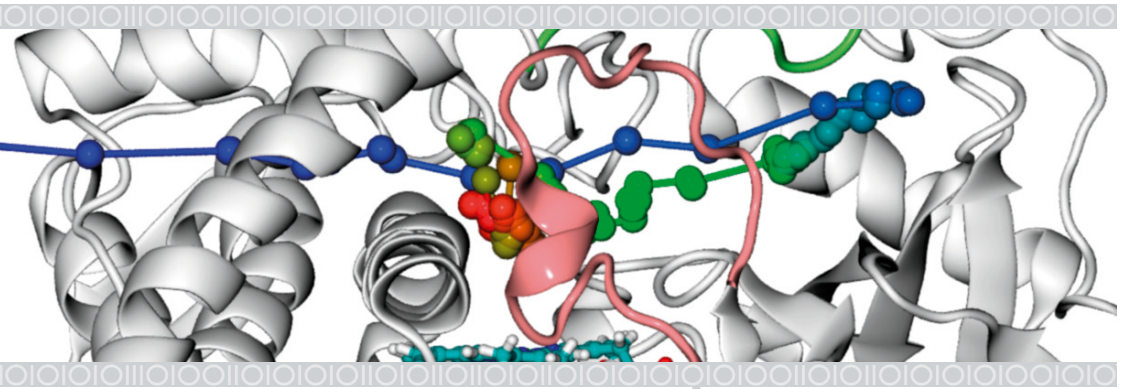
Annual Report



EML
Research

2006

Annual Report

EML Research gGmbH

2006

Edited by

EML Research gGmbH
Villa Bosch
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg
www.eml-r.villa-bosch.de

Our e-mail adresses have the
following structure:
Firstname.lastname@eml-research.de

Contact

Bärbel Mack
Phone: +49-6221-533 201
Fax: +49-6221-533 298

Editor

Dr. Peter Saueressig
Klaus Tschira Foundation
Public Relations
Phone: +49-6221-533 245
Fax: +49-6221-533 198

Layout and CD Design

Bernhard Vogel
Klaus Tschira Foundation

Pictures

EML Research gGmbH
(unless otherwise indicated)

ISSN 1438-4159

Table of contents

Andreas Reuter

2006 proved to be another very good year, both with respect to the fruits of our research and in terms of establishing EML Research as a brand name in the scientific community. This is another way of saying that we have managed to reconcile two rather different features typical of a scientific institution. One is participation in the ongoing pursuit of new ideas and the exploration of novel approaches - in short, the image that habitually comes to mind when the word "science" is mentioned. The other is much less prominent (and often overlooked). It centers around continuity and the provision of reliable services to the scientific community. We are particularly proud of having made two major services available to the research community in 2006, both of them connected with computational biology and related fields. One is the SABIO-RK reaction kinetics database made available to selected users as a beta release, the other the first full release of the Copasi simulator for complex pathways, developed in conjunction with the Virginia Biotech Institute (VBI). Together with the previously released software packages MMAX and PIPSA (the latter jointly developed with EMBL), this means that each of our research groups has now provided at least one service - and all these services are very popular in their respective communities.

In terms of research projects, work continued at a rapid pace. A number of projects were completed successfully, and at the same time new research contracts were concluded and ongoing projects continued. So EML Research grew quickly in 2006 – to the point that office space is becoming a serious issue. So far, we have managed to cope with all the problems arising in this respect, but if our groups maintain their current success rate we will soon need to come up with some new ideas about where to put our research staff members – and we certainly don't want to turn into a "virtual" institute. The new research contracts have come from the German Research Foundation (DFG), the European Union, and the Federal Ministry of Education and Research (BMBF).

The scientific results are as manifold as the projects they result from. Since many of them have drawn wide-spread interest from the scientific community, it would not be appropriate to single out one specific venture as a "highlight" of last year's work. But it is fair to say that the consistent level of high-quality research has put EML Research squarely on the map in at least two communities (a mere 3 years after its establishment): computational linguistics and systems biology. Members of the NLP group are actively involved in many of the key activities going on in computational linguistics at an international level, and the same is true of our three "bio-groups": BCB, MCM, and SDBV. It is particularly encouraging to see that these three groups, which started out from different positions and with different agendas, have now established very close working contacts and plan their new projects in such a way that each group's work can complement the research being done by the other two. These combined efforts have earned them a very good reputation, and this will be a sound basis on which to extend the agenda of EML Research in the life sciences.

In the past year, the rapidity with which new projects were initiated, the development of future activities orchestrated, workshops organized, etc, placed very high demands on our administrative staff, both technical and non-technical. Admittedly, we have almost come to take their efficient and timely support for granted, but that does not mean it can be overlooked. Making an organization run smoothly requires just as much dedication and professionalism as doing good research – and the research work would suffer quite considerably without that support. So a big "thank you" is due to our admin people. Your share in the overall success of EML Research is greatly appreciated.

As usual, we also wish to express our gratitude to the Klaus Tschira Foundation for providing us with the financial platform (and the beautiful environment) for doing the research we are so enthusiastic about. We appreciate the freedom we have to set our own priorities in our scientific endeavors. Thanks are also due to our colleagues in many countries for

Scientific and Managing Director

Prof. Dr.-Ing. Dr. h.c.
Andreas Reuter
Tel.: +49-6221-533200
Fax: +49-6221-533298

Managing Partner

Dr. h.c. Klaus Tschira
Tel.: +49-6221-533101
Fax: +49-6221-533199

Public Relations

Dr. Peter Saueressig
(Klaus Tschira Foundation)

Office

Bärbel Mack
(Office Manager)
Kornelia Gorisch
(Administration)
Benedicta Frech
(Administration)

Controllers

Christina Bölk-Krosta
Ingrid Kräling

their help, support, and criticism – whatever is needed most. It is good (in fact, absolutely essential) to be part of such a network. We are glad that more and more visiting scientists find it worth their while to come and spend some time at EML Research. Their contributions are extremely valuable, and we owe a debt of gratitude to all of them.

Much as we would like to say that the institute as it presented itself in 2006 is a model of what it will be like in the foreseeable future, that will not be the case (whether one would care to add "fortunately" or "unfortunately" is a question of taste). Panta rhei. Change is ubiquitous, and nowhere more than in science. So there will be no room for complacency. EML Research will do its best to meet all the challenges it is faced with. So be sure to visit us again one year from now at the latest. You may be in for a surprise!

Fig. 1: (from left to right): Klaus Tschira, Ingrid Kräling, Andreas Reuter, Bärbel Mack, Kornelia Gorisch, Benedicta Frech, Peter Saueressig, Christina Bölk-Krosta

2.1 Bioinformatics and Computational Biochemistry (BCB)

The Bioinformatics and Computational Biochemistry group at EML was established in autumn 1998. The main focus of research in the group lies in the development and application of computational methods for analyzing, modeling, and simulating signaling, metabolic, and genetic networks in the living cell.

In 2006, three projects were completed: Simulation of Biochemical Pathways, COPASI (Complex Pathway Simulator) with full release to the public, and SYCAMORE, of which a prototype is to be released in early 2007. Project management for COPASI was taken over in 2006 by Sven Sahle. Work was continued on BioSim and UniNet. Katja Wegner finished her PhD thesis in the early part of the year, and Tim Johann joined us as a new PhD student in the autumn. Finally, Gerold Baier, visiting scientist since summer 2005, completed his fruitful and successful sojourn at EML Research.

In July 2006 the group organized the 2nd UniNet Workshop: Data, Networks, and Dynamics in Heidelberg.

Group Leader	**Research Associates**
Dr. Ursula Kummer Tel.: +49-6221-533225 Fax: +49-6221-533298	Ralph Gauges, Femke Mensonides, Dr. Ulla Rost, Dr. Sven Sahle, Dr. Natalia Simus, Dr. Iulian Stoleriu, Dr. Irina Surovtsova, Dr. Andreas Weidemann
	Doctoral Students
	Tim Johann (since September 2006), Jürgen Pahle, Katja Wegner (until March 2006)
	Students
	Tim Johann (January – September 2006), Artjom König (since December 2006), Anton Ruff (since December 2006), Xin Yu (March – October 2006)
	Visiting Scientist
	Prof. Gerold Baier (January - August 2006)

Fig. 2: The BCB group in 2006 (from left to right): Iulian Stoleriu, Femke Mensonides, Ralph Gauges, Jürgen Pahle, Sven Sahle, Irina Surovtsova, Ursula Kummer, Tim Johann, Andreas Weidemann, Natalia Simus, Ursula Rost

2.2 Scientific Databases and Visualization

The Scientific Databases and Visualization (SDBV) group focuses on the development of methods and tools to support researchers in the analysis and management of scientific data, more specifically biochemical data. The group addresses problems such as the integration of different data types and formats, the storage of these data, and methods for querying and navigating them. In May 2006 we released the first Beta Version of the SABIO-Reaction Kinetics (SABIO-RK) database (see Section 3.6, SABIO-RK). The Systems Biology community in particular has shown great interest in this new resource. Its popularity and the use made of it are growing continuously. In 2007 the SDBV group will participate with SABIO-RK in two new projects funded by the BMBF (Federal Ministry of Education and Research) as part of the Hepatosys Network and the SYSMO (Systems Biology for Micro-organisms) initiative.

Apart from the web-based user interface we have developed a set of web services that enable researchers to include queries to SABIO-RK as part of their research work-

Fig. 3: The SDBV group in 2006 (from left to right): Martin Golebiewski, Andreas Weidemann, Renate Kania, Saqib Mir, Jasmin Saric (back row), Olga Krebs, Isabel Rojas, Ulrike Wittig (front row)

flows. Our objective in the future is to provide semantic web services containing information about the meaning of the data exchanged. In this context we are working on the development of methods for bootstrapping biochemical databases, based on biochemical ontologies.
As before, another main research area of the group is the development of methods to support the process of database population and curation. Here we are mainly working on the detection of biochemical compounds by their names or structures (see Section 3.7 BioReader).

Group Leader	**Research Associates**
Dr. Isabel Rojas Tel.: +49-6221-533231 Fax: +49-6221-533298	Martin Golebiewski, Renate Kania, Dr. Olga Krebs, Dr.Jasmin Saric, Dr. Andreas Weidemann, Dr. Ulrike Wittig
	Doctoral Students
	Saqib Mir (KTF fellowship holder), Stefanie Anstein (KTF fellowship holder)
	Students
	Xiaon Chen

2.3 Natural Language Processing (NLP)

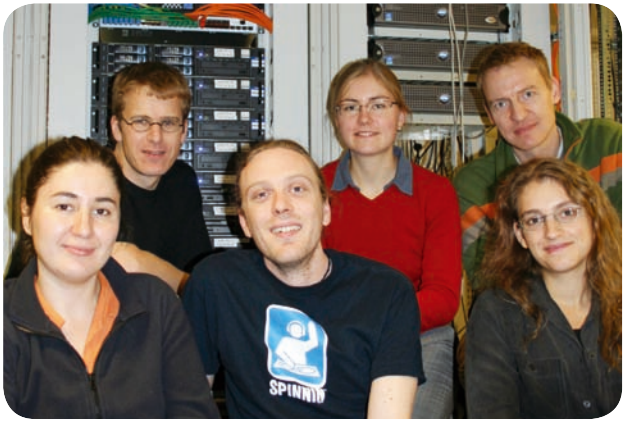


Fig. 4: The NLP group in 2006 (from left to right): Vivi Nastase, Simone Paolo Ponzetto, Margot Mieskes (front row), Michael Strube, Katja Filippova, Christoph Müller (back row)

2006 turned out to be a very successful year for the NLP group. First and foremost,we did some pioneering research in connection with the use of Wikipedia as a knowledge base for applications processing natural language. This work was very well received by reviewers and audiences at the HLT-NAACL and AAAI conferences [Ponzetto 2006a, Strube 2006]. In the future, this research may well qualify as a turning point in the rediscovery of knowledge-based NLP. Though we spent a lot of time and effort on completing this project to schedule , we also recorded at least two other highlights in the form of publications on pronoun resolution in spoken dialog at EACL [Müller 2006a] and discourse segmentation at EMNLP [Filippova 2006a].

Group Leader	Research Associates
Dr. Michael Strube Tel.: +49-6221-533243 Fax: +49-6221-533298	Margot Mieskes (until Oct. 2006), Christoph Müller, Dr. Vivi Nastase (since Sept. 2006)
	Klaus Tschira Foundation Scholarship Holders
	Katja Filippova, Margot Mieskes (since Nov. 2006), Simone Paolo Ponzetto
	Students
	Vanessa Doyle, Andrew Herd, Melissa Macellano, Violeta Sabutyte, Irina Schenk, Ganna Syrota, Grainne Toomey, Wolodja Wentland, Matthew White, Florian Winkelmeier

2.4 Molecular and Cellular Modeling (MCM)

Molecular recognition, binding, and catalysis are fundamental processes in cell function. The ability to understand how macromolecules interact with their binding partners and participate in complex cellular networks is crucial to prediction of macromolecular function and to applications such as protein engineering and structure-based drug design. The Molecular and Cellular Modeling (MCM) group develops and applies computational approaches to study the macromolecules of the cell: their structure, dynamics, interactions, and reactions. The central focus is on the interaction properties of proteins. An interdisciplinary approach is taken, entailing collaborations with experimentalists and a concerted use of informatics- and physics-based computational approaches. Techniques cover a wide spectrum from interactive, web-based visualization tools to atomic-detail molecular simulations.

Group Leader
Dr. Rebecca Wade Tel.: +49-6221-533247 Fax: +49-6221-533298
Research Assistants
Dr. Razif Gabdoulline, Dr. Stefan Henrich, Georgi Pachov, Dr. Matthias Stein, Dr. Peter Winn, Divita Garg
IT Specialist
Dr. Stefan Richter
Klaus Tschira Foundation Scholarship Holders
Dr. Vlad Cojocaru, Anna Feldman-Salit, Domantas Motiejunas
Fellowship holders
Dr. Outi Salo-Ahen, Sulaiman Faisal
Masters Student
Mai Zahran
Student Workers
Frederik Ferner, Matthias Janke
Guest Scientist
Prof. Arieh Ben-Naim

The research of the MCM group in 2006 is described in this report under five projects (see Sections 3.3, 3.8-11). We completed three externally supported projects this year:

- TASSFUN: Target-specific scoring functions. This project was carried out in collaboration with AstraZeneca, Sweden. COMBINE (COMparative BINding Energy) analysis was applied and developed to address the problem of designing drugs that are selective between related protein targets and to enable application in virtual screening (see Section 3.9).

- SYCAMORE: SYstems biology's Computational Analysis and MOdeling Research Environment (see Section 3.3). This collaboration between the BCB and MCM groups at EML Research is part of the Modeling Platform of the Federal Ministry of Education and Research's 'Hepatosys' Systems Biology program. The 3-year project finished at the end of 2006 and provides the basis for our new projects supported in the 'Hepatosys' programme starting in 2007. During the first funding period, a system of software tools and computational methods to support concerted computational and experimental approaches to systems biology problems has been developed. The MCM group worked on

Fig. 5: The MCM group in 2006 (from left to right): Sulaman Faisal, Georgi Pachov, Razif Gabdoulline, Domantas Motiejunas, Matthias Stein, Stefan Henrich, Stefan Richter, Rebecca Wade, Peter Winn, Anna Feldman-Salit, Vlad Cojocaru, Outi Salo-Ahen

the development and application of methods of harnessing protein structural information for systems biology projects. In particular, qPIPSA (quantitative Protein Interaction Property Similarity Analysis) was developed to estimate kinetic parameters from the molecular interaction fields of proteins.

• Optical Biosensors for Contaminant Monitoring (see Section 3.8). This NATO Collaborative Linkage project involving five groups in four countries focused on the engineering of haloalkane dehalogenases and their use for development of optical biosensors. Computational methods developed at EML Research were used to aid the design of enzyme mutants studied experimentally by the group of Jiri Damborsky (Brno, Czech Republic).

Three new externally supported projects were started during the year:

• Multiscale Approach to Biomolecular Interactions: From Molecular Dynamics to Brownian Dynamics Simulation of Chromatin Components supported by the German Research Foundation (DFG). This is a joint project with Jörg Langowski (German Cancer Research Center, Heidelberg) and Jeremy Smith (University of Heidelberg). The aim is to develop multiscale molecular simulation techniques and apply these to chromatin (see Section 3.11).

• Prosurf supported by the European Union. The project is coordinated by Stefano Corni and Elisa Molinari (University of Modena and CNR-INFM-S3, Italy) and also involves groups in Munich and Israel. The aim is to build a 'Computational Toolbox for Protein-Surface Docking.'

• LIGHTS (LIGands to interfere with Human TS) is also supported by the European Union. The project is coordinated by Maria Paola Costi (University of Modena, Italy) and involves six groups from five different countries. The aim is to design small ligands to interfere with thymidylate synthase dimer formation as new tools against resistance problems associated with anti-cancer drugs.

We welcomed two new research fellows this year:

- Dr. Outi Salo-Ahen, a postdoctoral fellow from Finland supported by the Alexander von Humboldt Foundation.

- Faisal Sulaiman, a predoctoral fellow from Pakistan supported by the German Academic Exchange Service (HEC-DAAD).

We were also fortunate to have Professor Arieh Ben-Naim (Department of Physical Chemistry, Hebrew University, Jerusalem, Israel) visiting the group as Guest Professor for several months. He is renowned for his research and books on molecular solvation phenomena. His stay in Heidelberg in the MCM group and in Jeremy Smith's group (University of Heidelberg) was supported by BIOMS.

Some of the other MCM activities are described in Section 5.

For more information on the group's activities, see: www.eml-research.de/english/research/mcm and http://projects.villa-bosch.de/mcm.

3.1 Simulating Biochemical Pathways

Simulating the Metabolism of Neutrophilic Leukocytes

Developing computational methods for the analysis and simulation of biochemical networks requires the application of those methods to real-world problems. It is also very rewarding to see the success of computational research in dealing with such problems. Accordingly, in this project we select interesting biochemical problems and improve our understanding of them by using both existing computational methods and new ones that we have devised ourselves.
Neutrophilic leukocytes are white blood cells that are important for the body's defense system, e.g. against bacteria. They detect invading bacteria, bear down on them, and kill them with oxygen radicals. This process has been subjected to detailed experimental observation by our collaborator, Howard Petty. Based on our previous results, which suggested that central glycolysis is crucially important for the observed metabolic oscillations - setting their frequency (whereas MPO sets the amplitude) we have now determined the likely biochemical origin of the switching between two very different metabolic states – activated vs unactivated . The origin appears to be the kinetic behavior of hexokinase in neutrophils. This specific enzyme is able to act as a glucose-sensing device and triggers a switch in metabolic behavior once glucose concentrations exceed a certain limit.
We are now awaiting experimental verification and preparing a corresponding publication.

Project Manager
Dr. Ursula Kummer
Project Members
Jürgen Pahle
Visitors
Jens Christian Brasen

Collaboration Partners

Prof. Dr. Lars F. Olsen, Jens Christian Brasen, Physical Biochemistry Group, University of Southern Denmark, Odense, Denmark (simulations)

Prof. Dr. Howard Petty, Dr. Andreij Kindzelskii, Wayne State University, Detroit, USA (experiments)

Dr. Ibrahim Coumbassa, University of Conakry, Guinea (simulations)

Sponsors

Klaus Tschira Foundation

European Science Foundation

Studying the Dynamics of Buffered Calcium Oscillations

Using the methodology and models previously studied in our group, we have investigated the effect of buffering (calcium-binding proteins) on the behavior of calcium ions in cells. This is important because calcium acts as a second messenger, and its specific dynamics carries information into the cell. The calcium concentration in the cytosol of hepatocytes may thus, for example, display complex dynamic behavior, including spiking and bursting oscillations and encoding different kinds of information.

We have studied the effects of stochastic fluctuations in models of calcium oscillations [Kummer (2000) Biophys. J. 79:1188] by comparing stochastic simulations [Gillespie (1976) J. Comp. Phys. 22:403] with the corresponding deterministic solutions. Stochastic effects like prolonged calcium bursts are more pronounced in models comprising low particle numbers, while the stochastic simulations show a transition to quasi-deterministic behavior as particle numbers increase. To facilitate the choice of the appropriate simulation method in our SYCAMORE software system (see below), it would be very valuable to be able to predict the particle numbers at which this transition actually occurs. We have established that this is highly dependent on properties of the system's attractor, such as the so-called divergence [Kummer (2005) Biophys. J. 89:1603].

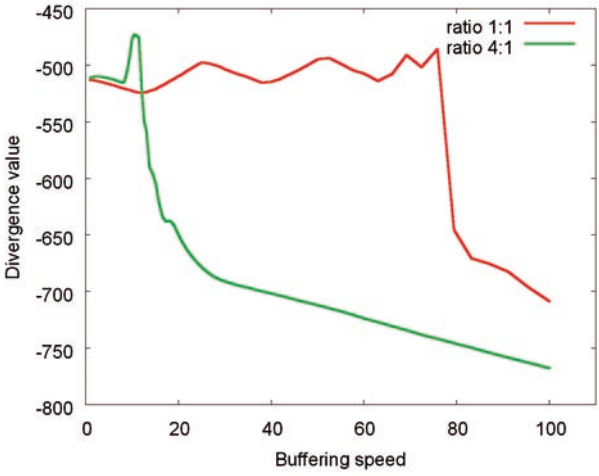


Fig. 6: Calculation of the divergence value (y-axis) for different protein buffering speeds (x-axis). The buffering speed is the sum of the calcium-binding and the calcium-release rates. The ratio of binding rate to release rate is 1:1 or 4:1, leading to about 50% or 80% of the total calcium bound to the buffer

Most of the calcium in cells (~80%) is bound to protein buffers. We included the buffering of calcium ions in the calcium oscillation models and studied how the dynamic behavior changes. Intuitively, one would expect buffering to dampen the stochastic effects. A mathematical indication for the correctness of this assumption would be that the local divergence of the system is decreased by buffering. The high frequency stochastic fluctuations are indeed diminished by the presence of the buffer. However, we have found that buffering can change the system dynamics in such a way that the divergence is not always decreased globally but can in fact even rise with increasing buffering speed (see Figure 6). This results in different behavior in stochastic and deterministic simulations (e.g. bursting oscillations and steady state, respectively) even if the particle numbers are large and the system is buffered.

3.2 COPASI

Project Manager

Dr. Sven Sahle

Project Members

Ralph Gauges, Artjom König, Dr. Ursula Kummer, Jürgen Pahle, Dr. Ulla Rost, Anton Ruff, Dr. Irina Surovtsova, Katja Wegner, Dr. Natalia Simus

COPASI, released for the first time as a test version in 2004, is a software package assembling various modeling, simulation, analysis, and visualization methods for biochemical pathways. It comes with a graphical user interface that allows users comparatively easy access to powerful simulation and analysis tools. The program is available for UNIX/Linux, Mac OS X, and Windows.

In 2006, we continued developing new features for COPASI and improving the existing ones. A full release was made available in spring 2006. Among the most significant new features are algorithms to calculate the divergence of the system and advanced sensitivity analyses and parameter estimation methods. In order to make the program more compatible, we have added a new export format – XPP. In addition, we have embarked on official collaboration with the CellDesigner team (Systems Biology Institute, Tokyo, Hiroaki Kitano) in order to integrate the tools in such a way that users can make seamless use of both tools.

COPASI is very closely associated with the SYCAMORE project, which involves the construction of models from different information sources. Accordingly, a number of methods for sensitivity analysis were added in 2006 that will mainly benefit users of SYCAMORE. In addition, a prototype version of a complexity reduction algorithm developed by our group has now been implemented.

COPASI is a joint project between our group and Pedro Mendes' research group at the Virginia Bioinformatics Institute in Blacksburg, USA. Test versions of COPASI have been published since 2004 and downloaded several thousand times. The latest version of COPASI can be downloaded at www.copasi.org.

