INTERGENOMICS
Towards a bioinformatic framework for "InterSystems Biology"

1 The scientific aims of "Intergenomics"

It was the scientific vision of the Intergenomics Center to elaborate bioinformatic methods and a bioinformatics infrastructure optimally suited for studying different aspects of infection mechanisms. The aim was to describe and analyze the processes that are encoded by the genomes of pathogens and hosts and their interaction during infection. A range of mostly prokaryotic pathogenic organisms were considered, with a focus on *Pseudomonas aeruginosa*, and both mammalian and plant organisms were the hosts to be studied in the Center.

2 Structure of the Center and its development

Altogether 11 groups of 4 organizations decided in 2001 to establish the Bioinformatics Competence Center Braunschweig, Intergenomics. The four institutions were the GBF (German Research Center for Biotechnology), TU (Technical University) Braunschweig, FH (Fachhochschule / University of Applied Sciences) Braunschweig/Wolfenbüttel and, as a commercial partner, BIOBASE GmbH in Wolfenbüttel. With the move of the coordinator, E. Wingender, to the University of Göttingen (Medical School), the research activities of the Center were geographically and institutionally extended, with the additional advantage of having one partner embedded in a medical environment.

INTERGENOMICS – Structure of the Competence Center

Project Coordination:
Prof. E. Wingender *(UKG / GBF)*

Project Management:
Dr. R.-J. Müller *(GBF)*

Scientific Advisory Board ***(external)***
Dr. R. Apweiler (EBI / Hinxton, UK)
Prof. R. Hofestaedt (University of Bielefeld)
Prof. D. Schomburg (Cologne University)
Dr. T. Werner (Genomatix Software GmbH, Munich)

Partners			
• GBF - German Research Centre for Biotechnology **• University of Goettingen UKG** 6 Groups Coordination: ***Dr. H. Blöcker***	**• Technical University of Braunschweig** 3 Groups (+2 being set up) Coordination: ***Prof. D. Jahn***	**• University of Applied Sciences Braunschweig / Wolfenbuettel** 1 Group Coordination: ***Prof. U. Klages***	**• Biobase GmbH** 1 Group Coordination: ***Dr. H. Saxel***

From the very beginning, the Bioinformatics Competence Center Braunschweig had a clear scientific focus on the bioinformatics characterization of infectious processes, i.e. on the interplay of the genome-driven interaction between hosts and pathogens. Consequently, mostly those groups that have a strong biological background started with the construction and population of databases of the relevant data for the pathogenic microorganisms and for the infected eukaryotes. Other participants focused on the development of software for the modeling and simulation of the processes. Thus,

having chosen a complex process as the main biological subject of the Center, the whole range of bioinformatic methods was to be applied.

Given the ambitious scientific goals of the center on the one side, the fact was to be acknowledged that only few of the participants had a proper bioinformatics structure in place when the project started in 2001. At the GBF, now HZI, the department of Genome Analysis (H. Blöcker) was involved in algorithmic and software development supporting genome-related research since many years, and the project and research group Bioinformatics was known for their work on different aspects of gene regulation (E. Wingender). From the latter, BIOBASE had been founded in 1997 as a company providing bioinformatics tools and services, mainly databases, and was considerably growing since then. Though much of the work at GBF was conducted in cooperation with the computer sciences at the TU Braunschweig (H. D. Ehrich), bioinformatics at the TU was practically non-existing at that time. This initial situation was certainly a difference to most of the other Competence Centers, but from this starting point, Intergenomics was quite successful in generating an excellent and attractive bioinformatics scene in Braunschweig.

As a consequence of the described starting conditions, the required methodological know how had to be acquired by the biological groups, which in turn helped the computer scientists in the consortium to develop the required understanding for the underlying biological mechanisms so that successful modeling and simulation could be tackled. Those members that had already established know-how in specific areas (e. g., BIOBASE in data handling) disseminated their expertise among the partners. As a result, a network of multiple cooperative contacts inside Intergenomics emerged quickly. Up to now, the partners have helped each other to establish a series of successful individual and joint research activities, as evidenced by a considerable number of peer-reviewed publications and numerous talks at international and national conferences.

In addition to the multiple common interests in their research, an information infrastructure was generated that comprised joint internal and external lectures, seminars, talks, thus enhancing the coherence among the partners and introducing external experts in a determined way. In particular, a number of cooperations with partners of the other Centers were established, and a Workshop involving all Centers of the NBCC as well as a number of DFG-funded bioinformatic centers was organized.

This was accompanied by substantial efforts to establish bioinformatics teaching at the TU Braunschweig. For this, two new chairs in bioinformatics were planned, but even long before the opening of these positions, remarkable teaching activities were organized by the partners with significant contributions from outside the Center.

Also as part of the comprehensive bioinformatics culture that was generated so far during this project, an internationally renowned peer-reviewed online journal was established which is re-printed by IOS Press, Amsterdam.

3 Scientific achievements of the Intergenomics Center

According to the basic concept of the Intergenomics Center, and acknowledging the need to build up many of the required resources from scratch, big efforts were undertaken to collect the required data and render them manageable in appropriate databases. Even where the needed database structures were already existing, it was necessary to update these resources with the corresponding contents.

Algorithms were developed and software packages were implemented to enable working with these knowledge bases and apply them on the full range of problems associated with the study of infection mechanisms. The required expertise was acquired so that state-of-the-art research in methodological bioinformatics could be performed. Since the scientific concept of the Center from the very beginning exceeded the “conventional” bioinformatics approach by including ideas of systems biology, a broad range of methods from decoding sequence-encoded signals up to analyzing network properties had to be developed and applied.

Finally, application of these tools on problems that were defined by the central subject of the Center as well as by the numerous external collaborations its members have been involved in revealed interesting results for a number of biological systems..

