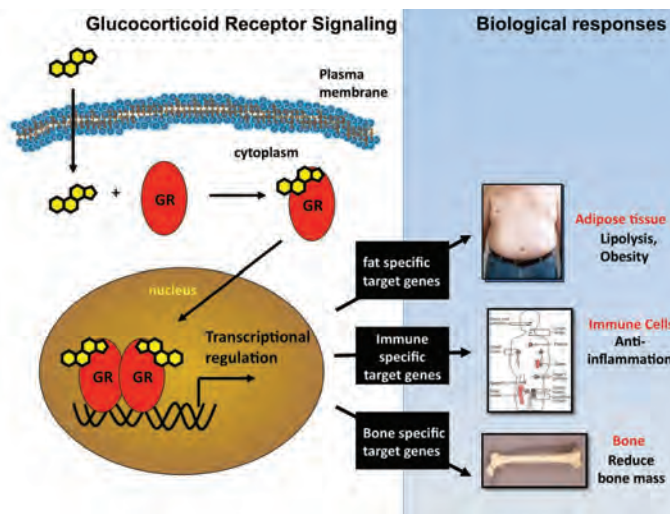


Figure 1: Transcriptional regulation by the Glucocorticoid receptor. Upon hormone binding, the receptor translocates into the nucleus where it activates a tissue specific set of target genes that in turn specify the biological response to hormone.

binding endogenous hormones called glucocorticoids, GR translocates from the cytoplasm, where it is inactive, to the nucleus where it orchestrates the expression of a cell- or tissue-specific subset of target genes (Fig. 1). For example, in immune cells, GR-target genes mediate the anti-inflammatory effects of glucocorticoids, whereas in liver glucocorticoids affect metabolism *via* a different set of target genes to stimulate gluconeogenesis. An attractive feature of GR to study gene regulation is that its activity can simply be turned on or off by the addition or removal of hormone. This on/off switch facilitates the identification of genes with changed expression levels upon hormone treatment by a single experiment using gene ex-



pression analysis such as microarrays or RNA-sequencing. In parallel experiments, Chromatin immunoprecipitation (ChIP) experiments reveal the genomic locations of GR DNA binding to identify primary target genes. Combining the expression and recruitment information in a single cell type aided by bioinformatical approaches yield leads for mechanisms that specify which genes are regulated and the magnitude of regulation. Similarly, by comparing cell types derived from different tissues, candidate mechanisms contributing to tissue specific regulation will be identified and studied.

Role of the DNA-sequence of binding site

The glucocorticoid receptor regulates target genes by associating with specific DNA binding sites, the sequences of which differ between genes. There is a paradox in how transcription factors, such as GR, bind DNA. On one hand binding sites can typically tolerate variation in its DNA-binding site, while still binding and activating genes (Fig. 2A). On the other, these sites are often conserved at individual genes, suggesting that the exact sequence is important for proper gene regulation (Fig. 2B). This paradox challenges the paradigm that transcription factor binding sites are simple docking sites. Using bioinformatical, structural, biochemical, and cell-based assays we found that GR binding sequences (GBSs), differing by as little as a single base pair, differentially affect GR conformation and regulatory activity. For example, we found GBS-specific requirements for cofactors and receptor domains, and GBS specific levels of transcriptional activation. Comparison of high resolution crystal structures we obtained, revealed GBS-specific GR conformations. Specifically, we observed alternative conformations of the lever arm, a domain connecting the DNA recognition helix and the dimer interface. In conclusion, we propose that DNA is a sequence-specific allosteric ligand of GR that tailors the activity of the receptor toward specific target genes.

Outlook

One goal is to understand how GBS sequences direct distinct modes of transcriptional regulation and their role in spacio-temporal control of gene regulation. We will approach this problem using several strategies. One approach will be done in collaboration with Ulrich Stelzl from the Otto Warburg Laboratories using high throughput yeast two-hybrid screening to identify GBS specific cofactors. Furthermore, previous studies were focused on the role of the GBS in isolation, however, for GR-target genes the GBS typically collaborates with other *cis*-acting elements. To study this interplay we will flank GBS sequence variants with naturally occurring *cis*-elements to determine how several signals are integrated to specify the transcriptional output.

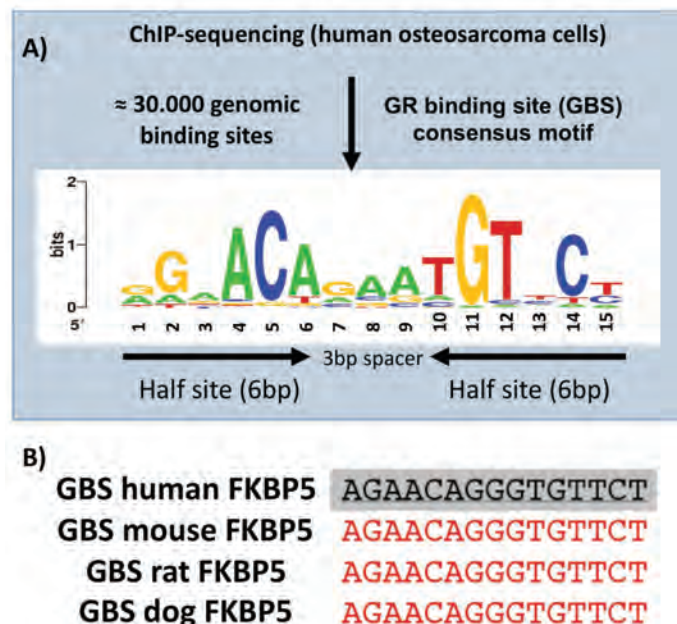


Figure 2: (A) Binding site motif for the glucocorticoid receptor binding site based on binding sites identified by ChIP-sequencing. Height of nucleotide indicates sequence conservation at that position. (B) GR binding sequences are often highly conserved. Shown is an example for a target gene associated GBS from human, mouse, rat and dog.

Alternative splicing generates a naturally occurring splice variant GR γ that inserts an arginine in the lever arm, the region that adopts alternative conformations depending on the sequence of the binding site. The inserted arginine does not affect DNA binding affinity and accordingly, its structure is superimposable on that of GR α the predominant splice isoform, except for the lever arm region. Hence, the extra amino acid specifically disrupts the lever arm while leaving the remainder of the receptor intact making it a great tool to understand the role of the lever arm in gene regulation. We assayed gene expression by microarray analysis and GR recruitment by ChIP-sequencing. Comparison of expression and recruitment data for GR α and GR γ indicates that insertion of the arginine influences transcriptional regulation at several steps and current efforts are aimed to identify the molecular mechanisms underlying the isoform specific regulation.

Furthermore, we plan to use the model organism *C. elegans* in unbiased approaches to identify novel factors and pathways that specify where and when target genes of the nuclear hormone receptor daf-12, are expressed and their role in physiology. The combination of these approaches allows us to determine how effectors of spacio-temporal gene regulation in whole animals act at individual genes and vice versa, how regulatory mechanisms acting at individual genes influence processes in whole animals.

Selected information

Selected publications

Meijsing SH, Pufall MA, So AY, Bates DL, Chen L, Yamamoto KR (2009). *DNA binding site sequence directs glucocorticoid receptor structure and activity*. Science 17:407-410.

Meijsing SH, Elbi C, Luecke HF, Hager GL, Yamamoto KR (2007). *The ligand binding domain controls glucocorticoid receptor dynamics independent of ligand release*. Mol Cell Biol. 27:2442-2451.

Wang JC, Shah N, Pantoja C, Meijsing SH, Ho JD, Scanlan TS, Yamamoto KR (2006). *Novel arylpyrazole compounds selectively modulate glucocorticoid receptor regulatory activity*. Genes & Dev. 15:689-699.

Selected invited talks

2008 Research in Progress series (UCSF).

2008 Keystone meeting: Nuclear Receptors: Steroid Sisters, Whistler, Canada.

2009 Otto Warburg International Summer School and Workshop on Regulatory (Epi)Genomics. Berlin, Germany.

2009 EMBO meeting: Nuclear receptors: From molecular mechanism to molecular medicine. Dubrovnik, Croatia.

Awards

2004 - 2007: Leukemia & Lymphoma Society Special Fellowship

2006 Keystone symposia Scholarship/Travel award

2008 Keystone symposia Scholarship/Travel award

Teaching activities

Teaching: Biochemistry/Pharmacology course for UCSF Medical students (2006).