Collaboration Partners

Dr. Pedro Mendes, Dr. Stefan Hoops, Virginia Bioinformatics Institute, Blacksburg, VA, USA

Prof. Hiroaki Kitano, Dr. Akira Funahashi, Systems Biology Institute, Tokyo, Japan

SBML Layout Extension

In previous years we developed an SBML Layout Extension to store and exchange important information for the graphical representation of biochemical systems in modeling and simulation software. SBML is now the worldwide recognized standard file exchange format for biochemical models.

In 2006, we completed the SBML Layout Extension and wrote implementations for reading SBML files with layout information and an XSLT stylesheet to convert layout information from SBML files into SVG drawings. A paper describing the layout extension has been published in Bioinformatics.

As the SBML layout extension can only provide the layout information for a reaction network graph, there is also great demand for ways to provide rendering information on the individual elements of the graph. Therefore we have cre-

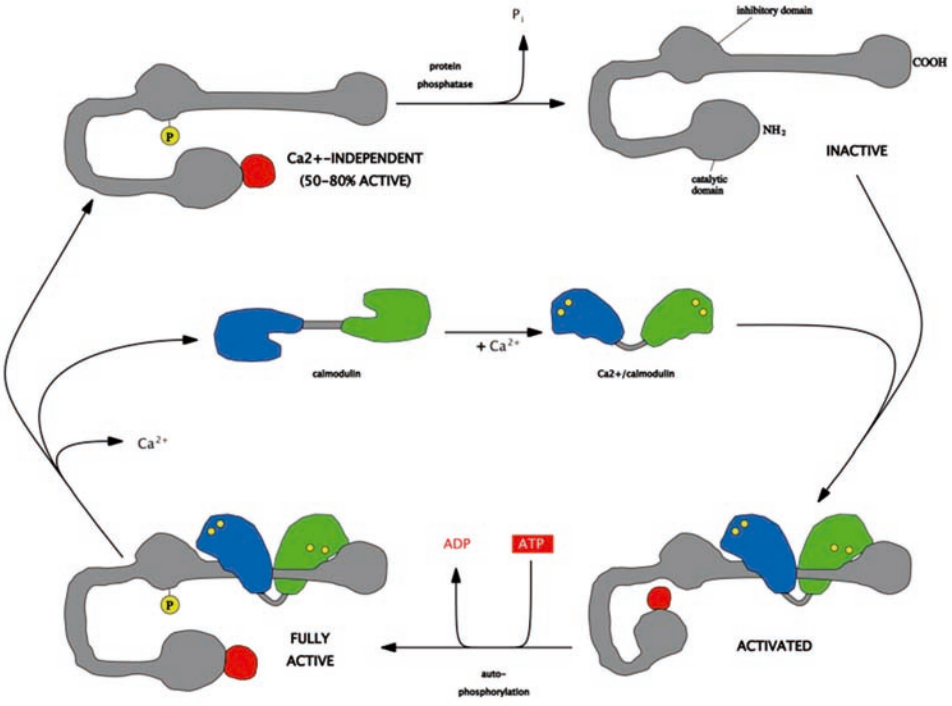


Fig. 7: Screenshot of SVG drawing created from layout and render information

ated a second extension to the SBML file format that works in conjunction with the layout extension and allows the storage of rendering information for the layout items. Together with the SBML layout extension, this enables programs to store very elaborate drawings of reaction networks within SBML files. This can be used for the graphic illustration and documentation of models, which sometimes enhances their understanding.

To demonstrate the features and capabilities of this new extension we have written a new XSLT stylesheet transforming layout + render information into an SVG drawing, together with some sample files.

This work has been presented at several international meetings. Future work will include writing an XSLT stylesheet that combines the features of the two current stylesheets. The new stylesheet will be able to use rendering information in the generation of an SVG drawing, if it is available, and resort to the behavior of the old XSLT stylesheet if no rendering information is provided in the file.

Collaboration Partners

Control and Dynamical Systems Group
at Caltech, Pasadena, USA

The SBML consortium
(incl. approx. 20 international groups)

New methods under development at EML Research for incorporation into COPASI include:

Graphical Visualization of Biochemical Pathways and their Simulation Results

The visual representation of biochemical pathways aids researchers in analyzing their system. Visualization can also help in studying the properties of these systems, e.g. by a comprehensive view of simulation results. Based on algorithms and tools created in previous years, a platform combining those approaches – SimWiz and SimWiz3D, plus others – has been created. This platform, ViPaSi, is now ready for integration into COPASI. Accordingly, some of

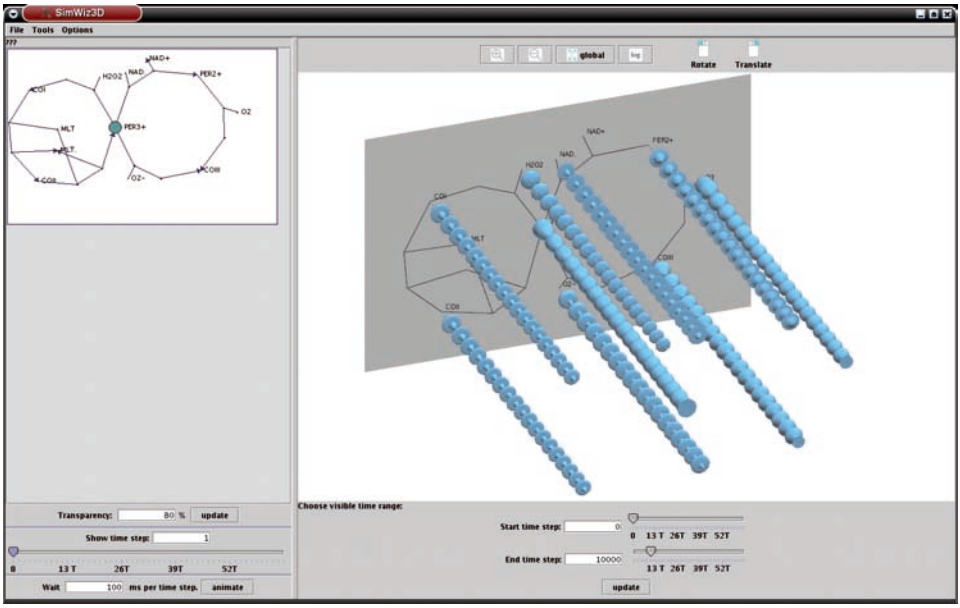


Fig. 8: GUI of SimWizD: The metabolic network is situated in a plane, the third dimension depicting time. The concentration of each metabolite is coded as the size of its corresponding node

the code has to be ported and adjusted to the respective requirements. We have been using C++, Qt, and OpenGL throughout the project. This work started in 2006 and will be continued in 2007.

In addition, the Motato tool (see annual report 2005) for the setup of models using a graphical representation of the network has been rewritten in C++ using the Qt4 toolkit. It is now ready for inclusion in COPASI once COPASI has also been converted to enable it to use the Qt4 toolkit. For the future, some updates to Motato have already been planned, such as improvements of the font rendering and inclusion of automatic layout algorithms.

Collaboration Partners

Prof. Dr. Markus Müller, Prof. Dr. Gerold Baier, Facultad de Ciencias, Universidad Autonoma del Estado de Morelos, Cuernavaca, Morelos, Mexico

Sponsor

Klaus Tschira Foundation

3.3 SYCAMORE

Project Managers

Dr. Ursula Kummer
Dr. Rebecca Wade

Project Members

Dr. Razif Gabdoulline, Ralph Gauges, Jürgen Pahle, Sven Sahle, Dr. Stefan Richter, Dr. Matthias Stein, Dr. Irina Surovtsova, Dr. Andreas Weidemann

SYCAMORE is a project aiming at the development of a system of tools and methods supporting users in employing and combining diverse computational approaches. It is a joint project of the Bioinformatics and Computational Biochemistry Group and the Molecular and Cellular Modeling Group at EML Research and is supported by the BMBF (Federal German Ministry of Education and Research). SYCAMORE is closely linked to COPASI (3.2), which provides some of the methodology for the SYCAMORE project.

The implementation of the SYCAMORE prototype has now been completed and, after some more testing and debugging, the system will be test-released in early 2007. The following methodological research has been done in this connection:

Complexity Reduction and Sensitivities in Large Biochemical Reaction Networks

Complexity reduction methods aim at a rational dissection of networks to learn about the interdependencies of subsystems. For the utilization of SYCAMORE we have already developed new complexity reduction methods able to automatically and dynamically dissect biochemical systems. After the testing of these methods, new and refined error criteria were introduced, and the algorithms were ultimately implemented in COPASI.

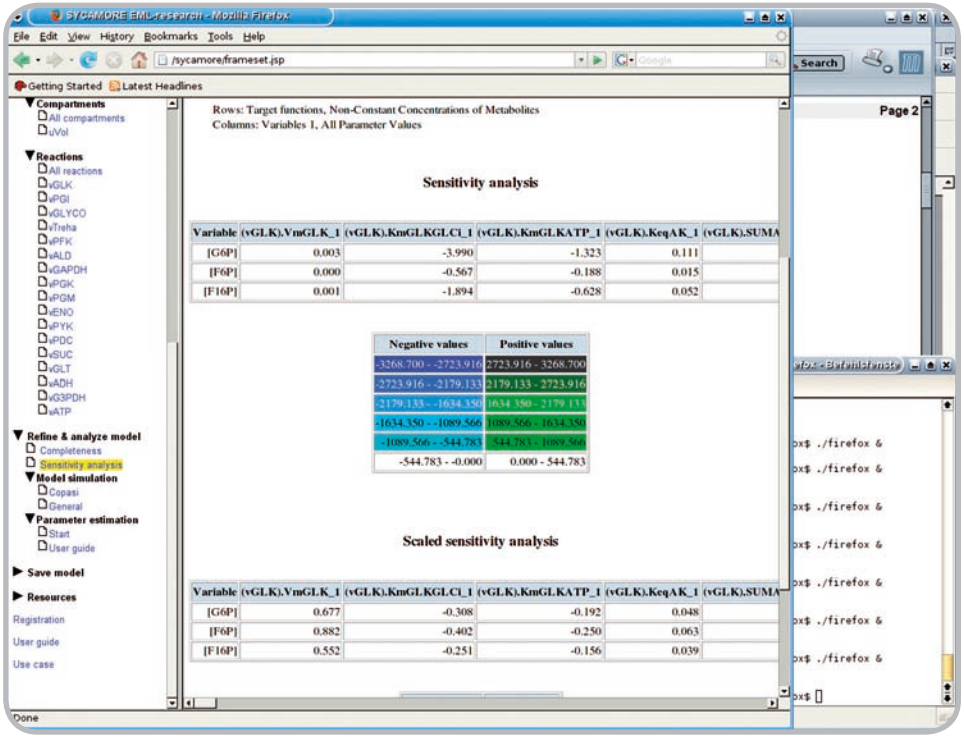
Advanced sensitivity analysis methods, which are useful both for other complexity reduction algorithms and for the general analysis of models, have also been integrated into COPASI. First of all, sensitivities are now derivable for any quantity with regard to any parameter in the system. In addition, the implementation of second-order sensitivities allows analysis of the local environment of the respective parameters to check whether for the robustness of a system with regard to a specific parameter changes drastically in the vicinity of parameter space.

Using Protein Structures in Systems Biology Projects.

In the SYCAMORE project, the MCM group is working on the development and application of methods of harnessing protein structural information for systems biology projects. One of the problems in creating mathematical models of

biochemical networks is either the absence of experimentally determined kinetic parameters or the incompatibility of experimental and simulation conditions. We are focusing on developing both methodology and software for protein-structure-based estimation of kinetic parameters. We have developed qPIPSA (quantitative Protein Interaction Property Similarity Analysis) to quantify differences in the molecular interaction fields, for example the electrostatic potentials, of related enzymes and to correlate these with enzyme kinetic parameters. The qPIPSA approach has been evaluated in studies of well-characterized enzymes. It was found necessary to adopt a conservative protein structural homology modeling protocol in which conserved amino acid residues occupy the same amino-acid side-chain orientation. qPIPSA was used to investigate the inter-organism variation of enzymes in a metabolic pathway. The enzymes of twelve model organisms for the glycolysis energy-converting pathway were characterized [Stein 2006] (see Figure 10).

Fig. 9: Screenshot of SYCAMORE showing sensitivities of concentrations with regard to parameter changes



We have implemented an automated workflow for the basic features of qPIPSA and integrated this into the SYCAMORE tool kit to allow users access to the approach for assistance in the choice of kinetic parameters in metabolic pathway simulations. This workflow includes retrieval of relevant data on kinetic parameters, protein sequences, and protein structures from several databases, including SABIO-RK, BRENDA, RCSB, and Swissprot; modeling of three-dimensional structures of proteins by homology; computation of protein electrostatic potentials; and tools for analysis of the similarity of electrostatic potentials and comparison with kinetic data.

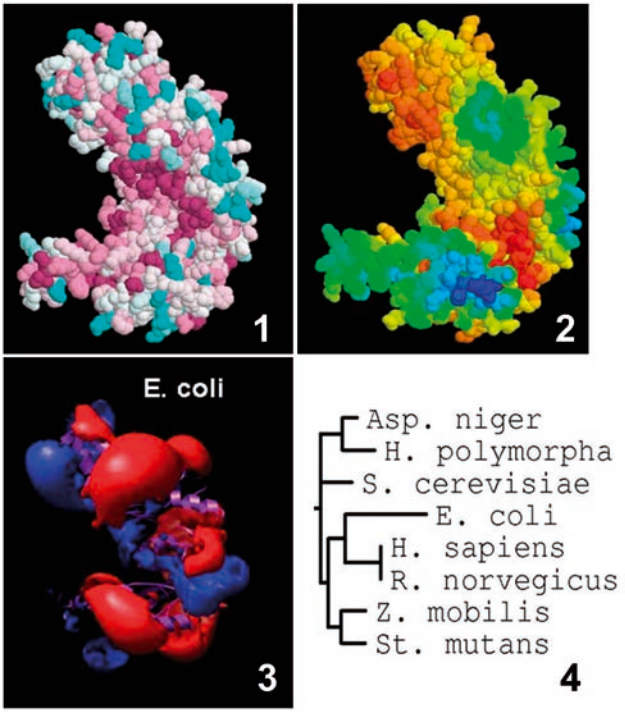
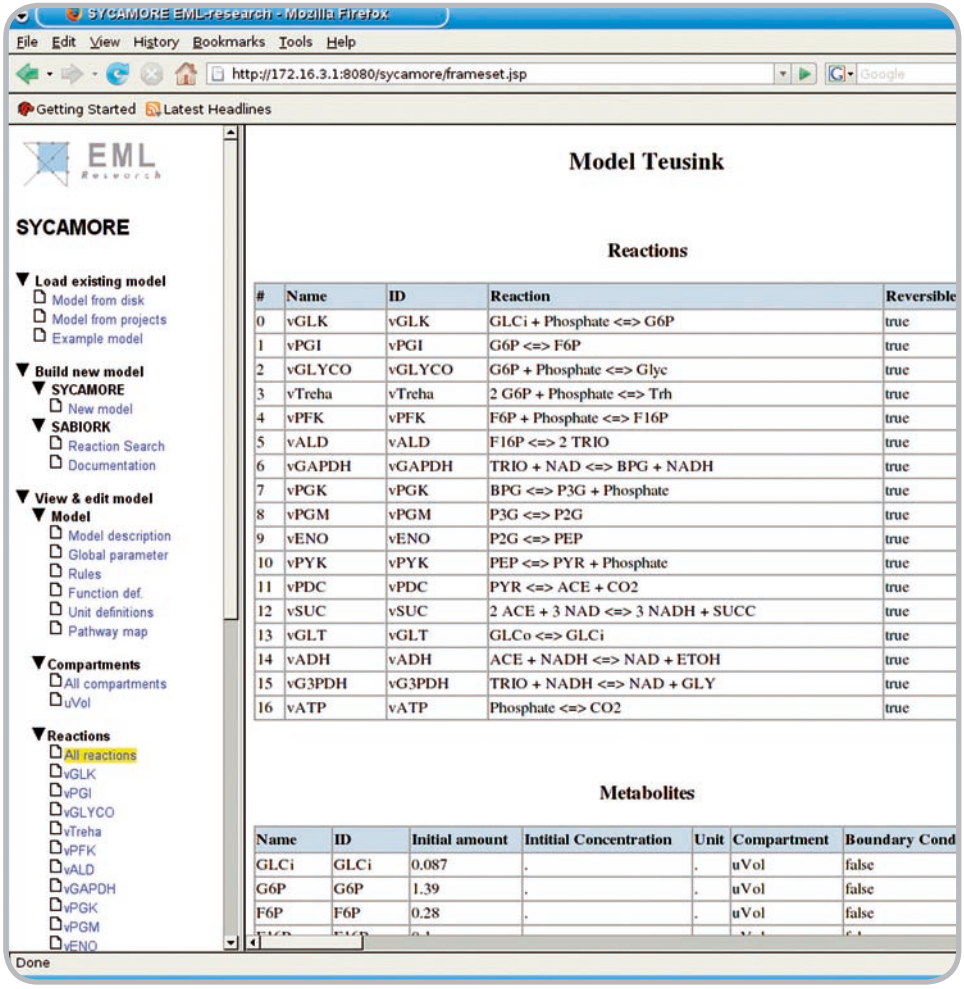


Fig. 10: qPIPSA analysis of the hexo/glucokinase glycolytic enzymes from eight different species. From left to right: conservation of the (1) amino acid multiple sequence alignment and (2) protein electrostatic potential shown on the three-dimensional protein structure (blue-less to red-more conserved); (3) electrostatic isopotential contours at ±0.6 kcalmol-1e-1 of E. coli hexokinase; (4) epogram showing comparison of the electrostatic potentials of the enzymes from different species

Implementation of SYCAMORE

As stated above, the implementation of the SYCAMORE prototype has now been completed. In addition to the GUI, wrappers for accessing different functionalities of COPASI

have been developed. Thus it is now possible to load or build a model and use COPASI to simulate it quickly via SYCAMORE. The model can be built by accessing SABIO-RK (see the corresponding chapter in this report) and refining it via methods provided in SYCAMORE to check the model for coherence and completeness, or to compute parameters based on the sensitivity analysis methods implemented in COPASI and the structure-based methods described above.



Model Teusink

Reactions

#	Name	ID	Reaction	Reversible
0	vGLK	vGLK	GLCi + Phosphate <=> G6P	true
1	vPGI	vPGI	G6P <=> F6P	true
2	vGLYCO	vGLYCO	G6P + Phosphate <=> Glyc	true
3	vTreha	vTreha	2 G6P + Phosphate <=> Trh	true
4	vPFK	vPFK	F6P + Phosphate <=> F16P	true
5	vALD	vALD	F16P <=> 2 TRIO	true
6	vGAPDH	vGAPDH	TRIO + NAD <=> BPG + NADH	true
7	vPGK	vPGK	BPG <=> P3G + Phosphate	true
8	vPGM	vPGM	P3G <=> P2G	true
9	vENO	vENO	P2G <=> PEP	true
10	vPYK	vPYK	PEP <=> PYR + Phosphate	true
11	vPDC	vPDC	PYR <=> ACE + CO2	true
12	vSUC	vSUC	2 ACE + 3 NAD <=> 3 NADH + SUCC	true
13	vGLT	vGLT	GLCo <=> GLCi	true
14	vADH	vADH	ACE + NADH <=> NAD + ETOH	true
15	vG3PDH	vG3PDH	TRIO + NADH <=> NAD + GLY	true
16	vATP	vATP	Phosphate <=> CO2	true

Metabolites

Name	ID	Initial amount	Intitial Concentration	Unit	Compartment	Boundary Cond
GLCi	GLCi	0.087	.	.	uVol	false
G6P	G6P	1.39	.	.	uVol	false
F6P	F6P	0.28	.	.	uVol	false

Fig. 11: Screenshot of the SYCAMORE user interface. The example model loaded describes glycolysis

Figure 11 shows a screen shot of the user interface. As with COPASI, users can navigate their way through the system by means of a tree in the left-hand panel.

Collaboration Partners

Various members of the HepatoSys initiative in Germany (www.systembiologie.de).

Dr. Isabel Rojas (EML Research)

Sponsors

Federal Ministry of Research and Education (BMBF)

Klaus Tschira Foundation

3.4 UniNet

UniNet is a European research initiative that aims at identifying and investigating common mathematical structures in various complex networks (genetic, metabolic (particularly at our site), neuronal, ecological, economic, etc.).

Stability in Systems with Unknown Parameters

Project Manager
Dr. Ursula Kummer
Project Members
Dr. Iulian Stoleriu, Tim Johann

The difficulties encountered in modeling biochemical systems include the wide variety of kinetic parameters involved, most of them unknown a priori, the large number of reactions in the network, which makes even numerical simulations very demanding, and the multiple time-scales in the system leading to stiff equation systems.
In order to find solutions notably for the first of these problems, we employ Stoichiometric Network Analysis (SNA). Stoichiometry is a qualitative relationship between the compounds of the network. To each metabolic network one can assign a stoichiometric matrix, where the rows represent the compounds of the reactions while the columns of the matrix correspond to the reactions themselves. Stoichiometric network analysis is a general approach to studying the qualitative dynamics of chemical networks. One of its advantages is that the kinetic parameters are not needed in the analysis.

Because of the high dimensionality of these networks their analysis can be cumbersome. The stability of steady states in biochemical networks is a very important issue. The steady states are usually difficult to find and, once found, depend upon unknown kinetic parameters. Accordingly, it is necessary to find methods for studying the equilibria without actually knowing them. One common method for doing this (it can also be employed to reduce the size of the network) is to use the extreme currents (or elementary flux modes, extreme pathways). The idea is to split the whole network into sub-networks that are easier to handle. Each of these sub-networks is generated by elements of a convex basis in the space of steady states. These elements are the extreme currents, and they are uniquely determined.

Our aim was to find conditions for the stability or instability of the steady states by starting from stability analysis of the extreme currents. In this way we have derived algo-

rithms that may lead to inequalities among the kinetic parameters in the system From these inequalities we can then deduce the stability or instability of the steady state of the entire network. The method is based on the study of the sign patterns of the principal leading determinants of the Jacobian matrix, written in terms of convex coordinates.
To establish the relations between the kinetic parameters we make use of Grőbner bases. This runs into difficulties when at least one of the determinants is equal to zero. The method may also be difficult to use when the number of extreme currents is too high. Despite this, our ultimate goal in this project is to improve the method and to implement it in COPASI.

Analysis of the Origin of Dynamics

Metabolic networks (MNs) can exhibit a multitude of different behaviors, notably steady, oscillating, or chaotic dynamics. This work started in September 2006 and aims at understanding which constituents or subsets of metabolic networks influence their qualitative behavior most effectively. Theoretical analysis and computer simulations of existing and/or artificial MNs are two equally important aspects the research will be based on.

Lyapunov exponents (LEs) will be used as a quantitative measure for qualitative behavior. LEs quantify the time-averaged convergence tendencies of different initial conditions and have to be calculated during simulation. Due to the vast number of system layouts, structural and parametrical considerations will guide the sampling process. We hope to derive inherent principles governing the dynamics of MNs and to establish how changes in their behavior are induced by variations in the concentrations of certain metabolites.

As a long-term goal our intention is to identify structural/parametrical dependencies indicating which part of a network affects its dynamics most significantly. The aim here is to minimize sampling space for MNs with unknown dynamics. These results could also be used to develop ways of controlling the overall qualitative dynamics of natural MNs.

Collaboration Partners

The UniNet consortium, especially Dr. Markus Kirkilionis (University of Warwick, UK)

Sponsors

EU

Klaus Tschira Foundation

3.5 BioSim

The project began in 2005 and aims at the development of models and methods supporting the prediction of the biochemical fate of drugs in an organism.

Project Manager

Dr. Ursula Kummer

Project Member

Femke Mensonides,
Tim Johann

In 2006 we refined the model of the central metabolism of yeast involved in xenobiotic transformations. Most metabolic models for yeast are defined for anaerobic conditions, but it is important to investigate aerobic conditions as these will be relevant for the corresponding lab work. Accordingly, we have continued with our endeavors to develop a model for the aerobic metabolism of yeast.