3.1 Data resources

Creating and populating databases for specific biological areas has been the major focus of the Research Group Bioinformatics at GBF when Intergenomics started. This expertise has been largely adopted by the spin-off company BIOBASE. Both groups made efforts to provide required contents to the Intergenomics partners: BIOBASE updated the databases TRANSFAC® and TRANSPATH® specifically for the tasks of the Intergenomics Center. Knowledge about transcription factors and their binding sites relevant in gene induction events during antibacterial and immune responses was accumulated and put into TRANSFAC®, specific positional weight matrices (PWM) were constructed for these factors to enable prediction of their binding sites. TRANSPATH was particularly enriched by data on networks that are relevant in these processes, such as the TLR4 pathway. The Research Group Bioinformatics at GBF (Wingender) maintained further the contents of the CYTOMER ontology on expressions sources (organs, tissues, cell types), the S/MARt DB (a database on scaffold/matrix attached regions), PathoDB and PathoSign (pathologically relevant mutations in transcription regulation or signaling components). More recently, and after the transfer of this group to the University of Göttingen, a new step towards more complex systems modeling was undertaken by establishing the database EndoNet on intercellular (firstly, endocrine) networks.

Serving as a seed, both groups gave initial help and advice in developing corresponding data resources at the Technical University of Braunschweig. Thus, TRANSPATH® was used as a kind of template to construct a database on signal transduction pathways that are relevant during plant infection (PathoPlant®; Hehl). TRANSFAC® provided the first paradigm for PRODORIC, a database on different aspects of gene regulation in prokaryotes (Jahn). Very soon, both groups successfully developed considerable own dynamics in generating additional data resources such as AthaMap, a genome-wide map of predicted transcription factor binding sites in *Arabidopsis* (Hehl), and SYSTOMONAS, a knowledge base on the systems biology of Pseudomonads (Jahn). This was complemented by an enzyme-reaction database that was constructed on the basis of KEGG contents after extensive revision (GBF/HZI, Zeng).

At the GBF/HZI, an integrated proteome database (LEGER) that can store, retrieve and analyze various information based on LC-MS data has been designed as a laboratory information platform that manages an entire set of experimental data focusing on *L. monocytogenes* (Wehland/Kärst). It also provides query functions, data mining tools, which analyze the statistical significance of experimental data, and links to other available public and non-public databases. The user interface is web-based, so that data acquisition and query can be performed on remote computers within the intranet. In particular, the subproteomes of different mutants or species can be examined against each other, resulting in lists of exclusively expressed, up/down regulated or equally expressed proteins. These lists can be linked to diverse genome annotation data, like InterPro, KEGG, etc. LEGER is accessible at http://leger2.helmholtz-hzi.de/cgi-bin/expLeger.pl. A new protein database for *K. pneumoniae* was also made available at the GBF/HZI (Zeng).

Also at GBF/HZI, a renowned database (VBASE2) was developed further and made available at http://www.vbase2.org. VBASE2 is an integrative V gene database, combining the V gene sequence resources from the EMBL database, Ensembl, VBASE, IMGT and KABAT. The VBASE2 dataset is generated automatically and is regularly updated.

3.2 Methods, algorithms and software tools

Mostly on the basis of the data gathered in the mentioned databases, new methods were developed for for data and text mining, for the detection of patterns in sequences, and for the analysis of regulatory and metabolic networks.

3.2.1 Text mining

MineBlast was developed to mine scientific literature for information relevant to identified novel target proteins based on amino acid sequences. Users can submit a simple list of protein sequences via

a web-based interface. MineBlast performs a BLASTP search in UniProt to identify all relevant names and synonyms based on homologous proteins and subsequently queries PubMed, using combined search terms in order to find and present peer-reviewed literature. MineBlast is accessible at http://leger2.helmholtz-hzi.de/cgi-bin/MineBlast.pl. Development and programming of this and some other tools mentioned above/below by (Wehland/Kärst) were done in collaboration with groups at the Technical University of Braunschweig. All mentioned activities of this group were integrated in national and international projects, e.g., Neue Methoden der Proteomforschung, Kompetenznetzwerk PathoGenoMik and the NoE Europathogenomics, and will also be part of the ERAnet SPATELIS. At the Ehrich/Eckstein group, the CaptionSearch tool was developed that allows specific search for pictures and tables, thus complementing the classic method of term search in abstracts by using a more refined layout based approach on the full text paper.

3.2.2 Application of signal theory to the analysis of biomolecules (Blöcker)

In this project it was intended to explore the application of signal theory (as established in image analysis and speech recognition) in bioinformatics for the function-oriented analysis of biomolecules. The general intent was to reveal similarities, homologies and analogies which are based on considerations of physico-chemical properties rather than on statistics of letters (or other global symbols) and confirm findings by wet lab data.

The detailed theoretical background was presented for applying signal theory (Kauer and Blöcker, 2003, 2004) to the analysis of biomolecules. Here, a key step is to use biochemical and biophysical properties of biomolecules to model an *n*-dimensional signal, which represents the entire information-bearing biomolecule. After a first pilot project, an implementation of the signal theory-based approach for the detection of novel types of DNA similarity, which are based on DNA physical properties (Deyneko *et al*., 2005) was presented. A systematic study of the sensitivity of the new similarity measure revealed qualitative differences to letter-based similarity. We showed that various DNA parameters and combinations thereof can be rather useful for revealing functional similarities of different DNA sequences.

The signal similarity measure was applied in order to reveal common structures shared by functionally related promoters of *E. coli*. Statistical analysis of cellular functions using the GenProtEC *E. coli* database identifies an enrichment of several functional categories of genes, the promoters of which showed high signal similarity of their enthalpy profiles. Comparing *E. coli* promoters, the DNA parameter melting enthalpy enabled one for the first time to recognize similarities, for example, between SOS response gene promoters. It is noteworthy, that in these cases the letter similarity (by BLAST) was only in the range of 40-55%, while signal identity (by FeatureScan) was more than 85%. In conclusion, the strong correlation was found between physico-chemical DNA similarity of promoters of *E.coli* genes and their transcription activity under various extra cellular conditions.

Comparative genomics was another application field of the developed approach. Based on the assumption that evolutionary important regions should preserve not just primary nucleotide sequence but rather features behind it, we carried out phylogenetic comparisons of promoter regions of genes from human, mouse, rat and dog. As a result we compiled a catalogue of motifs, many of which were similar to binding sites for different transcription factors. In comparison to results in the literature, our list of top-scoring motifs is not identical, but similar and more reliable scince it is related to conformational properties of DNA rather than to simple nucleotide conservation.

