



Department of Computational Molecular Biology



Head:

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

Co-ordination:

(Berlin Center for Genome Based Bioinformatics)
Dr. Patricia Béziat
Phone: +49 (0)30-8413 1716
Fax: +49 (0)30-8413 1671
Email: beziat@molgen.mpg.de

Secretary:

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

Introduction

The research of the Computational Molecular Biology Department focuses on the analysis of the data generated by today's sequencing and functional genomics programs. Numerous challenging questions can be posed based on these data concerning, e.g., the description of gene structure of human and mouse genes, gene regulation, the mapping of protein sequence space, whole genome comparison, the analysis of large scale gene expression data, and their utilization for disease diagnosis.

The department is structured into several smaller research units looking into certain of these questions. Where meaningful, groups interact very closely with each other. Depending on the methods required, some of the groups are more mathematical, while others apply existing methods to pursue their biological questions. Overall, department staff comes from various backgrounds including computer science, mathematics, genetics, biochemistry, biology, and physics.

Some significant research results of the last years are:

- the resolvent method for computing an amino acid exchange matrix (Müller, Spang);
- the "variance stabilization" normalization method for microarrays (von Heydebreck, in collaboration with W. Huber, DKFZ, Heidelberg)
- the online databases SYSTERS (protein families, Krause), GeneNest & SpliceNest (gene structure, Haas), and CORG (Comparative Regulatory Genomics, Dieterich)
- the establishment of the connection between protein-protein interactions of yeast transcription factors and the co-occurrence of their binding sites (Manke);
- collaborative projects on ancient genome duplications (Krause, with Panopoulou, Dept. Lehrach), and analysis of gene expression data on heart disease (von Heydebreck, with Sperling, Dept. Lehrach).

The department was founded in October 2000 with the new director taking office in October of that year. The rooms, then still at Harnackstraße, were quickly filled, partly with people coming along from Heidelberg to Berlin, and additionally with newly re-

cruited scientists. Jens Stoye, who at the time was a group leader for algorithmics on a C3 position, has meanwhile received an offer for a professorship in Bielefeld which he accepted. He left in spring 2002 and was succeeded by Alexander Schliep. More recently, Jörg Schulz, group leader protein function analysis, has received and accepted an offer for a professorship in Würzburg. In his place, Peter Arndt will build up a new group starting in October 2003. Early 2002 the department moved from Harnackstraße to the third floor of tower 2 of the main institute building. As of fall 2003, just above 30 people work in the department.

The computer equipment of the department comprises PCs under Linux or workstations on the desks, a central 16 processor compute server with large memory, a specialized computer for data base searching, and 2 TB of disk space. This set-up is maintained by the department system administrator 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 is bringing 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.

We are involved in a number of national and international projects and collaborations. Most prominently, we are part of the Berlin Center for Genome Based Bioinformatics (BCB), a large network made up of several Berlin bioinformatics groups and funded by the German Federal Ministry of Education and Research (BMBF). The “Computational Diagnostics” group headed by R. Spang is part of BCB. This group is also active in the German National Genome Research Network (NGFN) providing education and know-how in microarray data analysis.

With support from BCB and Max Planck Society, the department has organized a major international conference. RECOMB, the Annual International Conference on Computational Molecular Biology was held in Berlin in April 2003. This event brought more than 500 attendees to Berlin and presented a selection of highest quality up-to-date research in the field.

General information

(also see group reports for further information)

External funding

DFG, Vi 160/3: *Rechnergestützte phylogenetische Analyse großer genomischer Abschnitte*. Joint with Prof. Dr. A. von Haeseler, MPI für evolutionäre Anthropologie, Leipzig, ended in 2002, 1 position

DFG, SFB 1904 „Theoretische Biologie: Robustheit, Modularität und evolutionäres Design lebender Systeme“, subproject *Correlation between regulatory DNA sequences and gene expression data*, 1 position

BMBF-DHGP, 01KW9911/9: *Erstellung von Genexpressionsprofilen von Tumor- und Normalgewebe mittels komplexer Hybridisierung und mathematischer Analysen der Expressionsmuster*. Joint with Prof. Dr. Annemarie Poustka, Deutsches Krebsforschungszentrum, Heidelberg. Currently in its 2nd funding period, 1 position

BMBF-DHGP, 01KW9955/3: *Auswertung von EST-Daten in Hinblick auf Genstruktur, funktionale Annotation und Expressionsanalyse*. Joint with Dr. Bernhard Korn, Deutsches Krebsforschungszentrum, Heidelberg. Currently in its 2nd funding period, 1 position.

BMBF, (031U109/C): *Berlin Center for Genome Based Bioinformatics (BCB)*. Center grant to a consortium of Berlin research institutions and universities, 2 scientists, 1 administrative position

BMBF-NGFN Grant 031U117 (Optimierungsfond): *Bereitstellung von Ressourcen und Transfer von Know-how für die Analyse von Genexpressionsprofilen im NGFN*, 2 positions

BMBF: *Helmholtz Netzwerk Bioinformatik*. Consortium of German bioinformatics groups establishing a common, web-based infrastructure for bioinformatics. Ended in 02, 1 position



EU: *The European Molecular Biology Linked Original Resources* (TEMBLOR), 1 position.

EU: *BioSapiens*: Bioinformatics Excellence Network

PhD Theses

Antje Krause: *Large Scale Clustering of Protein Sequences*. PhD Thesis, University of Bielefeld, June 2002

Heiko Schmidt: *Phylogenetic Trees from Large Datasets*. PhD Thesis, University of Düsseldorf, July 2003

Appointments, scientific honors & memberships

Ina Koch, C2 Professorship at Technical University of Applied Sciences, Berlin (from 1.4.03)

Jörg Schulz, C3 Professorship at University of Würzburg (from 1.9.03)

Jens Stoye, C4 Professorship at University of Bielefeld (from 2002)

Sven Rahmann: Best paper Award, IEEE Computer Society Bioinformatics Conference, Palo Alto, USA, 2002

Anja von Heydebreck: Submission for Jahrestagung der Dtsch. Gesellschaft für Pathologie 2003 was awarded for one of four best research contributions

Organization of scientific events

Vingron M, organizer of a *MPG-Polish workshop on Bioinformatics*, Berlin, 2001

Vingron M & Freytag J-C, organizer of the *International BCB-Workshop on Data Bases and Data Integration in Genome Research*, Berlin, 7.+8.2.2002

Vingron M, local organizer of *The Seventh Annual International Conference on Research in Computational Molecular Biology - RECOMB 2003*, Berlin, 10.-13.4.2003

Indo-German workshop on Proteomics and Bioinformatics, Berlin, 2003

Co-operations

Detecting SNPs in regulatory regions, with Jörg Hoheisel, DKFZ, Heidelberg

Protein sequence analysis, reliability of multiple alignments, Andrei Lupas, Tübingen

Maximum Likelihood methods for computing phylogenetic trees, Arndt von Haeseler Düsseldorf

Gene structure and alternative splicing, experimental validation by PCR and by microarrays, with Annemarie Poustka, Bernhard Korn, Deutsches Krebsforschungszentrum, Heidelberg

Gene regulatory networks, with Ricardo Bringas-Perez, Habana, Cuba

Experimental verification of predicted SRF target genes, with Alfred Nordheim, Tübingen

Gene expression analysis, EST assembly and clustering, probe design, with Inge Jonassen, Eivind Coward, University of Bergen, Norway

Graph algorithms for the analysis of protein protein interactions, with David Sherman, Bordeaux, France

Gene regulation and microarray data in fruit fly, with Alvis Brazma, European Bioinformatics Institute, Hinxton

Pattern recognition in DNA sequences, with Jerzy Tiuryn, Warsaw, Poland

Tree models of chromosomal aberrations in tumors, with Simon Tavare, USC, Los Angeles, USA