External funding

2004 - 2007: Leukemia & Lymphoma Society Special Fellowship



Algorithmics

(Established: 05/2002 - 06/2009)

Head

Dr. Alexander Schliep (05/02-06/09)
Email: schliep@cs.rutgers.edu

PhD students

Wasinee Rungarityotin (03-07)
Ivan G. Costa (04-08)
Benjamin Georgi (05-09)
Ruben B. Schilling (07-09)



Scientific overview

Our research focuses on methods from mathematics and statistics which are crucial for answering relevant biological questions. An emphasis is put on analyzing high-dimensional and heterogeneous data, such as time-courses or imaging data.

Clustering heterogeneous data

Detecting relevant groups of co-expressed genes as a foundation for building more detailed regulatory networks is still one of the central unsolved problems. The massive amounts of data created in molecular biology, the high error rates and the question of how to combine several heterogeneous data sets, pose challenges for research. We have contributed clustering methods based on mixture models, novel component models for specific data types. We also were the first to propose the framework of partially supervised learning or constrained clustering to simultaneously analyze two different sets of complimentary data (data fusion).

We studied gene regulation during early *Drosophila* development. Gene expression measurements during the development of the fly *Drosophila melanogaster* are routinely used to find functional modules of temporally co-expressed genes. Complimentary large data sets of in situ RNA hybridization images for different stages of the fly embryo elucidate the spatial expression patterns. Using a semi-supervised approach, constrained clustering with mixture models, we can find clusters of genes exhibiting spatio-temporal similarities in expression, or syn-expression. The temporal gene expression measurements are taken as primary data for which pairwise constraints are computed in an automated fashion from raw *in situ* images without the need for manual annotation. We investigate the influence of these pairwise constraints in the clustering and discuss the biological relevance of our results. Spatial information contributes to a detailed, biological meaningful

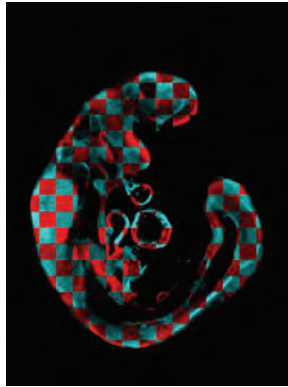


Figure 1: Two registered mouse embryos (cyan and red). Registration of in-situ stained mouse embryos facilitates the direct analysis of spatial gene expression patterns.

analysis of temporal gene expression data. In cooperation with the department of Developmental Genetics (B. Herrmann), we investigate the necessary image registration to deal with 3D *in situ* images of mouse embryos in a similar fashion (Figure 1).

Another model system we investigated is the lymphoid system. Gene expression measured in lymphoid cells in several distinguishable developmental stages helps in the elucidation of underlying molecular processes, which change gradually over time and lock cells in either the B cell, T cell or Natural Killer cell lineages. Large-scale analysis of these gene expression trees requires computational support for tasks ranging from visualization, querying, and finding clusters of similar genes, to answering detailed questions about the functional roles of individual genes.

We developed the first statistical framework designed to analyze gene expression data as it is collected in the course of lymphoid development through clusters of co-expressed genes and additional heterogeneous data. We introduce dependence trees for continuous variates, which model the inherent dependencies during the differentiation process naturally as gene expression trees. Such trees can have their structure estimated from the data or derived from expert knowledge. Several trees are combined in a mixture model to allow inference of potentially overlapping clusters of co-expressed genes. Computational results for several data sets from the lymphoid system demonstrate the relevance of this framework. We recover well-known biological facts and identify promising novel regulatory elements, including putative microRNAs, of genes and their functional assignments.

A complimentary technique for fusing heterogeneous datasets is based on a naive Bayes approach coupled with an innovative way of minimizing model complexity. CSIMixture (Context-specific independence mixture) modeling for sequence motifs provides a more physically realistic description transcription factor binding site motifs. Previous studies showed that for transcription factors which bind to divergent binding sites, mixtures of multiple PWMs increase performance. However, estimating a conventional mixture distribution for each position will in many cases cause overfitting. We circumvent this problem by employing a context-specific independence (CSI) framework. In CSI mixtures model complexity is automatically adapted to match the variability found in a given data set.

Another application of the CSI mixture framework is clustering of protein families for simultaneous inference of subgroups and prediction of specificity determining residues based on multiple sequence alignments of protein families. Furthermore, CSI mixtures can be used for the joint analysis of genotype and phenotype data from ADHD-patients, or other complex diseases, to simultaneously identify disease subgroups and relevant genotypic features. Typically, in such applications, none of the variables will be fully informative with respect to distinguishing between clusters, which makes an a-priori feature selection infeasible.

Optimizing experimental designs

We developed the first linear time algorithms for designing DNA tiling arrays which find global optima, balancing uniformity of hybridization conditions, avoidance of cross-hybridization and probe placement. This has been used by several experimental groups interested in custom tiling arrays, primarily for bacteria. We also extended the generation of candidate probes, to include more realistic aspects



of hybrid formation in a computationally feasible manner. This is in stark contrast to the Blast-based approaches dominantly used in the field.

The same underlying hybridization reactions underlying gene expression analysis using DNA-Microarrays can also be used to infer presence and absence of biological agents, say viruses or bacteria, in a sample from the hybridization pattern of oligonucleotide probes to genomic DNA of the agents. This detection is crucial for epidemiological studies and vaccine design when used for detection of viral subtypes in Influenza or HIV, in food and safety control, in studies of microbial diversity or, particularly in the USA, in bio-threat reduction.

Due to close evolutionary relationships between agents oligonucleotide probes uniquely identifying agents cannot be found in many applications; consider identifying all Influenza virus subtypes. Nevertheless, even non-unique probes can be used for the detection with great success - indeed, in simulations, we can correctly identify previously unknown agents in biological samples with a success rate of up to 70% - as there is a connection to the mathematical field of statistical group testing. It bridges across combinatorial design theory, Bayesian statistics and Markov Chain Monte Carlo methods. The distilled computer science, non-unique probe sets and generalized group testing, which we first formulated directly based on the biological application at hand, has received a great deal of interest in the theoretical computer science literature.

Selected information

Selected publications

I.G. Costa, S. Roepcke, C. Hafemeister, A. Schliep (2008). *Inferring differentiation pathways from gene expression*. Bioinformatics (Proceedings of Intelligent Systems for Molecular Biology, ISMB 2008) 24(13):i156-64, 2008.

A. Schliep, R. Krause (2008). *Efficient algorithms for the computational design of optimal tiling arrays*. IEEE/ACM Trans Comput Biol Bioinform. 2008 Oct-Dec; 5(4):557-67 (Invited paper selected from WABI 2007)

I.G. Costa, R. Krause, L. Opitz, A. Schliep (2007). *Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data*. BMC Bioinformatics 8(Suppl 10):S3, 2007

A. Schliep, S. Rahmann (2006). *Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree*. Bioinformatics (Proceedings of Intelligent Systems in Molecular Biology, ISMB 2006), 22 (14): e424-e430, 2006

Selected invited talks

Group testing DNA Micro-arrays for detection of biological agents. DIMACS workshop on combinatorial group testing, 19.05.06

Treeprobes: Group testing biological agents related by phylogenetic trees. Los Alamos National Laboratory, 24.05.06.

Mixture models for heterogeneous biological data. Symposium on Bioinformatics and Biomathematics, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 06.04.07

Detecting functional modules from heterogeneous mass data. College of Computing, Georgia Tech, 04.04.08

Detecting functional modules in heterogeneous biological data. Institute of Genetics and Molecular and Cellular Biology, Strasbourg, 02.07.08

Open access activities

Software packages implementing methods we develop are made available under open source licenses (GPL or LGPL), which guarantee open access and the freedom to extend the softwares: This includes GHMM, GQL, PyMix, Tileomatic, Gato and Mix DTrees (see <http://algorithmics.molgen.mpg.de/Software/>)

Work as scientific editor

- Associate Editor for Discrete Mathematics, Algorithms and Applications

Work as scientific referee

Alexander Schliep serves as scientific referee for the following journals and conference series: Bioinformatics, BMC Bioinformatics, Proteins, Functional and Integrative Genomics, Springer Verlag, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Pattern Analysis and Applications, Journal of the American Statistical Association, Bulletin of Mathematical Biology, Journal of Bioinformatics and Computational Biology, Journal of Computational Statistics and Data Analysis, RECOMB, WABI, ISMB, GfKI, IFCS, ECCB.

In addition, Alexander Schliep serves as scientific referee for the following institutions: European Commission (IST - Future & Emerging Technologies), Netherlands Organization for Scientific Research (NWO, Horizon programme), US-Israel Binational Science Foundation.