In a yet unpublished study the present signal theory-based technology was applied to the investigation of close genomes. In this pilot study (on chromosomes 21 from human and chimpanzee), it was found that nucleotide mismatches in promoters were apparently introduced in a correlated manner during the course of evolution, so that, for example, the DNA property “melting enthalpy” was retained. Such property conservation of promoters (in the presence of letter mismatches) is significantly different from nucleotide conservation, shows significant positional and functional biases and seems to represent a novel and fundamental basis of gene expressivity. Relating to these recent findings we coined the expression “compensatory mutations”.

It is stated in our publications that we suggest to substitute BLAST or similar algorithms by our tool when it comes to function-oriented comparisons of DNA. In general, both types of algorithms will develop independently and complement each other. To make the current implementation of the general technology available to a wider community, web sites and services are online (Deyneko *et al.*, 2006, http://genome.helmholtz-hzi.de/featurescan).

References

Kauer, G. and Blöcker, H. (2003). Applying signal theory to the analysis of bio-molecules. Bioinformatics **19**, 2016-2021.
Kauer, G. and Blöcker, H. (2004). Analysis of disturbed images. Bioinformatics **20**, 1381-1387.
Deyneko, I. V., Kel, A. E., Blöcker, H. and Kauer, G. (2005). Signal-theoretical DNA similarity measure revealing unexpected similarities of *E. coli* promoters. In Silico Biol. **5**, 547-555.
Deyneko, I. V. Bredohl, B., Wesely, D., Kalybaeva, Y. M., Kel, A. E., Blöcker, H. and Kauer, G. (2006). FeatureScan: revealing property-dependent similarity of nucleotide sequences, Nucleic Acids Res. **34**, W591-595.

3.2.3 Analysis tools for prokaryotic omics research

A number of analysis tools have been developed and made available that facilitate the interpretation of data coming from transcriptomics and proteomics studies. They predict transcription factor binding sites, regulons, identify split tRNAs, functionally interpret microarray data, predict certain protein features (signal peptides or membrane-spanning regions), simulate the appearance of 2D gel patterns, and help to quantify metabolomes (Jahn). The value of these tools has been proven in daily application in the laboratories next door and for some of these tools, lively interest has been expressed by industrial users.

Bacterial promoter sequences are very short and variable, making them very difficult to identify by pattern searches. Having access to several complete genomes sequences of closely related *Listeria* strains and species, the research group of Wehland/Kärst had decided to identify regulatory sequences by phylogenetic footprinting. The approach is based on suffix-trees that permit fast searches for repeats, palindromes and mismatches. The 'PromoS' (Promoter Search) tool currently offers repeat searches in one or two genomes and filter options to select, e.g., palindromes.

The same group also developed the visualization tool “LaneSpector”. This tool allows the systematic comparison between apparent and calculated protein masses. The detailed presentation of the LaneSpector plot promotes the recognition and dissection of data related to relevant biological issues such as posttranslational modifications. Investigating the data from the membrane-subproteome analysis stored in LEGER (s. above), a large proportion of the 301 membrane proteins identified based on a combination of SDS-PAGE and LC-MS showed an unexpected migration in SDS-PAGE (Wehland/Kärst).

Computational strategies for organizing and visualizing large-scale biological data, e.g., generated by LC-MS, were developed (LEGER, above). Exploratory visualization tools that combine experimental data and data from annotation databases facilitate the understanding of bacteria at the system level. Besides *Listeria* species other bacteria are also very important in infection research. A set of visualization tools were developed as web service that comprises three different views of genomic data for selected bacteria species that allow the visual exploration of annotation data from KEGG, Gene Ontology and the genome annotation with the same set of proteins. A list of proteins with or without regulatory data for two conditions (e.g. from microarray experiments) can be used for all three tools. VIS-O-BAC is accessible at http://leger2.helmholtz-hzi.de/cgi-bin/vis-o-bac.pl (Wehland/Kärst).

The group of Martins dos Santos at the HZI developed a computational platform for network comparatively representing new information using a proprietary ontology and integrating it in the data repository. The platform is a modular system where independently developed tools are plugged-in. The platform comprises so far modules based on structural network analysis, as well as on two-dimensional annotation and metabolic reconstruction. A module for dynamic modeling of small circuits is being implemented. Using this platform, a comparison of the metabolic space of the central metabolism of the Pseudomonads under study revealed that the number of elementary pathways representing the metabolic potential of PAO1 could be 2 to 6 times higher than in KT2440, despite the

fact that the former has only two more reactions than the latter in the considered reaction set, and indicates a higher flexibility and redundancy of the central metabolism of PAO1. Both kinetic and logical models (Boolean Networks, Petri Nets) were developed for the pWW0 plasmid, which is an ubiquitous, easily transferable plasmid coding for the metabolism of aromatics and that has been widely used as a model system in regulation studies. Comparative validation of the models (kinetic vs. logical) and in particular of their interplay (through markers of the metabolic state of the cell) with genome-wide metabolic models for the whole cell is being be done on the basis of the data available in literature and by carefully designed transcription experiments. These conceptual models are being extended for the description of key regulatory circuits in *Ps. putida* and or *Ps. aeuroginosa* (Martins dos Santos). A module for the unbiased inference of regulatory networks from microarray data using Bayesian networks has been developed and is currently being experimentally validated. Initial results on *E.coli* array data allowed identification of one "master" regulator controlling two circuits that are involved in copper resistance previously unknown to be connected. These findings are currently being tested experimentally with expression data from the pertinent knock-outs and ChIP-Chip measurements.

Several new and improved methods and algorithms have been developed for *in silico* reconstruction of metabolic networks, connectivity and modular network analysis and network decomposition. For example, the algorithm IdentiCS was developed for fast identifying protein coding sequences and reconstructing metabolic networks directly from unannotated genome sequences. This method was verified with different genome sequences datasets of *Klebsiella pneumoniae* and applied to other organisms (*E. coli* Nissle 1917, *Bacillus megaterium* and *Aspergillus niger)* (Zeng).