Given various limitations on the availability of experimental data to ground the models on, we also started to undertake experiments at Hans Westerhoff's department (Free University of Amsterdam). The resulting model will be used to compare xenobiotic metabolism under aerobic and anaerobic conditions.

Collaboration Partners

The BioSim consortium, especially Dr. Martin Bertau, Dr. Lutz Brusch (TU Dresden), and Prof. Hans Westerhoff (Free University of Amsterdam, The Netherlands).

Sponsors

EU

Klaus Tschira Foundation

3.6 SABIO-RK

Systems biology involves analyzing and predicting the behavior of complex biological systems like cells, organisms, or even whole ecosystems. This requires qualitative information about the interplay of genes, proteins, chemical compounds, and biochemical reactions. It also calls for quantitative data describing the dynamics of these networks. These data have to be collected, systematically structured, and made accessible for the set-up of biochemical model simulations.
To provide quantitative experimental data for systems biology, we have developed SABIO-RK, a database system offering information about biochemical reactions and their corresponding kinetics. It offers information about the reactions such as their substrates, products, enzymes, inhibitors and activators, as well as the description of reaction kinetics such as their mechanism types together with their respective equations defining the reaction rates with their corresponding parameters. The description of the environmental conditions used for parameter determination is also provided. At present, the content of the database mainly describes the steady-state kinetics of metabolic reactions. The specified kinetic parameters include rate constants, maximal velocities, and equilibrium constants like Michaelis, dissociation, or inhibition constants.

A new feature of SABIO-RK adds information about enzyme proteins to represent, for example, the composition of enzyme complexes. Another extension of SABIO-RK that we have recently started to work on is a detailed description of reaction mechanisms, including the specification of elementary reaction steps with their corresponding kinetic data.

SABIO-RK can be accessed in two different ways: via a web-based user interface to browse and search the data manually, and, more recently, via web-services that can be automatically called up by external tools, e.g. by other databases or simulation programs for biochemical network models. In both interfaces, reactions with kinetic data can be exported in SBML (Systems Biology Mark-Up Language), a data-exchange format widely used in systems biology.

Project Manager
Dr. Isabel Rojas
Project Members
Martin Golebiewski, Renate Kania, Dr. Olga Krebs, Dr. Andreas Weidemann, Dr. Ulrike Wittig
Klaus Tschira Foundation Scholarship Holder
Saqib Mir (PhD Student)
Student
Xioan Chen

The existence of a programmatic interface, together with the use of controlled vocabularies, ontologies, and links to external resources, makes SABIO-RK well suited for integration into different applications using or requiring kinetic data for biochemical reactions.

Database population, curation, and annotation

The SABIO-RK database is populated by merging general information about biochemical reactions and pathways mainly derived from databases like KEGG (Kyoto Encyclopedia of Genes and Genomes) with corresponding kinetic data manually extracted from literature. Using a web-based input interface (see Figure 12), students enter the kinetic data from relevant articles into a temporary database. The main objective of this user interface is to support a uniform format that students and curators can employ to include data found in the publications. The interface offers the possibility of choosing terms from predefined thesauri (allowing, of course, the introduction of new terms) and performs some consistency checks on the data introduced. Some of these vocabularies have been internally developed, e.g. to differentiate between kinetic law types or parameter roles, while others are controlled vocabularies or ontologies accepted as references by the scientific community, for example the organism taxonomy from the NCBI or descriptions of tissues and cellular locations from Brenda. The use of controlled vocabularies helps to avoid redundancies caused by aberrant notations or typing errors. Ideally, the students extract the following information for each reaction reported in a publication:

- reaction defined by substrates and products
- modifiers of the reaction (activators, inhibitors, catalysts, cofactors)
- cellular location of compounds
- enzyme classification number
- SwissProt accession number(s) (of the enzyme)
- variants of the enzyme (wild type or a certain isoenzyme or mutant)
- kinetic law type (e.g. Michaelis-Menten, Ping Pong Bi Bi)
- kinetic law formula
- kinetic parameters (e.g. Km, kcat, Vmax)

- concentrations used for reactants, enzymes, and modifiers
- experimental conditions (e.g. temperature, pH, buffer composition)
- biological source (e.g. cell type, tissue, organism, strain)
- information source (reference)

For most of this information, comment lines are available for adding information about such things as synthetic, labelled derivatives of physiological compounds, or host organisms for recombinant enzymes.

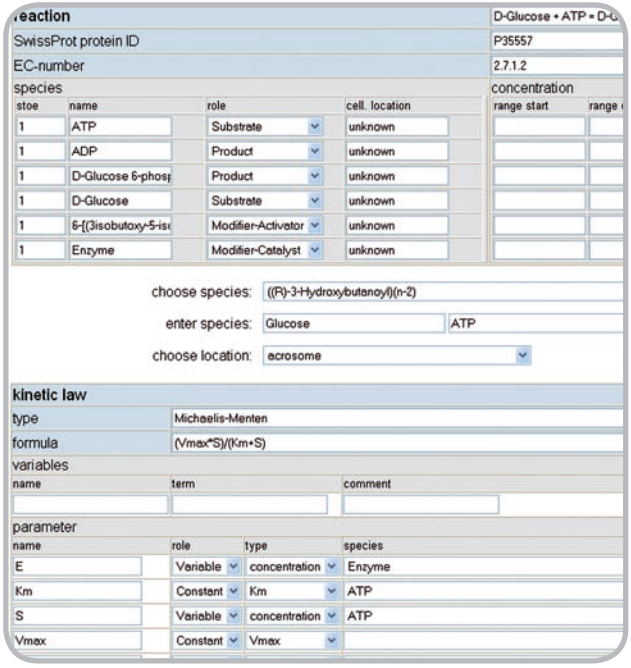


Fig. 12: Screenshot of input interface

Before the data is finally transferred to SABIO-RK, they are curated by a team of biological experts revising the data and correcting and complementing them if necessary. During the curation process, the data is consistently unified and structured to facilitate comparison of the kinetic data extracted from different sources. The reason for this is that

such data are usually obtained under different experimental conditions or from different organisms, tissues etc. The curators are frequently faced with problems like synonymous or aberrant notations of compounds and enzymes, multiplicity of parameter units, and missing information about assay procedures and experimental conditions. The problem of compound identification, and as a consequence the identification of reactions, is unfortunately a very frequent one. To address this problem, we have been working on methods for the normalization of compound names and on tools for deriving the chemical structure of a compound from its name (see Section 3.7).

Apart from providing resources aiding the student helpers and curators in processing of the data introduced, the input interface also supports curators in their administrative work (assignment of papers, statistics etc.). The publications to be revised have been obtained from PubMed by using various queries leading to papers that can be expected to contain information about biochemical reaction kinetics.

The information supported by this input interface covers most of the fields present in the STRENDA commission's recommendations for reports on reaction kinetics. Currently the input interface is only being used internally by the SABIO-RK development team. However we hope that in the future experimental partners will be able to introduce their data directly into the database and thus make it available via the SABIO-RK database interface.

In addition to the input interface we have used Oracle's Application Express to generate an application enabling curators to browse through the database without difficulty and to edit database entries directly. The application uses customized views presenting an integrated view of the different database tables storing the data.

To supplement the knowledge stored in SABIO-RK, we annotate entities and expressions to domain ontologies like SBO (Systems Biology Ontology) (for elements like parameter types and kinetic laws) or ChEBI (Chemical Entities

of Biological Interest) (for compounds), and to knowledge bases like KEGG (for compounds and reactions) or UniProt/Swiss-Prot (for proteins) (see Figure 14). These annotations provide a semantic basis for the data facilitating their interpretation, comparison, and integration. Annotations are also used to provide links from SABIO-RK to external resources providing supplementary information. Some of the annotations, such as those to SBO entities, are entered manually, whereas others, such as those to KEGG and ChEBI, are carried out automatically by creating mappings. In these cases, we have created mappings from SABIO-RK compounds to KEGG and CHEBI compounds based on the compound names (recommended names and synonyms).

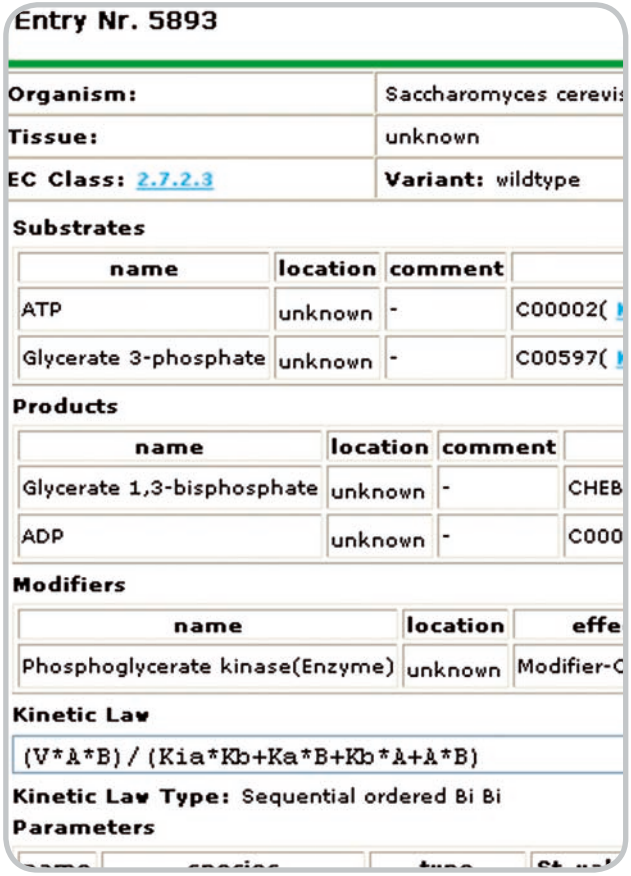


Fig. 13: Screenshot of database entry containing annotations to KEGG, ChEBI etc.

To ensure the integrity of these automatic annotations, the mappings are curated and regularly regenerated.

SABIO-RK data can be exported in SBML (Systems Biology Mark-Up Language) format. This requires specification of the data in particular formats. For example, parameter units have to be specified in scaled and composite SBML units, and kinetic law equations need to be described as required by SBML. These transformations are carried out in the data population process in order to comply with SBML standards. Due to the restrictions imposed by the SBML specification, it is in some cases necessary to exclude or summarize data and not merely change their format. One example is parameter values (though this should change in future versions of SBML): Although we store them as a range, it is only possible to give them one numerical value in SBML. Thus, we have decided to provide the mean value. Another example is the case of compound location. Here our temporary solution is to include all compounds at the same location, regardless of the location given in the papers (and stored in the database). The rationale behind this is that, given that a user can combine reactions and their kinetics from different papers, i.e. conditions, it is common for the compounds' locations not to be specified (i.e. unknown) in one paper, while they are specified in others. Accordingly, when generating the SBML files, the compounds with different locations would be created as different SBML species and there would not be any interfaces between the compounds of the reactions listed. In the generation of the SBML file we are attempting to implement all rules for the annotation of an SBML file described in the recently developed MIRIAM standard (Minimum Information Requested In the Annotation of biochemical Models).

As of December 2006, kinetic data from more than 1,230 publications have already been entered in the temporary database. More than 70% of them have already been curated and inserted in the SABIO-RK database. One publication can have multiple database entries caused by the different reactions, enzymes, kinetic laws, environmental conditions etc. described in it. By the end of 2006 SA-

BIO-RK contained about 9,300 curated single database entries describing some 1,500 different reactions for 420 distinct EC classes referring to about 280 different organisms. Each database entry comprises at least one kinetic parameter. 43% of all SABIO-RK entries are related to a reaction rate equation, making them especially useful for setting up simulation models.

SABIO-RK interfaces

SABIO-RK can be accessed in two different ways: via a web-based user interface for browsing and searching the data manually, or via web-services that can be automatically called up by external tools, e.g. by other databases or simulation programs for biochemical network models. In both interfaces, reactions with kinetic data can be exported in SBML.

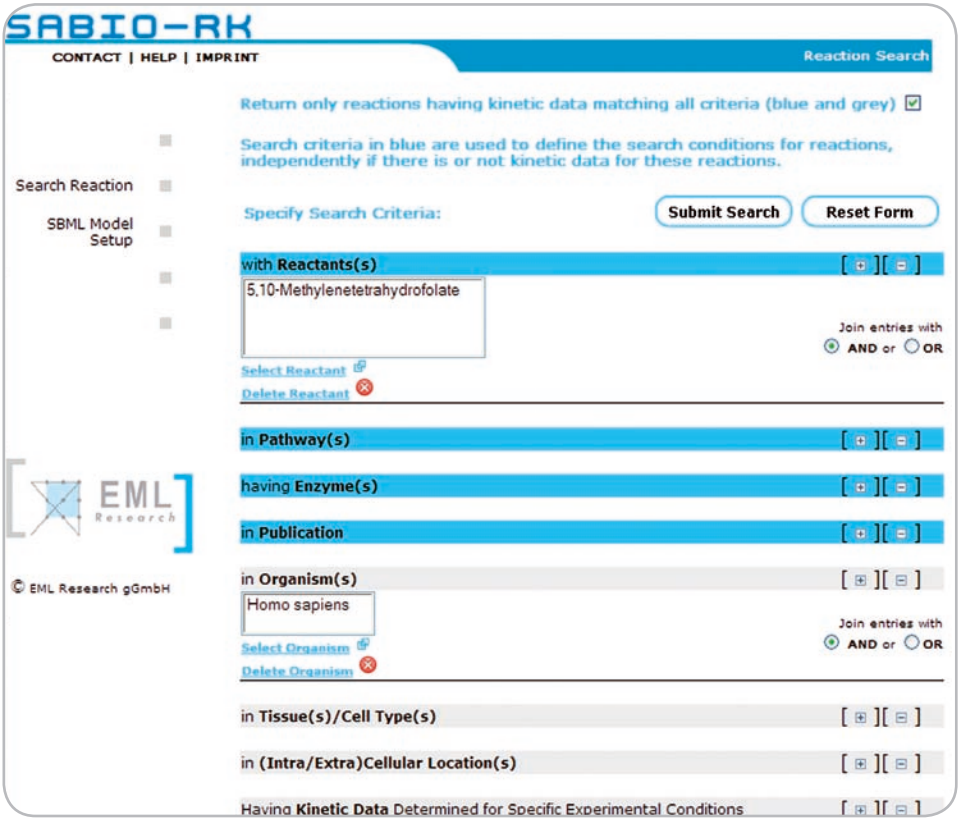


Fig. 14: SABIO-RK user interface (search criteria)

The current version of the SABIO-RK web interface allows users to perform searches for reactions by specifying their characteristics (one or many) and of the kinetic data searched (see Figure 14). For example, the user can specify a pathway in which the reaction participates (e.g. glycolysis) or one or more reaction participants (reactants or enzymes), and for these reactions search to see if there is any kinetic data available for, e.g., a particular organism, tissue, or cell type. Additional search terms include cellular locations, environmental conditions (pH and temperature), or publications in which kinetic data are reported. The system presents corresponding reactions, indicating whether there is kinetic information available for the criteria specified. This is done by employing a three-color code. Green means that there are kinetic data available for the associated reaction that match all the search criteria. For a search like "find all reactions within the glycolysis pathway for homo sapiens taking place in the liver," this would mean that for the given glycolytic reaction there are kinetic data reported on the human liver. Yellow means there are kinetic data available, but they do not match all the search criteria. For example, the kinetic data have not been determined for homo sapiens but for rattus sp, or not in the liver but in the heart. Red indicates that there are no kinetic data stored for the reaction.

Apart from showing the availability of kinetic data for the specified reactions, the system will also indicate whether there are kinetic data available for the enzymes catalyzing each of these reactions. We have taken this approach to offer complementary or alternative information about kinetic data for related reactions catalyzed by the same enzyme. The availability of kinetic data for the enzyme is shown using the same three color-code as the one used for the reactions. By clicking on a reaction, further information about it is displayed: reactants, pathways in which it parti-cipates, and enzymes catalyzing this reaction and reported with kinetic data in the database for a specific organism. Additional information about the enzyme (name, synonyms, classification, and reactions it catalyses) can be obtained by clicking on the EC number.

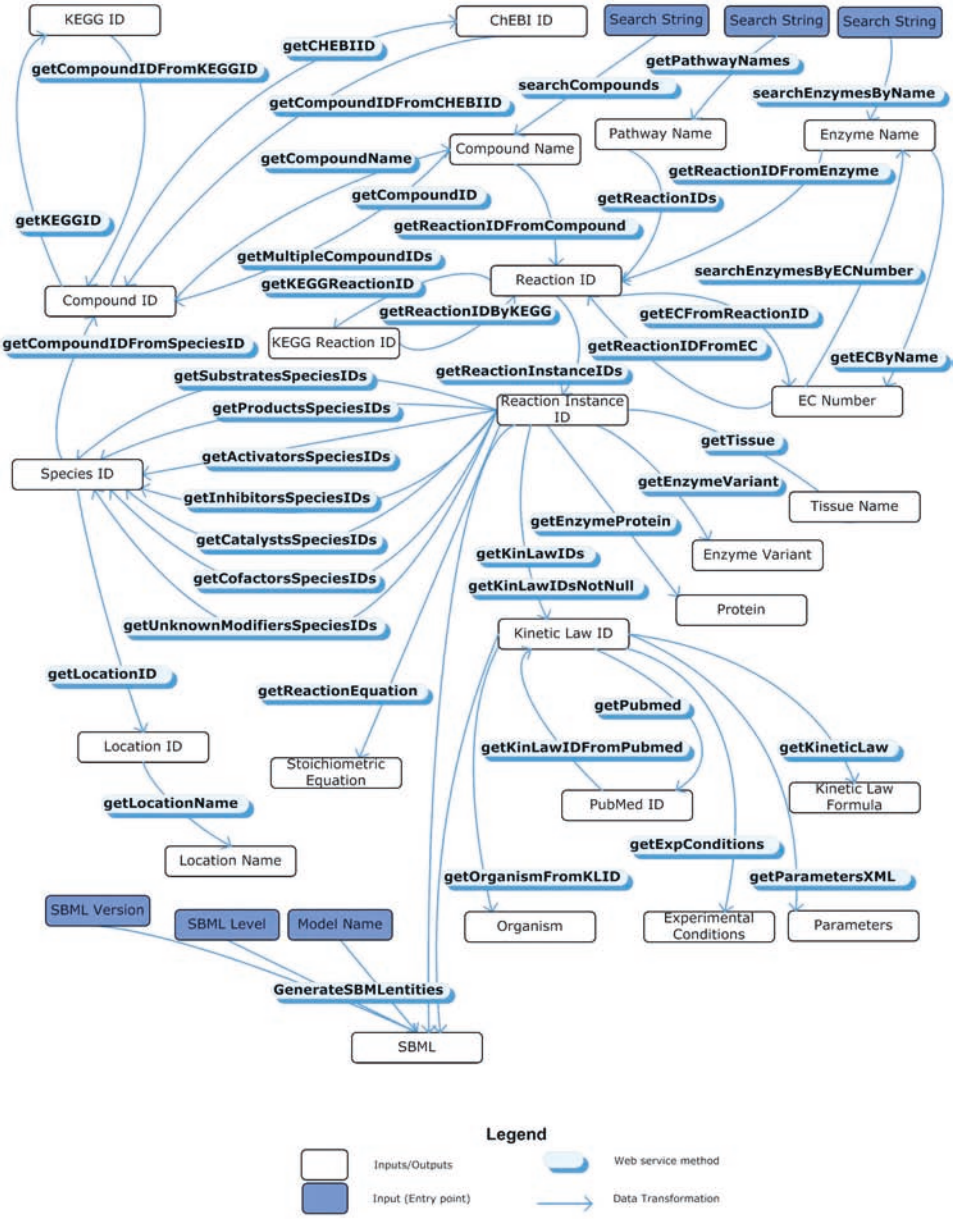


Fig. 15: SABIO-RK web services

From the result screen indicating the list of reactions retrieved, the user can view the kinetic data associated with a given reaction or all kinetic data available for an enzyme catalyzing this reaction. The kinetic data are presented in a new window in the form of entries, where each entry contains data on the reaction's kinetics (kinetic law formula, kinetic law type, parameters, etc.), on the context of the kinetics measured (i.e. organism, tissue/cell type, environmental conditions, etc), information about the enzyme (enzyme classification and variants), information about the reaction participants (substrates, products, and modifiers), and on the information source (provided by means of a link to the corresponding PubMedID, where applicable). Comment lines are also present at many different levels, providing additional information to the user that can be stored in a structured manner.

The SABIO-RK Web Service provides customizable points of entry into the SABIO-RK system in a language-independent fashion. Using this Web service, users can write their own clients to customize and automate access to SABIO-RK directly from their simulation software, tools, or databases, thereby dispensing with the need to parse information manually from Web pages. Currently SABIO-RK is being used by two systems biology platforms: CellDesigner and SYCAMORE (see section 3.3). Figure 15 shows the web services currently available for SABIO-RK. Our aim is to provide more semantics to the web services to facilitate their use and the exchange of data with other systems.

3.7 BioReader: Linguistic Tools for Analyzing Chemical Compound Names

Project Manager
Dr. Isabel Rojas
Project Members
Martin Golebiewski, Dr. Jasmin Saric
Klaus Tschira Foundation Scholarship Holder
Stefanie Anstein (until March 2006)

To support the curation process in SABIO-RK and to obviate the well-known problem of compound identification, we are working on the development of tools based on natural language processing (NLP) methods. The main focus of this work is on the systematic analysis of chemical compound names to identify synonymous notations of compounds and to distinguish between different chemical compounds based on variations in their names. One chemical compound can have many different names. It can have several trivial names, as well as several systematic names, although these uses may be compliant with naming recommendations such as those made by the International Union of Pure and Applied Chemistry (IUPAC). The different names assigned to a compound can be caused by a variety of factors, for example:

- trivial names vs. systematic names:
  e.g. valproic acid and 2-propylpentanoic acid

- different parts of the molecule being considered as lead structure (radical of the name): e.g. acetylphenol and phenylacetate

- aberrant order of the constituents of a lead structure:
  e.g. 2-amino-6-methyl-4-pyrimidol and 6-methyl-2-amino-4-pyrimidol

- reference to constituents either as a prefix (like amino-) or as a suffix (like –amine): e.g. 1-phenyl-2-aminopropane and 1-phenylpropan-2-amine

This makes τ name-based identification of chemical compounds a serious problem. To address this problem we are developing a program designed to generate synonyms for chemical compound names. With an extended list of synonyms we hope to improve the chances of matching chemical compounds with their names. Our approach combines a set of different strategies:

1) Normalization of variant spellings for compound names. This includes permutations of the order of constituents, trans-

formation of prefixes and suffixes, conversion of acids to their corresponding bases (e.g. –ic acid → -ate), and also simple transmutations concerning brackets, spaces, and hyphens, upper and lower case, etc.

2) Analysis of existing synonym lists (such as synonymous references to a chemical compound in comprehensive chemical databases like PubChem or ChEBI) using heuristic methods to identify synonymous parts in compound names. This results in the generation of a list of synonymous parts that can then be used for the generation of synonyms.

3) Generation of a dictionary of chemical morphemes describing single functional groups and skeletal structures of chemical compounds. Combining this with a compendium of trivial names then makes it possible to generate all variants of a compound name.