Service to the scientific community

A. Schliep has been a member of the following program committees: NIPS workshop on machine learning in bioinformatics (2004-2009), Whistler BC, APBC Asia-Pacific Bioinformatics conference (2006), IAPR Workshop on Pattern Recognition in Bioinformatics (2006, 2007), German Conference on Bioinformatics (2007)

Appointments of former members

Alexander Schliep: Associate Professor, Dept. of Computer Science and BioMaPS Institute for Quantitative Biology, Rutgers University, New Jersey

Ivan G. Costa: Assistant Professor, Center of Informatics, Federal University of Pernambuco, Recife, Brasil

Benjamin Georgi: Postdoc with Maja Bucan, Department of Genetics, University of Pennsylvania

External funding

DAAD: *Meta-Learning for Selection and Combination of Clustering Algorithms Applied to Temporal Series*. With Universidade Federal de Pernambuco, Recife, Brasil.

Teaching activities

Winter 05/06: *Algorithmen der Bioinformatik*. FU Berlin (lecture course, one unit only); Algebraic statistics. FU Berlin (seminar course).

Winter term 05/06; 06/07: *Angewandtes Data Mining*. FU Berlin (course design, two week full-time lectures and labs).

Organization of scientific events

Section chair Genome and DNA Analysis, GfKI meeting 2006 in Berlin.

Organizer of a Dagstuhl workshop on Group Testing in the life sciences, Juli 2008.



Transcriptional Regulation Group

(Established: October 2000)

Head

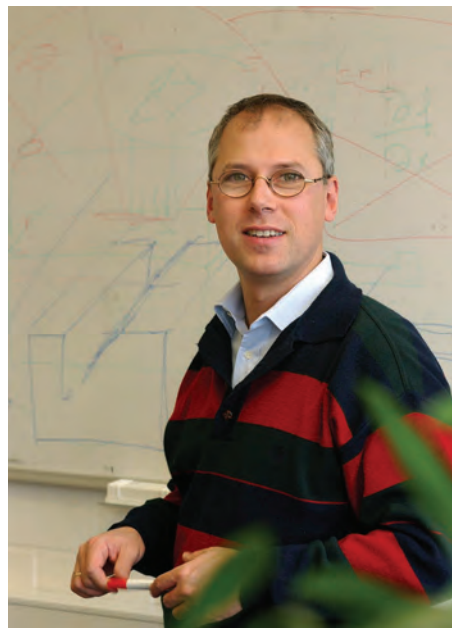
Prof. Dr. Martin Vingron
Phone: +49 (0)30 8413-1150
Fax: +49 (0)30 8413-1152
Email: vingron@molgen.mpg.de

Scientists

Thomas Manke (since 03/05)
Christine Steinhoff (since 08/05)
Ho-Ryun Chung (since 06/05)
Tomasz Zemojtel (since 02/04)
Szymon Kielbasa (since 03/05)
Julia Lasserre (since 04/08)
Anirban Banerjee (since 11/07)
Sarah Behrens (since 10/08)
Roman Brinzanik (since 11/07)
Andrew Hufton (since 11/06)
Morgane Thomas-Chollier (since 04/09)
Lloyd Demetrius (03-09)
Aditi Kanhere (09/05-05/08)

PhD Students

Marcel Schulz (since 09/05)
Ewa Szczurek (since 10/06)
Marta Luksza (since 10/09)
Akdes Serin (since 08/06)



Rosa Karlic (since 10/07)
Jonathan Goeke* (since 09/2007)
Holger Klein (05/03-10/09)
Hannes Luz (04/0 -12/06)
Abha Singh Bais (06/03-05/07)
Utz Pape (11/0109/04)

Visitors

Prof. Dr. Ina Koch (since 06)
Dr. Pawel Gorecki (10/06-09/07)

Introduction

The theoretical study of transcriptional regulation has entered a new era with the availability of many fully sequenced genomes in conjunction with a number of new, high-throughput experimental techniques for the study of protein-DNA binding. The gene regulation group focuses on the delineation of regulatory motifs and interactions based on the integration and analysis of this variety of information sources.

Scientific overview

From CORG to TRAP: Transcription factor binding site prediction

(Thomas Manke, Helge Roeder, Aditi Kanhere, Utz Pape)

A transcription factor tends to bind to particular DNA patterns which can be summarized by so-called Positional Weight Matrices (PWMs). After having explored the power and problems of matching PWMs to sequence in the CORG-database,

we have developed an alternative biophysics-inspired approach. This “TRAP” method (for Transcription Factor Affinity Prediction) transforms the match between a sequence and a pattern into a binding probability and integrates over the region of interest, say a promoter region. The TRAP method has been validated by comparison to large scale DNA binding experiments (ChIP-chip and ChIP seq experiments) and shown to be successful in predicting novel target genes of transcription factors. Based on a statistical normalization of the affinity scores, the most likely binding factors to a particular promoter can be inferred. Together with the group of Stefan Haas, TRAP has further been utilized to recognize transcription factor binding sites which are over-represented in co-expressed or tissue-specific groups of genes. Ongoing work applies TRAP for predicting possible effects of regulatory SNPs. It has also been applied in the context of deriving gene regulatory networks.

Estimating statistical significance of possible findings is crucial in the analysis of large data sets. To this end, we have derived probabilistic descriptions of the occurrence of hits to PWMs and of the distribution of TRAP scores. Significant efforts have gone into the development of measures of similarity among transcription factor binding sites, which in turn has proven instrumental in computing the probability of observing combinations of binding sites in a regulatory region, thus providing an instrument to study combinatorial regulation.

Epigenetic regulation

(Ho-Ryun Chung, Rosa Karlic, Julia Lasserre, Irit Gat-Viks)

Experimental progress over recent years has driven home the point, that in eukaryotes, and in particular in mammals, transcription factors are not the only regulators of gene expression. Chromatin structure, histone modifications, and DNA methylation also play vital roles or are at least correlated to expression status. Our own interest in this epigenetic level of regulation focuses on the question, in how far the DNA sequence can provide us with information not only about transcription factor binding sites, but also about, e.g., chromatin structure. While the sequence dependence of nucleosome positioning is still under debate, one does see a strong division of promoters into those with high vs. low contents of CpG dinucleotides. We have recently published that this distinction governs the localization and type of transcription factor binding sites, and are currently working on establishing that histone modification patterns also depend on these sequence features.

Evolution of regulation

(Tomasz Zemojtel, Szymon Kielbasa, Sarah Behrens, Andrew Hufton, Morgane Thomas-Chollier)

While protein evolution is nowadays generally described by a Markov process on the sequence positions with selection acting on the level of protein function, the evolution of regulatory DNA sequences is still badly understood. In collaboration with the Evolutionary Genomics Group of Peter Arndt, we have been studying the influence of the Cytosine deamination on the appearance of binding sites. A mutation of a C in the context of a CpG dinucleotide due to deamination is much more likely than other mutations. A careful inspection of Alu repeats has shown that these transposable elements carry possible predecessors of binding sites, which are moved around the genome in the course of evolution, bearing the potential to become functional binding sites upon deamination. In this context we are now systematically investigating the role of transposable elements and numerous types of transcription factor binding sites as a possible source of novel binding sites in evolution. In contrast to this mechanism, studies in the time it would take for a binding site to evolve purely by point mutations indicate that this is not flexible



enough a process to explain regulatory evolution. This theoretical work is further complemented by studies in ancient conserved elements in collaboration with the Poustka/Panoupoulou group (Dept. Lehrach) and studies in the evolution of Hox genes.

Gene networks

(Ewa Szczurek, Thomas Manke, Anirban Banerjee, Lloyd Demetrius, Roman Brinzanik, Utz Pape)

Several projects, many of them in collaboration with experimentalist, try to delineate regulatory networks or study the general features of biological networks. In collaboration with C. Sers (Charité) we are studying the regulatory cascade downstream of the ras-triggered MAP kinase pathway. Gene regulatory networks in heart development were the topic of a collaboration S. Sperling (Dept. Lehrach). In a collaborative project with other Max Planck Institutes we are working on the delineation of regulatory networks integrating data from metabolomics and transcriptomics. The analysis of gene networks has also led us to propose an algorithmic framework to predict most informative experiments to elucidate regulatory dependencies.

Following their earlier work on the evolution of complex networks, Lloyd Demetrius, Thomas Manke and Anirban Banerjee have continued to develop a graph-theoretical framework for the characterisation of biological networks. Applying an entropic formalism to large-scale protein interaction data, they investigated the relationships between the essentiality of a protein and its overall position in the molecular networks of yeast and nematode worm. In numerical studies on model networks they found that network entropy correlates positively with many heuristic measures of structural and dynamical robustness of networks, such as the percolation threshold and mixing rates. This approach has been complemented by work on the graph spectrum as an alternative characterisation of biological networks. The normalized graph Laplacian spectrum does not only provide insights into the modular organisation of networks, but also helps to measure the distance between networks with different sizes (cooperation with J. Jost, MPI-MIS Leipzig). Lloyd Demetrius has continued his work on ageing models, proposing that differences among organisms in the rate of ageing and life span are due to differences in metabolic stability, rather than differences in metabolic rate. J. Adjaye (Dept. Lehrach) provided experimental evidence in support of this theory.