In addition to method development, the metabolic and regulatory networks involved in the formation of virulence factors and stress responses of *P. aeruginosa* were studied both *in silico* and experimentally. A particular emphasis was the availability of iron and its interplays with the generation of oxidants (one of the innate immune responses of host cells upon infection by pathogens). Experimental data from transcriptomic and proteomic analysis are combined with bioinformatic tools to extend the networks for including other pathways and subsystems such as quorum sensing, arginine catabolism and general stress responses (rpoS, IHF etc.). This led to the identification of PA2384 as a new regulator involved in large-scale regulation of iron starvation and quorum sensing in *P. aeruginosa*. For *K. pneumopniae* a new, second triose phosphate isomerase was discovered and experimentally verified (Zeng).

3.2.4 Regulatory networks in eukaryotic cells

Gene regulation is mostly under transcriptional control. Since even for the best investigated genomes (e.g., human and mouse) only about 1% of all expected TFBS have been experimentally revealed (upper limit estimate), either reliable high-throughput methods or trustworthy prediction methods for detecting promoter properties are required. Therefore, significant efforts were done to provide tools for single site detection (e.g., Match^{TM}) and to optimize their parameterization (BIOBASE). These activities were complemented by the phylogenetic footprinting studies, which also yielded a number of tools that have been made available on the Intergenomics server (Wingender). The combinatorics of TFBS were intensively studied and used to derive predictive promoter models (Wingender; BIOBASE).

Eukaryotic regulatory networks and their structure were intensively studied in parallel projects, the results of which could be fed back into the Intergenomics work. Thus, and on the basis of TRANSPATH contents, a network of 742 genes could be constructed for the TLR4 pathway which showed small-world characteristics and was described best with a scale-free model. Its clustering properties indicate hierarchical modularity (Wingender).

The actin cytoskeleton is an essential organelle of eukaryotic cells. As a number of pathogenic microorganisms subvert its functions to promote invasion into host cells and cell-to-cell spreading, a complete understanding of these particular host pathogen-interactions is paramount in infection biology. Because the dynamic rearrangements of actin filaments involved in theses processes are very complex and depend on a remarkably fine tuned regulation we developed the software tool 'ActMoST' (Actin Motility Simulation Tool) to collect and integrate the available data and visualize the dynamic

interactions of the participating protein components. ActMoST covers essential components and scenes such as actin treadmilling, activation and branching permitting the building of dynamic models. The generated scenes can be exported as videos in AVI-format (Wehland/Kärst).

Functional analysis of transcriptomics data, in particular from microarrays, or from proteomics studies can now be analyzed in an integrated manner with pathway mapping and promoter analysis by a newly developed platform (ExPlain™; BIOBASE) into which many of the aforementioned concepts were subsumed.

Significant methodological contributions to the modeling of networks on the basis of, e. g., the TRANSPATH database were made. Tools were developed for the automatic creation of discrete models (Live Sequences Charts, Colored Petri Nets and UML-Statecharts) and their visualization (Ehrich/Eckstein).

4 Exploitation of the scientific results

Most noteworthy, concepts are presently developed for the exploitation of the achievements of the TU Braunschweig (subproject D. Jahn) by BIOBASE. In particular, the PRODORIC database and other resources established by this group may become an essential part of a comprehensive systems biological resource for prokaryotes which BIOBASE is planning to set up. Such a system should be particularly capable to generate hypotheses about the key regulators and their genes in pathogenic bacteria.

The University of Göttingen, Medical Faculty, has closed a cooperation contract with BIOBASE about the exploitation of results elaborated by the Department of Bioinformatics. This includes a right of first refusal for new achievements of this department. In 2007, negotiations will start about the transfer of EndoNet (a database on endocrinological networks) for expansion of its contents and distribution by BIOBASE.

For VBASE2, some negotiations about commercial exploitation were initiated but not yet closed. There is explicit interest by a bioinformatics company to distribute the database and the associated tools, a first informal agreement has been reached.

5 Teaching Bioinformatics in Braunschweig

5.1 Course offerings at TU and FH

In the framework of Intergenomics a solid Bioinformatics education was implemented at the TU Braunschweig. Bioinformatics modules are integral part of the Bachelor in Biology and Biotechnology education plans since 2004 (accredited with ZEWA). The Biology and Biotechnology Master education also contains Bioinformatics modules from their starts in 2006 on. The selection of a Bioinformatics major is possible. The modules consist of three different lectures (Bioinformatics 1 and 2, Systems biology) and two different computer pool-aided practical courses as well as various courses directly in the research groups. Moreover, Bioinformatics courses are part of the diploma education in Bioengineering and Informatics.

Bioinformatics graduate education (PhD programs) is exported by the TU to the Medizinische Hochschule Hannover in the Biomedical Research School (Graduate School in the excellence initiative) and to the graduate college “Pathogenicity and Biotechnology of Pseudomonades” (GRK 653/3).

At the TU a 30 place Bioinformatics CIP computer pool was established in 2004 with grants of the HBFG program. The TU financed infrastructure (room renovation, cooling system etc.) worth 100.000 Euro.

Due to these programs Bioinformatics education is permanently sustained at the TU Braunschweig.

At the University of Applied Sciences (FH) Wolfenbüttel, a bioinformatics literature seminar “*Informatik in der Biologie*“ was held by M. Scheer (summer term 2004 - winter term 2006/2007, in

total 6 times). This was complemented by a practical course in software ergonomics: *"Entwicklung von GUI-Protoypen für eine im Rahmen von Intergenomics entwickelte bioinformatische Java-Bibliothek"* by Prof. Dr. F. Klawonn & M. Scheer (winter term 2004/05, summer term 2005)

5.2 Bioinformatics professorships at the TU Braunschweig

In 2006 the first of two W3 professorships was established with Prof. Dr. Schomburg as new head of the Bioinformatics department in the Faculty of Life Science. This is a completely new Professorship with new staff and budget. The TU, the Land Niedersachsen, the Helmholtz center for infectious diseases (HZI) and the Medizinische Hochschule Hannover (MHH) invested several Mio Euro and created multiple new positions. The Intergenomics grant approval to the TU also contained approximately 500.000 Euro for this professorship which are not paid yet. The second professorship in the Faculty of Mathematics and Informatics is on the way.

