Intelligent Bioinformatics Systems (H0900 / B080)

Head: Dr. Roland Eils

Scientists
Dr. Benedikt Brors ( - 6/03)
Martin Bentele (IWR) (10/02 - 10/03)
Alia Bulashevskya (4/02 - 7/03)
Svetlana Bulashevskya (1/00 - 12/03)
Christian Conrad (7/00 - 6/04)
Prof. Dr. Werner Dubitzky (1/00 - 12/01)
Jürgen Eils (10/02 - 12/03)
Daniel Gerlich (1/00 - 6/02)
Dr. Martin Granzow (8/00 - 6/02)
Patrick Herde (11/02 - 6/04)
Dr. Thomas Kochmann (1/00 - 6/03)
Dr. Rainer König (2/00 - 6/03)
Dr. Olga Krebs (1/00 - 12/03)
Christian Conrad (7/00 - 6/04)
Dr. Kurt Sätzler (9/00 - 7/01)
Falk Schubert (4/01 - 2/04)
Dr. Wolfgang Tvarusko (1/02 - 12/03)
Dr. Dietmar Volz (5/01 - 4/03)
Marco Weismüller (2/00 - 12/03)
Dr. Jan Wiemer (1/02 - 9/02)

Guest Scientists
Dr. Vijayraghavan Sundararajan (8/01 - 10/02, India)

Postgraduate Students
Chaitanya Athale (5/00 - 4/03)
Daniel Berrar (7/00 - 1/02)
Gaëlle Dubios (10/01 - )
Juntao Gao (11/01 - 9/03)
Matthias Gebhard (5/00 - 4/03)
Joshua Allen Moore (9/02 - 3/03)
Ming Xu (11/02 - )

Graduate Students
Martin Eigel (5/02 - 3/03)
Armin Groll (3/01 - 7/02)
Jasmin Müller (11/01 - 12/02)

Technical Assistants
Johannes Fieres (10/01 - 9/02)
Karheinz Groß
Rolf Kabbe (10/01 - 12/03)
Kaan Saracoglu (1/00 - 12/01)

Students, Apprentices
Thomas Butler (1/02 - 3/02)
Bernd Kalbfuß (10/01 - 7/02)
Bernhard Tausch (11/02 - 8/03)
Luigi Vitulli (3/01 - 1/11)
Corinne Wacker (11/02 - 2/03)
Wolfgang Weiss (11/01 - 10/02)
Sven Wichter (9/02 - 11/02)

Secretaries
Christina Grosch (8/01 - 6/03)
Brigitte Pétillon (1/00 - 08/01)

The group “Intelligent Bioinformatics Systems” works on the development of technologies in bioinformatics and computational biology for the interpretation of complex data generated by analytic processes in molecular genetics and cell biology. In the recent years an increasing number of high throughput screening systems have been developed in molecular and cellular biology. While in the past decade most such techniques were devoted to sequencing the genome of humans and other organisms, more recently such techniques emerged to relating genomic structure to function. One example is the DNA chip technology, which allows the screening of the expression level of all genes in a given organism under different conditions in one experiment. With this technology the study of the genetic effects of certain drugs has become possible thus paving the way for optimised design of new drugs and therapies. High throughput screening techniques generate a huge amount of data, which is difficult or even impossible to analyse without computer assistance. The development of computer assisted methods for the analysis of complex data in molecular biology as well as the development of models and computational methods in cell biology is at the core of this division established in 2000 at the dkfz. The research areas of major interest include

- Integrated modelling of heterogeneous genomic and clinical data
- Knowledge based data mining in large genomic and clinical databases
- Highly automated computational methods for the interpretation of complex cell biological data
- Modelling and simulation of cellular processes
- Computer vision and multidimensional image analysis

The central achievements of the group include several patented software systems for knowledge based data mining in life sciences, for fully automated diagnostics in molecular cytogenetics and for the study of dynamic processes in living cells. The group is actively collaborating with several corporate partners to market these systems.