Collaborations

In addition to MPI-internal collaborations like the ones mentioned above, group members have collaborations within Berlin or Germany, as well as internationally. Frequently these collaborations will be in the context of a DFG-, BMBF- or EU-grant. While within Berlin we are closely cooperating with the FU bioinformatics group (K. Reinert) and with experimentalists at Charité (C. Sers), work together with A. Nordheim from Tübingen on SRF regulated genes has been very successful, too. On an international level, fruitful cooperation with B. Lenhard (Bergen, Norway), J. Tiuryn (Warsaw), E. Birney (Hinxton), and Fengzhu Sun (Los Angeles) need to be mentioned.

Selected information

Selected publications

Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA (2009). *PASTAA: identifying transcription factors associated with sets of co-regulated genes*. *Bioinformatics* 25(4): 435-442

Stritt C, Stern S, Harting K, Manke T, Sinske D, Schwarz H, Vingron M, Nordheim A, Knöll (2009). *Paracrine control of oligodendrocyte differentiation by SRF-directed neuronal gene expression*. *Nat Neurosci*. 12(4):418-27

Szczurek E, Gat-Viks I, Tiurnyn J, Vingron M (2009). *Elucidating regulatory mechanisms downstream of a signalling pathway using informative experiments*. *Mol Syst Biol* 5:287

Chung HR, Vingron M. (2008). *Sequence-dependent Nucleosome Positioning*. *J Mol Biol*. 386(5):1411-22

Manke T, Roider HG, Vingron M. (2008). *Statistical Modeling of Transcription Factor Binding Affinities Predicts Regulatory Interactions*. *PLoS Comput Biol*. 4(3): e1000039

Zemojtel T, Kielbasa SZ, Arndt PF, Chung HR, Vingron M. (2008). *Methylation and deamination of CpGs generate p53-binding sites on a genomic scale*. *Trends in Genetics* 25(2):63-66

Roider HG, Kanhere A, Manke T, Vingron M (2007). *Predicting transcription factor affinities to DNA from a biophysical model*. *Bioinformatics* 23(2): 134–141

Manke T, Demetrius L, Vingron M (2006). *An entropic characterization of protein interaction networks and cellular robustness*. *J R Soc Interface* 3(11): 843-50

Demetrius L. (2006). *Aging in mouse and human systems: a comparative study*. *Ann N Y Acad Sci*. 1067:66-82 (review)

Selected invited talks (Martin Vingron)

- Heidelberg Spring Workshop on Cancer Biology, 04/09
- Center for Algorithmic and Systems Biology, CASB-20 meeting, San Diego, 03/09
- First RECOMB Satellite Conference on Bioinformatics Education, 03/09

- Asia Pacific Bioinformatics Conference, Beijing, 01/09
- NGFN Plus and NGFN Transfer, München, 12/08
- Mini EURO Conference on Computational Biology, Bioinformatics and Medicine, Rome, Italy, 09/08
- 6th Georgia Tech-Oak Ridge National Lab, International Conference on Bioinformatics, Atlanta, GA, 11/07
- European Conference on Computational Biology (ECCB) 2006, Eilat, Israel, 01/07
- Basel Computational Biology Basel Conference [BC]², 03/06

Membership in journal editorial boards

- J Comput Biol (associate editor)
- Bioinformatics
- Briefings in Bioinformatics
- BMC Bioinformatics and BMC Genomics
- Naturwissenschaften
- J Experimental Zoology Series B
- Interface - Journal of the Royal Society
- Int J Data Mining and Bioinformatics

Membership in professional societies (selected)

- ACM – Association for Computing Machinery
- RSS – Royal Statistical Society
- ISCB – International Society of Computational Biology

Service to Scientific Community (selected)

- Chair of the Steering Committee of the International Conference on Computational Molecular Biology RECOMB
- Member of Scientific Steering Committee of the Isaac Newton Institute (07-09)
- Member of Bioinformatics Advisory Council, European Bioinformatics Institute EMBL-EBI



Teaching activities

All courses given at Freie Universität Berlin.

Algorithmic Bioinformatics, 4hrs per week plus tutorials, held every other year during winter semester (alternatingly with Prof. Knut Reinert)

Probability and Statistics for Bioinformatics, winter 2007/08, 2 hrs per week

Algorithms for the Computation of Phylogenetic Trees, summer 2006, winter 2009/10, 2 hrs per week

Various seminars and practical courses

Organization of scientific events

Co-organizer of the German Conferences on Bioinformatics, GCB, Potsdam 2007

Chair of Program Committee for RECOMB 2008 in Singapore

Regular member of Program Committees of the European Conference on Computational Biology (ECCB), ISMB, and RECOMB.

Organization of Otto-Warburg Summer Schools in 2006, 2007, 2009

General information about the whole Department

Complete list of publications (2006-2009)

2009

Baek, Y.S., Haas, S., Hackstein, H., Bein, G., Santana, M.H., Lehrach, H., Sauer, S. and Seitz, H. (2009). *Identification of novel transcriptional regulators involved in macrophage differentiation and activation in U937 cells*. BMC Immunology 10:18

Diella F, Chabanis S, Luck K, Chica C, Chenna R., Nerlov C, Gibson TJ (2009). *KEPE - a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors*. Bioinformatics 25(1): 1–5. Published online 2008 November 24. doi: 10.1093/bioinformatics/btn594.

Gat-Viks, I., Vingron, M. (2009). *Evidence for Gene-Specific Rather Than Transcription Rate-Dependent Histone H3 Exchange in Yeast Coding Regions*. PLoS Comput Biol 5(2): e1000282

Huang, X., Vingron, M. (2009). *Maximum Similarity: A New Formulation of Phylogenetic Reconstruction*. Journal of Computational Biology 16(7): 887-896

Kanhere, A., Vingron, M. (2009). *Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes*. BMC Evolutionary Biology 9:9

Ott CE, Bauer S, Manke T., Ahrens S, Rödelsperger C., Grünhagen J, Kornak U, Duda G, Mundlos S, Robinson PN (2009). *Promiscuous and Depolarization-Induced Immediate-Early Response Genes are Induced by Mechanical Strain of Osteoblasts Original Study*. J Bone Miner Res

Roider, H.G., Manke, T., O’Keeffe, S., Vingron, M., Haas, S.A. (2009). *PASTAA: identifying transcription factors associated with sets of co-regulated genes*. Bioinformatics 25(4):435-442

Steinhoff, C., Paulsen, M., Kielbasa, S., Walter J., Vingron, M. (2009). *Expression profile and transcription factor binding site exploration of imprinted genes in human and mouse*. BMC Genomics 10:144

Stritt C, Stern S, Harting K, Manke T., Sinske D, Schwarz H, Vingron M., Nordheim A, Knöll B. (2009). *Paracrine control of oligodendrocyte differentiation by SRF-directed neuronal gene expression*. Nat Neurosci 12(4): 418-27

Vingron, M., Brazma, A., Coulson, R., van Helden, J., Manke, T., Palin, K., Sand, O., Ukkonen, E. (2009). *Integrating sequence, evolution and functional genomics in regulatory genomics*. Genome Biology 10:202

2008

Bluethgen, N., Legewie, S., Kielbasa, S.M., Schramme, A., Tchernitsa, O., Keil, J., Solf, A., Vingron, M., Schaefer, R., Herzog, HP., Sers, C. (2008). *A systems biological approach suggests that transcriptional feedback regulation by dual-specificity phosphatase 6 shapes extracellular signal-related kinase activity in RAS-transformed fibroblasts*. FEBS J 276(4), 1024 - 1035

Chen, W., Kalscheu, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M.H., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G., Pajares, I.L., Goetz-Sothmann, M., Heinrich, U., Rost, I., Dufke, A., Grasshoff, U., Glaeser, B.G., Vingron, M., Ropers, H.H. (2008). *Mapping translocation break-points by next-generation sequencing*. Genome Research 18(7):1143-9. Epub 2008 Mar 7

Chung, H.R., Vingron, M. (2008). *Sequence-dependent Nucleosome Positioning*. J Mol Biol 386(5): 1411-22

Costa, I.G., Roepcke, S., Hafemeister, C., Schliep, A. (2008). *Inferring differentiation pathways from gene expression*. Bioinformatics 24(13):i156-i164