6 Networks of internal, national and international collaborations

6.1 Internal cooperations

INTERGENOMICS has intensive internal cooperation between the different projects and is also externally highly linked to other research groups, competence centers an funding programs. The intensive internal cooperations are documented by the fact that about on third of all publications (24/75) are with joint authorship of several groups that participate in the Center. In more detail, these cooperations cover the following topics:

- TU (Jahn) ⇔ HZI (Müller)
  Development of a web interface for VBASE2 database, joint publication
- TU (Jahn ⇔ Ehrich ⇔ Hehl)
  Regular meetings and joint supervision of study (Studienarbeiten) and diploma theses
- TU (Jahn) ⇔ BIOBASE (Saxel)
  PRODORIC database structure
- TU (Jahn) ⇔ HZI (Zeng / dos Santos)
  Integration and reconstruction of metabolic and regulatory networks of *P. aeruginosa* (database PRODORIC) and on proteomic and transcriptomic studies
- HZI (Wehland/Kärst) ⇔ TU (Ehrich/Jahn)
  Programming of software tools and interfaces
- HZI (Zeng) ⇔ TU (Ehrich/Weimar)
  Metabolic pathway analysis and teaching on a course Systems Biology
- HZI (Zeng ⇔ dos Santos)
  Metabolic network analysis, exchange of information and techniques
- HZI (Zeng ⇔ Blöcker)
  Reconstruction and analysis of metabolic networks of *E. coli* Nissle und Bordetella petrii from raw genomic data
- HZI (Zeng ⇔ dos Santos) ⇔ UKG
  Information exchange on regulatory network of *P. aeruginosa*
- TU (Hehl ⇔ Jahn ⇔ Ehrich) ⇔ BIOBASE
  PathoPlant® database structure and the web interface
- TU (Hehl) ⇔ BIOBASE
  TRANSFAC® annotation of plant transcription factors
- HZI (Müller ⇔ Blöcker)
  Analysis of the IgH locus of the 129/Sv mouse strain
- FH (Klages) ⇔ TU (Jahn)
  Close contextual and technical collaboration
- BIOBASE ⇔ UKG
  Database infrastructure; promoter, network and gene expression analyses

6.2 National networks

The Intergenomics Center has established numerous links to the other Bioinformatics Competence Centers of the NBCC. For instance, E. Wingender (UKG) and A. Kel (BIOBASE) had been appointed as members of the Scientific Advisory Boards of CUBIC (Köln) and BIC-GH (Gatersleben), respectively. On the other hand, D. Schomburg (CUBIC) and T. Werner (BFAM) were members of the Scientific Advisory Boards of Intergenomics. In addition, there is intense scientific exchange of the TU and FH groups with the BCB (Berlin), as well as of both BIOBASE and the UKG group with different groups at IPK Gatersleben.

Moreover, several members of the Center are (or have been) also participating in NGFN (e.g., Blöcker, Wingender) or other BMBF funded consortia such as FUGATO (Blöcker).

6.3 International visibility

Most of the Intergenomics participants brought their expertise to international collaborations and consortia, for instance projects that were/are funded by the EU. For instance, there is a highly synergistic effect between the network modeling attempts for Intergenomics done by the UKG group and BIOBASE with those made by these two participants within the EU-funded COMBIO consortium on the p53 network. The novel systems biological concept on intercellular networks pursued by the UKG group was a topic of a German-Japanese workshop held in March in Tokyo, with financial support from BMBF and the Tokyo Medical and Dental University. In addition to its participation in this and a number of other EU projects, BIOBASE (A. Kel) also coordinated a project proposal which was accepted among the last projects funded under framework 6 (Net2Drug). It brings together experts from bioinformatics, cheminformatics, and several omics fields for developing a novel drug development platform.

The HZI, through V. A. Martins dos Santos, coordinates two European consortia focused on the systems Biology of *Pseudomonas putida*, a non-pathogenic bacterium genetically and physiologically related to *P. aeruginosa*. These projects are (1) the 18-partners, Europe-wide ERA-NET SysMo Project (PSYSMO PD-01-06-12) on the Systems Biology of *P. putida* as cell factory of excellence for the production of fine chemicals.project, and; (2) a 7-partner EU NEST project (no. 029104) on the reprogramming of *P. putida* for high-value biocatalysis. Both projects rely, unto a certain extent on preliminary work done (in particular the computational infrastructure on metabolic networks) within this Intergenomics project.

7 Peripheral activities

In this section, some activities should be briefly reported which are not directly funded by the Intergenomics grant, but were significantly facilitated with the working Center as supporting background and, thus, Intergenomics had a major impact on.

In 1998, the association Bioinformation Systems e.V. has issued the online journal "In Silico Biology". It was the first journal in this field that was edited primarily as online journal, which was rather new at that time and required some time for the community to get used to. Since the foundation of the Center in 2001, the number of submissions increased significantly leading to a jump of published papers from 8 in 2000/2001 to 45 in 2002. The 63 articles published by ISB in 2006 compare with 137 submissions, indicating a rejection rate of about 50%. ISB has an internationally renowned board of editors and is indexed by PubMed. ISI is in the process of assigning an Impact Factor to ISB. ISB articles gained considerable visibility, some of the ISB publications are top-ranked when corresponding keywords are searched with Google.

Also not obviously related with the Intergenomics Center, but largely supported by the visibility it gained through the involvement in it, the commercial partner BIOBASE GmbH was able to initiate a considerable international development since 2001. Some of the most important milestones for this were: BIOBASE received Japanese investment (April 2002), managed to acquire the whole database business from Incyte Corp., in particular the Proteome Databases, and founded BIOBASE Corp. in Beverly, Massachusetts (January 2005), and established a production office in Bangalore, India

(October 2006). In the same time period, BIOBASE could consolidate further its international distribution network.

Flanking the establishment of bioinformatics courses at the TU Braunschweig, the move of the coordinator E. Wingender to Göttingen initiated the organization of several bioinformatics courses for students of the Medical, the Mathematical and the Biological Faculty at the Georg August University as well. The principal scientific idea of the Intergenomics Center also inspired a Systems Biology application to the Forsys funding program, in which 27 groups out of 6 faculties of the University, all three Max Planck Institutes in Göttingen and one commercial partner participated. Though at the end not elected for funding, this initiative generated already a network which will continue to cooperate on problems related to systems biological issues.