Research grants:
1998-2001 Deutsch-Israelische Gesellschaft (G.I.F. no. 6-491-112.13/96; approx. 30.000 €)
1998-2003 Deutsche Forschungsgemeinschaft (Ei358/2-1 and FOR 240; approx. 240.000 €)
2002-2002 Deutsches Humangenomprojekt (GF KW 01619098; approx. 200.000 €)
1999-2004 BMWF BioFuture (BEO/31/11880; approx. 1.2 Million €)
2002-2003 Deutsches Humangenomprojekt (GF KW 01619098; approx. 300.000 €)
2002-2003 Deutsches Humangenomprojekt (GF KW 01619098; approx. 300.000 €)
2001-2004 Nationales Genom-Forschungsnetzwerk, BMBF (approx. 1.5 Millionen €)
2000-2002 Deutsche Forschungsgemeinschaft (Ei358/1-1; approx. 125.000 €)
2001-2004 Nationales Genom-Forschungsnetzwerk, BMBF (approx. 1.5 Millionen €)
2001-2002 BMWF/EEF-Projekt (approx. 210.000 €)
2002-2005 Landesforschungsschwerpunkt Baden-Württemberg (approx. 200.000 €)
Data Mining of Molecular Genetic Data

B. Brors (head of project), G. Dubois, T. Kochmann, J. Müller, F. Schubert, J. Moore, A. Bulashevska, D. Berrar, W. Dubitzky, M. Granzow, J. Wiemer

Collaboration partners: Prof. Dr. T. Chakraborty, Institut für Medizinische Mikrobiologie, Universität Giessen; PD Dr. Torsten Haferlach, Leukämie-Diagnostiklabor, Universitätsklinikum Großhadern, München; Dr. Stefan Joos, Abt. Molekulare Genetik, DKFZ; Dr. Christoph Klein, Institut für Immunologie, Ludwig-Maximilians-Universität München; Prof. Dr. Lichter, Abt. Molekular Genetik, DKFZ; Prof. Dr. Manfred Löhr, Innere Medizin, Universitätsklinikum Mannheim; Prof. Dr. Christoph Niehrs, Abt. Molekulare Embryologie, DKFZ; Prof. Walter Pyerin, Abt. Biochemische Zellphysiologie, DKFZ; Prof. Dr. Guido Sauter, Universität Basel, Switzerland; Dr. Alexander Schramm, Pädiatrische Onkologie, Tumorzentrum, Universität Essen; phase-IT intelligent solutions AG, Heidelberg; Boehringer Ingelheim Austria

Modern molecular methods in biology and medicine are producing an unprecedented amount of data which are witnessing manual analysis by their mere volume. Computerized methods have so far mainly been developed for sequence data, while data from DNA chips, proteomic analyses by 2D gel electrophoresis or mass spectrometry, comparative genomic hybridization (CGH), multicolor fluorescent in-situ hybridization (FISH), loss-of-heterozygosity analysis (LOH), and single nucleotide polymorphism (SNP) analysis still present a major challenge to computational biology. In our project, we are adapting methods from machine learning to these data and are developing new methods for their analysis. In particular, we are seeking methods to find associations between such molecular data and clinical information like tumor stage or grade, disease progression, degree of malignancy or responsiveness to chemotherapy.

For this kind of analysis, supervised methods are best suited. These are statistical models whose parameters are fitted by training with data of known class membership. Once trained, these models are able to assign classes to new samples investigated by one of the above mentioned molecular methods. For biomedical research, such methods can be immediately applied in two areas. First, diagnostic systems can be constructed that, based on molecular data, are able to predict important diagnostic or prognostic subgroups. Second, the set of variables chosen by the classifier may be important to reveal part of the cellular pathomechanism.