Coulson, R., Manke, T., Palin, K., Roider, H., Sand, O., van Helden, J., Ukkonen, E., Vingron M., Brazma, A. (2008). *From gene expression profiling to gene regulation*. Modern Genome Annotation. The BioSapiens Network. Springer, Chapt. 2.3:105-115

de Souto, M.C.P., Costa, I.G., de Araujo, D.S.A., Ludermitz, T.B., Schliep, A. (2008). *Clustering cancer gene expression data: a comparative study*. BMC Bioinformatics 9:497



- de Souto, M.C.P., Prudencio, R.B.C.; Soares, R.G.F.; de Araujo, D.S.A.; Costa, I.G.; Ludermir, T.B.; Schliep, A. (2008). *Ranking and selecting clustering algorithms using a meta-learning approach*. IEEE International Joint Conference on Neural Networks (IJCNN) 2008, 1-8 June 2008, 3729 - 3735
- de Souto, M.C.P.; de Araujo, D.S.A.; Costa, I.G.; Soares, R.; Ludermir, T.B.; Schliep, A. (2008). *Comparative study on normalization procedures for cluster analysis of gene expression datasets*. IEEE International Joint Conference on Neural Networks (IJCNN) 2008, 1-8 June 2008, 2792 - 2798
- Duret, L., Arndt, P.F. (2008). *The Impact of Recombination on Nucleotide Substitutions in the Human Genome*. PLoS Genet 4(5): e1000071
- Giegerich, R., Brazma, A., Jonassen, I., Ukkonen, E., Vingron, M. (2008). *The BREW workshop series: a stimulating experience in PhD education*. Briefings in Bioinformatics 9(3), 250-253
- Grafahrend-Belau, E., Schreiber, F., Heiner, M., Sackmann, A., Junker, B.H., Grunwald, S., Speer, A., Winder, K., Koch, I. (2008). *Modularization of biochemical networks based on classification of Petri net t-invariants*. BMC Bioinformatics 9:90
- Grunwald, S., Speer, A., Ackermann J, Koch, I. (2008). *Petri net modelling of gene regulation of the Duchenne muscular dystrophy*. Biosystems 2008, Mar 10
- Hain, T., Hossain, H., Chatterjee, S.S., Machata, S., Volk, U., Wagner, S., Brors, B., Haas, S., Kuenne, C.T., Billion, A., Otten, S., Pane-Farre, J., Engelmann, S. and Chakraborty, T. (2008). *Temporal transcriptomic analysis of the Listeria monocytogenes EGD-e sigmaB regulon*. BMC Microbiol-ogy 8:20
- Hufton, A.L., Groth, D., Vingron, M., Lehrach, H., Poustka, A.J., Panopoulou, G. (2008). *Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement*. Genome Res 18(10), 1582—1591
- Jacob, J., Jentsch, M., Kostka, D., Bentink, S., Spang, R. (2008). *Detecting hierarchical structure in molecular characteristics of disease using transitive approximations of directed graphs*. Bioinformatics 24(7): 995-1001
- Kaluza, P., Vingron, M., Mikhailov, A.S. (2008). *Self-correcting networks: Function, robustness, and motif distributions in biological signal processing*. Chaos 18, 026113
- Kielbasa, S., Vingron, M. (2008). *Transcriptional Autoregulatory Loops Are Highly Conserved in Vertebrate Evolution*. PLoS ONE 3(9): e3210
- Koch, I. Fuellen, G. (2008). *A review of bioinformatics education in Germany*. Briefings in Bioinformatics Advance Access published on March 1, 2008
- Koch, I., Heiner, M. (2008). *Petri nets*. Analysis of Biological Networks, Wiley Book Series in Bioinformatics, chapter 7:139-180
- Koestner, U., Shnitsar, I., Linnemannstoens, K., Hufton, A.L., Borchers, A. (2008). *Semaphorin and neuropilin expression during early morphogenesis of Xenopus laevis*. Dev Dyn 237(12), 3853—3863
- Kostka D, Spang R (2008). *Microarray Based Diagnosis Profits from Better Documentation of Gene Expression Signatures*. PLoS Comput Biol 4(2): e22.
- Lee, J.S., Krause, R., Schreiber, J., Mollenkopf, H.-J., Kowall, J., Stein, R., Jeon, B.-Y., Kwak, J.-Y., Song, M.-K., Patron, J.P., Jorg, S., Roh, K., Cho, S.-N., Kaufmann, S.H.E. (2008). *Mutation in the Transcriptional Regulator PhoP Contributes to Avirulence of Mycobacterium tuberculosis H37Ra Strain*. Cell Host & Microbe 3(2), 97-103
- Lorena, A.C.; Costa, I.G.; de Souto, M.C.P. (2008). *On the Complexity of Gene Expression Classification Data Sets*. Eighth International Conference on Hybrid Intelligent Systems (HIS) 2008, 10-12 Sept. 2008 Page(s):825 - 830
- Manke, T., Roeder, H.G., Vingron, M. (2008). *Statistical Modeling of Transcription Factor Binding Affinities Predicts Regulatory Interactions*. PLoS Comput Biol; 4(3): e1000039
- Manke, T., Roeder, H.G., Vingron, M. (2008). *A biophysical approach to large-scale protein-DNA binding data*. Modern Genome Annotation. The Bio Sapiens Network. Springer, Chapter 2.2, 91-102

- Michael S, Trav G, Chenna R, Chica C, Gibson TJ (2008). *Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation*. *Bioinformatics* 24(4):453-7
- Mueller, F.J., Laurent, L., Kostka, D., Ulitsky, I., Williams, R., Lu, C., Rao, M.S., Shamir, R., Schwartz, P.H., Schmidt, N.O., Loring, J.F. (2008). *Regulatory networks define phenotypic classes of human stem cell lines*. *Nature* 455, 401-105
- Pape, U., Rahmann, S., Vingron, M. (2008). *Natural Similarity Measures between Position Frequency Matrices with an Application to Clustering*. *Bioinformatics*
- Pape, U.J., Klein, H., Vingron, M. (2008). *Statistical detection of co-operative transcription factors with similarity adjustment*. *Lecture Notes in Informatics: Proc. of German Conference for Bioinformatics (GCB)*, 2008, 96-105
- Pape, U.J., Vingron, M. (2008). *Statistics for Co-Occurrence of DNA Motifs*. *Proc. of International Workshop for Applied Probability (IWAP)*, 2008
- Pape, U.J., Rahmann, S., Sun, F., Vingron, M. (2008). *Compound Poisson approximation of DNA motif counts on both strands*. *Journal of Computational Biology* 15 (6): 547-564
- Polak, P., Arndt, P.F. (2008). *Transcription induces strand-specific mutations at the 5' end of human genes*. *Genome Res* 18:1216-1223
- Rausch, T., Emde, A.K., Weese, D., Doering, A., Notredame, C., Reinert, K. (2008). *Segment-based multiple sequence alignment*. *Bioinformatics* 24(16):i187-i192
- Rödelsperger C, Dieterich C. (2008). *Syntenator: Multiple gene order alignments with a gene-specific scoring function*. *Algorithms Mol Biol* 3:14
- A. Schliep, R. Krause (2008). *Efficient algorithms for the computational design of optimal tiling arrays*. *IEEE/ACM Trans Comput Biol Bioinform.* 2008 Oct-Dec; 5(4):557-67 (Invited paper selected from WABI 2007)
- Schulz, M.H.*, Bauer, S.*, Robinson, P.N. (2008). *The generalised k-Truncated Suffix Tree for time- and space-efficient searches in multiple DNA or protein sequences*. *International Journal of Bioinformatics Research and Applications* 4(1), 81- 95 // *shared first authorship
- Schulz, M.H., Weese, D., Rausch, T., Doering, A., Reinert, K., Vingron, M. (2008). *Fast and adaptive variable order Markov chain construction*. *WABI 2008*: 306-317
- Squartini, F., Arndt, P.F. (2008). *Quantifying the Stationarity and Time Reversibility of the Nucleotide Substitution Process*. *Molecular Biology and Evolution* 25(12): 2525-2535
- Sultan, M.* , Schulz, M.H.*, Richard, H.*, Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keefe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, ML (2008). *A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome*. *Science* 321(5891), 956-960 // *shared first authorship
- Tegha-Dunghu, J., Neumann, B., Reber, S., Krause, R., Erfle, H., Walter, T., Held, M., Rogers, P., Hupfeld, K., Ruppert, T., Ellenberg, J., Gruss, O.J. (2008). *EML3 is a nuclear microtubule-binding protein required for the correct alignment of chromosomes in meta-phase*. *J Cell Sci* 121: 1718-1726
- Toenjes, M., Schueler, M., Hammer, S., Pape, U.J., Fischer, J.J., Berger, F., Vingron, M., Sperling, S. (2008). *Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes*. *Mol. BioSyst.*4, 589 - 598
- Ulitsky, I., Gat-Viks, I., Shamir, R. (2008). *MetaReg: a platform for modeling, analysis and visualization of biological systems using large-scale experimental data*. *Genome Biology* 9:R1
- Weese, D., Schulz, M.H. (2008). *Efficient string mining under constraints via the deferred frequency index*. *Industrial Conference for Data Mining (ICDM 2008)*, LNAI 5077, 374-388