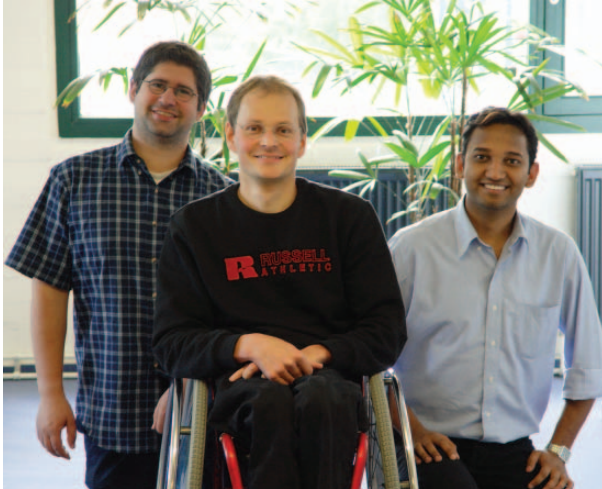
Pattern recognition in DNA sequences; orthology detection and syntenry, with Pavel Pevzner, UC San Diego, USA

Public relations

Organization of a public discussion *Hype oder Hoffnung - Podiumsdiskussion zur Rolle der Bioinformatik am Standort Berlin*, Berlin, 9.9.2002

Organization of the public seminar *Chancen der Bioinformatik in Berlin-Brandenburg - Die Biotechnologie und die Informatik lernen sich kennen*, Berlin, 1.4.2003

Gene Structure & Array Design Group



Head:

Dr. Stefan Haas

Phone: +49 (0)30-8413 1164

Fax: +49 (0)30-8413 1152

Email: stefan.haas@molgen.mpg.de

Scientists:

Dr. Eivind Coward (until 12/2001)

Dr. Ina Koch (until 3/2003)

Graduate students:

Shobhit Gupta

Sven Rahmann

Undergraduate students:

Marc Brüning (until 9/2002)

Stéphanie Boué (until 10/2002)

Scientific overview

The main interest of our group is the development of tools that enable the analysis of the exon-intron structure of genes with special emphasis on the evaluation of alternative splicing. In collaboration with experimentalists we are aiming to substantiate our *in silico* predictions by wet lab experiments.

ESTs and gene structure

Expressed sequence tags (ESTs) reflect semi-random parts of transcripts expressed in a defined tissue. Caused by the cost-efficient generation of ESTs the reliability of their sequence and annotation may vary strongly. However, the huge amount of EST sequences available in public databases comprises a valuable source for the reconstruction of so far unknown transcripts as well as for the analysis of gene expression. We developed the database GeneNest that represents genes by clusters of EST/mRNA sequences that share sequence similarities. Based on the subsequent sequence assembly these clusters are subdivided into contigs reflecting different transcripts of the respective gene. The consensus sequences derived from these contigs summarize the redundant EST sequence information and are usually of higher quality than the underlying ESTs. Therefore, our comprehensive set of consensus sequences can be efficiently used for database searches or further analysis of the respective transcripts.

By mapping these consensus sequences to the genome sequence using our SpliceNest software we derive potential exon-intron boundaries of the respective transcripts. Despite the improved sequence quality of the consensus sequences the predicted boundaries might still include artefacts, caused for instance by ESTs that originate from genomic DNA. In order to reliably detect real splicing events, we compute a confidence value for every exon based on the existence of splicing signals, the alignment quality, the redundant coverage by ESTs, etc. Similarly, we assign a confidence value to every potential splicing event thus prioritising the variants for validation in large-scale experiments.



EST tissue distribution

Besides the sequence information ESTs also provide details about the tissue or developmental stage from which these cDNAs descended. Despite the fact that these annotations are prone to errors they still provide a means to evaluate the expression of genes/transcripts. We simulated EST clusters with a random distribution of ESTs from different tissues in order to derive a p-value that describes the likelihood of observing the given number of ESTs present in a tissue by chance. Sorting EST-clusters according to their p-value provides us with a list comprising genes that are highly and/or specifically expressed in certain tissues. Since tissue-specific expression is more frequently observed for transcripts rather than genes we are currently focusing on the prediction of tumour/tissue-specific alternative transcripts. Such a collection of genes/transcripts showing tissue-specific expression will provide a basis for the analysis of diseases that are connected to a specific type of tissue such as many kinds of tumours. In a tight collaboration with the Resource Center (RZPD) we experimentally verify the predicted transcripts and their expression on a variety of tissues, aiming to define a reliable set of alternative transcripts for the design of a DNA microarray. We are especially focusing on splice isoforms for all genes on human chromosome 21 as well as for disease related genes on chromosome X that will be experimentally analysed by our in-house collaborators. A set of splice variants was also analysed on the level of protein structure in order to evaluate if alternative splicing leads to structural differences between isoforms.

Chip design

In the context of the construction of DNA-microarrays we developed algorithms based on a 'Longest Common Factor' approach aiming to generate microarrays that represent genes by a minimal set of short oligomers. In addition, we developed the software GenomePRIDE for the design of PCR- and long oligomer-based DNA-microarrays, which primarily computes PCR-primers or long oligomers that reflect unique parts of a set of genes. We successfully applied the software to the design of PCR-based whole transcriptome arrays for *Drosophila melanogaster*, *Schizosaccharomyces pombe*, *Listeria monocytogenes* etc. We also addressed further applications like the use of specific PCR-amplicons in RNAi experiments, the design of genomic tiling path arrays, and the design of splice isoform specific amplicons.

All tools developed in our group are either licensed (GenomePRIDE) or are presented via interactive WWW-interfaces (GeneNest, SpliceNest) that visualize all data related to EST-clusters, and the genomic mapping of the EST consensus sequences directly linking to in-house database as well as to external databases such as the EMBL- and the RZPD-database. These graphical web sites are particularly designed to support the efficient analysis of splice variants and/or tissue-specifically expressed transcripts covered by ESTs.

General information

Publications 2000-2003

Haas SA, Hild M, Wright APH, Hain T, Talibi D & **Vingron M** (2003). *Genome-scale design of PCR primers and long oligomers for DNA microarrays*. *Nucleic Acids Res* 31(19):5576-81

Kriventseva EV, **Koch I**, Apweiler R, **Vingron M**, Bork P, Gelfand MS & Sunyaev S (2003). *Increase of functional diversity by alternative splicing*. *Trends in Genetics* 19(3):124-128

Rahmann S (2003). *Fast large scale oligonucleotide selection using the longest common factor approach*. *J Bioinf Comp Biol* 1(2): 343-361

Rahmann S & Rivals E (2003). *On the distribution of the number of missing words in random texts*. *Combinatorics, Probability and Computing* 12:73-87

Xue-Franzen Y, **Haas SA**, Brino L, Gusnanto A, Reimers M, Talibi D, **Vingron M**, Ekwall K & Wright APH (2003). *A DNA microarray for fission yeast: minimal changes after a temperature shift to 36 C*. *Yeast* (in press)

Boué S, **Vingron M**, Kriventseva E & **Koch I** (2002). *Theoretical analysis of alternative splice forms using computational methods*. *Bioinformatics* 18(Suppl 2), eds. T. Lengauer, H.-P. Lenhof, 65-73

Coward E, Haas SA & Vingron M (2002). *SpliceNest: visualization of gene structure and alternative splicing based on EST clusters*. Trends Genet 18(1): 53-55

Krause A, Haas SA, Coward E, Vingron M (2002). *SYSTERS, GeneNest, SpliceNest: Exploring sequence space from genome to protein*. Nucleic Acids Res 30(1): 299-300

Boer JM, Huber W, Sultmann H, Wilmer F, **von Heydebreck A, Haas S**, Korn B, Gunawan B, Vente A, Füzesi L, **Vingron M** & Poustka A (2001). *Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500 element cDNA array*. Genome Res 11(11): 1861-1870

Petersohn A, Brigulla M, **Haas S**, Hoheisel J, Völker U & Hecker M (2001). *Global analysis of the general stress response of Bacillus subtilis*. J Bacteriol 183(19): 5617-5631

Invited talks

Stefan Haas: *EST clustering*, Dept. of Informatics, University of Bergen, Norway, 2002

Stefan Haas: *Primer design for DNA-microarrays*, Natural Sciences Section, Södertörns Höskola, Huddinge, Sweden, 2002

Sven Rahmann: *Rapid large-scale oligonucleotide selection for microarrays*. Research seminar Affymetrix Corp., Emeryville, CA, USA, 8/2002

Teaching

Sven Rahmann, Praktikum und Seminar *Sequence comparison*, 6 SWS, SS 03, Freie Universität Berlin