While a number of classification methods (e.g. linear discriminant analysis, k-nearest neighbors, support vector machines, or neural networks) have already been applied to molecular genetic data, all have to be adapted to the specific needs posed by these data. Usually, data sets suffer from the so-called “curse of dimensionality”, i.e. there are many more variables than cases under investigation. In gene expression analysis by DNA microarrays, for example, typical studies include less than 100 samples, but more than 10,000 variables (gene expression values). In this situation, it is easy to run into “overfitting”, i.e. the statistical model is perfectly adapted to training data but unable to deal with any new sample. Thus, validation becomes indispensable to estimate the generalization error, i.e. the misclassification rate on new data. Furthermore, it is desirable that any classifier makes transparent which of the variables investigated are important for the classification. This relates to the question of which important cellular processes influence the grouping into classes.

In our group, we have developed a classification based on aggregated decision tree classifiers [2]. This method is capable of dealing with high-dimensional data and yields transparent classifiers that show clearly the most important features (variables) for the classification analysis. It has been applied to classify gene expression data from three different subgroups of acute myeloid leukemia. The generalization properties on these data were excellent. One alternative approach was to apply variable selection first, and combine this with a neural network classifier [1]. This method performed very well on gene expression data from a breast cancer model, on proteomics data, and on CGH data from single metastatic cells (O. Schmidt-Kittler & C. Klein, unpublished data).

Other topics explored in our group include normalization and quality assessment of DNA microarray data, analysis of time series data in molecular genetics, and hypothesis-generating systems that are able to find interesting and highly significant molecular patterns in large ensembles of molecular databases [2-10].

Data Management and Analysis for Gene Expression Array

O. Krebs, R. Kabbe, K. Gross, J. Eils, P. Herde

Collaboration partners: Prof. Dr. Peter Lichter, Div. Molecular Genetics, DKFZ; Prof. Dr. Annemarie Pousetka, Abteilung Molekulare Genomanalyse, DKFZ; Prof. Dr. rer. nat. Sándor Suhai, Abteilung Molekulare Biophysik, DKFZ; Prof. Dr. rer. nat. Christof Niehrs, Abteilung Molekulare Embryologie, DKFZ; Prof. Dr. Werner Mewes, Institute for Bioinformatics/MIPS, GSF, Neuherberg; Prof. Dr. Alexander Schramm, Pädiatrische Onkologie, Universität Essen; Prof. Dr. Christoph Klein, Institut für Immunologie, Ludwig-Maximilians-Universität München; O. Krebs, R. Kabbe, K. Gross, J. Eils, P. Herde

Using microarray technology, research groups at DKFZ and other partner institutions produce a huge amount of data that has to be stored and managed. To address these needs we have designed and implemented an array informatics system which integrates data management and analysis. This system is intended to support and integrate RNA expression data with other kinds of functional genomics data [12-14].

Its functionality ranges from the storage of the data in relational data base management systems (Oracle RDBMS running on Unix) and Data Warehouse to frontend tools for the presentation and maintenance of the data. We have developed an Oracle database schema iCHIP which stores information including a detailed description of the biological samples and clones, image data, hybridization conditions, experiment data , gene expression values (raw data), quality indicators, and descriptive information. The Data Warehouse gathers and transforms consistent data obtained from a variety of sources and stores it in a separate server; this can be seen as computing a materialized view (subject to delays in propagating source updates) to simplify their subse-

quent analysis. It can provide easy access to the data on SQL-level for online analytical processing (OLAP) and datamining tools.

Initially the fully annotated experimental data is stored in iCHIP, CGH (Comparative Genomic Hybridisation) and Gene Annotation databases. During the upload process the data is combined with other sources and assembled in the staging area. From there the necessary information is uploaded in the warehouse and is now available for the analysis tools, deployed on an innovative integration platform "mine-it". This platform offers a quite general framework for the systematic processing of data and is especially valuable for high-throughput data-mining tasks.

A fundamental concept of the main data structure is the notion of underlying data types that model the data entities and their behavior. The data types support among other things conversions, defaults, imputations and determine the mathematical functionality of the data like distance measures, ordering, basic math operations. The set of data types is extendible, making it easy to develop custom data types suiting special needs (e.g. time series, composite data types). Driven by the data type objects, the tabular data structure itself provides much functionality, like basic statistics, histograms, automatic data type detection, or missing value imputation.