8 Future perspectives

Whereas in the very beginning of the Intergenomics Center, bioinformatics at the TU Braunschweig largely benefited from the stimuli exerted by GBF and BIOBASE, the Intergenomics successfully caused a powerful bioinformatics center being built now at the University from which HZI may benefit in near future, and which is now going to develop a new kind of partnership with BIOBASE. With the successful establishment of a new chair of Bioinformatics at the TU Braunschweig and the acceptance of the call by D. Schomburg, the previous coordinator of the CUBIC Center, it has been ensured that Braunschweig will continue to be place with a highly significant contribution to German bioinformatics research and development in the future. It also guarantees that the tedious but tireless and successful efforts of the past five years to implement profound bioinformatics teaching at the Technical University will yield fruits and be further expanded. It was also a remarkable and successful experiment to integrate the expertise of the Computer Scientists at the FH Braunschweig-Wolfenbüttel into the Center. The joint work of TU and FH in research and teaching will certainly be prolonged since life-science oriented topics will be of increasing importance at the FH as well.

The concepts of Systems Biology were already inherent in the original scientific idea of the Intergenomics Center. Within the last five years, it was possible to develop and establish an infrastructure of tools, databases and a network of relations and know-how which has made the local community ready for the next step, i. e. an integrated attempt to tackle problems of infectious diseases by means of a systems biological approach. This matches the strategic alignment of the HZI as well as the long-term scientific interest of involved groups from TU Braunschweig. Also the commercial partner of the Center, BIOBASE, is developing further towards this goal.

Database of Gene Regulation in Pathogenic Microorganisms

Project leader: D. Jahn **TechnicalUniversity Braunschweig**

Summary of Scientific Results

Aim of the project

Establishment of systems biology databases on gene regulation, signal transduction and metabolic networks in prokaryotes with focus on pathogenic bacteria. Development of software tools for the deduction, visualization and prediction of gene regulatory and metabolic networks in prokaryotes.

Major results: Online databases and systems biology tools (2001-2006)

PRODORIC 2.0	Largest database on gene regulation in bacteria (www.prodoric.de).
SYSTOMONAS	Systems biology database of Pseudomonads (www.systomonas.de)
JVirgel 2.0:	Calculation and simulation of virtual 2D protein gels (www.jvirgel.de)
PrediSi:	Prediction of signal peptides (www.predisi.de)
JCat:	Codon usage adaptation tool (www.prodoric.de/jcat).
Split-tRNA-Search:	Search of split tRNAs in whole genomes (www.prodoric.de/sts/)
Virtual Footprint:	Regulon prediction in prokaryotes (www.prodoric.de/vfp)
JCaMelix:	Prediction of membran-spanning regions (www.jvirgel.de)
JProGO:	Functional interpretation of microarray data (www.jprogo.de)
Metaquant:	Quantification of metabolome data (bioinformatics.org/metaquant)

PathoPlant®: A Database on Plant-Pathogen Interactions

Project leader: R. Hehl **Technical University Braunschweig**

Summary of Scientific Results

Aim of the project

The subproject's goals was the establishment of an interactive database on signal transduction components and reactions related with plant-pathogen interactions, allocation of the database via Internet, and development of software to enable modeling, simulation and prediction of plant-pathogen interactions.

Major results (2001-2006)

The PathoPlant® database was created and a client software was developed for annotation. An interactive web interface was implemented for accessibility of the database via Internet. AthaMap has been established as another database to display gene regulatory elements in the *Arabidopsis thaliana* genome. The PathoPlant® database was extended with microarray data from *Arabidopsis thaliana* treated with phytopathogens and elicitors. The functionality of AthaMap was expanded to display and calculate the occurrence of combinatorial regulatory elements. Both databases were linked for the analysis of co-regulated genes.

PathoPlant® http://www.pathoplant.de
AthaMap http://www.athamap.de

Databases of genome-driven infection processes

Project leader: H.-D. Ehrich / S. Eckstein **Technical University Braunschweig**

Summary of Scientific Results

Major results (2002-2006)

Our Project has the aim to support the building and annotation of biological databases. Three problems in this area seemed most important.

Literature annotation is one of the most time-consuming steps while establishing new databases. So, the first goal was to provide meaningful tools to support this process, making it more effective. The CaptionSearch tool allows specific search for pictures and tables. This complements the classic method of term search in abstracts by using a more refined layout based approach on the full text paper.

Second, in database integration many different syntactic and especially semantic conflicts between the database schemas and the data occur. The rate of evolutionary change of the schemas is the source of many of those problems. Ontologies support integration by describing the schemas and data semantically. We developed a coevolution approach to minimize the user involvement by keeping the ontological annotation of the schemas up-to-date.

Finally, in direction of systems biology, dynamic process models were developed for simulation on basis of existing databases like TRANSPATH. Modeling manually is very time-consuming and error-prone. Thus, we developed tools for the automatic creation of discrete models (Live Sequences Charts, Colored Petri Nets and UML-Statecharts) and for their visualization: the GUI-editor, the TRANSPATH-Connector and the Pathway Modeler.

Development of Methods and Frameworks for Bioinformatic Applications and Internet Webfrontends of Biological Databases

Project leader: U. Klages **FH-Braunschweig / Wolfenbüttel**

Summary of Scientific Results

Aim of the project

Development of methods and frameworks in context with web front ends for biological databases and bioinformatics applications. For this purpose, integration of resources from student practical courses.

Major results (Nov. 2002 - Dec. 2006)

A new Java-based program suite (JProGO) for the automated functional interpretation of high-throughput gene expression data from bacteria such as microarray data was developed including a proper statistical validation of the results. In addition, an interactive web interface with several visualization capabilities was developed. JProGO is based on the comprehensive Prodoric database on biological networks in prokaryotic organisms which was established by the Intergenomics partner group of Prof. Dr. Jahn at TU Braunschweig. In this course, the Prodoric database was expanded by integrating the function classification system *Gene Ontology* and appropriate protein annotations. In addition, student resources were used in software development courses where prototypes for a enhanced graphical user interface for JProGO were created. Besides JProGO, support was given in the development of other web-based tools and database systems of our Intergenomics partner (Prof. Jahn).

Planned further research (until June 2007)

The JProGO software suite (see 'Major results') provides several starting points for expansion towards the integration of other types of biological networks such as regulatory networks. Hence, it shall be expanded to include groups of genes regulated by the same transcriptional regulator (regulons).