Zemojtel, T., Kielbasa, S.Z., Arndt, P.F., Chung, H.R., Vingron, M. (2008). *Methylation and deamination of CpGs generate p53-binding sites on a genomic scale.* Trends in Genetics 25(2), 63-66

2007

Arndt, P.F., Vingron, M. (2007). *The Otto Warburg International Summer School and Workshop on Networks and Regulation.* BMC Bioinformatics 8(Suppl 6):S1

Bais, A.S., Grossmann, S., Vingron, M. (2007). *Simultaneous alignment and annotation of cis-regulatory regions.* Bioinformatics 23(2):e44-e49

Bais, A.S., Grossmann, S., Vingron, M. (2007). *Incorporating evolution of transcription factor binding sites into annotated alignments.* Journal of Biosciences 32, Suppl.1, 841-850

Beisel, C., Buness, A., Roustan-Espinoza, I.M., Koch, B., Schmitt, S., Haas, S.A., Hild, M., Katsuyama, T., Paro, R. (2007). *Comparing active and repressed expression states of genes controlled by the Polycomb/Trithorax group proteins.* Proc. Natl. Acad. Sci. 104(42):16615-16620

Bozek, K., Kielbasa, S., Kramer, A., Herzel, HP. (2007). *Promotor Analysis of Mammalian Clock Controlled Genes.* Genome Informatics Series Vol. 18, 65-74

Costa, I.G., Roepcke, S., Schliep, A. (2007). *Gene expression trees in lymphoid development.* BMC Immunology 8:25

Costa, I.G., Krause, R., Opitz, L., Schliep, A. (2007). *Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data.* BMC Bioinformatics 8(Suppl 10):S3

Costa, I.G., de Souto, M.C.P., Schliep, A. (2007). *Validating Gene Clusterings by Selecting Informative Gene Ontology Terms with Mutual Information.* Lecture Notes in Computer Science, Advances in Bioinformatics and Computational Biology 4643/2007, 81-92

Chung, H.-R., Kostka, D., Vingron, M. (2007). *A physical model for tiling array analysis.* Bioinformatics 23(13):i80-i86

de la Chaux, N., Messer, P.W., Arndt, P.F. (2007). *DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage.* BMC Evol Biol. 7: 191

Demetrius, L., Gundlach, V.M., Ziehe, M. (2007). *Darwinian fitness and the intensity of natural selection: Studies in sensitivity analysis.* Journal of Theoretical Biology 249(4), 641-653

Demetrius, L., Ziehe, M. (2007). *Darwinian fitness.* Theoretical Population Biology 72(3), 323-345

Dieterich C, Franz MW, Vingron M. (2007). *Developments in CORG: a gene-centric comparative genomics resource.* Nucleic Acids Res 35(Database issue):D32-5

Georgi, B., Spence, M.A., Flodman, P., Schliep, A. (2007). *Mixture model based group inference in fused genotype and phenotype data.* Studies in Classification, Data Analysis, and Knowledge Organization, 2007

Georgi, B., Schultz, J., Schliep, A. (2007). *Context-Specific Independence Mixture Modelling for Protein Families.* Lecture Notes in Computer Science, PKDD 2007, Vol. 4702/2007, p.79-90

Georgi, B., Schliep, A. (2007). *Partially-supervised context-specific independence mixture modeling.* Workshop on Data Mining in Functional Genomics and Proteomics, ECML 2007

Grossmann, S., Bauer, S., Robinson, P.N., Vingron, M. (2007). *Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.* Bioinformatics 23(22):3024-31

Guo, G., Bauer, S., Hecht, J., Schulz, M.H., Busche, A., Robinson, P.N. (2007). *A short ultraconserved sequence drives transcription from an alternate FBN1 promoter.* The International Journal of Biochemistry & Cell Biology

Haesler, S., Rochefort, C., Licznarski, P., Georgi, B., Osten, P., Scharff, C. (2007). *Knockdown of FoxP2 in Songbird Basal Ganglia Impairs Song Learning.* PloS Biology 5(12), e321

Hooper, S.D., Boué, S., Krause, R., Jensen, L. J., Mason, C.E., Ghanim, M., White, K.P., Furlong, E. E. M., Bork, P. (2007). *Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis.* Molecular Systems Biology 3 Article number: 72

- Kaluza, P., Ipsen, M., Vingron, M., Mikhailov, A.S. (2007). *Design and statistical properties of robust functional networks: A model study of biological signal transduction*. PHYSICAL REVIEW E 75:015101 (R)
- Kielbasa, S., Herzel, HP., Axmann, I.M. (2007). *Regulatory Elements of Marine Cyanobacteria*. Genome Informatics Series 18, 1-11
- Klau, G.W., Rahmann, S., Schliep, A., Vingron, M. and K. Reinert (2007). *Integer linear programming approaches for non-unique probe selection*. Discrete Applied Mathematics 155(6-7), 840-856
- Klein, H., Vingron, M. (2007). *Using Transcription Factor Binding Site Co-Occurrence to Predict Regulatory Regions*. Genome Informatics 18, 109-118
- Lichtenauer U.D., Duchniewicz M., Kolanczyk M., Hoeflich A., Hahner S., Else T., Bicknell A.B., Zemojtel, T., Stallings N.R., Schulte D.M., Kamps M.P., Hammer G.D., Scheele J.S., Beuschlein F. (2007). *Pre-B-Cell Transcription Factor 1 and Steroidogenic Factor 1 Synergistically Regulate Adrenocortical Growth and Steroidogenesis*. Endocrinology 148(2):693-704
- Mao, L., Zabel, C., Herrmann, M., Nolden, T., Mertes, F., Magnol, L., Chabert, C., Hartl, D., Heralut, Y., Delabar, J.M., Manke, T., Himmelbauer, H., Klose, J. (2007). *Proteomic Shifts in Embryonic Stem Cells with Gene Dose Modifications Suggest the Presence of Balancer Proteins in Protein Regulatory Networks*. PLoS ONE 2(11): e1218
- Markowitz, F., Kostka, D., Troyanskaya, O.G., Spang, R. (2007). *Nested effects models for high-dimensional phenotyping screens*. Bioinformatics 23(13):i305-i312
- Markowitz, F., Spang, R. (2007). *Inferring cellular networks - a review*. BMC Bioinformatics 8(Suppl 6):S5
- Messer, P.W., Bundschuh, R., Vingron, M., Arndt, P.F. (2007). *Effects of Long-Range Correlations in DNA on Sequence Alignment Score Statistics*. Journal of Computational Biology 14(5): 655-668
- Messer, P.W., Arndt, P.F. (2007). *The Majority of Recent Short DNA Insertions in the Human Genome Are Tandem Duplications*. Molecular Biology and Evolution 24(5):1190-1197
- Opitz, L., Schliep, A., Posch, A. (2007). *Analysis of fused in-situ hybridization and gene expression data*. Advances in Data Analysis, p. 577-584
- Roider, H.G., Kanhere, A., Manke, T., Vingron, M. (2007). *Predicting transcription factor affinities to DNA from a biophysical model*. Bioinformatics 23(2), 134-141
- Rungsarityotin, W., Krause, R., Schodl, A., Schliep, A. (2007). *Identifying protein complexes directly from high-throughput TAP data with Markov random fields*. BMC Bioinformatics 2007, 8:482
- Schliep, A., Krause, R. (2007). *Efficient Computational Design of Tiling Arrays Using a Shortest Path Approach*. Algorithms in Bioinformatics, p. 383-394
- Zemojtel, T., Penzkofer, T., Schultz, J., Dandekar, T., Badge, R., Vingron, M. (2007). *Exonization of active mouse L1s: a driver of transcriptome evolution?*. BMC Genomics 2007, 8:392
- 2006**
- Ayerdi-Izquierdo, A., Stavrides, G., Selles-Martinez, J.J., Larrea, L., Bovo, G., de Munain, A.L., Bisulli, F., Marti-Masso, J.F., Michelucci, R., Poza, J. J., Tinuper, P., Stephani, U., Striano, P., Striano, S., Staub, E., Sarafidou, T., Hinzmann, B., Moschonas, N., Siebert, R., Deloukas, P., Nobile, C., & Perez-Tur, J. (2006). *Genetic analysis of the LGI/Epitempin gene family in sporadic and familial lateral temporal lobe epilepsy*. Epilepsy Res. 70, 118-126
- Behzadi, B. and Vingron, M. (2006). *An Improved Algorithm for the Macro-evolutionary Phylogeny Problem*. CPM 2006, LNCS 4009, 177-187
- Behzadi, B., Vingron, M. (2006). *Reconstructing Domain Compositions of Ancestral Multi-Domain Proteins*. Comparative Genomics RECOMB 2006, LNBI 4205, 1-10
- Costa, I.G., Schliep, A. (2006). *On the feasibility of Heterogeneous Analysis of Large Scale Biological Data*. Proceedings of ECML/PKDD 2006 Workshop on Data and Text Mining for Integrative Biology, 55-60
- Dohm, J. C., Vingron, M., Staub, E. (2006). *Horizontal Gene Transfer in Aminoacyl-tRNA Synthetases Including Leucine-Specific Subtypes*. J. Mol. Evol. 63(4), 437-447