Results and discovered knowledge are stored back in the warehouse. The data in the original database remain untouched.

An application server is utilized to provide a Web interface for researchers to display their data. A commercial middleware product, Oracle Application Server, is used for interactive queries of several internal and public databases.

Links to services
http://www.dkfz-heidelberg.de/ibios/services

Biomedical Computer Vision Using Geometrical and Dynamical Models

J. Mattes (head of project), M. Gebhard, J. Gao, J. Fieres


The project focuses on the comparison of spatial data sets as it is required, for instance, for the analysis of movement and deformations in spatial data sequences. On the one hand, we study the quantification, visualization and motion correction in dynamic processes using methods of computer vision and we apply statistical methods to evaluate the quantified data. All steps are investigated, from image segmentation, rigid and non-rigid registration, graphical visualization up to the modelling of surface dynamics.

On the other hand, in cases where the different data sets do not correspond to temporal sequences but to a family of objects of same type the preservation of anatomical or topological features is studied and the differences are characterized statistically. We are working on several application-specific projects with a focus on the dynamics of the cell nucleus and the preservation of its architecture but covering also the study of cell membrane movement and applications in metallurgy. A software platform has been developed integrating the registration, quantification, motion compensation and visualization modules [15-17].

Modern imaging devices make it possible to study in vivo the mechanisms of dynamics in living systems. Intra-operative MRI devices, for instance, producing several images per second, offer physicians nowadays the possibility to follow the heart-cycle in real-time. Tumour growth can be supervised by imaging systems, which also guide the focus of a radiation source. In cell biology, fluorescent confocal microscopy permits to observe cellular structures such as chromosomes or the nuclear envelope over time.

To provide researchers of these different fields with new technologies for a detailed insight into the mechanisms of such dynamic processes we developed an image analysis system for quantification and visualization of motion and deformation. In particular, we are developing algorithms to reconstruct, to visualize, and to quantitatively analyse the movements of objects in spatial data series. We also investigate the conception of models, which allow the simulation of quantified motion.

Besides the development of image analysis concepts, we are also specialising our methods to specific applications. The focus lies on the investigation of the dynamics of the cell nucleus [18] and the preservation of its architecture during its replication over several generations. Another project is to characterize the movement of cell membranes and to distinguish different kinds of movement.

Methods:

The development of techniques currently focuses on the following topics (in key words):

- Object segmentation using deformable models [19,23]: active contours, snake splines, gradient vector flow field, tensor product B-spline surfaces, NURBS surfaces, level sets, optimization techniques, active shape models, surface reconstruction

- Object tracking and visualization of the trajectories [24]: fuzzy logic, cubic B-splines, spatial-temporal reconstruction, motion analysis

- Rigid and non-rigid image registration, for the comparison of spatial data sets. In particular, we estimate and analyse motion and deformation in subsequent image (data) stacks, in order to quantify and visualize movement [17,20] and to assess the topological and geometrical similarity of spatial configurations: thin-plate splines, optimization techniques, 3D-graphics and visualization, geometric algorithms (nearest neighbor), similarity measures, clustering, landmarks

- Statistical analysis for the concise description of quantified values [18,15]: density based cluster analysis, confinement tree analysis, principal component analysis
An Integrative Toolbox for the Cell Nucleus on a Portal Basis Using Grid-Resources

M. Bentele, A. Groll, V. Sundarajan, W. Tvarusko, M. Xu

Collaboration partners: Vincent Breton, LPC/CNRS.; Chris Cooper, IBM.; Dominique Haussler, University Hospital of Geneva, Switzerland; Chris Jones, CERN.; Guy Londsdale, NEC.; Isabelle Magnin, CREATIS/CNRS; Marcel Sobermann, CNRS.; Hernandez Vicente, UPV.; Francis Wray; EPCC.