These methods for the generation of synonyms for compound name are to be combined with our work on the development of chemical structures from chemical compound names (CHEMorph) [Anstein 2006a]. CHEMorph analyzes systematic and semi-systematic names, class terms, and

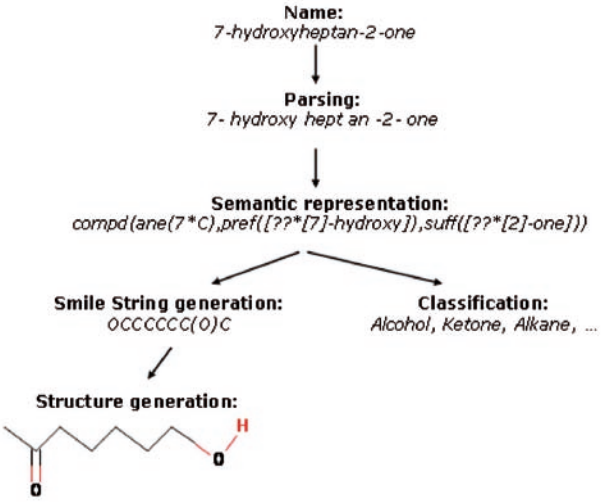


Fig. 16: Processing biochemical compound names

otherwise underspecified names by using a morphosyntactic grammar developed in accordance with IUPAC nomenclature. It yields an intermediate semantic representation describing the information encoded in a name. The tool provides SMILES strings for the mapping of names to their molecular structures and also classifies the terms analyzed. The combination of CHEMorph and the methods for the creation of synonyms of compound names (see Figure 16) will constitute an optimal platform for the identification of compounds, either by the comparison of names or by the comparison of chemical structures.

3.8 Modeling and Simulation: From the Molecule towards the Cell

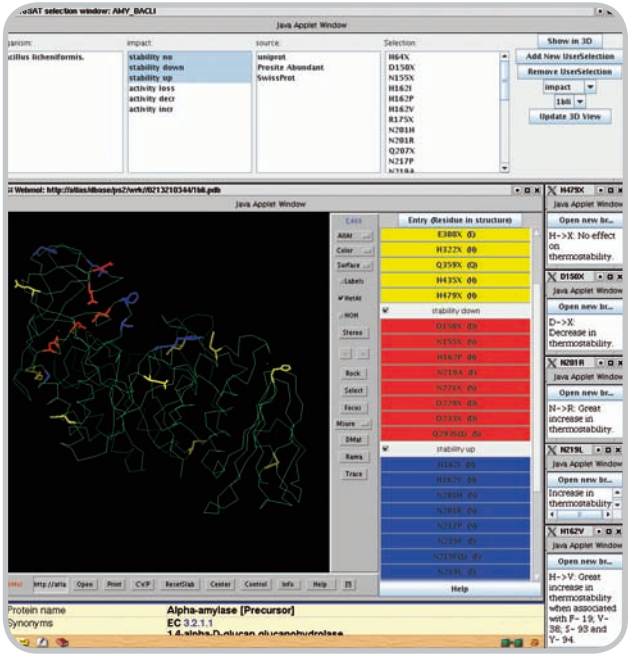
Bioinformatics Resources on the Web

This year we released a new web-server, the ProSAT2 [Gabdoulline, 2006]. ProSAT2 is a server that selects and groups protein residue-based annotations (such as the effect of a point mutation on enzyme function) and explores them interactively on a 3D structure of a protein. ProSAT2 includes features from ProSAT version 1 (Protein Structure Annotation Tool) displaying SwissProt and ProSite annotations. The residue-based information can be extracted from databases (e.g. BRENDA or UniProt) or user-defined. Visualization is based on the WebMol Java Protein viewer (see Figures 17 and 18).

In addition, the following web-based resources and software have been further maintained and made publicly available at http://projects.villa-bosch.de/mcm:

- DSMM: a Database of Simulated Molecular Motions.
- MolSurfer: a Macromolecular Interface Navigator
- ProSAT: PROtein Structure Annotation Tool 1.
- The Ubiquitin and Ubiquitin-like Protein Web Resource
- PIPSA (Protein Interaction Property Similarity Analysis)
- SDA (Simulation of Diffusional Association).

Fig. 17: Screenshot showing a ProSAT2 session



Project Manager

Dr. Rebecca Wade

Project Members

Dr. Vlad Cojocaru, Anna Feldman-Salit, Sulaiman Faisal, Frederik Ferner, Dr. Razif Gabdoulline, Matthias Janke, Domantas Motiejunas, Dr. Stefan Richter, Dr. Outi Salo-Ahen, Dr. Matthias Stein, Dr. Peter Winn, Mai Zahran

Guest Scientist

Prof. Arieh Ben-Naim

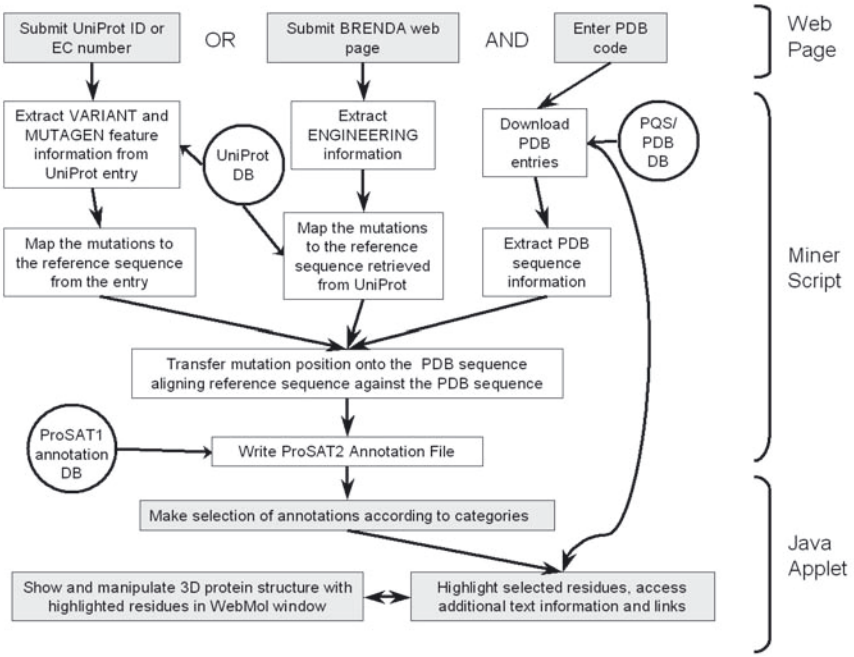


Fig. 18: Flowchart illustrating how ProSAT2 extracts residue-specific data from databases and permits the user to sort them and highlight residues on 3D protein structures, along with information on their function or the impact of their mutation

Methodological Developments

Incorporating sequence and experimental information into protein-protein docking procedures

Macromolecular interactions play a crucial role in various biological processes. Examples include the interactions involved in signal transduction, the control of gene expression, the inhibition of enzymes, antibody antigen interactions, maintenance and regulation of the cytoskeleton, and others. The structure of a protein-protein complex can be determined at the atomic level with X-ray crystallography or NMR. However, it is a difficult, time-consuming task and limited to particular types of complexes. Therefore computational techniques pose an attractive alternative for addressing the macromolecular complexation problem, notably due to increasing computational power and the wealth of available biochemical information relevant to protein-protein interactions, which can be incorporated to aid the computational methods.

Computational modeling of macromolecular interactions and protein-protein complex prediction is a challenging task [Schleinkofer 2006]. A very large number of variables are involved in the intermolecular interactions that take place in living organisms. Overall, the main challenges in protein-protein docking can be summarized as the sampling problem and the scoring problem. The sampling problem arises from the need to generate a large number of potential binding arrangements, while the scoring problem arises from the need to evaluate all these configurations. Due to these problems, it is common for docking methods to go through different stages: first the initially efficient, but approximate, sampling and evaluation of potential complexes, and later the slower, but more detailed, refinement and filtering stages. This approach has been taken in the docking protocol we have developed (see Figure 19).

The main purpose of the first stage of our protocol is to perform extensive sampling in a computationally efficient way and thus generate a large number of reasonable configurations for the two proteins. This is achieved by rigid-body protein-protein docking with a modified version of the SDA (Simulation of Diffusional Association) program. It employs Brownian Dynamics simulations to perform the sampling, together with a simple but efficient energy function comprising electrostatic force, a shape exclusion term, and incorporated biochemical data. Biochemical data relevant to complex formation are collected from various resources including sequence conservation data, side-directed mutagenesis, mass spectrometry, cross-link experiments, and others. The structures thus generated are so-called Encounter Complexes representing the early stage of protein-protein complexation. The task of the subsequent steps is to select and refine the structures into plausible bound complexes. The most representative structures are selected from the pool of orientations generated with SDA by employing hierarchical clustering procedures with an average linkage rule for inter-cluster distance calculation. The number of clusters giving the best representation of the data set is identified, and the representatives of these clusters are then used in further steps. In the final stage, the selected

representative structures are subject to flexible refinement via short molecular dynamics simulation runs with the Amber program and the NPSA implicit solvent model [Wang 2006]. This computational protocol has been validated on the basis of a set of structurally and functionally diverse test cases. The docking procedures are being applied to several protein-protein complexes in conjunction with experimentalists (see Section *Applications*).

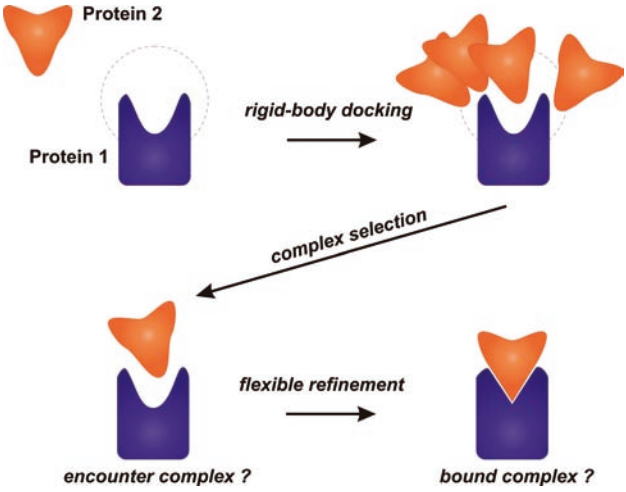


Fig. 19: Schematic diagram showing the stages of our protein-protein docking protocol

Simulation of biomacromolecular diffusion

Improvements to the methodology in our SDA (Simulation of Diffusional Association) software are ongoing in this project and described in Sections 3.10 and 3.11. Aspects studied include computation of free energy landscapes for protein-protein diffusion and the effects of mutations on the shape of these free energy landscapes [Spaar 2006], the modeling of short-range, non-polar interactions in Brownian dynamics simulations, and the estimation of diffusional and activation barriers for inter-protein electron transfer.

Applications

Optical biosensors for contaminant monitoring

This application focuses on the engineering of haloalkane dehalogenases and their use for the development of optical biosensors. Haloalkane dehalogenases are bacterial enzymes that cleave the carbon-halogen bond of halogenated aliphatic compounds by means of a hydrolytic mechanism. These enzymes have a potential application in detoxification of subsurface pollutants, recovery of industrial side-products, and biochemical sensing for the presence of halogenated contaminants in the environment. Modification of the substrate specificity and activity of these enzymes is required to optimize their properties for biotechnological applications. Our role was to employ modeling and simulation techniques to guide the engineering of haloalkane dehalogenases, alter their substrate specificity, and optimize their activity. Mutants were designed on the basis of COMBINE analysis to affect substrate specificity [Kmunicek 2005], and we continue to use the RAMD simulation method to investigate product release mechanisms in haloalkane dehalogenases. Preliminary results of the RAMD simulations and mutational experiments show that mutations can be introduced to affect the kinetics of haloalkane dehalogenases limited by product release.

Collaboration partners: Prof. Jiri Damborsky, Martin Klvana (Brno, Czech Republic), Dr. Federico Gago (University of Alcala de Henares, Spain), Prof. Ken Reardon (Colorado State University, USA), Prof. Thomas Scheper (University of Hanover, Germany).

Sponsor NATO Collaborative Linkage Grant

Modeling and simulation of cytochrome P450 dynamics and interactions

Over 100 different isoforms of P450 cytochromes (P450) form a class of enzymes that catalyze the mono-oxygenation of a wide range of compounds using molecular oxygen as a substrate. Their catalytic function is performed by the iron-containing heme prosthetic group deeply buried inside the protein. They play an essential role in drug metabolism and contribute to the biosynthesis of critical signaling molecules used for the control of development and homeostasis. Despite their function in detoxification, P450s

can produce toxic entities with carcinogenic effects. The genetic polymorphism of P450 cytochromes is responsible for different rates of drug metabolism in humans, while drug-drug interactions occur frequently because one drug may interfere with another's metabolic pathways.

We are using a range of computational techniques to study P450-electron transfer protein interactions and P450-substrate and product interactions Our primary focus this year was on substrate access and product egress from the deeply buried active site of P450s, which is dependent on the dynamic opening of channels in the protein. The conformational flexibility required for these processes confers part of the enzyme specificity.

In March 2006 we performed a systematic analysis of active site access channels in all cytochrome P450 crystal structures available at that time. The crystal structures demonstrate how some of the channels can merge when the protein structure opens up, resulting in a wide cleft to the active site caused largely by movements of two distinct secondary structure elements, the F-G helix-loop-helix and the B-C loop. This study revealed significant differences between the number and position of open channels in membrane-bound mammalian P450s when compared to the soluble bacterial enzymes.

By applying Random Acceleration Molecular Dynamics (RAMD) simulations to cytochrome P450 2C9 (a hepatic endoplasmic reticulum membrane-bound mono-oxygenase that contributes to the metabolism of about 20% of all drugs), we identified three main channels opening to the protein surface and allowing a product or substrate molecule to egress from the active site either to a lipid membrane or to aqueous solvent. The predominant egress pathway was dependent both on the nature and shape of the substrates and products and on the conformation of the F-G loop. To investigate how the anchoring in the lipid bilayer influences the conformational changes of the enzymes, we built several models of a membrane-bound cytochrome P450, and we are performing molecular dynamics simulations on these models.

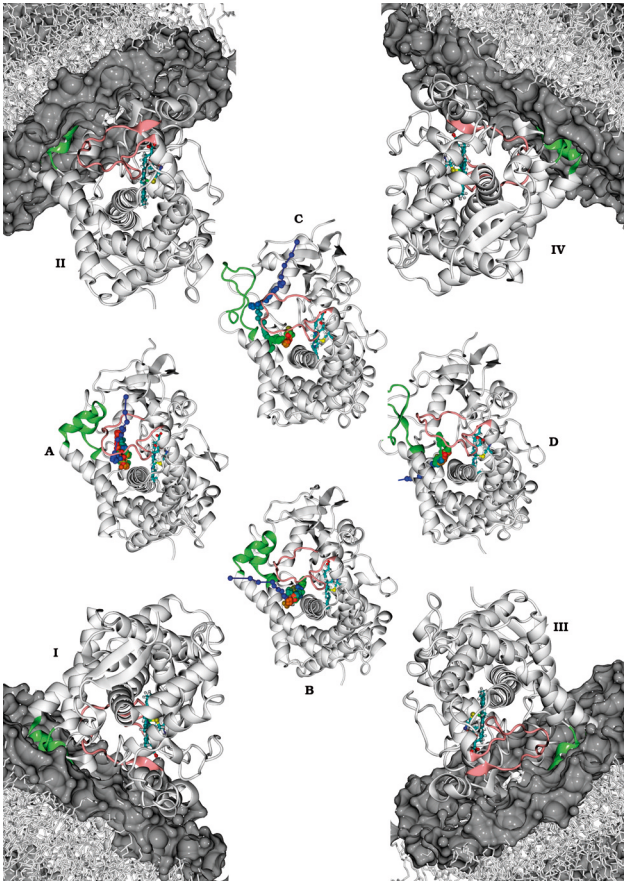


Fig. 20: Panels I-IV show two different views of the model of cytochrome P450 2C9 bound to the lipid bilayer, while panels A-D show the 4 major ligand egress pathways from the active site found for cytochrome P450 2C9 using RAMD methodology. The lipid bilayer headgroups are shown by the gray solvent-accessible surface, while the tails are shown as white sticks. The protein is shown in a white cartoon representation with the F-G loop highlighted in green and the B-C loop in pink. The heme is colored by atom type. The pathways are shown as traces of the ligand's center of mass during RAMD trajectories, colored according to time (red – beginning; blue – end)

Mechanisms of cellular regulation by ubiquitin

Ubiquitin is a small protein essential for the regulation of cellular function. Malfunctions in ubiquitination pathways lead to abnormal cellular regulation, which is involved in many diseases, notably cancer and neurodegeneration. Moreover, some viruses can hijack the body's ubiquitination machinery, allowing them to spread more easily. We have been studying the ubiquitin and ubiquitin-like protein conjugating pathways in conjunction with Amit Banerjee (Wayne State University Detroit, USA). The physical and biochemical properties of the enzymes involved in ubiquitination are presented on our newly updated website www.ubiquitin-resource.org. The website primarily provides protein models

and analysis of the structures of the ubiquitin-conjugating enzymes. At the time of writing, the structures of 1,580 ubiquitin-conjugating enzymes are available on the server, with more being added as their sequences are included in public databases. In particular, the user can search for the proteins that display the greatest electrostatic similarity to a query protein and view amino-acid mutations in a structural context via a link to ProSAT.

Protein-protein interactions in the cysteine synthase complex

Plants can assimilate and incorporate inorganic sulfur into organic compounds such as the amino acid cysteine. They thus make sulfur available to animals and humans. Sulfur is required for the synthesis of essential compounds, including vitamins and metal clusters. Cysteine biosynthesis in plants and bacteria involves a bienzyme complex, the "cysteine synthase complex" made up of the enzymes Serine Acetyl-Transferase (SAT) and O-Acetyl-Serine-(Thiol)-Lyase (OAS-TL). The biological function of this complex and the

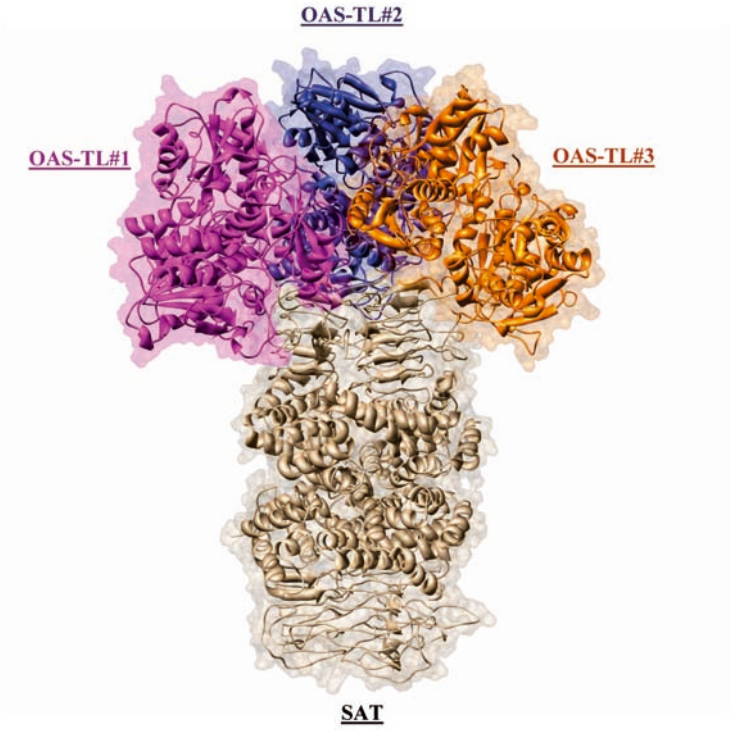


Fig. 21: A model of a plant cysteine synthase complex

reciprocal regulation mechanism of the constituent enzymes are still poorly understood. In conjunction with Rüdiger Hell and Markus Wirtz (Faculty of Plant Science, University of Heidelberg), we are investigating the SAT and OAS-TL enzymes from Arabidopsis thaliana mitochondria and their complexation. We have applied computational techniques to model the three-dimensional structures of the enzymes and are using our protein-protein docking techniques (see Section *Methodological Developments*) to model their complexes. The rigid-body docking protocol, based on Brownian dynamics, was first applied to obtain the encounter complexes. To test the stability of the complexes in an aqueous environment, the complexes were then refined via molecular dynamics. A model of the complex between three OAS-TL dimers and one SAT hexamer is shown in Figure 21; due to its symmetry, a further three OAS-TL dimers may also bind in a similar way to the other side of the SAT.

Signal recognition particle–receptor interactions

The signal recognition particle (SRP) and its receptor (SR) are part of the machinery that mediates the co-translational targeting of membrane and secretory proteins to the membrane of the endoplasmic reticulum (ER) or the bacterial plasma membrane. The targeting of proteins is an essential process and is therefore highly conserved in all organisms. We have used molecular modeling techniques and our protein-protein docking protocol (see Section *Methodological Developments*) to investigate the complex formation between SRP and SR from the thermophilic archea Sulfolobus solfataricus (proteins SRP54 and FtsY respectively). These two proteins have a similar two-domain structure comprising G and N domains. Investigation of the electrostatic properties of these proteins and docking studies suggest that N domains play an important role in the formation of the SRP54 / FtsY complex, even though, in the crystal structure of the bound complex, the main interface is formed by the G domains. Further computational studies with protein mutants have confirmed the possible role of N domains in initial stages of protein-protein complexation and also identified the potential hot-spot residues with a significant impact in the association of these proteins. These predictions are supported by the results of site-directed

mutagenesis experiments performed by Irmi Sinning and Gert Bange (Biochemistry Center, University of Heidelberg).

Thymidylate synthase

Thymidylate synthase (TS) is an essential enzyme for DNA synthesis as it catalyzes a step in the synthesis of the nucleotide thymine. Since all rapidly growing cells require large amounts of nucleotides, TS has proved to be a critical target in cancer therapy. Current TS inhibitors are either substrate or folate cofactor analogs. Unfortunately, resistance to this type of inhibitor occurs, with TS inhibition resulting in increased levels of TS. One reason for this is that the binding of inhibitors to TS disrupts the regulatory binding of TS to its own mRNA, resulting in increased protein synthesis. TS is a homodimeric enzyme, and RNA binding is thought to occur at the interface between the two monomers (Figure 22). We have therefore analyzed the TS dimer interface to identify potential „hot spot" residues that are particularly important for TS dimerization. We are also carrying out molecular dynamics simulations of TS and its mRNA to investigate determinants of TS dimerization and TS-RNA binding. In the LIGHTS project (see Section 2.4) starting toward the end of 2006, the results of these studies will be subjected to experimental investigation and will provide a basis for the design of small molecules to overcome resistance problems of anti-cancer agents.

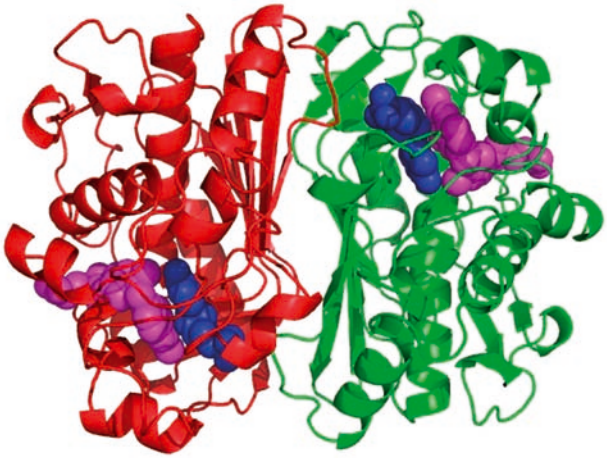


Fig. 22: Structure of the thymidylate synthase dimer with the protein in cartoon representation and a substrate (blue) and an inhibitor (cofactor analogue)(pink) in each of the active sites

3.9 TASSFUN Target-specific Scoring Functions

Project Manager

Dr. Rebecca Wade

Project Member

Dr. Stefan Henrich

The family of trypsin-like serine proteases plays a central role in blood clotting, fibrinolysis, immune response, and digestion. These proteins are thus important targets in drug design. Due to the high structural similarity of the members of this protein family, it is difficult to discover or design inhibitors that bind specifically to one protease but not to others. This specificity is necessary to reduce potential side-effects. In the TArget-Specific Scoring FUNctions (TASSFUN) project, the structures of thrombin, trypsin and urokinase bound to small inhibitors were analyzed with respect to their quantitative structure-activity relationships (QSAR). We applied COMparative BINding Energy (COMBINE) analysis, a receptor-based 3D QSAR analysis method, and developed procedures to aid the design of specific compounds.