- Duchniewicz M and Zemojtel T, Kolanczyk M, Grossmann S, Scheele JS, Zwartkruis FJ (2006). *Rap1A-deficient T and B cells show impaired integrin-mediated cell adhesion*. Mol Cell Biol. 26(2), 643-53
- Groene, J., Mansmann, U., Meister, R., Staub, E., Roepcke, S., Heinze, M., Klaman, I., Brummendorf, T., Hermann, K., Loddenkemper, C., Pilarsky, C., Mann, B., Adams, H. P., Buhr, H. J., & Rosenthal, A. (2006). *Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III*. Int. J. Cancer 119, 1829-1836
- Hecht, H., Kuhl, H., Haas, S.A., Bauer, S., Poustka, A.J., Lienau, J., Schell, H., Stiege, V., Seitz, V., Reinhardt, R., Duda, G.N., Mundlos, S. and Robinson, P.N. (2006). *Gene Identification and Analysis of Transcripts Differentially Regulated in Fracture Healing by EST Sequencing in the Domestic Sheep*. BMC Genomics, 7:172
- Hummel, M., Bentink, S., Berger, H., Klapper, W., Wessendorf, S., Barth, T.F.E., Bernd, H.-W., Cogliatti, S.B., Dierlamm, J., Feller, A.C., Hansmann, M.L., Haralambieva, E., Harder, L., Hasenclever, D., Kühn, M., Lenze, D., Lichter, P., Martin-Subero, J.I., Möller, P., Müller-Hermelink, H.K., Ott, G., Parwaresch, R.M., Pott, C., Rosenwald, A., Rosolowski, M., Schwaenen, C., Stürzenhofecker, B., Szczepanowski, M., Trautmann, H., Wacker, H.-H., Spang, R., Loeffler, M., Trümper, L., Stein, H., Siebert, R. (2006). *A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling*. NEJM 354:2419-2430
- Lipatov, M., Arndt, P.F., Hwa, T., Petrov, D.A. (2006). *A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution*. Journal of Molecular Evolution 62: 168-75
- Luz H, Vingron M (2006). *Family specific rates of protein evolution*. Bioinformatics 2(10):1166-1171.
- Luz H, Staub S, and Vingron M (2006). *About the interrelation of evolutionary rate and protein age*. Genome Informatics 17(1): 240-250.
- Manke, T., Demetrius, L., Vingron, M. (2006). *An Entropic Characterization of Protein Interaction Networks and Cellular Robustness*. Royal Soc. Interface 3(11):843-850.
- Messer, P.W., Arndt, P.F. (2006). *CorGen-measuring and generating long-range correlations for DNA sequence analysis*. W692-W695 Nucleic Acids Research 34
- Messer, P.W., Bundschuh, R., Vingron, M., Arndt, P.F. (2006). *Alignment Statistics for Long-Range Correlated Genomic Sequences*. RECOMB 2006, LNBI 3909:426-440
- Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Koenig, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F., Fischer, M. (2006). *Gene-expression based classification of neuroblastoma patients using a customized oligonucleotide-microarray outperforms current clinical risk stratification*. J. Clin. Oncol. 24:5070-5078
- Pape, U.J., Grossmann, S., Hammer, S., Sperling, S., Vingron, M. (2006). *A new statistical model to select target sequences bound by transcription factors*. Genome Inform 17(1):134-40
- Qin, Y., Polacek, N., Vesper, O., Wilson, D. N., Staub, E., Einfeldt, E., & Nierhaus, K. H. (2006). *The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome and is essential for viability at high ionic strength*. Cell 127(4): 721-733
- Roepcke, S., Zhi, D., Vingron, M., Arndt, P.F. (2006). *Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters*. Gene 365: 48-56
- Schoenhuth, A., Costa, I.G., Schliep, A. (2006). *Semi-supervised Clustering of Yeast Gene Expression*. Japanese-German Workshop on data analysis and classification, Springer
- Schulz, W.A., Steinhoff, C., Florl, A.R. (2006). *Methylation of Endogenous Human Retroelements in Health and Disease*. CTMI 310:211-250
- Schwecke T, Göttling K, Durek P, Duenas I, Käufer NF, Zock-Emmenthal S, Staub E, Neuhof T, Dieckmann R, von Döhren H (2006). *Nonribosomal peptide synthesis in S. pombe and the architecture of ferrichrome-type siderophore synthetases in fungi*. ChemBioChem 7(4): 612-622

Staub, E., Mackowiak, S., Vingron, M. (2006). *An inventory of yeast proteins that are associated with nucleoli and ribosomal components*. Genome Biology 7(10): R98

Staub, E., Grone, J., Mennerich, D., Roepcke, S., Klamann, I., Hinzmann, B., Castanos-Velez, E., Mann, B., Pilarsky, C., Brummen-dorf, T., Weber, B., Buhr, H. J., & Rosenthal, A. (2006). *A genome-wide map of aberrantly expressed chromosomal islands in colorectal cancer*. Mol. Cancer 5, 37

Steinhoff C., Vingron M. (2006). *Review: Normalization and Quantification of Differential Expression*. Briefings in Bioinformatics, Advance Access published on March 7, 2006

Yang, X., Bentink, S., Scheid, S., Spang, R. (2006). *Similarities of Ordered Gene Lists*. Journal of Bioinformatics and Computational Biology 4(3), 693–708

Yeang, Chen-Hsiang, Vingron, Martin (2006). *A joint model of regulatory and metabolic networks*. BMC Bioinformatics 7:332

Zemojtel T, Kolanczyk M, Kossler N, Stricker S, Lurz R, Mikula I, Duchniewicz M, Schuelke M, Ghafourifar P, Martasek P, Vingron M, Mundlos S. (2006). *Corrigendum to "Mammalian mitochondrial nitric oxide synthase: Characterization of a novel candidate" [FEBS Lett. 580 (2006) 455–462]*. FEBS Lett. 581(10):2072-2073

Zemojtel T, Penzkofer T, Duchniewicz M, Zwartkruis FJT (2006). *hRap1B-retro: a novel human processed Rap1B gene blurs the picture?* Leukemia 20:145-146

Zemojtel T and Kolanczyk M, Kossler N, Stricker S, Lurz R, Mikula I, Duchniewicz M, Schuelke M, Ghafourifar P, Martasek P, Vingron M, Mundlos S. (2006). *Mammalian mitochondrial nitric oxide synthase: Characterization of a novel candidate*. FEBS Lett. 580(2), 455-462

Zemojtel T, Frohlich A, Palmieri MC, Kolanczyk M, Mikula I, Wyrwicz LS, Wanker EE, Mundlos S, Vingron M, Martasek P, Durner J. (2006). *Plant nitric oxide synthase: a never-ending story?* Trends Plant Sci. 2006 Nov, 11

PhD theses

2009

Benjamin Georgi: *Context-specific Independence Mixture Models for Cluster Analysis of Biological Data*. PhD Thesis, Bioinformatics, Freie Universität Berlin, 06/09 (supervisor: Alexander Schliep)

2008

Helge Roeder: *Eukaryotic promoter analysis by means of a biophysical model for DNA transcription factor interactions*. PhD Theses, Freie Universität Berlin, 11/08 (supervisor: Stefan Haas)

Utz Pape: *Statistics for Transcription Factor Binding Sites*. PhD Thesis, Freie Universität Berlin, 10/08 (supervisor: Martin Vingron)

Ivan Gesteira Costa Filho: *Mixture Models for the Analysis of Gene Expression: Integration of Multiple Experiments and Cluster Validation*. PhD Thesis, Bioinformatics, Freie Universität Berlin, 06/08 (supervisor: Alexander Schliep)