Sven Rahmann, teaching assistant and tutor for the lecture *Algorithmische Bioinformatik*’, WS 2002/03, Freie Universität Berlin

Invited lectures

Stefan Haas, *Biologische Sequenzanalyse*, Akademie für Weiterbildung, Universities of Heidelberg/Mannheim, 2001-2003

Sven Rahmann, *Spezielle Methoden der Statistik*, and *Molekulare Evolution*, Akademie für Weiterbildung, Universities of Heidelberg/Mannheim, 2002

Diploma Theses

Stéphanié Boué, *Computational investigation of alternative splicing*. Master’s thesis in Bioinformatics at Max Planck Institute for Molecular Genetics Berlin & ESBS École supérieure de biotechnologie Strasbourg at ULP - Université Louis Pasteur Strasbourg, France, August 2002

Marc Bruning, *Genomweite Analyse von Expressed Sequence Tags zur Identifizierung gewebsspezifisch exprimierter Gene*, Diploma Thesis, Technical University of Berlin, 2002

Co-operations

Design of a whole genome microarray of Drosophila melanogaster, with M Hild, R Paro, J Hoheisel, ZMBH+DKFZ, Heidelberg (2001-2003)

Design of a whole genome microarray of Schizosaccharomyces pombe, with A Wright, Södertörns Höskola, Huddinge, Sweden (2002-2003)

Design of a flexible DNA-microarray for parallel use of PCR fragments in expression cloning and RNAi (Anopheles gambiae), with G Christophides, F Kafatos, EMBL, Heidelberg (2003)

Representing genes on DNA-chips by a minimal set of short oligonucleotides, with M Beier, FeBit AG, Mannheim

Analysis of conserved intronic sequences in the context of alternative splicing, with A Bindereif, University of Giessen

Prediction and experimental analysis of alternative splice variants based on ESTs, with B Korn, Resource Centre, Heidelberg

Protein Families & Evolution Group



Head:

Dr. Antje Krause

Phone: +49 (0)30-8413 1155

Fax: +49 (0)30-8413 1152

Email: antje.krause@molgen.mpg.de

Scientists:

Thomas Meinel

Dr. Eike Staub (starting 09/03)

Graduate student:

Hannes Luz

Undergraduate student:

Ralf Mehle (10/03-02/04)

Protein Families

With the overwhelming growth of biological sequence databases, handling these amounts of data has increasingly become a problem. Protein sequences constitute one such data type for which the databases have grown to an impressive size. A protein family contains sequences that are evolutionarily related. Generally, this is reflected by sequence similarity. Therefore, one aims at organizing the set of all protein sequences into clusters based on their sequence similarity. Clustering a large set of sequences as opposed to dealing only with the individual sequences offers several advantages. A frequent problem is the identification of sequences that are similar to a new query sequence. This task can be executed much quicker when only one comparison to an entire cluster has to be performed rather than one comparison per database sequence. Another important application lies in the possibility of analysing evolutionary relationships among the sequences in a cluster and the species they come from.

We designed a collection of graph-based algorithms to hierarchically partition a large set of protein sequences into homologous families and superfamilies (see Figure 1). The methods unified now under the name SYSTEMS (short for SYSTEMatic Re-Searching) are based on an all-against-all database search and run fully automated. Using these methods, we clustered a non-redundant union of the SWISS-PROT and TrEMBL databases as well as of the predicted protein sequence sets of several completely sequenced organisms into families and superfamilies.

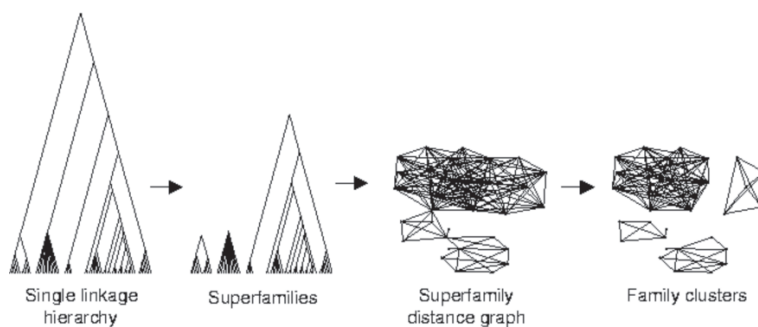


Figure 1: Overview of the graph-based SYSTEMS clustering procedures

Due to the huge amount of data (in 2003 about 1,000,000 non-redundant sequences) the computational requirements for processing are constantly growing. Therefore, only few such initiatives worldwide exist.

For optimal utility the clustering was postprocessed by multiply aligning the families, computing trees for them, annotating domain information, and extracting consensus sequences descriptive for groups of sequences. Based on either the multiple alignments or the consensus sequences a user can search the database, thus using it, e.g., for annotation of new sequences. The database and the associated services are available at: <http://systems.molgen.mpg.de/>.

Taxonomical analysis

Every protein sequence in the sequence set underlying the SYSTERS database originates from one species. On the other hand, most protein family clusters contain sequences from several different species. Thus, querying the protein family database, one is often interested in:

- the taxonomical complexity of a protein family,
- all protein families a specific taxon belongs to,
- protein families specific for one taxon, or
- protein families shared by several different taxa.

With taxon we do not only mean a species, but any arbitrary taxonomical level as given by the NCBI taxonomy. To facilitate phylogenetic studies, the SYSTERS web server now provides an interface to select protein family clusters satisfying a user defined set of taxa.

Evolutionary analysis

In collaboration with the Dept. Lehrach of the MPI for Molecular Genetics, we have derived the COPSE database for evolutionary analyses. COPSE (short for Clusters of Orthologous and Paralogous SEquences) is a clustering of invertebrate and vertebrate sequence data as a prerequisite for the analysis of vertebrate evolution and functional annotation. Major duplication events are assumed to have happened during vertebrate evolution. To prove this hypothesis one depends on well separated vertebrate gene families having only one orthologous representative in the invertebrates.

Algorithms developed for this project were applied to other data sets, e.g., for the comparison of the complete sequence sets of man and mouse. The resulting groups of orthologous sequences were used in the CORG (COmparative Regulatory Genomics) project (Ch. Dieterich / M.Vingron, MPI) for the detection of conserved non-coding blocks in the upstream regions of orthologous human and mouse genes.

The GenomeMatrix (A.Hewelt, RZPD / H.Lehrach, MPI) is a platform integrating data from several completely sequenced organisms thus simplifying multi-gene cross-species analyses. Orthology relationships are used to combine information originating from different organisms. The calculation of orthology relationships in the GenomeMatrix was done in our group.

Another effort in the group focuses on the estimation of the speed of evolutionary changes within specific protein families. Results will be presented in the near future.

Sequence analysis platform

SYSTERS is together with GeneNest (S.Haas, MPI) and SpliceNest (S.Haas, MPI / E.Coward, formerly MPI) integrated now into one framework. This allows the user an over-all exploration of the whole sequence space covering protein, mRNA and EST sequences, as well as genomic DNA. The databases are available for querying and browsing at: <http://cmb.molgen.mpg.de>.

Future work

The SYSTERS data set will be regularly updated, thus providing an up-to-date resource for the scientific community.

The work of the group will further continue in the direction of evolutionary analyses. It will be extended in the direction of the analysis of protein domain composition and domain arrangement and will serve as a basis for the detection of new domains.