COMBINE analysis relies on a training set of structures of protein-inhibitor complexes with accompanying bioactivity values, such as inhibitor constants. For these complexes, the electrostatic and van der Waals interaction energies are calculated between the inhibitor and each protein residue. In addition, electrostatic desolvation energy terms are computed by solving the Poisson-Boltzmann equation for the protein and the inhibitor in each complex. The decomposed interaction energies and the electrostatic desolvation energy terms are correlated by Partial Least Squares (PLS) to bioactivity or binding free energy values. Subsequently, the derived target-specific scoring function is used to predict the bioactivity of inhibitors for which no experimental values are available.

In this project, an automatic workflow suitable for the modeling and analysis of a large number of receptor-ligand complexes was constructed. Target-specific scoring functions were generated for each of the three serine proteases studied. These provided insights into favored and disfavored receptor-ligand interactions and highlighted the importance of certain amino-acid residues. The target-specific scoring functions were used to predict the binding affinity of a test set of ligands that were computationally docked into the binding sites of the three protein targets.

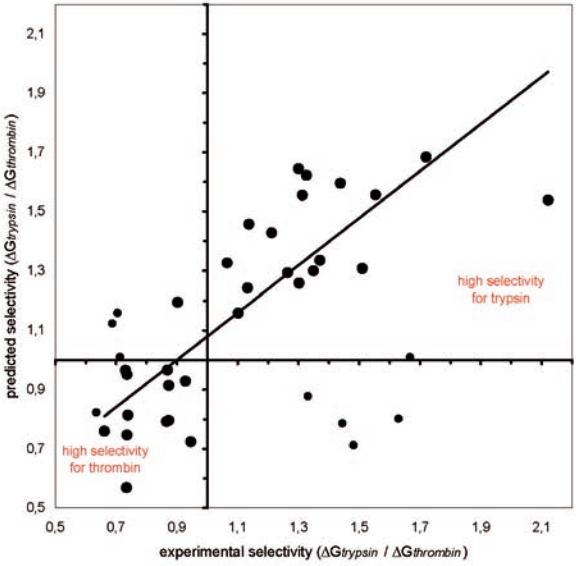


Fig. 23: COMBINE analysis of protease-inhibitor complexes. Predicted vs.experimental selectivity of inhibitors for trypsin vs. thrombin. Outliers (small dots) are due to compounds for which the binding affinity prediction deviates by more than 3 log units from the experimental value. The regression line is for the large dots (R=0.69)

Criteria were derived to rank the predictions for different docking poses and optimize the reliability of the predicted binding affinity. A comparison of the COMBINE models permitted prediction of binding selectivity (Figure 23).

Collaboration Partners

Dr. Niklas Blomberg, Dr. Isabelle Feierberg
(AstraZeneca R&D, Mölndal, Sweden)

Sponsor

AstraZeneca

3.10 Modeling Macromolecular Motions in the Cell by Brownian Dynamics Simulations

Project Manager
Dr. Razif Gabdoulline

Macromolecular motions and interactions in the cell are essential events in cellular life and occur on a variety of time scales. Processes like macromolecular diffusion and transport, many types of protein-protein interactions, and protein domain rearrangements occur on timescales of milliseconds and longer. These processes cannot be described by standard molecular dynamics (MD) simulation methods. Brownian dynamics (BD) simulation is one of the methods that permit simulation of macromolecular motions on the millisecond time-scale while preserving atomic-level accuracy in the representation of the molecules, albeit usually neglecting the internal dynamics of macromolecules.

SDA, a software suite for the Simulation of Diffusional Association (http://projects.villa-bosch.de/mcm/software/sda), permits the simulation of the relative diffusional motion of two atomically detailed macromolecules to compute association rate constants and study encounter-complex formation. The goal of this project is to develop this software and the methodology so that the diffusional motion of many macromolecules or large proteins consisting of rigid domains connected by flexible linkers can be simulated with a force field based on the atomically detailed protein structure. In order to obtain realistic protein interaction times in these simulations, it is necessary to model forces resulting from hydrophobic interactions.

In addition to investigating how to correctly implement nonpolar interactions, two other topics were studied in detail this year. One is the development of efficient reduced models for simulating long timescale phenomena. In these models, the interaction between proteins has a simplified form, with parameters reproducing the interaction properties derived from all-atom representation of proteins with a tunable degree of accuracy. The other topic was the optimization of the sampling of protein configurations. A biasing method was developed to reduce simulation time by avoiding repeated sampling of configurations.

Collaboration Partner Claudia Mühle-Goll (EMBL/University of Mannheim)

Sponsor BIOMS Center for Modeling and Simulation in the Biosciences, Heidelberg

3.11 Multiscale Approach to Biomolecular Interactions: From Molecular Dynamics to Brownian Dynamics Simulation of Chromatin Components

Project Manager
Dr. Rebecca Wade
Project Member
Georgi Pachov

A major challenge in the simulation of biomolecular interactions is to find ways of transferring the information obtained from atomic-detail molecular dynamics (MD) simulations, which can be performed on time-scales up to $\sim 10^{-9}$-10^{-7} s and for studying conformational changes on the $\sim 10^{-10}$-10^{-9} m length scale, to models that can be used to understand biological processes occurring on longer length and time scales (e.g. 10^{-9}-10^{-6} m and up to 10^{-3}-1s). This strategy is of particular importance for understanding the dynamic organization of the genome in the cell nucleus, where it is known that phenomena occurring on the local atomic scale (e.g. protein and DNA modifications) can influence global genome organization. The aim of this project is to develop multiscale simulation methods by combining molecular dynamics and Brownian dynamics techniques, and to apply them to studying chromatin. Chromatin is a complex of DNA and protein found in cell nuclei. It has a hierarchical structure enabling the genomic DNA to be packed into a small volume in the cell (see Figure 24).

The project started in April 2006 and this year's efforts have been devoted to studying electrostatic interactions within the nucleosome and to simulating the docking of linker histone proteins to the nucleosome by Brownian dynamics techniques. The results of electrostatic calculations based on the atomic-detail structure of the nucleosome and an implicit model of the solvent are being used to derive suitable descriptions of electrostatic interactions for coarse-grain simulations of DNA wrapping and unwrapping from the histone core.

Collaboration Partners

Prof. Jörg Langowski
(German Cancer Research Center, Heidelberg)

Prof. Jeremy Smith
(IWR, University of Heidelberg)

Sponsors

German Research Foundation (DFG)

Klaus Tschira Foundation

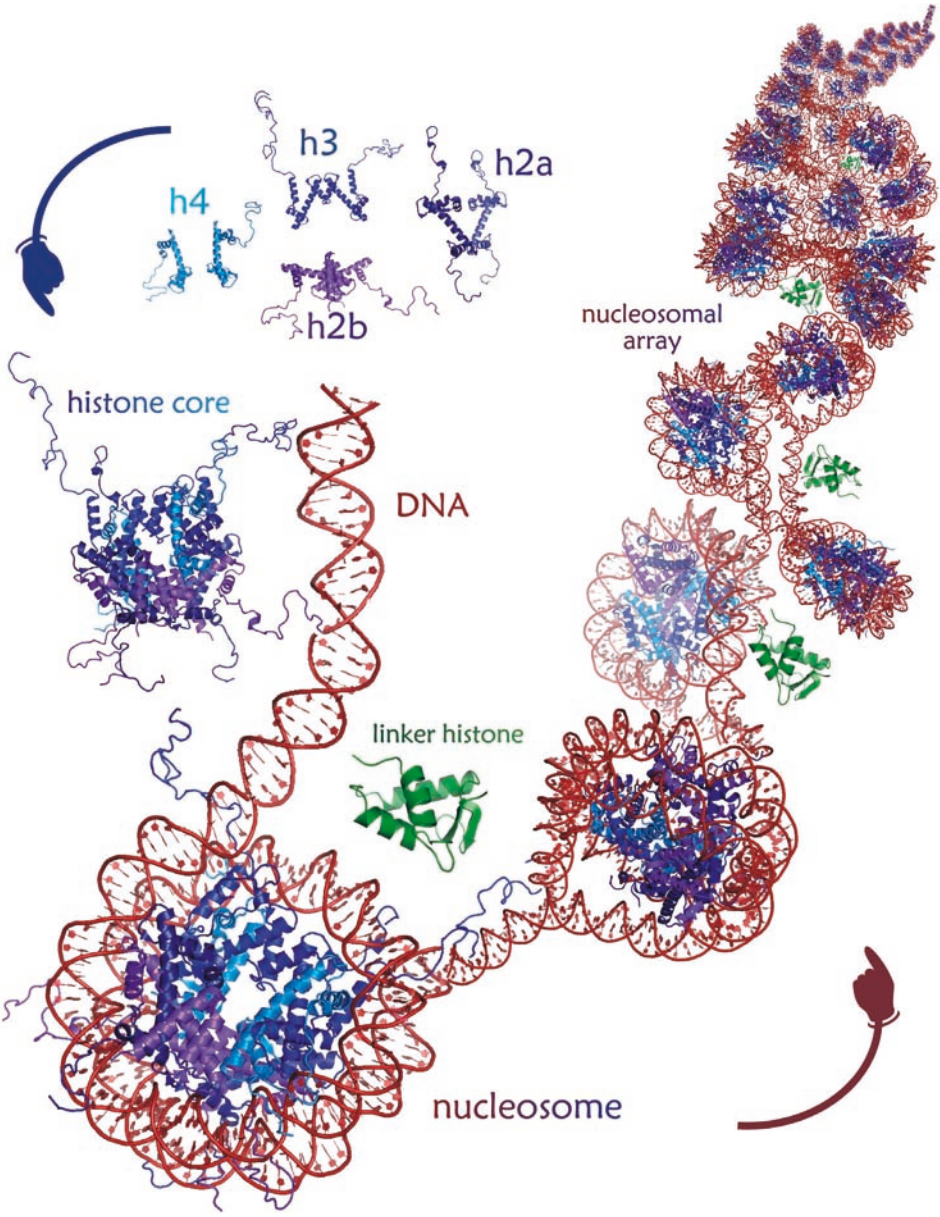


Fig. 24: Schematic picture showing the hierarchical structure of the chromatin fiber

3.12 DIANA-Summ (NLP)

The aim of DIANA-Summ is to automatically generate the minutes of meetings. We consider this task to be similar to that of automatic summarization. Instead of developing a summarizer for meetings, i.e. for verbal exchanges between different participants, we focus on preprocessing techniques designed to enable meetings to be dealt with by existing summarization techniques.

Deutsche Forschungsgemeinschaft
DFG

DIANA-Summ is funded by the DFG (German Research Foundation). We have just received confirmation of funding for a third year (Nov. 2006 - Oct. 2007). This will enable us to complete the project and notably to finalize the anaphora resolution component, one of the core research topics addressed by DIANA-Summ.

Automatic Disfluency Detection

After completing the manual annotation of disfluencies in 2005, we were able to explore methods for automatically detecting disfluencies and categorizing them. To this end we used combinations of pattern-matching algorithms and machine learning methods (decision trees). The pattern-matching algorithm was used to cover non-lexicalized filled pauses (nlfp), verbatim repetitions (repet), and abandoned words (abw). The manually annotated data served to train a decision tree to automatically annotate the remaining categories of disfluencies (lexicalized filled pauses (lfp), repairs (repai), and abandoned utterances (abutt)). We achieved the best results by using features

Fig. 25: Results of automatic disfluency detection and annotation

	total (in data)	correctly detected by the system	wrongly detected by the system	precision	recall	f-measure
nlfp	8110	7404	2506	74.71	91.92	82.18
lfp	8489	6881	1737	79.84	81.06	80.45
abutt	4568	2198	1541	58.7	47.94	52.78
abw	2531	2275	989	69.7	89.89	78.52
repai	6620	3949	2494	61.29	59.65	60.46
repet	5156	3677	1347	73.19	71.19	72.18

based on speaker information, position of the word in the utterance, and information on preceding disfluencies in addition to POS. Our best configuration produced the results presented in Figure 25.

Automatic Detection of Topic Boundaries

Project Manager

Dr. Michael Strube

Project Members

Margot Mieskes
Christoph Müller

Students

Vanessa Doyle, Andrew Herd, Melissa Macellano, Violeta Sabutyte, Irina Schenk, Ganna Syrota, Grainne Toomey, Wolodja Wentland, Matthew White, Florian Winkelmeier

Meetings normally deal with several topics. In summarizing them, it is important to segment the meetings into coherent parts dealing with a single topic. Building on a manual annotation for topic boundaries we developed a method for automatically detecting topic boundaries. In a next step we tested several baselines for annotating topic boundaries. The baselines were random, even, and all. A certain number of boundaries in the meetings were placed randomly, at equal distances, or at all possible junctures (in this case, between all segments). Additionally, we evaluated a state-of-the-art system (g03). The results obtained by using the evaluation metric Pk are shown in Figure 26.

Pk takes into account the fact that topic boundaries can be correct even if they do not match exactly. This means that topic boundaries situated within a certain range should still be considered correct. In the state-of-the-art system (g03) the likelihood of a boundary being wrong is 20%.

In our work we experimented with a method that looks for connections between two windows of meeting data in order to determine topic boundaries. The size of the windows were determined experimentally. Additionally, we tried to find out which kind of preprocessing method produced the best results. The preprocessing methods applied were segmentation (either segments or spurts, for more details see last year's report), morphological analysis (stemming or lemmatization), and filtering for stopwords (stopword lists for speeches or texts). Stopword lists contain words that occur very frequently either in texts or in speeches or both (e.g. "the") but do not carry information. We found that the best results were achieved with a window size of 40 seg-

	random	even	all	g03
Pk	48.9	55.1	82.3	20.4

Fig. 26: Results of common baselines and current state-of-the-art method

ments, using stemming and filtering with a stopword list for texts. These results are significantly better than the results obtained with the state-of-the-art method cited above.

Summarization Annotation

Three human annotators were asked to mark the elements (segments) of the meetings they considered relevant for a summary. The annotators examined sections of the meetings between two topic boundaries, based on manual topic annotation. The annotations of all three annotators were then used to create a gold standard for the summary. The gold standard contained all items marked by at least two annotators. These data serve as the basis for the automatic identification of items relevant for the summaries of the meetings. These automatic methods will be further explored in the near future.

EvalMeasure	Result
Pk	17.4

Fig. 27: Results of our topic boundary detection method

Improving Summarizations Using Human Interaction

We have developed a plug-in for the MMAX2 tool (see Section 3.12) that makes it possible to read and browse the summarizations. The plug-in is shown in Figure 28.

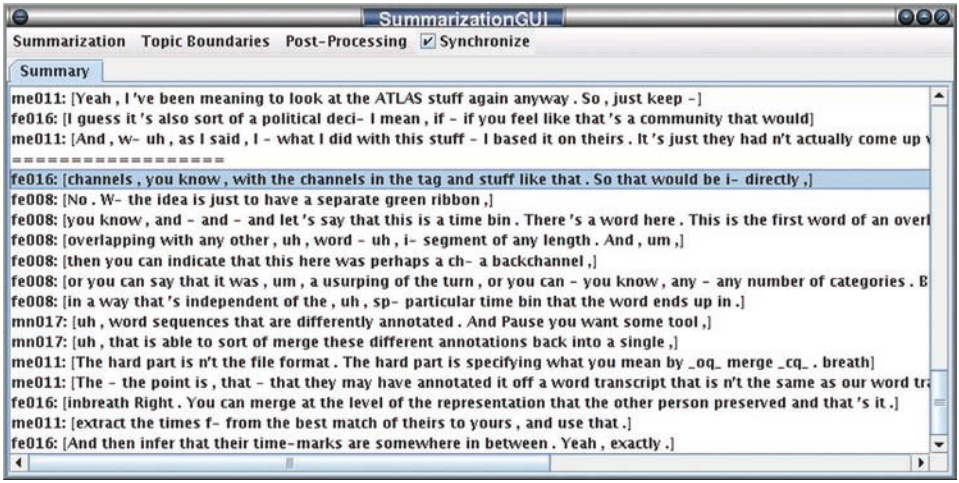


Fig. 28: MMAX2 summarization plug-in

Figure 29 shows both the information extracted manually as being relevant for a summary and the speaker and topic boundaries.

Users can mark certain segments and then read the context of the specified utterance in the main MMAX2 window. They can also change the categorization of utterances from relevant to non-relevant or vice versa. This is very similar to the way in which the annotators selected the summary-relevant items in their original annotations. The aim is to give users the opportunity to train the summarization system to the specific environment the system is used in. The system will contain a basic summarization system that is trained on our data and can be improved by everyday usage.

Anaphora Resolution

Anaphora resolution in the DIANA-Summ project focuses on the personal pronoun *it* and the demonstrative pronouns *this* and *that*. One reason for this is the high frequency of these words in the corpus. Figure 30 shows the absolute and relative frequencies of the most frequent words in the ICSI Meeting Corpus.

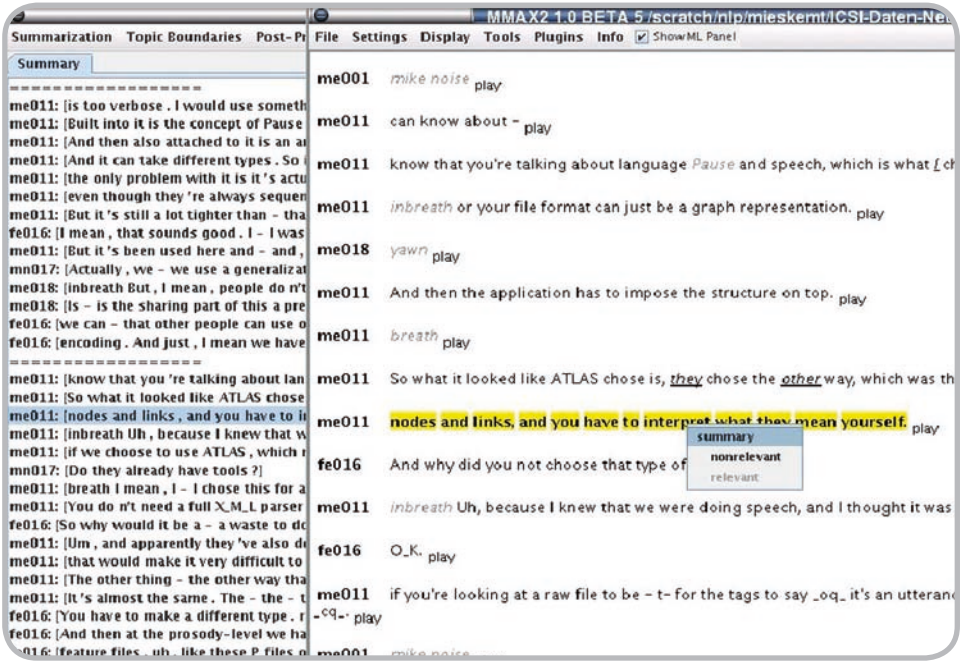


Fig. 29: MMAX2 summarization plug-in, speaker and topic boundaries

Rank	Token	Abs. Freq.	Rel. Freq.
1	the	33107	3.99
2	I	24583	2.96
3	**that**	**22220**	**2.68**
4	**it**	**21245**	**2.56**
5	and	19675	2.37
6	you	18881	2.27
7	´s	16969	2.04
8	to	16320	1.97
9	a	13984	1.69
...	...	...	...
21	**this**	**7131**	**0.86**

Fig. 30: Most frequent words in the ICSI Meeting Corpus

It, *this*, and *that* figure very frequently in Figure 30 Together, they account for more than 6% of all words in the corpus. It needs to be noted, however, that the counts for *this* and *that* also include cases where they are not demonstrative pronouns but determiners or relative pronouns. The percentage of demonstrative pronouns is however still considerable.

Another reason for concentrating on pronouns rather than, say, anaphoric noun phrases is the type of application into which anaphora resolution is embedded within the DIANA-Summ project, i.e. extractive summarization. In the context of extractive summarization, it is reasonable to make individual utterances in the corpus less context-dependent by replacing the pronouns that figure in them by some explicit representation of the antecedent (e.g. a noun phrase or a proper name). This can be expected to have a positive influence both on the extraction of relevant utterances and on the readability of the final summary.

Not all instances of *it*, *this*, and *that* as anaphoric references can be resolved in this way. Pronominal instances of *this* and *that* have to be distinguished from determiners and relative pronouns. This task can normally be performed by automatic part-of-speech (POS) taggers. For *it*, the situation is different, as the distinction between so-called referential (i.e. resolvable) and non-referential it requires more sophisticated methods. We have developed a machine learning system for the detection of non-referential *it* in spoken dialog [Müller 2006a]. The system is based on a corpus of more than 1,000 manually classified instances of *it*. In a first step, the classifier was trained on a subset of the annotated data in order to learn the distinction between both types of *it*. Each instance of *it* was represented as a vector of features describing the context in which the instance of *it* occurs. Among the features used were the syntactic construction in which *it* occurs, the verb that governs *it*, and distances to various types of words in the immediate proximity. In addition, each feature vector was also tagged as either referential or non-referential.

The resulting classifier consisted of a set of automatically acquired rules for detecting non-referential instances of *it*. One of these rules is that an instance of *it* should be treated as non-referential if the word *to* or an adjective occur within a certain range of words following *it*.

```
dist_to_next_to <= 8 and
dist_to_next_adj <= 4
        ==> class = nonref (53.0/16.0)
```

This rule correctly identifies instances like the following as non-referential:

... *it's very easy to whip up something quickly* ...

The performance of the classifier can be evaluated by comparing its automatic predictions with the manual annotation. The best result achieved in the context of the work described in [Müller 2006a] was a precision of 80.0%, a recall of 60.9%, and a resulting F-measure of 69.2%. This

level of performance, and notably the high degree of precision, makes it possible to employ the system as a preprocessing filter in spoken dialog pronoun resolution.

While pronoun resolution for written (especially newspaper) texts is now an established discipline in computational linguistics, it is still much less common for spoken dialog. One of the reasons may be that pronoun resolution in spoken dialog is a more formidable task because spoken language is generally more difficult to process. More importantly, it tends to contain a higher proportion of vague and ambiguous pronouns.

In the DIANA-Summ project we have employed four naïve, project-external annotators (two non-native and two native speakers of English) to annotate anaphoric relations for referential instances of *it*, *this*, and *that*. One reason for employing more than one annotator was to check inter-annotator agreement, i.e., we wanted to investigate how reliably this type of annotation can be done in spoken dialog. A simple estimation of agreement can be made just by looking at the number of anaphors and antecedents identified by all four annotators. The following Figure 31 contains the percentage of anaphors and antecedents in each of the five annotated dialogs that were identified by all four annotators. The first column contains the overall figure, the second the figure for pronouns only, and the third the figure for non-pronouns only.

	all	pronouns	non-pronouns
Bed017	27.46%	54.34%	6.70%
Bmr001	31.50%	57.37%	5.21%
Bns003	24.76%	49.78%	6.07%
Bro004	20.20%	39.75%	4.15%
Bro005	24.91%	43.95%	8.16%

Fig. 31: Inter-annotator agreement on resolving anaphoric relations

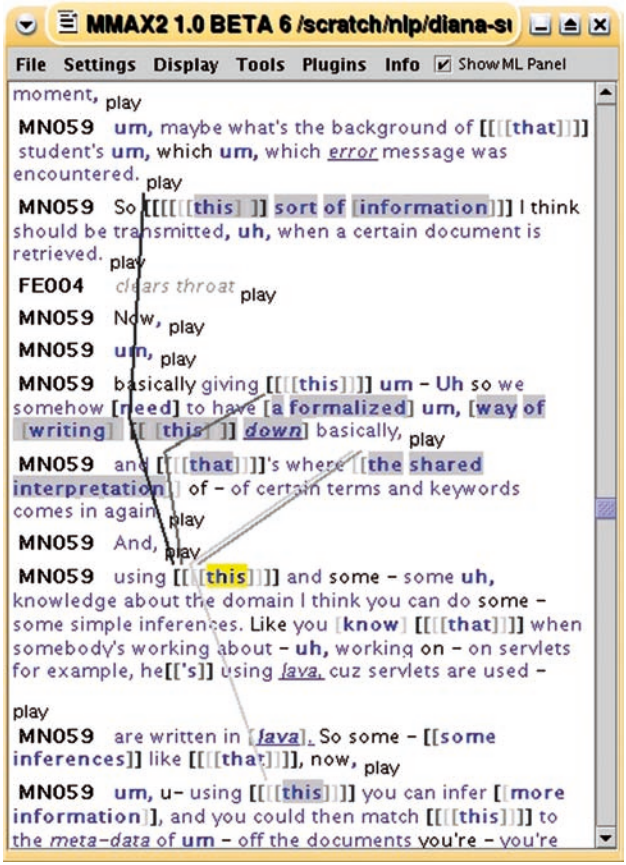


Fig. 32: A case of disagreement in anaphoric annotation

It is apparent that the overall agreement on even the mere identification of expressions is rather low. For non-pronouns in particular, the annotators agreed in only about 6% of all cases. This in itself is a useful indication of relative unreliability in the identification of noun phrases and verb phrases as antecedents of anaphoric pronouns. Finally, the merely quantitative analysis can be supplemented by a qualitative impression. Figure 32 is a screen shot of the annotation tool MMAX2 used for all annotations in the context of the DIANA-Summ project. It shows a currently selected pronoun (*this*). The differently colored lines show the different anaphoric relations annotated for this pronoun by the four annotators individually.