Modelling of the interaction between regulatory networks

Project leader: E. Wingender **University of Göttingen**

Summary of Scientific Results

Aims of the project

Extraction of signaling networks relevant for infection mechanisms out of the corresponding partner databases and the analysis of these networks; analysis of the regulatory regions of the genes involved in these mechanisms and development of predictive models.

Major results (2001-2006)

To reliably identify single transcription factor binding sites (TFBS), the major breakthrough beyond the application of positional weight matrices (PWMs) is still outstanding. One significant improvement could be achieved by a subclassification approach of degenerate sets of experimentally determined TFBSs. As an additional and independent criterion, comparative genome analysis (phylogenetic footprinting) was systematically applied and methodologically expanded. We could demonstrate to what extent TFBS are conserved, e. g. between human and rodents, and that non-conserved sites in most cases have “surrogates” at other places in the orthologous promoters. We could also show that the well-known variability of TFBS does not point to arbitrariness in their composition, but rather to biologically meaningful subtle functional differences in the instances of the binding sites recognized by one and the same factor. On the basis of these studies, the conventional PWM approach has been combined with phylogenetic footprinting into one hidden Markov model which shows significantly improved reliability compared with the pure PWM method.

Beyond the detection of single TFBS, it is of crucial relevance for analyzing the specificity of promoters to recognize the combinatorics of their composition, i. e. typical combinations of TFBS. For this purpose, new methods of establishing promoter models were developed to identify characteristic pairs of TFBSs and to analyze their distance distributions. The applicability to sets of individual orthologous genes as well as to sets of coregulated genes were proven.

The gene regulatory processes that are triggered in host cells by pathogenic bacteria are part of a complex signaling network, mainly comprising the TLR4 pathway. This and related pathways have been subject to thorough investigation by graph theoretic approaches. Doing so, the generally accepted view of most biological networks as small-world and scale-free networks could be mainly confirmed, but have clear limits. Moreover, graph parameters have been identified that are most suitable to identify the most crucial keynodes in the network.

Application of Signal Theory to the Analysis of Biomolecules

Project leader: H. Blöcker **GBF-Braunschweig**

Summary of Scientific Results

Aim of the project

In this project it was intended to explore the application of signal theory (as established in image analysis and speech recognition) in bioinformatics for the function-oriented analysis of biomolecules. The general intent was to reveal similarities, homologies and analogies which are based on considerations of physico-chemical properties rather than on statistics of letters (or other global symbols) and confirm findings by wet lab data.

Major results (2001-2006)

In the DNA field, our approach is fully established as tool which will replace BLAST and similar programs when it comes to functional analysis of DNA stretches. Already based on a pure software implementation it could be shown that the system works as in theory. Single nucleotide resolution could be demonstrated. Hardware implementation (DSP card) allows for searching large eukaryotic genomes in acceptable time – all this was done with artificial and natural sequences. Systematical selectivity and sensitivity investigations have been carried out with 38 different encoding schemes. We have been looking for similarities of heat shock motifs, promoter regions (*E. coli*), entry sites for the gene dosage complex (*Drosophila*), promoter similarities between man and chimp as well as property-conserved regulatory elements in several mammalian genomes. The results of the latter experiments confirm that our system is able to spot property-dependent similarities where letter code-based systems fail. Further, we continue to investigating genomic elements, which will particularly result in alternate property-related classification systems (currently: *E.coli* promoter). On the route to applying signal theory to problems in systems biology we recently finished the modeling of a fair part of the peroxisome from yeast following principles of OOD. This represents a first step to setting up a framework for applying signal theory to the modeling of infection processes/dynamics.

Products

"FeatureScan" is available as a web service at http://genome.helmholtz-hzi.de/featurescan

Characterisation of the Immunoglobulin Heavy Chain Locus of the Mouse

Project leader: W. Müller **HZI-Braunschweig**

Summary of Scientific Results

Aim of the project

To elucidate the molecular base of antibody formation, the sequence providing the genes for the immunoglobulin heavy chain of the 129/Sv mouse strain was elucidated and analyzed. As a tool for the analysis, a dynamic V gene database was developed.

Major results (2001-2006)

The database VBASE2 is available at http://www.vbase2.org. VBASE2 is an integrative V gene database, combining the V gene sequence resources from the EMBL database, Ensembl, VBASE, IMGT and KABAT. The VBASE2 dataset is generated automatically and is regularly updated.

The 3 Mb genomic sequence of the 129/Sv *Igh* locus was assembled and annotated (EMBL-AccNo. AJ851868, AJ972403, AJ972404, AM041147, AM084333). The annotation implies functional and non-functional V, D, J and C genes as well as additional elements of the non-coding sequence. As the murine *Igh* locus is known to show distinct polymorphism, further sequence analyses focused on the comparison of the *Igh* locus of the mouse strains 129/Sv, C57BL/6 and BALB/c. As the strains 129/Sv and BALB/c were previously shown to share the same *Igh-V* haplotype it was an interesting new finding that there are considerable differences in the D_H region of these strains. For 129/Sv and C57BL/6, representing the Igh^a and Igh^b haplotype, respectively, an assignment of V_H alleles was performed. Further comparison of these haplotypes gained new insights into the evolution of the immunoglobulin genes. A phylogenetic analysis of the V gene family V_H7183 showed that expansion of functional V genes seems to occur by duplication of larger sequence blocks, thereby propagating adjacent pseudogenes. As a similar duplication pattern was previously observed for the human *Igh-V* lambda region it seems to represent a general mechanism for the evolution of large immunoglobulin loci, explaining the existence of a high number of pseudogenes by a positive selection for the duplication of the adjacent functional genes (Retter *et al*., in preparation).