General information

Publications 2000-2003

Panopoulou G, Hennig S, Groth D, Krause A, Herwig R, Vingron M & Lehrach H (2003). *New evidence for genome wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes*. *Genome Research* 13(6a):1056-1066

Dieterich C, Wang H, Rateitschak K, Luz H & Vingron M (2003). *CORG: a database for Comparative Regulatory Genomics*. *Nucleic Acids Research* 31 (1): 55-57

Dieterich C, Cusack B, Wang H, Rateitschak K, Krause A & Vingron M (2002). *Annotating regulatory DNA based on man-mouse genomic comparison*. *Bioinformatics* 18 (Suppl 2): S84-S90

Krause A, Haas SA, Coward E & Vingron M (2002). *SYSTERS, GeneNest, SpliceNest: Exploring sequence space from genome to protein*. *Nucleic Acids Research* 30(1): 299-300

Haas S, Beissbarth T, Rivals E, Krause A & Vingron M (2000). *GeneNest: automated generation and visualization of gene indices*. *Trends in Genetics* 16(11): 521-523

Krause A, Stoye J & Vingron M (2000). *The SYSTERS protein sequence cluster set*. *Nucleic Acids Research* 28(1): 270-272

Talks

Krause A, Stoye J, Vingron M. *Large scale hierarchical clustering of protein sequences*. 26th Annual Conference of the Gesellschaft für Klassifikation, Mannheim, 22.– 24.7.2002

Krause A. *Clusterung großer Proteinsequenzdatenmengen*. Universität Bielefeld, 19.6.2002 (Disputation)

Krause A, Panopoulou G, Hennig S, Vingron M. *Determination of vertebrate gene families*. 8th Congress of The European Society for Evolutionary Biology, Aarhus, Denmark, 20.– 25.8.2001

Krause A. *Determination of protein families with special interest in vertebrate evolution*. Universität Bielefeld, 21.5.2001

Krause A, Panopoulou G, Hennig S, Vingron M. *Determination of vertebrate gene families*. DIMACS Workshop on Whole Genome Comparison, Rutgers University, Piscataway, NJ, USA, 28.2.– 2.3.2001

Krause A, Stoye J, Vingron M. *Clustering in processing of nucleotide and protein sequence databases*. 7th Conference of the International Federation of Classification Societies, Namur, Belgium, 11.– 14.7.2000 (invited talk)

Patents

Verfahren zur Eingruppierung von Sequenzen in Familien.

Patent-Nr.: 197 45 665 C1

Patentinhaber: Deutsches Krebsforschungszentrum Heidelberg

Erfinder: M.Vingron, A.Krause

Teaching

Bioinformatics, Studiengang Biosystemtechnik / Bioinformatik, WS 2003, 2x 4 SWS, Technische Fachhochschule Wildau

Assistent Bioinformatik, lectures during an advanced training (Assistent Bioinformatik, 8 days full-time, 2002, 2003, Berufsbildungszentrum C&Q

Interns

Ralf Mehle (10/03 – 02/04)

Andrea Y. Weiße (12/02 – 02/03)

Thomas Meinel (10/01 – 12/01)

Work as scientific referee

Referee for *Bioinformatics*

Sub-reviewer for

- RECOMB
- WABI
- ECCB
- GCB

Co-operations

Integration of SYSTERS, GeneNest and SpliceNest into a sequence analysis platform, with SA Haas, MPI for Molecular Genetics, Berlin

Comparative Regulatory Genomics, with Ch.Dieterich / M.Vingron, MPI for Molecular Genetics, Berlin

GenomeMatrix, with A. Hewelt, RZPD Berlin and H.Lehrach, MPI for Molecular Genetics, Berlin

A Platform for reconstructing vertebrate phylogeny, with G. Panopoulou / H.Lehrach, MPI for Molecular Genetics, Berlin

The SYSTERS protein family database is part of the "Helmholtz Netzwerk für Bioinformatik" (HNB)

Algorithms Group



Head:

Dr. Alexander Schliep (since 5/02)

Phone: +49 (0)30-8413 1166

Fax: +49 (0)30-8413 1152

Email: alexander.schliep@molgen.mpg.de

Graduate students:

Martin Oksrlar (until 7/03)

Harindar Singh Keer (until 8/03)

Ivan Gesteira Costa Filho (since 10/03)

Wasinee Rungsaritoyotin (since 10/02)

Undergraduate students:

Benjamin Georgi (since 9/03)

Jonas Heise (since 7/03)

Research themes

Our research focuses on novel machine learning methods and algorithms which we apply to a range of biological problem settings. An emphasis is put on analyzing high-dimensional, heterogeneous and time-series data.

Algorithmics and machine learning

One key area, to which we apply machine learning techniques, is the annotation of protein sequences. We developed a cluster-based approach for the detection of remote homologs which exceeds the sensitivity of PSI-Blast, the most widely used tool for finding homolog sequences, by 40%. Currently we are employing Support-Vector-Machines for deciding significance of sequence similarity (thus circumventing the problems with the typically used statistics), information-theoretic methods for detecting key residues *ab initio*, and a decision-tree variant for classifying Kinases. The main focus in the future will be an integration of the learning process of the underlying statistical models and the classifier, to improve overall performance.

Hidden-Markov-Models (HMMs)

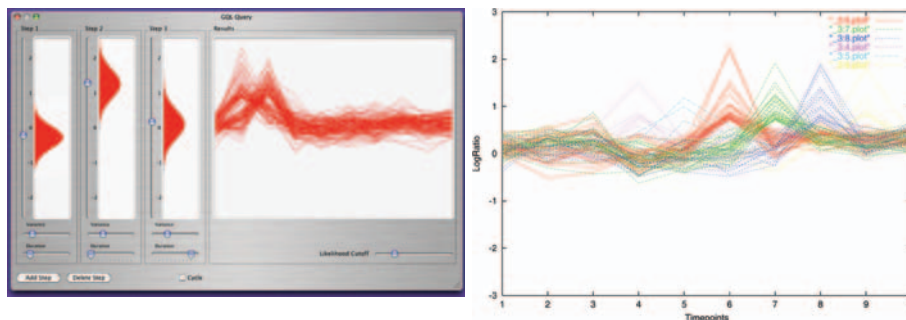
Hidden-Markov-Models, originally developed for speaker-independent speech recognition, have been widely used in their basic form as so-called Profile HMMs for the detection of remote homologs, or in the slightly more complex form of labeled HMMs, for finding eukaryotic genes. The basic framework supports a number of extensions; they can also be used for either classification or clustering.

On one hand our work with HMMs concerned itself with learning HMM topology and different training methods. On the other hand, we investigated novel applications using HMMs as qualitative time-series models and, among others, non-Markovian HMM extensions applied to the detection of circular permutations. Also, we are the first to propose the framework of partially supervised learning for both clustering and mixture modeling. This has been employed with great success for gene expression time-series data. Furthermore, we develop the only free (licensed under the LGPL) library for HMMs, the General Hidden Markov Model Library (GHMM), which is widely used both in industry and academia. To the best of our knowledge, we also introduced the first XML-format for HMMs as well as a graphical tool to edit HMMs with discrete or continuous emissions. We developed a number of novel extensions to the HMM formalism - non-homogeneous Markov chains, clustering and mixture modeling - and implemented them in the library.



We make use of the well-known HMMer-package, which implements profile HMMs and loosely collaborate with the well-established groups developing HMMs, Anders Krogh, Soren Brunak, Kevin Karplus and David Haussler, on file formats and the graphical editor. Our work complements the existing body of work uniquely.

Future work planned includes development of new learning algorithms geared towards discriminative classification, mixed-domain multi-variate emissions, and a hierarchical HMM framework supporting for example protein-domain combinations or custom gene finders.



Besides their use in finding remote homologue proteins, Hidden Markov Models can for example be used in time-course analysis. One important application area is analyzing gene expression over time. A graphical user interface (left) allows to specify a HMM encoding a particular qualitative behaviour of time-courses. This can be used to query large data sets interactively or for clustering. Other HMM algorithms such as computation of the Viterbi-path allow to decompose similar time-courses according to their phase (right).

Group Testing

A prototypical problem in molecular biology is the screening of a large collection of samples with some specific test. If a positive test outcome is a rare event, analyzing several samples simultaneously — this is also called multiplexing or pooling — can provide substantial savings of experimental effort, for example in screening clonal DNA-libraries. The mathematical formulation of this problem is known as statistical group testing, and bridges across a number of mathematical fields such as combinatorial design theory, Bayesian statistics and Markov Chain Monte Carlo methods for the design and analysis of experimental protocols. The same problem arises for — superficially unrelated — problems such as picking oligos which detect and differentiate between the presence of closely related species, e.g. virus subtypes, in a sample, or in picking Single Nucleotide Polymorphisms (SNP) markers to infer haplotypes.

Based on prior work at Los Alamos National Laboratory we have implemented a method to select oligos for DNA chips in situations where due to a high degree of sequence similarity unique oligos cannot be found. This will be applied to the analysis of meiobenthos samples, as well as to HIV subtyping, where the very high incidence of multi-viral HIV infections such as in populations in Southern Africa make analysis difficult. Further work on the theoretical side includes optimization of the underlying combinatorial designs and modeling more of the underlying biology — e.g. phylogenetic information in the analysis step.