Software tools for the quantitative data analysis and computer graphical visualization should be combined with modelling and simulation applications.

Software tools for the quantitative data analysis and computer graphical visualization should be combined with modelling and simulation applications. Applications will be supported by a private grid in DKFZ. The integrative presentation of the web-based approach addresses researchers investigating dynamic problems in the cell nucleus.

A Portal Framework:
The traditional model of software applications will be replaced by a new modern, web-oriented processing-model. This approach allows the integration of tools running on diverse platforms in the LAN. In the long term other groups should have the possibility to provide their tools to this single-point of access. On the other hand the integrated applications will be present on desktops of registered users only using a standard web-browser independent of their geographical region. The maintenance of the tools will be more efficient. New updated versions of the tools can be integrated fast and made immediately accessible to biologists as a user group.

Grid Resources:
In addition necessary processing power can be made available centrally by a grid-computing infrastructure. The accurate simulation of complex biological processes can require high spatial and temporal resolution in order to resolve fine-scale detail. The coupling of distributed resources into an integrated private grid can be used in such situations to overcome resolution barriers and hence to obtain qualitatively new scientific results.

We are currently setting up a test bed with the globus toolkit to run a "test" application in the field of simulation of transduction networks. This application will be adapted to make use of grid-services provided by the globus toolkit. The application needs not be entirely rewritten before it can operate in such an environment: Services will be introduced incrementally, with functionality increasing at each step. On the other hand we will extend the test bed and supporting IT-infrastructure step by step. In the midterm the resources provided by the grid will be available to support applications being accessed through the portal.

In the long term the iBioS group is involved in a European wide network called HealthVenture (HEAVEN). This network contributes to the building of an environment where information at five levels (molecule, cell, tissue, individual, population) can be associated to provide individualised healthcare. This environment will lead to the advancement of ambient intelligence and medical knowledge leading to a new generation of eHealth systems assisting in the individualisation of disease prevention, diagnosis and treatment. The grid is central to the realisation of this vision. The iBioS group will participate with several european-wide grid-enabled pilot applications and, through that early use, contribute to the development, usability and wide applicability of such applications.

Dynamics and Constitution of the Cell Nucleus: Modeling and Image Processing I) Transport Phenomena in Interphase Nuclei

C. Athale, M. Eigel, D. Volz


The principles governing nuclear architecture have been recently approached in the group by studying the transport-dynamics of fluorescent molecules in the nucleus. The most recent modeling approaches have shown that the "well-mixed" compartmental assumption fits the re-
spective quantitative dynamics data, but contradicts long
held models of nuclear organization. In order to overcome
this contradiction, we try to improve these models e.g. by
taking into account anisotropy effects. Also the use of im-
proved models based on partial differential equations in-
cluding the influence of the cell boundary are investigated.
These endeavors together yield better predictability of the
dynamics of respective molecules inside the cell nucleus
and, concerning the inverse problem, the composition of
interphase nuclei.

Therefore, data with a higher spatial resolution is used.
Our approach [25] yields two problem settings which can
be described as follows:

a) The forward problem. Based on biological knowledge
about chromatin distribution in interphase nuclei, a few
scenarios for recovery from FRAP are being simulated
and compared to experimental data. Approximately, the
mathematical model that triggers the recovery is given by
a diffusion equation. In this approach, this equation is
handled by a tool called ‘UG’ which has been developed
by the Interdisciplinary Center for Scientific Computing
(IWR) in Heidelberg. UG effectively solves transport equa-
tions numerically in a multigrid and (optionally) parallel
setting. Applying UG, models of transport phenomena in
the cell nucleus also incorporate boundary conditions,
which are not treatable by pure analytical equation solving.
The improvement (refinement) of data acquisition by new
experimental techniques therefore faces appropriate data
evaluation using improved mathematical models.

b) The inverse problem. In this approach, the diffusion
dynamics of intra-nucleus molecules has been investigat-
ged using bleaching data of high resolution in order to
estimate an appropriate diffusion tensor. The inverse prob-
lem of parameter fitting has been solved using the funda-
mental solution of a time-dependent diffusion equation
within MATLAB. The solution of an inverse problem pro-
vides for parameters describing the composition of the
medium and therefore marks a first step towards diffusion
tomography of the cell nucleus.