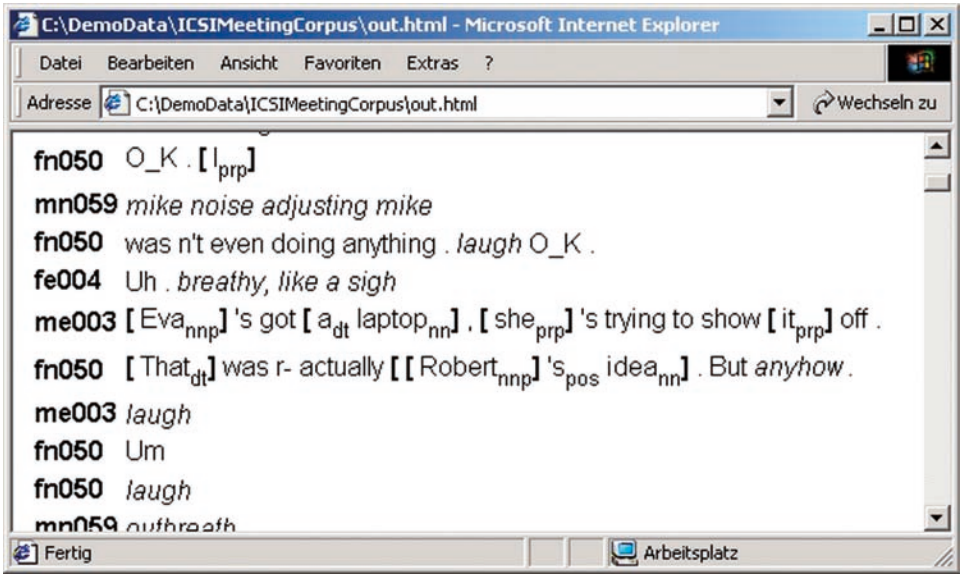
3.13 MMAX2 (NLP)

Project Manager
Dr. Michael Strube
Project Member
Christoph Müller

The MMAX2 annotation tool has been developed further in two respects. First, a mechanism has been added for the XSL-based transformation of MMAX2 corpora [Müller 2006b]. Corpus transformation may be necessary if annotations have to be converted from their original source format into a different target format. This target format may be the data format required by another annotation tool or by some other program. The tool's style sheet engine now provides various special functions for handling stand-off annotation as realized in MMAX2. These functions are integrated into the style sheet engine in such a way that they can be used like normal XSL functions. Figure 33 shows an HTML document created from a MMAX2 annotation by means of an XSL style sheet.

Another improvement to MMAX2 is the newly added support for plug-ins. A plug-in is a user-specified Java class that can be accessed from within the tool via the Plug-ins menu. For a Java class to be usable as a MMAX2 plug-in, it has to be derived from the class org.eml.MMAX2.MMAX2Plugin. A MMAX2 plug-in has access to the data currently loaded in the tool. In particular, the plug-in can

Fig. 33: HTML view of an MMAX2 annotation



modify the data to make automatic annotation possible. To illustrate this usage we have developed a Named-Entity recognizer plug-in. This plug-in can be used to automatically find and classify all occurrences of so-called named entities (e.g. persons, companies, place names, etc.) in a text. The plug-in stores its output (i.e. all identified entities) in the form of markables at a user-specified annotation level. This format is the same as the one used in the manual annotations, thus making automatic annotation by means of plug-ins fully transparent. This also facilitates subsequent manual correction.

A MMAX2 plug-in was also used to integrate the MMAX2 tool into the summarization component of the DIANA-Summ project. By making the summarization component executable from MMAX2, the tool's GUI can be utilized for the manual correction phase (see Section 3.12).

3.14 Natural Language Generation (NLP)

Researcher

Katja Filippova

To look at a text the best way to understand which lacks this property what coherence means is and confuses everything: the logical structure, pronominalization and word order. By ordinary people is coherent produced every text.

Every text produced by ordinary people is coherent. The best way to understand what coherence means is to look at a text which lacks this property and confuses everything: the logical structure, pronominalization, and word order. Just like the paragraph above.

Coherence has many facets. Some of them are text-external, others are text-internal. For example, it is our non-linguistic knowledge about the simple routine of waking up first and having breakfast afterwards that makes us perceive one text as normal and the other one as weird: "Jane woke up early this morning and had only a cup of coffee for breakfast" as opposed to "Jane had only a cup of coffee for breakfast and woke up early this morning." The linguistic properties ensuring coherence include the use of pronouns, the order of words in sentences, and splitting a text into discourse units. Textual coherence is crucially important because deficiencies in this respect can not only lead to difficulty in understanding texts but also interfere with their intended meaning.

We are developing methods and algorithms that can improve the coherence of texts generated both by computers and by humans. Our investigations largely relate to German, but most of our findings can be applied to other languages as well. The research done (a) on word order and pronominalization, and (b) on discourse segmentation is reported in [Filippova 2006a] and [Filippova 2006b]. The experiments described there involved judgements by native speakers of German and suggested an implementation of our ideas that utilizes machine learning techniques.

Our other main research focus is on automatic multi-document summarization performed on a corpus of biographies. Suppose you would like to know more about Albert Einstein and the early period of his life. Where was he born? Who

were his parents? Where did he study? Although you have a fairly good idea of what you are interested in, it might still be hard to sift out the relevant information from the huge number of biographies available and the welter of facts they contain about Einstein's life. Even if you know that the information you require must be in there somewhere, it may well be scattered over various different texts, thus making the search extremely tedious and time-consuming. Some facts figure in all of the sources, others are dealt with more or less exhaustively in the various texts available. For example, one biography may tell you where Einstein studied, while another tells you the exact period in question. Ideally, you might want to have a plug-in built into the search engine you use that would provide a brief summary of the information you are interested in taken from the whole set of relevant documents. This is the configuration for the summarization system we are working on at EML Research. First, we calculate the importance of facts found in biographies by looking at their relative frequency. Some information, like the date of birth or death, is ubiquitous and therefore highly important. Other information is person-specific and can be detected by its high frequency in biographies about the same person. Second, we classify sentences according to the aspect of the person's life they deal with – for example family, work or social activity – and the period they refer to. These classifications simplify the search required to process a user's query. We also align different texts and investigate how different aspects of the same fact can be assembled in one sentence. This merging or fusion saves space and provides relevant information with a minimum of words.

Before the summary is given to the user its overall coherence needs to be improved. This is where our two lines of research converge.

3.15 Word Sense Disambiguation (NLP)

Researcher

Dr. Vivi Nastase

Because language is such an integral part of our identity as humans, we fail to realize how much processing is required to comprehend a simple sentence. Part of the problem is relating the words we hear or read to concepts we are familiar with. Within the domain of Natural Language Processing, this task is referred to as "word sense disambiguation", i.e. distinguishing between the different meanings of homonyms like "king" ("male monarch" or "the most vulnerable but most important chess figure") or more subtle differences between "book in the sense of a piece of published written work" or "book as a physical object, e.g. used as a doorstop". Teaching computers how to select the appropriate meaning in a given case is a formidable task. Researchers use different methods to address this problem, but the underlying idea is common to them all: „You shall know a word by the company it keeps", (Firth 1957), meaning that the context in which a word appears provides the clues that tell us what the word means. In our experiments we use a method that collates knowledge about the meanings of words, i.e. word senses: examples of language usage listed in the British National Corpus, syntactic information enabling us to identify the contextual elements that influence the intended meaning, and a PageRank style algorithm that enables us to determine the different meanings of each word in a text by highlighting those that combine best with others.

3.16 Exploiting Wikipedia for Research in NLP

Researcher

Simone Paolo Ponzetto

How many times have you interacted with a machine designed to support some intellectual task and ended up totally frustrated, complaining that, this machine is really stupid'? The problem is simply that machines that process information need knowledge, a fact that has been insisted upon since the very dawn of Artificial Intelligence research. John McCarthy, for instance, in his influential paper "Programs with Common Sense" (1958) pointed out that machines need knowledge to simulate intelligent behavior in much the same way as humans exhibit common sense. This includes knowledge of facts such as "birds are creatures with feathers and wings", "birds can fly", "glass breaks easily" and so on. Early AI research concentrated on mak-

ing knowledge explicit in the form of manually encoded, machine-readable knowledge bases (KBs). Unfortunately, KBs are expensive and time-consuming to build and maintain. Also, their coverage is both limited and arbitrary. This eventually gave rise to the widespread use of statistical and machine learning techniques in AI and NLP. But while advances towards robust statistical inference methods will certainly improve the computational modeling of natural language, we believe that crucial advances will also come from the rediscovery of the use of symbolic knowledge, i.e. the deployment of large-scale KBs.

In our work we attempt to overcome the shortcomings of KBs by drawing upon Wikipedia, a wide-coverage, online encyclopedia developed by a large number of users. More precisely, we have addressed the question of whether and how Wikipedia can be integrated into NLP applications as a knowledge base. We started on the assumption that the articles in the encyclopedia represent concepts. We then took advantage of the fact that Wikipedia allows for structured access by means of categories: articles can be assigned to one or more categories, which are then further categorized to provide a so-called "category tree" that can be used as a (collaboratively generated) taxonomy. We use this taxonomy to compute how many linguistic expressions (e.g., nouns such as "kitchen", or proper names such as "George W. Bush") are closely related. We compute the relatedness of words by retrieving the Wikipedia articles that describe them and computing how distant from one another the articles are in the taxonomic categorization (Figure 34). This level of information, referred to in the literature as *semantic relatedness*, is known to be useful for many NLP tasks.

We have demonstrated not only that Wikipedia-based measures of semantic relatedness are competitive with the ones induced from a hand-crafted, widely used standard resource such as WordNet, but also that including semantic knowledge mined from Wikipedia into an NLP system dealing with coreference resolution - that is, a system identifying different textual expressions as referring to the same

entity in the world, e.g., “George W. Bush” and “the president of the United States” - is in fact useful. Our work is one of the first attempts in Natural Language Processing to make extensive use of the “encyclopedia that anyone can edit.” Empirical findings have shown that Wikipedia can be considered a semantic resource in its own right and that it can be used equally profitably in NLP applications as hand-crafted taxonomies. In fact, Wikipedia is a resource that will repay further exploitation for NLP research. Apart from its very wide coverage, the information it contains is always up-to-date. For instance, the “North Korea” entry contains information about the 2006 North Korean nuclear tests. This information was entered at almost exactly the same time that the news spread though the standard mass-media channels. In the near future we plan to investigate how to automatically induce a fully-fledged ontology from a semi-structured encyclopedic resource such as Wikipedia, thus proposing Wikipedia as a direct competitor of artificially-designed taxonomic resources like WordNet.

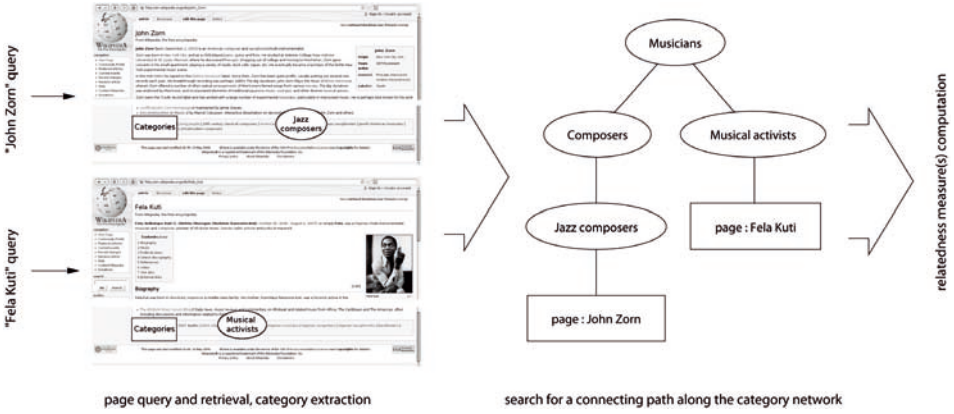


Fig. 34: Computing semantic relatedness using Wikipedia

3.14 MORABIT

Project Manager

Prof. Dr. Andreas Reuter

Acting Project Manager

Dr. Rainer Malaka, European Media Laboratory (January-October 2006)

Project Member

Matthias Merdes

Student Worker

Dirk Dorsch

In the MORABIT project we have investigated new methods for testing component systems at run-time. Special attention has been given to properties found in mobile systems, i.e. the inherent scarceness of resources such as computing power and network bandwith and the dynamic nature of such systems. While the latter restricts the application of traditional forms of testing, the former places special requirements on run-time testing.

To gain new insights in this area it was necessary to develop both new concepts and a special kind of component middleware. Conceptual advances include the improvement of the understanding of run-time tests in mobile and ad-hoc systems and the influence of limited resources on these tests. The new component middleware is essentially a run-time infrastructure for software components that enables resource-aware execution of run-time tests. This infrastructure plays an important role in validating the conceptual and methodological contributions.

A software component is essentially an encapsulation of some functionality serving as a building-block for composing larger software systems. The components work together in well-defined ways to achieve a certain functionality. Generally speaking, components can play two roles. They either provide services to other components (server role) or they use the services of other components (client role). Ideally, these building-blocks can be exchanged easily without compromising the composite system, as the dependencies between the individual components are clearly and formally specified.

In addition to this traditional characterization, MORABIT components also possess test definitions and metadata describing the relationships between the core components and the tests. These tests are intended to check whether a potential service-providing component has some important property or behavior for the component requiring the test. Traditionally, test definitions take the form of well-known programmed test cases, e.g. in the popular “JUnit” testing framework. Such defect tests are performed at develop-

ment time and serve to uncover problems in the software before it is actually used in real life. By contrast, run-time tests like the MORABIT tests do not aim at fixing bugs in the software but at letting individual components assess the reliability of other components during the execution of the program. As the test cases are bundled with the core components, they are often called "built-in tests" (BIT). The necessary component metadata includes information both on when and how to execute the test cases of a component and on how to react to the outcome of the execution of these tests.

After the development of an initial prototype for the MORABIT infrastructure in the first half of the project, the work in 2006 concentrated mainly on the following goals:

- improving the understandability of MORABIT technology for non-experts,
- increasing the stability and usability of the MORABIT infrastructure,
- applying the basic principles of run-time testing [Suliman 2006a] to special areas such as reliability of ubiquitous software systems [Merdes 2006a] and verification of component-based systems [Brenner 2006], and communicating the results to the research community.

These activities serve the purpose of making the research results and technology accessible to the general public, software developers interested in applying the MORABIT technology, and the scientific community.

To communicate the benefits of the MORABIT approach to a wider audience, we have developed a small number of demonstration scenarios. These are very simple examples of MORABIT applications with a clear focus on one particular aspect of the technology. They illustrate the kind of problems MORABIT is intended to solve and how these problems have been addressed. These examples have been implemented as MORABIT components and deployed to the run-time infrastructure. With the help of the various visualization tools included in the infrastructure, it becomes

feasible to illustrate how the respective problem is solved by the infrastructure. The first scenario shows how errors in the business logic can be avoided by having the infrastructure select a superior server component based on the results of run-time tests. This demonstrates the key benefit of run-time testing. The second example deals with a problem that is specific to run-time testing: the isolation of the test execution from the "normal" functionality. It shows how this issue can be approached by cloning (copying) components in an appropriate way before testing. In the last example, we demonstrate how the "normal" functionality of an application can be protected from the effects of test execution by means of resource-aware test execution. Demonstrations of these examples, e.g. to project reviewers, indicate that examples with a focus on a single aspect of the technology are better suited to improving the understanding of MORABIT than large complex applications showing many features at the same time.

A number of steps have been taken to increase the usability of the MORABIT infrastructure for software developers. These include maintenance of the infrastructure in the form of bug fixes and many small improvements, such as better configuration definition and validation support. Also, new features have been introduced, most notably the new Component Interaction View. This new view is realized by integrating the previously developed S4J sequence diagram visualization tool [Merdes 2006b] into the MORABIT Runtime Inspector. With the help of this feature it is possible to visualize the business logic of the current application in the form of a dynamic UML sequence diagram. As such diagrams are widely used in common software engineering activities, such as requirements analysis, design, and documentation, typical users of the infrastructure can be expected to be familiar with this form of visualization. An example of such a visualization showing an interaction within the auction house demonstration domain can be seen in Figure 1. In addition to these enhancements, comprehensive technical documentation of the run-time infrastructure has been compiled. This developer handbook describes many technical aspects of the infrastructure and its use in great detail [Merdes 2006c].

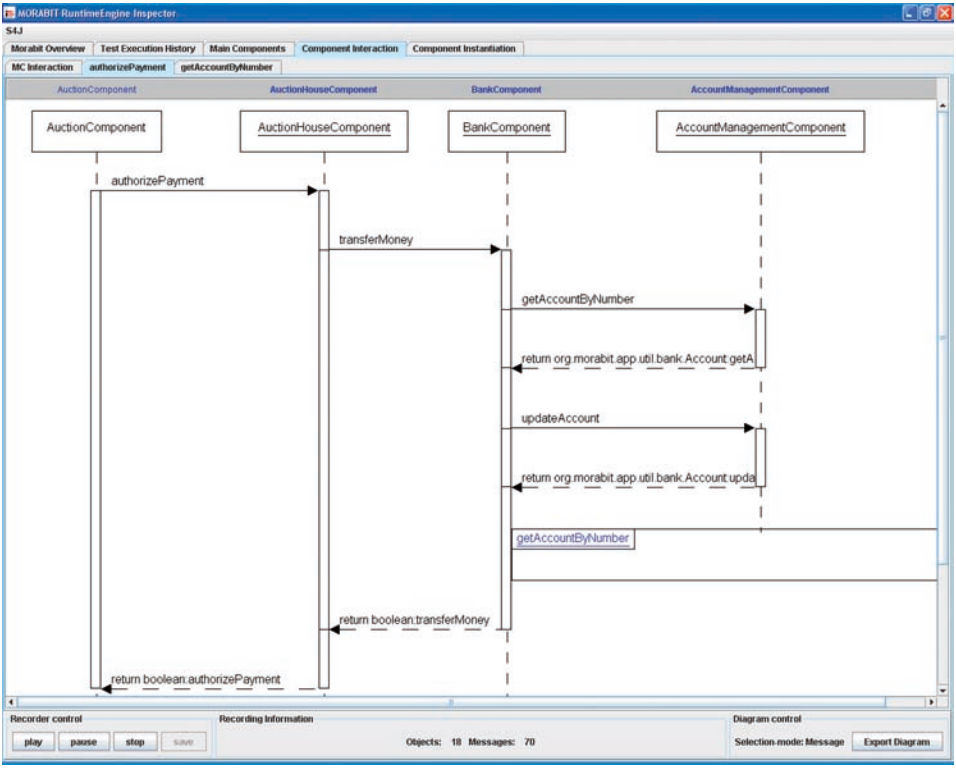


Fig. 35: Component Interaction View of the MORABIT Runtime Inspector showing an example interaction from the auction house domain

The remaining goal of the project was to communicate the technical and scientific results to the research community. This not only helps to make the MORABIT approach better known but also provides us with valuable peer feedback. To this end we applied the MORABIT principles to two special areas. In [Merdes 2006a] we describe this application to the domain of ubiquitous software systems. Such systems are highly dynamic in nature and thus cannot be subject to traditional integration testing. Furthermore, they are typically deployed to devices with fairly limited resources. These properties make ubiquitous software systems an ideal application area for the MORABIT principles and technology. While the selection of possible servers based on run-time test results improves the reliability of the resulting system, great care must be taken during the execution of these run-time tests in order not to consume too many resources. This is achieved by utilizing one of a number of resource-aware

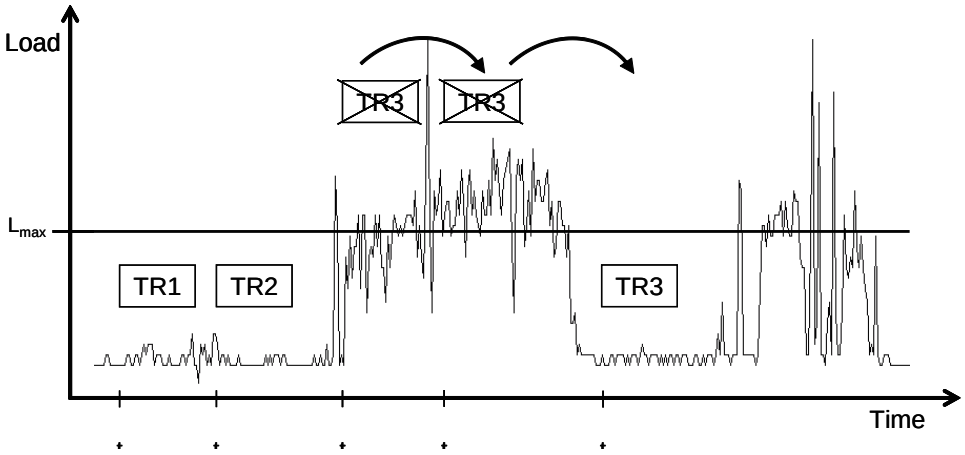


Fig. 36: Visualization of a resource-aware test execution strategy with three test requests (TR1–TR3) and load threshold (Lmax) under changing load conditions

test execution strategies. As an example, the application of one such strategy is visualized in Figure 36. It shows how the execution of test requests occurring at different points in time can be optimized with respect to the changing resource situation. In this example, the postponement of the test is triggered when the given resource (here: the processor load) exceeds a certain threshold value.

In [Brenner 2006], the MORABIT technology was also applied to quite a different situation, the area of component-based enterprise software systems. Here we argue that many of the benefits of component-based development are lost during the system integration and verification phases. With the application of MORABIT built-in test technology, this verification process can be partly automated, resulting in higher reliability and lower effort and thus cost. The successful specialization of the core technology in these two very different areas demonstrates the versatility of the MORABIT approach. These applications in special areas help to communicate the MORABIT approach to the technical community. This is complemented by the improvement of the run-time infrastructure described above and its accompanying technical documentation for software developers, as well as the development of pedagogical demonstration examples for a wider audience.