Research to use this approach for selecting most informative SNPs for haplotype detection, simultaneously for several individuals, is underway. We closely collaborate with David Torney, who first proposed and implemented group testing to reduce the experimental work in the context of the first physical map of Human chromosome 16.

Visualization

Both teaching and research in algorithms are accelerated by computer tools which allow to experience the dynamic nature in a rich multi-medial environment. Gato, the Graph animation toolbox, provides such an environment. Due to its flexibility and (semi-) automated visualization of user-implemented graph algorithms, it surpasses the capabilities of existing products. A Springer textbook, covering an introduction to combinatorial optimization, is forthcoming. As an extension, visualization of bioinformatics algorithms is under research as well as a graphical tool for working with Hidden-Markov-Models.

General information

Publicatons 2000 - 2003

Schliep A, Torney DC & **Rahmann S** (2003). *Group Testing With DNA Chips: Generating Designs and Decoding Experiments*. Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)

Schliep A, Schönhuth A & **Steinhoff C** (2003). *Using Hidden Markov Models to Analyze Gene Expression Time Course Data*. Proceedings of the ISMB 2003. Bioinformatics 19(Suppl 1): I255-I263

Knab B, **Schliep A**, Steckemetz B & Wichern B (2003). *Model-based clustering with Hidden Markov Models and its application to financial times series data*. Proceedings of the GfKI 2002 conference. In M. Schader, W. Gaul, M. Vichi (eds). Between Data Science and Applied Data Analysis. Springer, 2003

Pipenbacher P, **Schliep A**, Schneckener S, Schönhuth A, Schomburg D & Schrader R (2002). *ProClust: Improved clustering of Protein Sequences with an extended graph-based approach*. Proceedings of the ECCB 2002. Bioinformatics 18 (Suppl 2): S182-S191

Kaderali L & **Schliep A** (2002). *An algorithm to select target specific probes for DNA chips*. Bioinformatics 18(10):1340-9

Schliep A & Hochstättler W (2002). *Developing Gato and CATBox with Python: Teaching graph algorithms through visualization and experimentation*. Multimedia Tools for Communicating Mathematics. Springer Verlag, 2002, 291-310

Talks

Gato & CATBox: Teaching Graph Algorithms through visualization and experimentation. Workshop on Visualization and Mathematics, Berlin, 23.5.2002

Model-based Clustering with Hidden Markov Models and its Application to Financial Time-series Data. Jahrestagung der deutschen Gesellschaft für Klassifikation, Mannheim, 24.7.2002

Selecting target-specific probe sets for DNA chips. University of California at Irvine, 26.7.2002 (invited talk)

GHMM & HMMed: A comprehensive HMM toolkit. Bioinformatics Open Source Conference (BOSC), Edmonton, Canada, 2.8.2002

Experimenting on Algorithms: Teaching Bioinformatics methods visually. Workshop on Education in Bioinformatics (WEB 02), Edmonton, Canada, 8.8.2002

Group testing and DNA chips. Center for Non-linear Studies, Los Alamos National Laboratory, 21.8.2002 (invited talk)

Proclust: Improved Clustering of Protein Sequences with an Extended Graph-Based Approach. European Conference on Computational Biology (ECCB), Saarbrücken, 9.10.02

Proclust: Graph-Based Clustering of Protein Sequences. Berlin Center for Genome Based Bioinformatics, Berlin, 13.11.2002 (invited talk)

Dealing with Non-Unique Probes: DNA Chips and Group Testing. Dagstuhl Seminar "Computational Biology", Schloß Dagstuhl, 20.11.2002 (invited talk)

Using Hidden Markov Models to Analyze Gene Expression Time Course Data. Intelligent Systems in Molecular Biology (ISMB), Brisbane, Australien, 30.6.2003

A Model-based framework for time-course analysis. Center for Non-linear Studies, Los Alamos National Laboratory, 25.9.2003 (invited talk)

Analyzing gene expression over time using a mixture approach. University of California at San Francisco, 3.10.2003 (invited talk)

Mixtures of Hidden-Markov-Models. University of California at Berkeley, 8.10.2003 (invited talk)

Teaching

Vorlesung *Algorithmische Bioinformatik*, WS02/03, 4 SWS, Freie Universität Berlin

Seminar *Markov Ketten*, WS02/03, 2SWS, Freie Universität Berlin

Vorlesung *Statistische Mustererkennung in der Bioinformatik*, SS03, 2SWS, Freie Universität Berlin

Softwarepraktikum zur Vorlesung *Statistische Mustererkennung in der Bioinformatik*, SS03, 2SWS, Freie Universität Berlin

Vorlesung *Algorithmische Bioinformatik*, WS03/04, 4SWS, Freie Universität Berlin

Seminar *Clusteranalyse Heterogener Daten*, WS 03/04, 2SWS, Freie Universität Berlin

Seminar *Classification: Contrasting Statistics with Machine learning*, WS 03/04, 2SWS, Freie Universität Berlin

Kompaktkurs *Angewandtes Data Mining*, WS 03/04, 4SWS, Freie Universität Berlin

Lecture *Statistical Pattern Classification in Bioinformatics*. Ringvorlesung Bioinformatik, Berlin Center of Genome Based Bioinformatics and FU Berlin, 29.1.2003



Theses

Benjamin Georgi: *A graph-based approach to Clustering of Profile HMMs*, Bachelor Thesis, Bioinformatik, Freie Universität Berlin

Olaf Wendisch: *Klassifizierung entfernt homologer Proteinsequenzen mit Support Vector Maschinen*, Diploma Thesis, Mathematisches Institut, Universität zu Köln

Andrea Weiße: *Detection of circular permutations in proteins*, Bachelor Thesis, Bioinformatik, Freie Universität Berlin

Jonas Heise: *Using phylogenetic information in the design of DNA chips*, Bachelor Thesis, Bioinformatik, Freie Universität Berlin

Interns

Holger Meyer, bioinformatics student, Freie Universität Berlin (3 month internship)

Melanie Kaspar, bioinformatics student, Universität des Saarlandes (3 month internship)

Work as scientific referee

Referee for

- Bioinformatics
- BMC Bioinformatics
- Proteins
- Functional and Integrative Genomics
- Jahrestagung der Gesellschaft für Klassifikation
- Springer Verlag

Sub-reviewer for

- RECOMB
- WABI

Academical co-operations

Analysis of bacterial MLST data, with Mark Achtman, Max-Planck-Institut für Infektionsbiologie, Berlin

Machine learning approaches for detection of remote homolog proteins, with Lars Arvestad, Stockholm Bioinformatics Center, KTH, Stockholm

HIV virus subtyping with DNA chips, with Winston Hide, South African National Bioinformatics Institute, University of the Western Cape

Visualisation of graph algorithms - multimedia for computer science education, with Winfried Hochstättler, Institut für Mathematik, BTU Cottbus

Combinatorial optimization methods for group testing designs, with Knut Reinert, Fachbereich Mathematik und Informatik, Freie Universität Berlin

Genotyping ADHD and Autism, inference of complex phenotypes, with Anne Spence, College of Medicine, Human Genetics, University of California at Irvine

Analysis of gene expression time-series data, with Alexander Schönhuth, Zentrum für Angewandte Informatik Köln (ZAIK), Universität zu Köln

Clustering protein sequences, with Dietmar Schomburg, Institut für Biochemie, Universität zu Köln

Identifying members of microbacterial communities with DNA chips, with Diethard Tautz, Institut für Genetik, Universität zu Köln

Group testing DNA chips, Genotyping ADHD and Autism, Statistical methods for analysis of sequence data with long-range correlations, with David C. Torney, Los Alamos National Laboratory

Industrial co-operations

Clustering protein sequences, with Sebastian Schneckener, Science Factory, Cologne

Clustering heterogenous data, Olav Zimmermann, Science Factory, Cologne

Organization of scientific events

Member of the local organizing committee of of *The Seventh Annual International Conference on Research in Computational Molecular Biology - RECOMB 2003*, Berlin, 10.-13.4.2003

Public relations

A primer in Bioinformatics: theory and practice, one-day workshop for high-school students (25.6.2002, Heinrich-Hertz Gymnasium)