Other products

VBASE2 database: http://www.vbase2.org
EMBL entries : AJ851868, AJ972403, AJ972404, AM041147, AM084333

Modelling of Metabolic and Genetic Networks of *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*

Project leader: A. Zeng **HZI-Braunschweig**

Summary of Scientific Results

Aim of the project

Development of methods for the reconstruction, analysis and modeling of genome-based metabolic and regulatory networks of organisms, especially for the selected pathogens

Major results (2001-2006)

- A new algorithm (IdentiCS) for fast identifying protein coding sequences and reconstructing metabolic networks directly from unannotated genome sequences. This method was verified with different genome sequences datasets of *K. pneumoniae* and applied to other organisms.
- We extensively revised the KEGG gene-enzyme-reaction database. Based on this database metabolic networks of >140 organisms including *K. pneumoniae* and *P. aeruginosa* were *in silico* reconstructed and analyzed in terms of network topology, key enzymes and metabolites, metabolic capacity and evolution.
- A new method for the decomposition of metabolic networks into functional modules.
- A comparative and phylogenetic genomic analysis of the synthesis and signal transduction pathways for autoinducer-2 in *P. aeruginosa* and other 138 sequenced organisms was conducted.
- The metabolic and regulatory networks involved in the formation of virulence factors and stress responses of *P. aeruginosa* under iron deprivation have been studied both *in silico* and experimentally. This led to the identification of PA2384 as a new regulator involved in large-scale regulation of iron starvation and quorum sensing in *P. aeruginosa*.
- A new method is developed to reveal the regulation hierarchy and functional modules in the transcriptional regulatory network of *E. coli*. The method is being extended to *P. aeruginosa***.**
- A new, second triose phosphate isomerase was discovered in *K. pneumoniae* which was bioinformatically and experimentally studied comprehensively.
- To infer functional linkages between genes from expression data, we presented a new method, termed as trend correlation (TC), realized by calculating a maximal local alignment of change trend score by dynamic programming and a change trend correlation coefficient between the maximal matched change.

Integration of Metabolic and Regulatory Network Models to Describe Key Cellular Processes in *Pseudomonads*

Project leader: V.A.P. Martins dos Santos **HZI-Braunschweig**

Summary of Scientific Results

Aim of the project

To elucidate, by developing appropriate mathematical descriptions, relevant aspects of the interplay between cell metabolism and key regulatory networks in *P. putida* and *P. aeruginosa*.

Major results (2002-2006)

- We developed a computational platform to retrieve heterogeneous data published under different formats, representing new information using a proprietary ontology and integrating it in the data repository. The platform is a modular system where independently developed tools are plugged-in.
- Based on the platform developed, in-silico genome-wide, constraint-based metabolic models describing genotype-phenotype relationships for *P. putida* KT2440 and *P. aeruginosa* strain PAO1 have been developed. A preliminary comparison of the metabolic space of the central metabolism of these bacteria revealed a higher flexibility and redundancy of the central metabolism of PAO1 as compared to KT2440, which was not foreseeable from comparison of the respective gene lists.
- A detailed kinetic model for the TOL plasmid (including three regulatory loops) was developed and integrated with the constraint-based, genome-wide metabolic model of *P. putida* described above.
- By genome-wide comparative analyses of pseudomonads, we generated a list of 345 genes potentially involved in mammalian pathogenesis. The involvement of these genes in is now being tested in mammalian models by large-scale expression profiling.
- Based the analysis of nucleotide composition of 16rRNA of 280 prokaryotes, we found that the uracil content in 16rRNA of thermo- and psychrophilic prokaryotes inversely correlates with the optimal growth temperature, providing thereby a simple method for the inference of the temperature settings for the cultivation of uncultured prokaryotes. We show as well that genome shrinkage during reductive evolution in prokaryotes decays exponentially and provide a method to predict the extent of this decay on an evolutionary time-scale. We validated predictions by comparison with estimated extents of genome reduction known to have occurred in mitochondria and *B. aphidicola*, through comparative genomics and by drawing on available fossil evidence.

Analysis and visualisation tools for cell biology and functional genomics

Project leader: U. Kärst/J. Wehland **HZI-Braunschweig**

Summary of Scientific Results

Aim of the project: The intention of this project was the development of visualization and database tools to promote the analysis of data from cell biology and proteomics experiments focused on the eukaryotic cytoskeleton and host-pathogen interactions with *Listeria monocytogenes* as the model system for human pathogens.

Major results (2001-2006): Several software tools for data visualization and analysis have been developed. ActMoST collects and integrates available data on cytoskeletal rearrangements and visualizes the dynamic interactions of the participating protein components such as actin treadmilling, activation and branching permitting the building of dynamic models. LEGER, LaneSpector, VIS-O-Bac, and Mineblast together form a package dedicated to data organization and integration, analysis and visualization from post-genomic investigations in infection biology. LEGER is the first post-genome database for *Listeria* research, offering sequence and expression analyses for functional genome studies. LEGER also integrates VIS-O-BAC and Mineblast for expression data visualization and literature mining based on protein sequencing, respectively.

Future work: We develop a database and software tool for quantitative phosphoproteome analyses by mass spectrometry in cooperation with FH Wolfenbüttel and SOAP-based database networks together with TU Braunschweig.

Bioinformatic Modelling of the Interaction of Genomes

Project leader: H.Saxel **BIOBASE GmbH**

Summary of Scientific Results

Aim of the project

The aim of the project is to model the genome-driven molecular interactions between pathogens and host organisms (animal or plant cells/organisms) during infection.

Major results (2001-2006)

The content of the BIOBASE databases TRANSFAC®, TRANSCompel, and TRANSPATH® has been highly extended by manual curation of full-text publications and automatic data import from the HumanPSD® database. The emerging broad data basis helps to understand the reaction of eukaryotic cells to pathogenic infection against the background of healthy physiological processes. The integration of database structures and data of the originally separated databases improved the efficiency of curation and enabled the construction of comprehensive regulatory networks involving signal transduction and gene regulation processes.

Bioinformatic tools (PathwayBuilder, ArrayAnalyzer) have been developed further to visualize networks and to be able to use TRANSPATH® data for interpreting the results of microarray experiments. The ArrayAnalyzer allows identifying key molecules, important subnetworks (cluster) or functional groups (associated with pathway, disease…) for a given input list of genes or proteins.

To perform searches for potential binding sites for immune response-related transcription factors, we have constructed an immune cell-specific and a cell cycle-specific profile that can be applied in conjunction with the analysis tools MATCH and Composite Module Analyst (CMA).

The analysis capabilities were applied on relevant microarray data and the results published as proof of principle.

Free-of-charge access for users from non-profit entities has been granted at http://www.gene-regulation.com.