Project Partners

Prof. Dr. Barbara Paech, Chair for Computer Science, University of Heidelberg

Prof. Dr. Colin Atkinson, Chair for Computer Science, University of Mannheim

Sponsors

Landesstiftung Baden-Württemberg

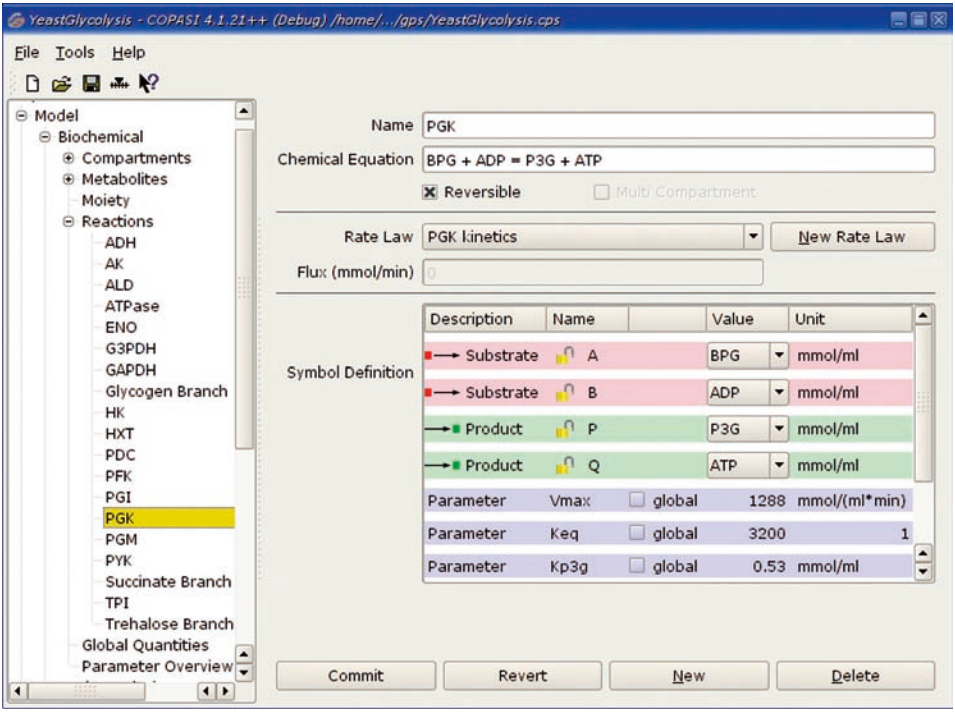
Klaus Tschira Foundation

Copasi (BCB Group)

Copasi is a software for the modeling, simulation and analysis of biochemical networks.
It is developed as a joint project with the group of Pedro Mendes at the Virginia Bioinformatics Institute, VA, USA.

- URL: http://www.copasi.org
- Downloads (2006): ca. 2300
- Users all over the world (Americas, Asia, Europe, Australia, Africa)

Fig. 37: Screenshot of the Copasi interface



MMAX2 (NLP Group)

MMAX2 is a multi-level annotation tool for creating, browsing, visualizing and querying linguistic annotations on multiple levels. It uses an XML-based stand-off annotation format to allow for the trouble-free coexistence of any number of annotation levels in a single document.

Different forms of licenses (for academic teaching and academic and commercial research) are available.

- URL: http://mmax.eml-research.de
- Academic and commercial research licenses (2006):
  3 single licences,
  4 multiple licences, 2 teaching licences
- Users in Germany, France, UK, the Netherlands, Italy, Russia, USA, Brazil, Singapore, and Japan

PIPSA (MCM group)

PIPSA (Protein Interaction Property Similarity Analysis) 2.0 may be used to compute and analyze the pairwise similarity of 3D-interaction-property fields for a set of proteins. The interaction properties, such as electrostatic potential, are computed from the 3D-coordinates of a set of superimposed proteins. PIPSA is available at: http://projects.villa-bosch.de/mcm/software/pipsa.

- New licences distributed in 2006: 31

MolSurfer (MCM Group)

MolSurfer is a graphical tool that links a 2D-projection of a macromolecular interface to a 3D-view of the macromolecular structures. MolSurfer can be used to study protein-protein and protein-DNA/RNA interfaces. The 2D-projections of the computed interface aid visualization of complicated interfacial geometries in 3D. Molecular properties, including hydrophobicity and electrostatic potential, can be projected onto the interface. MolSurfer is available for use through a web server at http://projects.villa-bosch.de/dbase/molsurfer/index.html. It can also be downloaded for local installation from this web site.

- New licences for local use distributed in 2006: 69

SDA (MCM Group)

SDA (Simulation of Diffusional Association) is a software to perform Brownian dynamics simulations of the diffusional association of macromolecules. SDA is available at: http://projects.villa-bosch.de/mcm/software/SDA.

- New licences distributed in 2006: 34

5.1 Workshops and Courses

5.1.1 3rd EMBO Practical Course on Biomolecular Simulation

June 28 - July 4, 2006

Pasteur Institute, Paris, June 28 - July 4, 2006

Organizers: Michael Nilges (Pasteur Institute, Paris) and Rebecca Wade (EML Research, Heidelberg)

Course faculty: David Beck (University of Washington, USA), Konrad Hinsen (CEA, Saclay, France), Leslie Kuhn (Michigan State University, USA), Richard Lavery, (IBPC, Paris, France), Alan Mark, (University of Queensland, Brisbane, Australia), Adrian Mulholland (Bristol University, UK), Michael Nilges (Institut Pasteur, Paris, France), Tom Simonson (Ecole Polytechnique, Palaiseau, Paris, France), Anna Tramontano (University of Rome, Italy), Rebecca Wade (EML Research Heidelberg, Germany), Tru Huynh (Institut Pasteur, Paris, France), Arnaud Blondel (Institut Pasteur, Paris, France)

Molecular simulation techniques constitute an important component of the biologist's toolbox. The function of biological macromolecules is determined by their three-dimensional structure and dynamics. Even with structural genomics projects, the structure-sequence gap is widening at increasing speed. Modelling techniques are thus a major source of structural information, and simulation techniques allow dynamic features to be explored to levels of detail rarely possible experimentally. The EMBO Practical Course addressed three types of simulation appropriate for studying biomolecules at different temporal and spatial scales: quantum mechanics, molecular mechanics/dynamics and Brownian dynamics. These topics were supplemented by practical introductions to protein homology modelling, macromolecular electrostatics, protein-ligand docking, structure-based drug design, as well as programming techniques.
The course aimed to provide the basic theory and practical hints for using these methods so that the students would know how to begin to put them into practice when they

returned to their laboratories. Each topic was addressed by 1-2 lectures, followed by practical sessions. The course was oversubscribed with the level of the applicants being such that we had to reject the applications of many good candidates. We accepted 16 students and several additional students were able to sit-in on the course. The students came from all over Europe and their background was broad and included biology, physics, mathematics, and physical and organic chemistry. There were both experimentally and theoretically oriented participants, ranging from doctoral student to assistant professor level. Overall, the participants found the course very worthwhile, even if the intensity of the course and the accompanying Parisian summer heat and the excitement of the football world cup were exhausting.

Apart from Rebecca Wade, other MCMers participating were Anna Feldman-Salit and Georgi Pachov, who assisted with the practical on electrostatics and Brownian dynamics. More details can be found at http://projects.villa-bosch.de/mcm/conferences.

Sponsor European Molecular Biology Organisation (EMBO)

5.1.2 Workshop: ‚Data, Networks & Dynamics' (2nd UniNet Workshop)

July 3-4, 2006

The second UniNet Workshop was designed to exchange current network theories in the different application nodes (genetic, metabolic, neuronal, ecological, and economic networks) and to provide essential information on state-of-the-art data acquisition. Accordingly, there were not only presentations from the principal investigators reviewing their respective fields and EML post-docs talking about their research, but also talks on data acquisition by four experts on the subject:

- Sven Bergmann (Lausanne)
- Jildau Bouwman (Amsterdam)
- Imre Vida (Freiburg)
- Roland Amann (Konstanz)

5.2 Colloquium Presentations

Prof. Dr. Daniel Zajfman
Director, Max Planck Institute of Nuclear Physics, Heidelberg
Designated Director of the Weizmann Institute, Tel Aviv
January 25, 2006: An Overview of the Activities at the Weizmann Institute of Science

Prof. Dr. Manuel Peitsch
Global Head of Informatics and Knowledge Management at the Novartis Institutes for BioMedical Research, Basel
May 29, 2006: Computational Knowledge Management in Drug Discovery

Prof. Dr. Wolfram Koch
CEO, Gesellschaft Deutscher Chemiker e.V.
(German Chemical Society), Frankfurt
June 19, 2006: Wissenschaftliche Fachgesellschaften im Wandel der Zeiten

Prof. Dr. Helmut Schwarz
Technische Universität Berlin
November 13, 2006: From Bare FeO+ to Cytochrome P-450: New Insights on the Intriguing Mechanisms of C-H Bond Oxygenation

5.3 Heidelberg Innovation Forum 2006 (May 8-9)

The first Heidelberg Innovation Forum held in November 2005 was so successful that, together with the main organizer MFG, we decided to repeat the event in more or less the same format: short 10-minute pitches focusing on either a product idea or a project proposal. The aim is to facilitate the translation of research ideas into real products.

Since the first Forum was devoted to topics the European Media Laboratory is working on, the 2nd Forum, which took place in May 2006, was targeted at the life sciences and related fields of research.

The sessions in which new products and innovative ideas were presented covered subjects such as bioinformatics, image processing, novel instruments, diagnostic aids, and assisted living, to mention only a few. There were 42 (short) presentations altogether, 6 of them from EML Research staff members. The feedback we and the organizers from MFG have received in the meantime indicates that quite a number of the ideas presented at the Forum have resulted in hands-on cooperation with industry – which is exactly what the Heidelberg Innovation Forum is all about.

The keynote speaker at the 2nd Forum was Prof. Dr. Klaus P. Schäfer from Altana Pharma, who gave a very detailed overview of the interplay between research and development in the pharmaceutical industry. He emphasized the

Fig. 38: Turning science into business: Andreas Reuter (right) and Klaus Haasis (MFG) at the press conference

Fig. 39:Sven Sahle demonstrating the Copasi software to a participant

need for carefully orchestrating the complex and year-long process of creating and introducing a new drug. Otherwise, he said, neither the costs nor the time involved can be kept within acceptable limits.

The after-dinner speaker at the end of the first day was Prof. Dr. Heinz-Otto Peitgen from Bremen University, who is both a highly renowned scientist and a successful entrepreneur. He delivered a very lively, humorous talk in which he reflected on how the "tension" between the two sides of the innovative process (research and business) can be turned into a productive force rather than being perceived as a disruptive factor.
The whole event was extremely well received by the participants, which both EML and MFG see as an encouragement to continue with this series of events in the future.

5.4 Miscellaneous: Visit from the Molecular Modeling Group of Kuopio University, Finland

Villa Bosch, Heidelberg, May 17, 2006

Organizer: Dr. Outi Salo-Ahen

Professor Dr. Antti Poso's Molecular Modeling Group from the University of Kuopio, Finland (http://www.uku.fi/farmasia/fake/modelling/), visited the MCM group on May 17, 2006. In the morning, we first held a scientific seminar in the Studio. All the MCM group members briefly introduced their research projects, and some of the visitors gave more extended talks on the ongoing projects in Kuopio. Then the visitors were introduced to the software being developed by the MCM group with tutorials presented by the MCM group members. The social program included an afternoon excursion to the Castle and dinner at a typical German restaurant.

Fig. 40: The participants in the Villa Bosch garden

6.1 Publications

[Anstein 2006a] Stefanie Anstein, Gerhard Kremer and Uwe Reyle: Identifying and Classifying Terms in the Life Sciences: The Case of Chemical Terminology. In: Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odijk and Daniel Tapias (editors), Proceedings of the Fifth Language Resources and Evaluation Conference (LREC 2006), Genoa, Italy, May 24-26, 2006, p. 1095

[Anstein 2006b] Gerhard Kremer, Stefanie Anstein and Uwe Reyle: Analysing and Classifying Names of Chemical Compounds with CHEMorph. In: Sophia Ananiadou and Juliane Fluck (editors), Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006), JULIE Lab, Friedrich-Schiller-Universität Jena, Germany April 9-12, 2006, pp. 37-43

[Brenner 2006] Daniel Brenner, Colin Atkinson, Barbara Paech, Rainer Malaka, Matthias Merdes and Dima Suliman: Reducing Verification Effort in Component-Based Software Engineering through Built-In Testing. In: Proceedings of the 10th IEEE International Enterprise Distributed Object Computing Conference (EDOC'06), IEEE Computer Society, Hong Kong, China, October 16-20, 2006, pp. 175-184

[Filippova 2006a] Katja Filippova and Michael Strube: Improving Text Fluency by Reordering of Constituents. In: Proceedings of the ESSLLI Workshop on Modelling Coherence for Generation and Dialogue Systems Malaga, Spain, July 31- August 11, 2006, pp. 9-16

[Filippova 2006b] Katja Filippova and Michael Strube: Using Linguistically Motivated Features for Paragraph Segmentation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing Sydney, Australia, July 22-23, 2006, pp. 267-274

[Gabdoulline 2006] Razif R. Gabdoulline, Stefan Ulbrich, Stefan Richter and Rebecca C. Wade: ProSAT2 - Protein Structure Annotation Server. In: Nucl. Acids. Res. 2006, 34, W79-W83

[Gauges 2006] Ralph Gauges, Ursula Rost, Sven Sahle, and Katja Wegner: A model diagram layout extension for SBML. In: Bioinformatics, Vol. 22 (15), 2006, pp.1879-1885

[Grinfeld 2006] Michael Grinfeld, Iulian Stoleriu: Truncated gradient flows of the van der Waals free energy. In: Electronic Journal of Differential Equations No. 145, 2006, pp. 1-10

[Hallingbaeck 2006] Henrik R. Hallingbaeck, Razif R. Gabdoulline, and Rebecca C. Wade: Comparison of the binding and reactivity of plant and mammalian peroxidases to indole derivatives by computational docking. In: Biochemistry 2006, 45, pp. 2940-2950

[Jalkanen 2006] Karl J. Jalkanen, Vibeke Würtz-Jürgensen, Anetta Claussen, Abdoul Rahim, G. M. Jensen, Rebecca C. Wade, Frederico Nardi, Christiane Jung, Ivan M. Degtyarenko, Risto M. Nieminen, Frank Herrmann, Michaela Knapp-Mohammady, Thomas A. Niehaus, Kenneth Frimand, and Sandor Suhai: Use of Vibrational Spectroscopy to Study Protein and DNA Structure, Hydration, and Binding of Biomolecules: A Combined Theoretical and Experimental Approach. In: Intl. J. Quant. Chem. 2006, 106, 2006, pp. 1160-1198

[Hoops 2006] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, Ursula Kummer: COPASI - a complex pathway simulator. In: Bioinformatics, 2006, Vol. 22, pp. 3067-3074

[Kania 2006] Renate Kania, Ulrike Wittig, Martin Golebiewski, Olga Krebs, Andreas Weidemann, Saqib Mir and Isabel Rojas: A Curated Database for Reaction Kinetics. In: Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses, Fundación Cajamurcia, Murcia, 2006, pp. 3-14

[Kummer 2006] Ursula Kummer: An Overview of Computational Approaches to Metabolic Networks. In: Proceedings of the 2nd UniNet Workshop: Data, Networks and Dynamics, Heidelberg, July 3, 2006 July 4, 2006, pp. 72-76

[Malaka 2006] Rainer Malaka, Jochen Häußler, Hidir Aras, Matthias Merdes, Dennis Pfisterer, Matthias Jöst, and Robert Porzel: SmartKom-Mobile: Intelligent Interaction with a Mobile System. In: Wahlster, W. (ed.) SmartKom - Foundations of Multimodal Dialogue Systems, Cognitive Technologies, Springer Verlag, Berlin, pp. 505-522

[Merdes 2006a] Matthias Merdes, Rainer Malaka, Dima Suliman, Barbara Paech, Daniel Brenner and Colin Atkinson: Ubiquitous RATs: How Resource-Aware Run-Time Tests can improve Ubiquitous Software Systems. In: Proceedings of the Sixth International Workshop on Software Engineering and Middleware (SEM 2006), ACM Press, New York, NY Portland, Oregon, USA November 10, 2006, pp. 55-62

[Merdes 2006b] Matthias Merdes and Dirk Dorsch: Experiences with the Development of a Reverse Engineering Tool for UML Sequence Diagrams: A Case Study in Modern Java Development, In: Proceedings of the 4th international Symposium on Principles and Practice of Programming in Java (PPPJ 2006), ACM Press, New York, NY, Mannheim, Germany, August 30 - September 1, 2006, pp. 125-134

[Strube 2006] Michael Strube and Simone Paolo Ponzetto: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, Mass., July 16-20, 2006, pp. 1419–1424

[Mieskes 2006] Margot Mieskes and Michael Strube: Part-of-Speech Tagging of Transcribed Speech. In: Proceedings of the 5th International Conference of Language Resources and Evaluation Genoa, Italy, May 22-29, 2006, pp. 935-938

[Müller 2006a] Christoph Müller: Automatic Detection on Nonreferential It In Spoken Multi-Party Dialog. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics Trento, Italy, April 3-7, 2006, pp. 49-56

[Müller 2006b] Christoph Müller: Representing and Accessing Multilevel Annotations in MMAX2. In: Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing, Trento, Italy, April 4, 2006, pp. 73-76

[Müller 2006c] Christoph Müller and Michael Strube: Multi-Level Annotation of Linguistic Data with MMAX2. In: Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214

[Müller, M. 2006] Markus Müller, Katja Wegner, Ursula Kummer and Gerold Baier: The Quantification of Cross-correlations in Stochastic Spatio-temporal Systems. In: Phys. Rev. E, Vol. 73, id. 046106

[Pu 2006] Calton Pu, Jim Johnson, Roderio de Lemos, Andreas Reuter, David Taylor, and Irfan Zakiuddin: Break Out Session on Guaranteed Execution, in: Atomicity: A Unifying Concept in Computer Science, Dagstuhl Seminar Proceedings, March 19-24, 2006, 2006:641

[Sahle 2006] Sven Sahle, Stefan Hoops, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes and Ursula Kummer: Simulation of Biochemical Networks Using COPASI - a Complex Pathway Simulator. In Proceedings of the 2006 Winter Simulation Conference, Monterey, USA, December 3-6, 2006, pp. 1698-1706

[Saric 2006] Lars Jensen, Jasmin Saric and Peer Bork: Literature mining for the biologist: from information retrieval to biological discovery. In: Nature Reviews Genetics 7 (2), February 2006, pp. 119-129

[Schleinkofer 2006] Karin Schleinkofer, Ting Wang and Rebecca C. Wade: Molecular Docking. In 'Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine', Eds: Ganten, D., Ruckpaul, K., Springer, ISBN-10: 3-540-29623-9, 2006, pp. 1149-1153.

[Stein 2006] Matthias Stein, Razif R. Gabdoulline, Rebecca C. Wade: Integrating Structural and Kinetic Enzymatic Information in Systems Biology. In: From Computational Biophysics to Systems Biology 2006, Hansmann, Ulrich H.E.; Meinke, Jan; Mohanty, Sandipan; Zimmermann, Olav (Eds.), NIC Series 2006, 34, pp. 129-132.

[Ponzetto 2006a] Simone Paolo Ponzetto and Michael Strube: Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-06) New York, N.Y., June 4-9, 2006, pp. 192–199

[Ponzetto 2006b] Simone Paolo Ponzetto and Michael Strube: Semantic role labeling for coreference resolution. In: Companion Volume of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy, April 3-7, 2006, pp. 143–146

[Stoleriu 2006] Iulian Stoleriu: Stability analysis of metabolic networks. In: Proceedings of the Proceedings of the 2nd Workshop: Data, Networks and Dynamics, pp. 77-87

[Suliman 2006a] Dima Suliman, Barbara Paech, Lars Borner, Colin Atkinson, Daniel Brenner, Matthias Merdes and Rainer Malaka: The MORABIT Approach to Runtime Component Testing. In: Proceedings of the 30th Annual International Computer Software and Applications Conference (COMPSAC'06), IEEE Computer Society, Chicago, Illinois, USA, September 17-21, 2006, pp. 171-176

[Suliman 2006b] Dima Suliman, Lars Borner, Matthias Merdes and Daniel Brenner: Laufzeitttest mobiler und komponentenorientierter Software. In: Proceedings of doIT Software-Forschungstag, 2006, Mannheim, Germany, July 13, 2006

[Surovtsova 2006a] Irina Surovtsova and Jürgen Zobeley: Focusing on Dynamic Dimension Reduction for Biochemical Reaction Systems. In: Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses, Fundacion Caja Murcia (Spain), 2006, pp. 31-47

[Surovtsova 2006b] Irina Surovtsova, Sven Sahle, Jürgen Pahle and Ursula Kummer: Approaches to Complexity Reduction in a System Biology Research Environment (SYCAMORE). In: Proceedings of the 2006 Winter Simulation Conference Monterey, USA, December 3, 2006 December 6, 2006 pp.1683-1689

[Wang 2006] Ting Wang and Rebecca C. Wade: Force Field Effects on a β-sheet Protein Domain Structure in Thermal Unfolding Simulations. In: J. Chem. Theory Comput, 2006, 2, pp. 140-148

[Wittig 2006] Ulrike Wittig, Martin Golebiewski, Renate Kania, Olga Krebs, Saqib Mir, Andreas Weidemann, Stefanie Anstein, Jasmin Saric and Isabel Rojas: SABIO-RK: Integration and Curation of Reaction Kinetics Data. In: Data Integration in the Life Sciences, Lecture Notes in Bioinformatics, Springer LNCS Series, Volume 4075, 2006, pp. 94-103

6.2
Guest Speaker Activities

"Modeling macromolecular motions by Brownian dynamics simulations", MMS06: Methods of Molecular Simulation 2006, Interdisciplinary Center for Scientific Computing (IWR), Ruprecht-Karls-University of Heidelberg, Heidelberg, Germany, September 20-22, 2006	**Razif Gabdoulline**
„SABIO-RK: Making Reaction Kinetics Data Accessible". First BioModels.net training camp, European Bioinformatics Institute, Hinxton, UK, April 10-12, 2006	**Martin Golebiewski**
A Dynamic View On Neutrophil Biochemistry, VBI, Blacksburg, VA, USA, February 22, 2007	**Ursula Kummer**
Lassen sich Körperfunktionen berechnen? Hygiene-Museum, Dresden, Germany, April 25, 2006	
Copasi – A Complex Pathway Simulator, Institut für Biochemie, Universität zu Köln, Köln, Germany, October 23, 2006	**Jürgen Pahle**
„Transactions – Past, Present and Future", Talk at the Dagstuhl-Workshop „Atomicity", March 19-24, 2006	**Andreas Reuter**
Invited Talk, Institutstag am Institut für Informatik, University of Heidelberg, July 7, 2006	
Gastvortrag anlässlich des 100-jährigen Jubiläums der Stadtbücherei Heidelberg, September 23, 2006	
Keynote, TPF Users Group Conference, Atlanta, USA, October 23, 2006	
Panel Talk, Forum Biotechnologie Baden-Württemberg, Stuttgart, October 24, 2006	

Isabel Rojas SABIO RK (System for the Analysis of Biochemical Pathways - Reaction Kinetics), 2nd International Symposium on „Experimental standard conditions of enzyme characterizations“, Jagdschloss Niederwald, Rüdesheim, Germany, March 20-22, 2006

Ursula Rost Biologie + Informatik = Bioinformatik? Heidelberg, Germany, May 19, 2006

Jasmin Saric Lessons learned from collecting kinetic data Mini-Symposium „Semantic Enrichment of Scientific Literature“, European Bioinformatics Institute (EBI), Hinxton, UK, February 20, 2006

Literatur Mining for the Biologists nuGO Workshop, German Institute of Human Nutrition, Postdam, Germany, June 1, 2006

Extraction of Protein Interactions from Medline Institut for Computer Linguistics, University of Zürich, Switzerland, July 22, 2006

Literature Mining in the life Science Pharma Documentation Ring, Bracknell, UK, September 27, 2006

Matthias Stein “Integrating Enzymatic Structural and Kinetic Data in Systems Biology”, ESCEC2 Symposium, Rüdesheim, Germany, March 19-23, 2006

Michael Strube „Meeting minutes at the push of a button“: Automatically summarizing spoken multi-party dialogue Informatisches Kolloquium, Department Informatik, Universitaet Hamburg, Hamburg, Germany, January 30, 2006