A primer in Bioinformatics: theory and practice, one-day workshop for high-school students (17.12.2002, Walter-Rathenau Oberschule)

Co-organization of a public discussion *Hype oder Hoffnung - Podiumsdiskussion zur Rolle der Bioinformatik am Standort Berlin*, Berlin, 9.9.2002

Protein Function Analysis Group



Head:

Dr. Jörg Schultz (until 8/03)

Email: joerg.schultz@molgen.mpg.de

Graduate student:

Birgit Pils (since 2/02)

Research

The focus of the group is the understanding of protein function and evolution using genomic, structural and proteomic data. Central to this question is the concept of the domain: a structurally conserved, genetically mobile unit. When viewed at the three-dimensional level of protein structure, a domain is a compact arrangement of secondary structures connected by linker polypeptides. It usually folds independently and possesses a relatively hydrophobic core. The importance of domains is that they cannot be divided into smaller units – they represent a fundamental building block that can be used to understand the evolution and function of proteins. In collaboration with the group of Dr. P. Bork, we are developing the SMART (Simple Modular Architecture Research Tool) domain database, which, to date, allows the identification of more than 600 divergent domain homologues in user supplied sequences and provides rich manual and automatic annotation for each domain. Furthermore, we are active in the hunting of novel domains to further complete the description of evolutions domain repertoire (Schultz, submitted). Having access to the whole set of building blocks used by protein evolution, we now can start to analyse domains in their protein context. In a recent project, we have used co-occurrence of domains to predict their cellular localisation and, following, the localisation of whole proteins (Mott et al., 2002).

Ongoing experimental characterisation of domain families revealed, that the ‘one domain – one function’ concept does not hold true. On the contrary, function can diverge heavily within a single, homologous domain family. Prediction of a protein’s domain architecture will be sufficient to roughly characterise it; it will not give insights into molecular details of its function. To overcome this strong hindrance in function prediction, we are working on more advanced methods with the goal to make the step from domain to function prediction. One direction is the identification of functional sites within domains to use these for a more detailed function prediction. We have applied this approach to predict functional regions and catalytic sites of N-acetyl-b-D-glucosaminidase (O-GlcNAcase) (Schultz and Pils, 2002). This protein, which is linked to different diseases as diabetes and cancer, is involved in the reversible, intracellular modification of proteins by O-linked N-acetylglucosamine. Our hypotheses are currently tested experimentally in collaboration with Prof. Dr. Schmidt, Universität Bonn.

The anecdotal description of tyrosine phosphatases with mutations in catalytic sites raised our interest in the evolution of these so-called ‘anti-phosphatases’. A genome



wide analysis of tyrosine as well as dual specific phosphatases revealed, that these mutations are more frequent than expected. Using phylogenetics, we could show, that these mutations occurred multiple times independently and are conserved within evolution. Site-specific analysis of the evolutionary rates allowed a functional subclassification of this large protein family and gave insights into the evolution of these subtypes (Pils and Schultz, submitted).

As a member of the protein analysis group of the mouse genome project, we compared proposed functional sites of human disease proteins with the corresponding mouse sequences. This analysis revealed a small but significant number of cases where the mouse wildtype equals the human disease mutation, either caused by differences in the function of the corresponding proteins or filtered out by corresponding mutations (Mouse Genome Sequencing Consortium, 2002).

In summary, the development and application of more advanced methods for function prediction will further increase the amount of information we can get by sequence and structure. This new level of information raises new problems. Currently, the function of a protein is stored mainly as free text in databases. This blocks the integration of data from genomic projects with data from e.g. proteomic and gene expression projects. Therefore, we are developing methods to represent the function of a protein in a 'computer-understandable' way. Currently, we are focussing on signalling and protein interaction networks, as their features are largely determined by the activity and not by the expression level of the involved proteins (Ratsch et al, 2003). In a pilot project, we combined domain based function prediction with protein interaction data to reconstruct and annotate bacterial signalling networks (Schultz, submitted). This project underlined the value of this method, as it even allowed to predict the influence of mutations to pathways and the whole network.

General information

Publications 2002-2003

Ratsch E, **Schultz J**, Saric J, Cimiano Lavin P, Wittig U, Reyle U & Rojas I (2003). *Developing a protein interactions ontology*. *Comp Func Genomics* 4:85-89

Mouse Genome Sequencing Consortium (2002). *Initial sequencing and comparative analysis of the mouse genome*. *Nature* 420: 520-562

Schultz J & Pils B (2002). *Prediction of structure and functional residues of O-GlcNAcase, a divergent homologue of Acetyltransferases*. *FEBS Letters* 529:179-182

Mott R, **Schultz J**, Bork P & Ponting CP (2002). *Predicting Protein Cellular Localisation Using a Domain Projection Method*. *Genome Res* 12:1168-1174

Interns

Jonas Heise (2 month internship)

Co-operations

Functional characterisation of GlcNAcase, with Prof. Dr. B. Schmitz, Universität Bonn

Protein Analysis Group of the Mouse genome project, with Prof. C.P. Ponting, MRC Functional genetics unit

Development of the SMART domain database, with Dr. P. Bork, EMBL Heidelberg

Developing an ontology for protein interaction, with Dr. I. Rojas, EML Heidelberg

Computational Diagnostics Group



Head:

Dr. Rainer Spang

Phone: +49 (0)30-8413 1175

Fax: +49 (0)30-8413 1152

Email: rainer.spang@molgen.mpg.de

Scientist:

Dr. Claudio Lottaz

Graduate students:

Jochen Jäger

Dennis Kostka

Florian Markowetz

Stefanie Scheid

Undergraduate student:

Jörn Tödling

The focus of this group is on developing statistical methodology for the use of gene expression profiles in medical diagnostics. We aim to identify pattern in expression profiles that improve or facilitate diagnosis, help to predict clinical outcome or refine common diagnostic schemes. Our work includes both theoretical projects in which we aim to develop novel analysis methodology and applied data analysis projects with clinical cooperation partners.

Methods development

In a project on molecular symptoms we combine gene expression data with functional annotations of the genes on the microarray. Statistical models for microarray data produce lists of genes that are up/down regulated, that constitute clusters of genes with correlated expression or that form predictive signatures for microarray based diagnosis of diseases. It is common practice to use the functional annotations of the identified genes for further biological interpretation. In this posterior use of annotations the gene functions have no influence on the statistical model at all. The expression levels exclusively drive the models. In this project we explore the a priori use of functional annotations for model building and structuring. Our aim is the identification of molecular subtypes of a disease. In order to exploit functional annotations, we structure the variable space (genes) using a functional grid, provided by the biological processes branch of the gene ontology graph.

Another project aims at detecting the loss of co-regulation mechanisms. The task is to identify sets of genes that display strongly correlated expression in a control group of patients but loose this pattern of co-regulation in the disease group. We have developed a score for loss of co-regulation that we can apply to any subset of genes in a microarray study. We (heuristically) optimize this score over all possible such subsets and have developed a permutation-based test to check for the significance of such a result in this extreme multiple testing setting.

In view of future developments in the context of genome wide RNAi screens, we started investigating how expression profiles from RNAi assays can improve network reconstruction using Bayesian networks. We have recently started to collaborate with Dr. Michael Boutros from the German Cancer Research Center for investigating possibilities of applying our theoretical results on signaling network reconstruction to real large scale RNAi data.



Further projects include multiple testing problems when screening for differentially expressed genes, the analysis of learning curves to decide on an optimal time point for switching from genome wide expression screening to expression analysis with a smaller and cheaper customized chip with which diagnostic signatures can be fine tuned, significance testing of groups of genes for weak but consistent up and down regulation, and cross platform analysis of microarray data.

Co-operative projects

In co-operation with biologists and clinicians from the Charite medical school in Berlin we have started analyzing expression profiles from childhood ALL relapse patients. The goal of this project is the identification of molecular risk factors characteristic for all or part of the patients with a poor treatment response. In co-operation with pathologists from the UBK Berlin, the medical school of Free University, we started analyzing a data set of expression profiles from Hodgekin lymphomas and B-cell cell lines, with the goal of defining subtypes of Hodgekin lymphomas corresponding to developmental stages of B-cells. In co-operation with the group of Patricia Ruiz from the Department of Hans Lehrach and a group of cardiologists from the university of Heidelberg we work on the design of a gene expression chip for cardiac diseases with a focus on cardiomyopathies. Further collaborations include the analysis of breast cancer profiles, gene expression in neural development, and heart failure in mice.