II) Dynamics of Nuclear Architecture Through
Mitosis and Chromosome Topology in
Interphase

J. Gao, M. Gebhard, D. Gerlich, J. Mattes,
K. Saracoglu, W. Tvarusko, M. Xu

www.embl-heidelberg.de/ExternAllInfo/ellenberg/; Peter Lichter,
DKFZ, http://www.dkfz-heidelberg.de/komp_genome/
Start_e.html; David Spector, Cold Spring Harbour Laboratories,
NY, USA. http://www.cshl.org/labs/spector/index.html; Tom
Misteli, NCI, MD, USA. http://rex.nci.nih.gov/RESEARCH/basic/
lrbge/cbge.html; Kevin Sullivan, Scripps Research Institute, CA,
USA. http://www.scripps.edu/cb/sullivan/

One of the most revolutionary recent technical develop-
ments in cell biology was the cloning of the green fluores-
cent protein (GFP), enabling visualization of virtually any
structure in living cells. Two dimensional time-lapse confo-
cal live cell recordings are widely used for the investigation
of dynamic fluorescently labelled structures in living cells
with high spatial resolution. A crucial step for the inter-
pretation of dynamic imaging data was the development of
computational methods for the automated quantitative
analysis and time-space visualization [24]. However, a de-
tailed understanding of complex dynamic processes such
as disassembly and reassembly of the nucleus through
mitosis can only be achieved when microscopy is carried
out in three spatial dimensions over time (4-D imaging),
which requires advanced imaging approaches. For this
project, novel techniques have been developed for the
fully automated quantitative analysis and visualization of
surfaces from dynamic three-dimensional fluorescent
structures in living cells [16-24,26] and applied to investi-
gate the breakdown and reassembly of nuclear architec-
ture during mitosis [18,21,22,26].

Multi-colour Fluorescence in Situ
Hybridisation (M-FISH) is
a combinatorial colour labelling technique allowing the rec-
ognition and distinction of different chromosome territories
(CTs) in nuclei of fixed cells [36]. This technique makes it
possible to identify structural and quantitative aberrations
in cancer cells. In order to provide a highly precise distinc-
tion of the different territories, a sensitive colour segmen-
tation is necessary. In this project, an automatic colour
classification approach called goldFISH has been devel-
oped combining spatial information with the different direc-
tions in the colour space and allowing such a sensitive
colour segmentation [35,37]. Based on the segmented
CTs the preservation of the topological arrangement in dif-
frent cell nuclei can be investigated.

To compensate for low image quality typically inherent
in live cell video microscopy, a highly sensitive anisotropic
diffusion filtering is applied that only smooths in areas
with homogeneous image information without perturbing
overall morphology. After thresholding, interpolated 3-D
surface models are reconstructed from segmented slices
[23]. This approach has proven to deal particularly ad-
equate with the great anisotropy commonly inherent in 4-D
live cell recordings, due to low z-resolution. Interpolation
between 3-D surfaces over time then gives a continuous
reconstruction of the entire 4-D data set (morphing). Glo-
bal rotational and translational movements in a time series
are automatically corrected by a rigid registration ap-
proach. The animated surface reconstruction is embedded
in a powerful multi-dimensional graphical viewer that al-
lows real time user interaction. The binarised object repre-
sentation is used to quantitate object volume and fluores-
cence intensity over time [19,23].