Ho-Joon Lee: *Computational Genomic Analysis of Transcriptional Regulation*. PhD Thesis, Freie Universität Berlin, 04/08 (supervisor: Martin Vingron)

Philipp Messer: *Tandem Duplications in the Human Genome*, PhD Thesis, Freie Universität Berlin, 03/08 (supervisor: Peter Arndt)

2007

Wasinee Rungsarityotin: *Algorithms to identify protein complexes from high-throughput data*. PhD Thesis, Bioinformatics, Freie Universität Berlin, 11/07 (supervisor: Alexander Schliep)

Abha Singh Bais: *Annotated Alignments*. PhD Thesis, Freie Universität Berlin, 07/07 (supervisor: Martin Vingron)

2006

Hannes Luz: *Family Specific Rates of Protein Evolution*. PhD Thesis, Freie Universität Berlin, 12/06 (supervisor: Martin Vingron)

Dennis Kostka: *Methodology for exploring and communicating molecular characteristics of disease*. PhD Thesis, Freie Universität Berlin, 12/06 (supervisor: Rainer Spang)

Stefanie Christina Scheid: *Novel Concepts for the Significance Analysis of Microarray Data*. PhD Thesis, Freie Universität Berlin, 10/06 (supervisor: Rainer Spang)



Jochen Jäger: *Deriving small diagnostic biomarker panels from genome wide, clinical microarray studies*. PhD Thesis, Freie Universität Berlin, 07/06 (supervisor: Rainer Spang)

Florian Markowetz: *Probabilistic Models for Gene Silencing Data*. PhD Thesis, Freie Universität Berlin, 04/06 (supervisor: Rainer Spang)

Student theses 2008

Christina Bianca Heitzer: *Prediction of Alternative Operons using an HMM approach with Transcription Factor Binding Sites and Intergenic Distances*. Diploma Thesis, Philipps-Universität Marburg, 2008 (supervisor: Roland Krause)

Sebastian Dominik Mackowiak: *Combined analysis of genome wide expression and copy number data of human tumors*. MSc Thesis, Freie Universität Berlin, 2008 (supervisor: Martin Vingron)

Marko Briesemann: *Evaluation and extension of a community detection approach using linear programming*. MSc Thesis, Freie Universität Berlin, 2008 (supervisor: Martin Vingron)

Christopher Hardt: *Evolutionary Rate Dynamics of Protein Families*. MSc Thesis, Freie Universität Berlin, 2008 (supervisor: Hannes Luz)

Christian Hoffmann: *Detection of Chimeric Small-Subunit rRNAs*. MSc Thesis, Freie Universität Berlin, 2008 (supervisor: Martin Vingron)

Christoph Hafemeister: *Efficient Computation of Probe Qualities*. MSc Thesis, Freie Universität Berlin Berlin, 2008.

Annekatriin Wiedenhoef: *The regulatory power of 3'UTRs: The analysis of the 3'UTR sequences of human cytosolic RP genes*. BSc Thesis, Freie Universität Berlin, 2008 (supervisor: Martin Vingron)

Peter Hansen: *Kategorisierung von Aminosäuren und genomweite Zusammenfassung von Proteindomänen zur niederparametrischen Modellierung von Protein-Evolution*. BSc Thesis, Freie Universität Berlin, 2008 (supervisor: Hannes Luz)

Katharina Schmidt: *A Phylogenetic Parsimony Method Considering Neighbored Gaps*. BSc Thesis, Freie Universität Berlin, 2008 (supervisor: Hannes Luz)

Jevgeni Erehman: *Theoretische Untersuchungen von Proteinstrukturen alternativ gespleißter Proteine*. BSc Thesis, Freie Universität Berlin, 2008 (supervisor: Martin Vingron)

2007

Ruben Schilling: *Elastische Registrierung in 3D Volumen Daten*. Diploma Thesis, Universität Freiburg, 2007 (supervisor: Alexander Schliep)

M. Turewicz: *Lernen von CSI Mixturen mit MCMC Methoden*. Diploma Thesis, Martin Luther University, Halle, 2007 (supervisor: Alexander Schliep)

Nicole de la Chaux: *The Evolution of the Human Genome: Insertions and Deletions in Protein Coding Regions*. MSc Thesis, Freie Universität Berlin, 2007 (supervisor: Peter Arndt)

Christopher Hardt, MSc Thesis, Freie Universität Berlin, 2007 (supervisor: Hannes Luz)

Thomas Engleitner: *Insertions and Deletions in the Human Genome*. BSc Thesis, Freie Universität Berlin, 2007 (Supervisor: Peter Arndt)

Ricardo Raspe, BSc Thesis, Freie Universität Berlin, 2007 (supervisor: Hannes Luz)

J. Li: *Methoden für das Design von DNA Tiling arrays*. BSc Thesis, Freie Universität Berlin, 2007 (supervisor: Alexander Schliep)

M. Ruegen: *Optimale Probenauswahl fuer die Targeterkennung mit DNA microarrays*. BSc Thesis, Freie Universität Berlin, 2007 (supervisor: Alexander Schliep)

2006

Petko Fiziev, Diploma Thesis, Freie Universität Berlin, 2006 (supervisor: Martin Vingron)

Michael Seifert: *Analyzing microarray data using homogeneous and inhomogeneous Hidden Markov Models*; Diploma Thesis, Martin Luther University, Halle, 2006 (supervisor: Alexander Schliep)

Lennart Opitz: *Analyse von Bildern der mRNA- in Situ-Hybridisierung*. Diploma Thesis, Martin Luther University, Halle, 2006 (supervisor: Alexander Schliep, joint supervision with Dr. S. Posch, Halle)

Max Flöttmann, BSc Thesis, Freie Universität Berlin, 2006 (supervisor: Roland Krause)

Sebastian Mackowiak, BSc Thesis, Freie Universität Berlin, 2006 (supervisor: Roland Krause)

Christoph Hafemeister: *Learning topologies of conditional trees*. BSc Thesis, Freie Universität Berlin, 2006 (supervisor: Alexander Schliep)

Moritz Wade, BSc Thesis, Freie Universität Berlin, 2006 (supervisor: Aditi Kanhere)

Guest scientists since 2006

Shen Lin, Chinese Academy of Sciences, Wuhan, P.R. China, 11.01.2009 – 10.01.2010

Ina Koch, Technical University of Applied Sciences Berlin, since 2006

Pierre Nicodeme, Laboratoire LIX, Ecole Polytechnique, Palaiseau cedex, Frankreich, 24.08. – 05.09.2009; 10.09. – 01.10.2007; 25.08. – 05.10.2005

Peter Clote, Boston College, Chestnut Hill, MA, USA; 01.08. – 03.09.2009; 01.05. – 30.06.2007

Oliver Eulenstein, Iowa State University, Ames, Iowa, USA, 02.01. – 30.06.2009

Guillaume Bourque, Genome Institute of Singapore, Singapore, 10.06. – 12.06.2009

Martin Frith, AIST Tokyo, Japan, 22.04. – 13.05.2009

Roland Dosch, Universität Genf, Zürich, 07.01. – 10.01.2009

Lev Levitin, Boston University, Boston, MA, USA, 12.07. – 14.08.2008; 10.07. – 29.08.2007; 18.06. – 31.08.2006

Jerzy Tiuryn, Warsaw University; Polen; 06.-08.03.2008

Xiaoqiu Huang, Iowa State University, Iowa, USA; 15.08. – 14.11.2007

Ricardo Bringas, Centro de Ingenieria, La Habana, Kuba, 02.07. – 27.09.2007; 02.04. – 09.07.2005; 01.09. – 31.10.2004

Irit Gat-Viks, Tel-Aviv University, Israel; 29.05. – 01.06.2007

Marcilio Carlos Pereira de Souto, University of Tio Grande do Norte, Natal, Rn, Brasil, 01.10.2006 – 30.09.2007

Morgan Bishop, University SUNY Geneseo, New York, NY, USA, 30.4.-4.5.2007

Dmitri Petrov, Stanford University, 08/2007

Norbert Doyer, Warsaw University, Polen, 19.06. – 07.07.2006

Serdar Cakici, Sabanci University, Istanbul, Türkei, 28.07. – 28.08.2006

Pawel Gorecki, Warsaw University, Polen, 01.10.2006 – 30.09.2007

Abhinab Ray, Juni Line Faculty, India, 17.05. – 16.08.2006

Jacub Pas, Warsaw University, Polen, 03.11.2005 – 31.01.2006

Jing Zhang, Yunnan University, P.R. China, 01.09.2005 – 28.02.2006