Standing and future plans

Microarray data analysis both in its theoretical and applied form is a highly competitive field worldwide. Our strength is that we have brought together people with different backgrounds into one group. These people work together in the same office, discuss their different points of view and hence develop adequate data analysis strategies, that are backed up by a solid understanding of both their biological and theoretical foundations.

Our theoretical projects have been, on average, running for a little more than one year. They all have produced first promising results, but none of them is finished. Several publications are in preparation. In terms of applied work we plan to focus on cancers of the immune system. In this field we plan to extend the scope of our collaborators, from ALL and Hodgekin lymphoma to leukemia and lymphoma in general. We jointed in the grant application on nationwide network for the analysis of gene expression analysis in malignant lymphoma (Funding: Deutsche Krebshilfe) and found the support of an also nation wide leukemia research network (Funding: NGFN).

General information

Publications 2000-2003

Scheid S & Spang R (2003). *A False Discovery Rate Approach to Separate the Score Distributions of Induced and Non-induced Genes*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (accepted)

Markowitz F & Spang R (2003). *Evaluating the Effect of Perturbations in Reconstructing Network Topologies*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (accepted)

Grzeskowiak R, Witt H, Drungowski M, Thermann R, Hennig S, Perrot A, Osterziel KJ, Klingbiel D, **Scheid S, Spang R, Lehrach H & Ruiz P** (2003). *Expression profiling of human idiopathic dilated cardiomyopathy*. Cardiovascular Research (to appear)

Lottaz C, Iseli C, Jongeneel CV & Bucher P (2003). *Modeling Sequencing Errors by Combining Hidden Markov Models*, to be published in The 2nd European Conference on Computational Biology ECCB'03, Paris, France

Jäger J, Sengupta R & Ruzzo WL. *Improved Gene Selection for Classification of Microarrays*. Biocomputing - Proceedings of the 2003 Pacific Symposium, 53-64

Spang R, Blanchette C, Zuzan H, Marks JR, Nevins J & West M (2002). *Prediction and uncertainty in the analysis of gene expression profiles*. In *Silico Biol 2*

Markowitz F & von Heydebreck A (2002). *Class discovery in gene expression data: characterizing splits by support vector machines*. Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation 2002: 662-669

Spang R, Rehmsmeier M & Stoye J (2002). *A Novel Approach to Remote Homology Detection: Jumping Alignments*. *J Comput Biol* 9(5): 747-760

Müller T, Spang R & Vingron M (2002). *Estimating Amino Acid Substitution Models: A Comparison of Dayhoff's Estimator, the Resovent Approach and a Maximum Likelihood Method*. *Mol Biol Evol* 19(1): 8-13

West M, Blanchette C, Dressman H, Huang E, Ishida S, **Spang R**, Zuzan H, Olson JA Jr, Marks JR & Nevins JR (2001). *Predicting the clinical status of human breast cancer by using gene expression profile*. *PNAS USA* 98(20): 11462-7

Ishida S, Huang E, Zuzan H, **Spang R**, Leone G, West M & Nevins JR (2001). *Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis*. *Mol Cell Biol* 21(14):4684-99

Spang R & Vingron M (2001). *Limits of homology detection by pairwise sequence comparison*. *Bioinformatics* 17(4): 338-342

Book contributions & reviews

Spang R (2003). *Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine* (Review). *Biosilico* 1(2): 64-68

Spang R, Béziat P & Vingron M (eds.). *Currents in Computational Molecular Biology 2003*. RECOMB 2003, Berlin

Teaching

Practical Microarray Data Analysis Courses
Rainer Spang: Vorlesung *Genomische Datenanalyse*, 4 SWS, SS 2003, FU Berlin

Stefanie Scheid, Dennis Kostka, Florian Markowitz: Übungen *Genomische Datenanalyse*, 2 SWS, ss 2003, FU Berlin

Theses

Jörn Tödling: *Cross-Platform Assessment of Microarray Experiments on Gene Expression Profiles*. Bachelor Thesis in Bioinformatics, FU Berlin

Stefan Bentink: *Gene ontology as a tool for the systematic analysis of large-scale gene-expression data*. Master Thesis in Bioinformatics, TFH Berlin

Internships

Joern Toedling, Bachelor student, Freie Universität Berlin, 8 weeks internship

Martin Held, Bachelor student, Freie Universität Berlin, 8 weeks internship

Julie Floch, Student, post graduate degree in bioinformatics, Université Evry-Val d'Essonne, France, 6 month internship

Guest scientists

Dr. Nicola Armstrong, ESF supported visitor from EURANDOM, Eindhoven, Netherland, will be visiting 2003-2004 (6 month)

Xinan Yang, DAAD supported visitor from the Southeast University, China, will be visiting 2003-2005 (2 years)



Co-operations

Identification and functional characterization of molecular risk factors in acute leukemias, with Prof. Dr. Christian Hagemeier, Prof. Dr. Wolf-Dieter Ludwig, Prof. Dr. Karl Seeger, Prof. Dr. Leonid Karawajew, Dr. Renate Kirschner, Charité, Humboldt-Universität Berlin

Gene expression analysis in Hodgekin lymphoma, with Prof. Dr. Harald Stein, Dr. Michael Hummel, Universitätsklinikum Benjamin Franklin, Freie Universität Berlin

Deriving Signaling Networks by Integrating Genome-wide RNAi, Expression Profiling and Computational Analysis, with Dr. Michael Boutros, Deutsches Krebsforschungszentrum, Heidelberg

Design of a diagnostic cardio chip, with Dr. Patricia Ruiz, Max-Planck-Institut für molekulare Genetik, and Dr. Boris Ivandic, Dr. Dieter Weichenhan, Universitätsklinikum Heidelberg

Courses in practical microarray data analysis, with Dr. Wolfgang Huber, Deutsches Krebsforschungszentrum, Heidelberg, Dr. Ulrich Mansmann, Universitätsklinikum Heidelberg, Dr. Jörg Rahnenführer, Max-Planck-Institut für Informatik, Saarbrücken

Predictive Bayesian modeling using microarray data with applications to breast cancer, with Prof. Dr. Mike West, Prof. Dr. Joe Nevins, Duke University and Duke medical center, USA

Jumping Alignments, with Prof. Dr. Jens Stoye, Universität Bielefeld

Organization of scientific events

Member of the organizing committee of *RECOMB 2003* and editor of the *Currents in Computational Molecular Biology 2003*, Berlin, 10.-14.4.2003

Organizer of the “*International BCB-workshop on statistics and cancer genomics 2003*”, Berlin, 21.8.2003

Transcriptional Regulation Group



Head:

Prof. Dr. Martin Vingron

Phone: +49 (0)30-8413 1150

Fax: +49 (0)30-8413 1152

Email: vingron@molgen.mpg.de

Scientists:

Dr. Thomas Manke

Dr. Stefan Röpcke

Dr. Christine Steinhoff (joint with Dept. Ropers)

Dr. Steffen Grossmann

Dr. Lloyd Demetrius

Graduate students:

Christoph Dieterich

Holger Klein

Haiyan Wang (until 07/2003)

Affiliated researcher:

Anja von Heydebreck

The availability of complete genomes for several organisms has opened up new possibilities of studying gene regulatory mechanisms and in particular cis-regulatory elements. The gene regulation group focuses on the delineation of regulatory motifs and interactions based on an integration of a variety of information sources. In yeast, where extensive protein-protein interaction data have been generated, this information can serve to aid in the identification of regulatory modules. In mammalia, comparison of non-coding, upstream sequences of orthologous genes can pinpoint regions that are likely to have a regulatory role. This can be extended by comparing sequences to binding site descriptors that have been collected in publicly available databases. Microarray generated gene expression data may further serve to understand regulatory interactions between genes.