These imaging modules were utilized to analyse nuclear
dynamics in interphase and mitotic cells. Especially, chro-
mosome dynamics throughout mitosis [27] and the reas-
sembly of the nuclear envelope [26] were investigated in
detail. Further applications of the developed imaging tech-
niques include the analysis of nuclear envelope break-
down [18,21,22], quantitative analysis of promylocylic body
motion [18], dynamics of nuclear speckles [24], localiza-
tion of methyl-DNA binding proteins [29], secretory vesicle
dynamics, and movements of labelled centromeres in liv-
ing cells.
Quantitative data obtained from live cell experiments can be used to compare with biological models. Therefore, mathematical models have to be generated, which allow quantitative testing of the biological hypothesis. 4-D imaging experiments of labelled subsets of chromosomes showed that interphase chromosome positions are inherited throughout mitosis, thereby possibly maintaining a specific pattern of gene expression [27]. To test different possibilities for a mechanism that maintains or re-establishes chromosome positional order, a mathematical model of mitotic chromosome movements was set up. Graphical visualization and quantification of chromosome configurations were used to compare simulation runs with experimental data. Based on perturbation experiments and such numerical simulations a mechanism is suggested that can stably transmit chromosomal positions from one cell generation to the next.

An approach based on FISH-Imaging to study the inheritance of chromosome positions throughout mitosis is to investigate the similarity of the chromosome positions in the daughter cells (and grand-daughter cells) with respect to arbitrary cells. A measure to quantify this similarity has been defined and is currently evaluated.

In this project, a set of tools has been designed that allows an automated quantitative analysis of multi-dimensional live cell and FISH imaging data. A wide range of applications addressing the dynamics of nuclear architecture has shown the invaluable use of these approaches in the field of cell biology.

**Modeling and Simulation of Signal Transduction**

M. Weismüller, G. Dubois

Collaboration partner: Biobase, Germany (http://www.biobase.de)

Signal transduction is one of the most important functions to cause abnormal cell behaviour. Therefore signal transduction is one focus of cancer research. The task is to model signal transduction networks in silico to gain more information about cell behaviour.

The approach is to model whole signal networks. As quantitative experimental data is available only for some well-studied signal pathways in form of concentration or activity rates, we try to do semi-qualitative modelling. One very comprehensive data source on qualitative signal transduction data is the signal transduction database TRANSPATH(R) [Schacherer, F. et al. Bioinformatics 17 (2001) 1053-1057] professional database. We import the TRANSPATH(R) data into a local signal transduction database. We are using this data to generate model descriptions, which are put into the Swarm system [http://www.swarm.org], a general purpose agent-based simulator implemented in Objective-C [Paton et al. BioSystems 50 (1999) 159-171].

**Modelling of Cancer Pathogenetic Processes and Genetic Regulatory Processes Based on Molecular Genetics Data.**

S. Bulashevska, B. Brors

Collaboration partners: Dr. Stefan Joos, Abt. Molekulare Genetik, DKFZ; Prof. Dr. Gyula Kovacs, Abt. Urologie und Poliklinik, Chirurgische Universitätsklinik Heidelberg

Recent advancements of molecular genetics increased the amount of experimental data allowing the researchers to gain insight into the complex function of normal or tumor genome. The development of cytogenetics has made possible the comprehensive detection of chromosomal abnormalities in the whole genome by means of one experiment. These methods include comparative genomic hybridization (CGH), various in situ hybridization techniques like fluorescence in situ hybridization (FISH) and multicolor fluorescence in situ hybridization (MFiSH), methods for detection of allelic instabilities like loss of heterozygosity (LOH). It was shown that chromosomal abnormalities are related to the initiation and progression of tumor. The challenge of our work was to reconstruct the possible flow of progression of genetic abnormalities on the long way from normal to malignant cell from the single “snapshot” of abnormalities provided by cytogenetic experiments.

We use a Bayesian network technology that enables to uncover and quantify the multivariate probabilistic dependencies between stochastic events. We applied the Bayesian Network learning to the CGH data of gastrointestinal tumors (GISTs) and to the allelic-loss data (LOH) of Papillary Urothelial Cancer (UC). Exploiting the mechanism of probabilistic reasoning in Bayesian Networks we could reveal primary and secondary genetic abnormalities and suggest the possible tumor pathogenic pathways. This “high-level” genetic information can give an insight into the underlying mechanisms of pathogenic gene regulation resulting in cancer formation. Our approach is another step on the way of mathematical modelling of tumorgenesis.

A part of this project was the implementation of a software tool for parsing the cytogenetic data which is conventionally notated in a special notation ISCN95 [30]. The concept of our ISCN Parser, briefly, is that it can interpret the ISCN as a formal language. We developed the grammar rules for the ISCN in Extended Backus-Naur Form (EBNF) (see [30])

The microarray technology allows to measure the expression levels of thousands of genes simultaneously, as they change over time and react to external stimuli. The great challenge of functional genomics is to uncover the mechanisms of gene regulations.

The main topic of our present research is the investigation of new probabilistic graphical models (especially boolean logic-based) and the development of learning algorithms for them in order to apply them for the reconstruction of the processes of genetic regulation.
Modeling of Metabolic Networks

R. König, S. Wichert

The topologies of metabolic networks are extracted from the publically available databases KEGG [Goto et al. Bioinformatics 14 (1998) 591-599] and Ecocyc (Karp et al., Nucl. Acids. Res., 30 (2002) 56-58). Ecocyc includes a sufficient comprehensive dataset for the Escherchia coli model organism. The network topology is adapted to construct a bi-partite graph with enzymes and metabolites as alternating vertices [31]. Publically available data of high throughput gene expression profiling experiments are taken (e.g., [Khodursky et al. PNAS 97 (2000) 12170-5]) to define distances within the graph. The topology of the net is investigated by a variety of multivariate data analysis tools. The network is observed under a variety of treatments such as hunger stress and tryptophan deficiency. As an example case, the signal transduction of MKP-1 on farnesyl transferase was studied [32].

Protein Folding

R. König, V. Sundararajan

Determining the structure of proteins is one of the most fascinating subjects in sturctural biology. One main goal is to discover the relationship between protein sequence, structure, stability and finally its function. Wie developed in silico simulation methods for the structural formation of proteins and polypeptides by using evolutionary based techniques such as genetic algorithms [33, König & Dandekar Biosystems 50 (1999) 17-25].

The method to obtain the structure of a protein based on two hypotheses:

a) The protein structure evolves by adding one or a few amino acids.

b) Initially the protein structure evolves quickly within segments, then it establishes a combined, global native structure.

Two strategies are followed to test these hypotheses:

a) Molecule evolution by adding amino acids, done with the genetic algorithm technique.

b) During the optimisation the molecule is decomposed into fragments. These fragments evolve by the use of genetic algorithms on several processes and are evaluated globally.

The main torsion angles of the molecule chains are taken as basis variables to calculate the energy terms.

Cellular Localisation

C. Conrad, V. Sundararajan

Collaboration partners: J. Ellenberg and Nathalie Daigle (EMBL); R. Pepperkok and Holger Erfle (EMBL); Thomas Lörch, MetaSystems, Germany

In living cell assays, proteins can be fused to spectral mutants of the green fluorescent protein (GFP) producing a fluorescent product when expressed in cells [Chalfie et al. Science 263 (1994) 802-805]. GFP-fusion proteins as well as fluorescence labeled antibodies can be used to monitor gene expression and protein localisation in living organisms, but screening of thousands of proteins is a labour-intensive manual screening approach and so far dependent on a manual evaluation by trained personal.

For this task a new software system for the detection and classification of different cellular structures by a machine learning approach was developed. After automatic detection of cells and particular image preprocessing, specific patterns of proteins in the cell are recognised by classifiers, which are trained by neural network learning approach [Mitchell MIT-Press (1997)]. The software automatically assigns the labelled proteins to functional classes with high recognition accuracies. For example, the functional classes can be related to cellular functional entities (such as organelles, cytoplasm, chromatin, or focal adhesion sites) or to functional states of cells (such as apoptosis) according to cellular morphology. Current target applications are cDNA cell arrays to enlarge the throughput of tested clones.

Publications and patents (* = external co-author)