Gene regulation in yeast

In yeast, only a small number of transcription factor binding sites have been described. The focus of our work lies on the combinatorial interplay of transcription factors as they interact with regulatory DNA regions. We have, for the first time, fully integrated the protein-DNA binding data into the larger network of protein-protein interactions. This allowed for the identification of modules of transcription factors which co-regulate sets of functionally related genes. From this information we construct regulatory motifs, and computationally search for further targets in a genome-wide scan. Our results also demonstrate that, despite the inherent noise in large-scale data sets, there are significant commonalities which can be exploited to increase the reliability of network predictions (Manke et al., 2003).

Predicting regulatory elements in the human genome

Comparative sequence analysis of two or more genomes is an appropriate tool to investigate gene structure and surrounding functional elements in the vast sequence space of non-coding DNA. This assumption is validated by the observation that experimentally verified transcription factor binding sites map to highly conserved regions in man-mouse sequence comparisons. An initial large-scale *in silico* study on sequence conservation



upstream of the translational start site demonstrated the power of the comparative approach (Dieterich et al., 2002). Our principal repository for annotated conserved blocks (CNBs) in homologous upstream regions of man and mouse is CORG, the database for Comparative Regulatory Genomics (Dieterich et al., 2003a). CORG contains a precomputed set of CNBs for the upstream regions of more than 12,000 orthologous gene groups. The origin of sequence conservation is often explained by the functional annotation of the CNBs. We distinguish untranslated exons from other conserved regions by screening all CNBs with pre-assembled EST clusters. Here, an important part of our research concerns the reliable annotation of transcription factor binding sites within CNBs.

Microarray data and the identification of target genes

An imminent subsequent step is to associate evolutionarily conserved predicted binding sites with complementary biological data like time-course microarray data. In an on-going collaboration, suggested downstream genes of the transcription factor SRF have been scanned for evolutionarily conserved SREs (the SRF binding site). These hypothetical direct target genes are currently under investigation in the laboratory of A. Nordheim (Tübingen). A further study was performed on another much-studied biological process: the response of dendritic cells to LPS, a component of the cell wall of gram-negative bacteria (Dieterich et al., 2003b). An analysis of the upstream regions of genes that appear to be co-regulated in the respective microarray experiment allows to identify the endpoints of Toll-receptor signalling which is involved in this pathway. Likewise, regulation of the cell cycle in human (HeLa) cells has been studied. Some transcription factor binding sites, like those of the E2F family, show a strong enrichment in the upstream regions of genes that fall into particular cell cycle phases. We have now initiated a co-operation with the research group of Constance Scharff at the MPIMG to assess the impact of selected transcription factors on cell cycle progression using RNA-interference technology.

Evolution of binding sites

Binding sites evolve and certainly play a role in phenotypic diversity and species diversity. Although an understanding of the evolution of regulatory elements may be far away, this is a key question in understanding cellular processes. Multi-species comparisons are key to elucidate patterns of appearing and disappearing functional elements over time. Multiple alignments facilitate to trace the history of individual binding sites. Good examples to study are developmental processes, which are remarkably conserved in vertebrates. We have teamed up with the research group of Stefan Mundlos at the MPIMG to detect potential binding sites of RUNX2, which are conserved in many vertebrate genomes. The promoter regions of interest are being sequenced in the laboratory.

Microarrays: Data normalization and statistics

A line of work that goes back to the time of the group at DKFZ, Heidelberg, is concerned with gene expression profiles in renal carcinoma and within the department is mostly pursued by Anja von Heydebreck. Starting with the comparison of renal carcinoma to healthy kidney tissue, her research has resulted in a number of data analysis papers (e.g., Boer et al., 2001) as well as some significant methodological advances (Huber et al, 2003a, 2003b). The method of normalization that was developed in this context is called “normalization by variance stabilization” and has found wide acceptance. The major achievement of this new method is that it simultaneously solves the treatment of the technical features of an array with the basic problem that expression level changes in low intensity genes appear much more dramatic than in highly expressed genes. The variance stabilization step maps all changes to a common interval and thereby allows for comparison of changes across intensities. In that, it is superior to the commonly used fold-change measure.

Spawned by the research on kidney carcinoma, a new problem has caught our attention: tracing the development of chromosomal aberrations in tumors. Novel mathematical methods have now been developed to tackle this question and have successfully been applied to cytogenetic data on renal carcinoma (von Heydebreck et al, 2003; Gunawan et al., 2003).

General information

Selected publications 2001-2003

Dieterich C, Wang H, Rateitschak K, Luz H & Vingron M (2003a). *CORG: a database for COMparative Regulatory Genomics*. *Nucleic Acids Res* 31(1):55-7

Dieterich C, Herwig R & Vingron M (2003b). *Exploring potential target genes of signalling pathways by predicted conserved transcription factor binding sites*. *Bioinformatics* (to appear)

Gunawan B, **Heydebreck A v**, Fritsch T, Huber W, Ringert R-H, Jakse G & Füzesi L (2003). *Cytogenetic and Morphologic Typing of 58 Papillary Renal Cell Carcinomas: Evidence for a Cytogenetic Evolution of Type 2 from Type 1 Tumors*. *Cancer Research* (to appear)

Heydebreck Av, Gunawan B, Huber W, **Vingron M** & Füzesi L (2003). *Mathematical tree models for cytogenetic development in solid tumors*. *Proceedings of the German Society for Pathology* (to appear)

Huber W, **Heydebreck A v**, Sültmann H, Poustka A & **Vingron M** (2003a). *Parameter estimation for the calibration and variance stabilization of microarray data*. *Statistical Applications in Genetics and Molecular Biology* 2(1), Art 3 (online publication)

Huber W, **Heydebreck A v** & **Vingron M** (2003b). *Analysis of microarray gene expression data*. In *Handbook of Statistical Genetics*, Bulding DJ, Bishop M, Canning C, eds., Wiley, Chichester, West Sussex, Vol 1 (2nd edition):162-187

Manke T, Bringas R & **Vingron M** (2003). *Correlating Protein-DNA and protein-protein interactions*. *J Mol Biol* 333:75-85

Marchfelder U, **Rateitschak K** & **Ehrenhofer-Murray AE** (2003). *SIR-dependet repression of non-telomeric genes in Saccharomyces cerevisiae?* *Yeast* (to appear)

Steinhoff C, **Müller T**, **Nuber UA** & **Vingron M** (2003). *Gaussian Mixture Density Estimation applied to Microarray Data*. LNCS (Lecture Notes in Computer Sciences) 2810:418-429

Dieterich C, **Wang H**, **Rateitschak K**, **Krause A** & **Vingron M** (2002). *Annotating regulatory DNA based on man-mouse genomic comparison*. *Bioinformatics* 18(Suppl 2): S84-90

Boer JM, Huber W, Sültmann H, Wilmer F, **Heydebreck A v**, **Haas S**, Korn B, Gunawan B, Vente A, Füzesi L, **Vingron M** & Poustka A (2001). *Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500 element cDNA array*. *Genome Res* 11(11): 1861-1870

Martin Vingron: Selected Invited talks
UC San Diego, USA (8/2003)

Joint Statistics Meeting, San Francisco, USA (8/2003)

Royal Statistical Society Topic Meeting *Genetics and Statistics*, Belgium (8/2003)

University of Gießen (6/2003)

Conference on the occasion of M. Water-man's 60th birthday, Los Angeles, USA (3/2003)

DMV Tagung 2002, Leipzig (2002)

LMU München (2/2002)

Universität Dortmund (12/2002)

Eurandom, TU Eindhoven (12/2002)

ETH+Univ. Zürich joint statistics colloquium

IWR, Universität Heidelberg (11/2001)

Teaching

Martin Vingron, within bioinformatics curriculum at Free University:

- Seminar *Biological Sequence Analysis*, SS 2001
- Course *Algorithms for phylogeny construction*, WS 2001/02
- Seminar *Molecular evolution*, SS 2002
- Course *Algorithmic bioinformatics*, WS 2002/03
- *Bioinformatics Software Exercises*, SS 2003
- Seminar *Sequence comparison algorithms*, WS 2003/04:

Lectures at ESF International Genetics and Bioinformatics School, Portofino, 10/2001

Lectures at the International Summer School Computational Biology, Warsaw, Poland, 9/2003

Martin Vingron: Others

Associate editor of *J Comp Biol*

Editorial Board Member of

- *Bioinformatics*
- *Briefings in Bioinformatics*
- *J Mol Med*