„Meeting minutes at the push of a button“: Automatically summarizing spoken multi-party dialogue, NLP Seminar, INRIA-LORIA, Nancy, France, March 9, 2006

Gesprächsprotokolle auf Knopfdruck: Die automatische Zusammenfassung von Dialogen, Jahrestagung des Instituts für deutsche Sprache, Mannheim, Germany, March 14-16, 2006

World Knowledge Induced From Wikipedia: A Prospect For Knowledge-based NLP Dept. of Computer Science and Engineering, Korea University, Seoul, Korea, November 14, 2006

Rebecca Wade

"Computational approaches to biomolecular recognition", SFB-Workshop on 'Recognition and adsorption processes of biomolecules', Physics Faculty, University of Bielefeld, March 6, 2006

"Cytochrome P450 monooxygenases: Insights from computer simulations", 'Katalytische Selektivoxidationen von C–H-Bindungen mit molekularem Sauerstoff' SFB-Kolloquium, University of Stuttgart, March 20, 2006

"Simulation of protein-ligand interactions" 367. WE-Heraeus-Seminar on 'Biomolecular Simulation: From Physical Principles to Biological Function' Physikzentrum Bad Honnef (Germany), May 22–24, 2006

"Cytochrome P450 monooxygenases: Insights from computer simulations", Department of Pharmacochemistry, Vrije Universiteit, Amsterdam, June 12, 2006

"Enzyme binding processes and kinetics: molecular interaction fields and simulations", ISQBP 2006 President's Meeting, University Louis Pasteur, Strasbourg, France, June 26, 2006

"Exploring Biomolecular Recognition Mechanisms by Modeling and Simulation", Theoretical Physics Department, University of Heidelberg, July 20, 2006

"Biomolecular recognition: Exploring protein interactions using protein structures", "Biomolecular recognition: Dynamic mechanisms for protein-ligand binding specificity", "Biomolecular recognition: Modeling and simulation of protein-protein binding", 3 lectures as a Visiting Scientist, A*Star Bioinformatics Institute, Singapore, August 8-14, 2006

"Exploring Biomolecular Recognition Mechanisms by Modeling and Simulation", MPG-Koç Workshop on Protein Bioinformatics, Koç University, Istanbul, Turkey, September 6-8, 2006

Peter Winn "The ins and outs of the cytochromes P450: understanding key drug metabolizing enzymes"; "Structure Based Drug Design: GRID and Anti-influenza Agents."; "Quantitative Comparisons of Protein Electrostatic Potentials as a Means to Understanding Function." 3 lectures at the 'Protein Folding and Drug Design' Course, International School of Physics, "Enrico Fermi", Varenna, Italy, July 4-14, 2006

"The ins and outs of the cytochromes P450: understanding key drug metabolizing enzymes", MGMS Conference: "Quantum Pharmacology – 30 years on." Oxford, UK, September 17-20, 2006

6.3 Presentations

Talks

Razif Gabdoulline "Software for Analysis of Biomolecular Interaction Properties", Heidelberger Innovationsforum, Villa Bosch, Heidelberg, Germany, May 8-9, 2006

Stefan Henrich "COMBINE-Analyse: Vorhersage von Bindungsstärken zwischen Rezeptor und niedermolekularen Substanzen", Heidelberger Innovationsforum, Villa Bosch, Heidelberg, May 8-9, 2006

Olga Krebs, Martin Golebiewski, Renate Kania, Saqib Mir, Jasmin Saric, Andreas Weidemann, Ulrike Wittig and Isabel Rojas SABIO-RK: INTEGRATING AND SHARING REACTION KINETICS DATA, International Workshop on NanoBioTechnologies, 2006, Saint-Petersburg, Russia, November 25-29, 2006

SABIO-RK presentation Meeting of German-Russian network, Bielefeld, Germany December 11-12, 2006	**Olga Krebs**
Copasi - a complex pathway simulator, Workshop on Systems Biology, MPI for Molecular Genetics, Berlin, Germany, March 2-3, 2006	**Ursula Kummer**
Simulation and Parameter Estimation on the Basis of Experimental Enzymatic Data, ESCEC, Rüdesheim, Germany, March 20-22, 2006	
The Implications of Choosing a Specific Simulation Method - Exemplified on a Model of Calcium Signal Transduction, Towards Molecular Systems Biology , Bielefeld, Germany, June 6-9, 2006	
Computational research on metabolic networks, 2nd Uni-Net Workshop: Data, Networks and Dynamics, Heidelberg, Germany, July 3-4, 2006	
Integrating experimental and computational approaches to unravel the mechanisms of neutrophil activation, BTK 2006, Trakai, Lithuania, September 14-17, 2006	
WikiRelate! and Semantic Knowledge Sources for Coreference Resolution, Institutsversammlung of the Institute for Natural Language Processing, University of Stuttgart, May 24, 2006	**Simone Paolo Ponzetto**
Modeling, simulating, and analyzing biochemical systems with COPASI, ICSB 2006, Pacifico Yokohama, Yokohama, Japan, October 8, 2006	**Sven Sahle**
Advanced model analysis with COPASI, ICSB 2006, Yokohama Pacifico, Yokohama, Japan October 8, 2006	
COPASI and SBML Layout, SBML Forum, Miraikan, Tokyo, Japan October 12-13, 2006	
„Tackling the cellular drug resistance of thymidylate synthase“, Introductory Meeting of Alexander von Humboldt Foundation, University of Köln, Germany, October 13, 2006	**Outi Salo-Ahen**

Natalia Simus Simulation of Biochemical Networks Using COPASI - a Complex Pathway Simulator, 2006 Winter Simulation Conference, Monterey, USA, December 3-6, 2006

Matthias Stein "Integrating Enzymatic Structural and Kinetic Data in Systems Biology" From Computational Biophysics to Systems Biology, Jülich, Germany, June 6-9, 2006

"Identification of the Intermediates in the Reaction Cycle of Hydrogenases", 3rd International Symposium on Bioorganometallic Chemistry, Milan, Italy, July 5-8, 2006

Iulian Stoleriu Mathematical analysis of metabolic networks, 2nd UniNet Meeting, Villa Bosch, Heidelberg, Germany, July 3-4, 2006

Mathematical analysis of metabolic networks, Mathematics Institute, University of Warwick, UK, January 31, 2006

Irina Surovtsova Two approaches on dynamical dimension Reduction for Biochemical Systems, 1st International Symposium on System Biology. From genomes to In silico and back, Murcia, Spain, June 1-2, 2006

Approaches to Complexity Reduction in a System Biology Research Environment (SYCAMORE), 2006 Winter Simulation Conference, Monterey, USA, December 3-6, 2006

Ulrike Wittig SABIO-RK: Integration and Curation of Reaction Kinetics Data, 3rd International Workshop on Data Integration in the Life Sciences 2006 (DILS'06), Hinxton, UK, July 20-22, 2006

Posters

Vlad Cojocaru "The Ins and Outs of Cytochrome P450s", Vlad Cojocaru, Peter J. Winn and Rebecca C. Wade. Presented at: Workshop on "Computer Simulation and Theory of Macromolecules 2006", Hünfeld, Germany, May 19-20, 2006

ISQBP 2006 President's Meeting, University Louis Pasteur, Strasbourg, France, June 24-27, 2006

"Exploring the routes out of the active site of cytochrome P450 2C9", Vlad Cojocaru and Rebecca C. Wade. Gordon Research Conference on "Computational Chemistry", Les Diablerets, Switzerland, October 8-13, 2006

„Protein Recognition Processes: The 'Cysteine Synthase' Complex" Anna Feldman-Salit, Domantas Motiejunas, Razif Gabdoulline, Markus Wirtz, Rudiger Hell, Rebecca Wade. Presented at: EMBO practical course on „Biomolecular simulation", Pasteur Institute, Paris, France, June 28 - July 4, 2006 — **Anna Feldman-Salit**

„MMS'06: Methods of Molecular Simulation", University of Heidelberg, Germany, September 20-22, 2006

The SBML Layout Extension, International Conference of Systems Biology, Yokohama, Japan, October 9-11, 2006 — **Ralph Gauges**

A Relational Database for Biochemical Reactions and their Kinetics: SABIO-RK, 20th IUBMB International Congress of Biochemistry and Molecular Biology and 11th FAOBMB Congress of Biochemistry and Molecular Biology: „Life: Molecular Integration & Biological Diversity", Kyoto, Japan, June 18-23, 2006 — **Martin Golebiewski, Renate Kania, Ulrike Wittig, Andreas Weidemann, Olga Krebs, Saqib Mir, Isabel Rojas**

"The Application of COMBINE Analysis to Generate Target-Specific Scoring Functions", S. Henrich, I. Feierberg, T. Wang, N. Blomberg, R.C.Wade, Presented at: Keystone Symposia, Structure Based Drug Discovery, Whistler, Canada, April 4-9, 2006 — **Stefan Henrich**

Methods of Molecular Simulation (MMS06), Heidelberg, Germany, September 20-22, 2006

2. German Conference on Chemoinformatics, Goslar, Germany, November 12-14, 2006

European Science Forum, Heidelberg, November 28, 2006

Renate Kania

An Integrative Database for Biochemical Reaction Kinetics: SABIO-RK, The Seventh International Conference on Systems Biology, Pacifico Yokohama, Yokohama, Japan October 9-13, 2006

Femke Mensonides

PharmBiosim - Simulation of Drug Metabolism, 12th Meeting of the International Study Group for System's Biology, Trakai, Lithuania, September 14-17, 2006

PharmBiosim - Simulation of Drug Metabolism, 2nd BioSim conference, Mallorca, Spain, October 18-21, 2006

Domantas Motiejunas

"Protein-Protein Docking Guided by Biochemical Data", Domantas Motiejunas, Ting Wang, Anna Feldman-Salit, Tim Johann, Razif Gabdoulline, Peter Winn and Rebecca C. Wade. Presented at: 367. WE-Heraeus Seminar on "Biomolecular Simulation: From Physical Principles to Biological Function", Bad Honnef, Germany, May 22-24, 2006

"Quantum Pharmacology - 30 Year On" A Special Conference in honour of Prof. W. Graham Richards, MGMS, Oxford, UK, September 17-20, 2006

Georgi Pachov

„Chromatin: a multiscale approach to biomolecular interactions" Georgi V. Pachov, Rebecca C. Wade. Presented at: EMBO practical course on „Biomolecular simulation", Pasteur Institute, Paris, France, June 28 - July 4, 2006

"MMS'06: Methods of molecular simulation 2006", University of Heidelberg, Germany September 20-22, 2006

Jürgen Pahle
Sven Sahle
Ursula Kummer

Dynamic behavior of buffered calcium ions in stochastic and deterministic simulations, EMBO Workshop on Principles of Self-Organization in Living Matter, Heidelberg, Germany, June 2-4, 2006

Hybrid simulation of biochemical reaction networks, 3rd BMBF-Status Seminar of HepatoSys, Heidelberg, Germany, July 12, 2006

Dynamic behavior of buffered calcium ions in stochastic and deterministic simulations, SBMC 2006, Conference on Systems Biology of Mammalian Cells, Heidelberg, Germany, July 12-14, 2006	
Dynamic behavior of buffered calcium ions in stochastic and deterministic simulations ICSB 2006, International Conference on Systems Biology, Yokohama, Japan, October 9-11, 2006	
Semantic role labeling for coreference resolution, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy, April 3-7, 2006	**Simone Paolo Ponzetto**
COPASI, SBMC 2006, Heidelberg, Germany, July 12-14, 2006	**Sven Sahle**
COPASI, ICSB 2006, Yokohama Pacifico, Yokohama, Japan, October 9-11, 2006	
„Searching for the hot spots in the dimer interface of the human thymidylate synthase“. Salo-Ahen, O.M.H., Costi, M.P. and Wade, R.C. Presented at: German Conference on Bioinformatics, Tuebingen, Germany, September 20-22, 2006	**Outi Salo-Ahen**
Workshop on Molecular Targets for Cancer, Luxembourg, October 6-7, 2006	
European Science Forum, Heidelberg, Germany, November 28, 2006	
COPASI - a Complex Pathway Simulator, 2006 Winter Simulation Conference, Monterey, USA, December 3-6, 2006	**Natalia Simus**

Matthias Stein "Comparison of Molecular Interaction Fields of Enzymes in Molecular Systems Biology", Stein,M., Gabdoulline, R.R. and Wade R.C. Conference on Systems Biology of Mammalian Cells, Heidelberg, Germany, July 12-14, 2006

Demos

Ralph Gauges COPASI Beginners and Advanced Tutorial, International Conference of Systems Biology, Yokohama, Japan October 9-11, 2006

Matthias Merdes Demonstration of the Morabit Runtime Infrastructure, Evaluation Meeting with Project Sponsor DLR/Landesstiftung Baden-Württemberg, Mannheim, Germany, July 17, 2006

Demonstration of the Morabit Runtime Infrastructure, Final Project Evaluation Meeting at Landesstiftung Baden-Württemberg, Stuttgart, Germany, November 2, 2006

Christoph Müller MMAX2 - ein Werkzeug für die „tiefe" Annotation relativ kleiner Korpora auf mehreren linguistischen Ebenen, 42. Jahrestagung des Instituts für Deutsche Sprache: Sprachkorpora - Datenmengen und Erkenntnisfortschritt, Mannheim, Germany, March 15-16, 2006

6.4 Memberships

Ursula Kummer Member of the Steering Committee and Coordinator of BIOMS

Andreas Reuter Scientific Member of Max-Planck-Gesellschaft (Max Planck Institute of Computer Science, Saarbrücken)

Member of the Scientific Committee, BIOMS, Heidelberg

Member of the Scientific Committee of BIOTEC, Dresden

Member of the Advisory Board of Fraunhofer Gesellschaft Informations- und Kommunikationstechnik (IuK)

Member of the Advisory Board “First Ventury AG”

Member of the Advisory Board „Beratungsforum Information, Telekommunikation und Software“ (BITS Baden-Württemberg)

Member of the Heidelberg Club International

Member of the Board of Trustees of the Wissenschaftspressekonferenz, Bonn

Member of the Advisory Board of Parallel Computing Journal

Co-editor “Database Series”, Vieweg-Verlag

Mitglied Fachgremium “Landesstiftung Baden-Württemberg“

Member of the Board of Trustees of Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Mitglied des Forschungsverbundes Wissenschaftliches Rechnen Baden-Württemberg „WiR“

Vorsitzender der Gutachtergruppe für das Förderprogramm „D-Grid“ beim BMBF.

Mitglied im Herausgeberrat der DISBIS-Reihe, AKA-Verlag, Berlin

Mitglied im e-Science Kuratorium des BMBF, Bonn.

Mitglied der High-Performance Computing in Europe Task Force (HET).

Kuratoriumsmitglied des Internationalen Begegnungs- und Forschungszentrums für Informatik (IBFI), ab Mai 2007

Vorsitzender des Gutachterausschusses für das Forschungsprogramm BW-FIT, Stuttgart

Michael Strube Editorial Board: Journal of Dialogue Systems

Editorial Board: Research on Language and Computation

Rebecca Wade Editor: Journal of Molecular Recognition

Editorial Board: Journal of Computer-aided Molecular Design; Journal of Medicinal Chemistry, Journal of Molecular Graphics and Modelling; Biopolymers

Member of "Faculty of 1000" for "Theory and Simulation" section

6.5 Contributions to the Scientific Community

Program Committee Memberships

Jasmin Saric International Symposium on Semantic Mining in Biomedicine, Jena, Germany, April 9-12, 2006

Workshop for Integrative Bioinformatics Rothamsted Research, Harpenden, Hertfordshire, UK, September 4-6, 2006

Workshop „Data and Text Mining for Integrative Biology", within the „European Conference on Machine Learning / European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKKD), Berlin, Germany, September 18-22, 2006

Michael Strube Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL ´06). Senior Program Committee Member for the Area Discourse, Dialogue, and Multimodality, New York, USA, June 4-9, 2006

Workshop on Sentiment and Subjectivity in Text, Workshop at the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (COLING-ACL ´06) Sydney, Australia, July 22, 2006

IEEE and ACL 2006 Workshop on Spoken Language Technology (SLT ´06) Palm Beach, Aruba, December 10-13, 2006

External Reviewer Ph.D.: Yang, Xiaofeng. A twin-candidate model for learning based coreference resolution. National University of Singapore, School of Computing

Rebecca Wade

Co-organizer with Michael Nilges of EMBO Practical Course on "Biomolecular Simulation", Pasteur Institute, Paris, France, June 28 - July 4, 2006

PhD external referee, "Cytochrome P450-Drug Interactions: Computational Binding Mode and Affinity Predictions in CYP2D6", Chris De Graaf, Department of Chemistry and Pharmaceutical Sciences, Free University, Amsterdam, Netherlands, November 6, 2006

Participant: ESF 'Forward Look on European Computational Science', Lifescience Community Level Workshop, Chilworth Manor, UK, November 20-21, 2006

Organization Committee Memberships

Workshop Organization

Ursula Kummer

2nd UniNet Workshop: Data, Networks and Dynamics Villa Bosch, Heidelberg, Germany, July 3-4, 2006

Referee Work

Christoph Müller

ESSLI Workshop on Ambiguity in Anaphora, Malaga, Spain, August 7-11, 2006

Michael Strube

Computational Intelligence Journal

Computational Linguistics Journal
Journal of Artificial Intelligence Research

Speech Communication Journal

12th Conference of the European Chapter of the Association for Computational Linguistics (EACL ´06) Trento, Italy, April 3-7, 2006

9th Workshop on the Semantics and Pragmatics of Dialogue (BranDial ´06) Potsdam, Germany, September 10-13, 2006

Jasmin Saric

BioMed Central

Isabel Rojas

BioMed Central

6.6 Award

Georgi Pachov

„MMS '06, Methods of Molecular Simulation", Institute for interdisciplinary scientific computing (IWR), University of Heidelberg, Heidelberg, Germany, September 20-22, 2006 - 1st poster prize

7 Teaching

Lectures

Statistical Approaches to Coreference Resolution: Data, Representation, Knowledge. ACL/HCSNet Advanced Program in Natural Language Processing, University of Melbourne, Melbourne, Australia, July 10-14, 2006	**Michael Strube**
CUBIC Postgraduate Programme in Bioinformatics, Module: „Biochemical Simulations: Stochastic Methods and Copasi“, Universität zu Köln, October 24, 2006	**Jürgen Pahle**
Einführung in die Bioinformatik, Universität Freiburg,(Aug 2006), Freiburg, Germany	**Ursula Rost**
Financial Mathematics, Faculty of Mathematics, Al. I. Cuza” University, Iasi, Romania (March - May 2006)	**Iulian Stoleriu**
Partial Differential Equations, Faculty of Mathematics, “Al. I. Cuza” University, Iasi, Romania (March - May 2006).	
Sub-languages of the Life Sciences, Summer Semester 2006, Stuttgart, Germany	**Jasmin Saric**

Practicals

Tutorial: Modeling, simulating and analyzing biochemical systems with Copasi, ICSB 2006, International Conference on Systems Biology, Yokohama, Japan, October 8, 2006	**Jürgen Pahle**
Mathematical Statistics, Faculty of Mathematics, “Al. I. Cuza” University, Iasi, Romania (March - May 2006)	**Iulian Stoleriu**
Data bases in Visual FoxPro, Faculty of Mathematics, “Al. I. Cuza” University, Iasi, Romania (March - May 2006).	
Introduction to Prolog, Institute for Natural Language Processing, Summer Semester 2006, University of Stuttgart, Germany	**Jasmin Saric**

Degrees

[Zahran 2006] Mai Zahran: "Analyse à haut debit des propriétés électrostatiques des proteins de la famille des ubiquitines", Masters Thesis, Bioinformatics, University Paris VII and EML Research, Rebecca Wade and Peter Winn, 2006

Courses

R.R. Gabdoulline, M. Stein, P.J. Winn (with M. Ullmann) "Modeling and Simulation of Biomolecular Interactions", Practical Course in Bioinformatics, University of Bayreuth, February 21-24, 2006

R.C. Wade, R.R. Gabdoulline, M. Stein, P.J. Winn, D. Motiejunas (with J.C. Smith)"Modeling and Simulation of Biomolecular Interactions", Practical Course, University of Heidelberg, June 19-23, 2006

R.C. Wade, "Brownian Dynamics Simulations", EMBO Practical Course on „Biomolecular simulation„, Pasteur Institute, Paris, France, June 28 – July 4, 2006

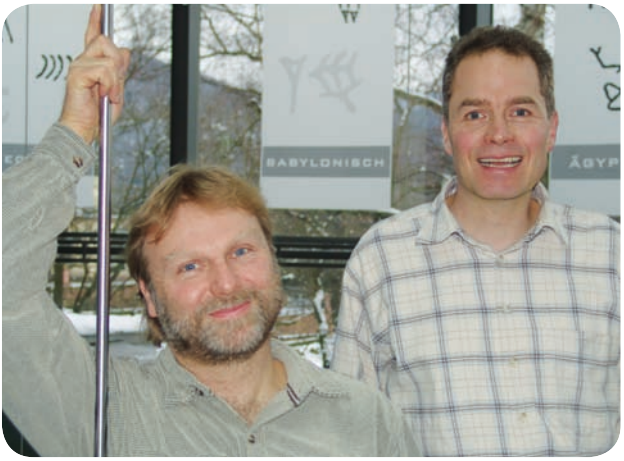


Fig. 41:
Norbert Rabes (left), Achim Beck

The EML Research computer network is currently based on Microsoft Windows 2000 and makes up 85% of the IT backbone. In realizing the active directory structure particular importance has been attached to security, simple and flexible administration, scalability, extensibility, and support for open-standard applications.
UNIX and LINUX account for the remaining part of the backbone. The network is divided into 6 virtual LANs.
The LAN Backbone has been expanded to GigaBit. Optical (LWL) GigaBit lines connect the two buildings, while within the buildings ordinary Cu-GigaBit is used for the network.
The internet connection is controlled by a Cisco Router with a dynamic data-transfer rate of up to 4 Mbits/s and a firewall connected to it. External access is granted by means of a virtual private network (VPN) consisting of a Cisco PIX 501 device.

System Administrators
Norbert Rabes Phone: +49-6221-533-265
Achim Beck Phone: +49-6221-533-268

In 2006, EML´s former Checkpoint firewall was replaced by a Linux server with multiple network interface cards running netfilter software (iptables). The firewall also serves as a central router to EML´s networks.

EML Research has 22 Windows 2000/2003 servers (MS Exchange2000, MS SQL 2000, Oracle 9), 14 Linux servers (NIS, LDAP, DNS, DHCP, Web (apache, php), CVS/Subversion, MySQL, PostgreSQL, File (NFS/Samba), Print (cups), Kerberos, NTP), and various application servers.

In 2005 an e-mail filtering server ("spam server") was put into service running on a FreeBSD system and open source ‚spamassasin' software.

Furthermore, EML Research operates a Linux compute cluster consisting of 21 compute nodes (8 dual Xeon, 7 dual Opteron, and 6 dual Dual Core Opteron servers) and one master node (dual Xeon) running under Debian Linux. The Dual Core Opteron Systems were deployed in 2005. This was the second upgrade of the cluster since it was purchased in 2003. The first upgrade took place in 2004.

A common SAP R/3 is used for financial accounting, project control, travel management, and human resources management – together with EML, KTF, and KTA. The Libero library management software with an underlying Cache-5 database is used to operate KTF's library, which is also widely used by the aforementioned parties.

Another server for common use is a media server for video and audio data with a storage capacity of 1TB.

A backup system including a 40-slot LTO2 tape library running Veritas backup software under Windows Server 2003 ensures data protection. A new 6TB RAID storage system was acquired in 2006 serving as a (disc) backup medium for the Linux servers. An additional backup onto tape is made from this RAID system.

The complete storage capacity of the backup tapes before replacement amounts to 8 TByte (uncompressed). Both Windows- and Linux-operated servers and selected clients are backed up regularly.

OS Variant	Servers	Workstations
Windows (div)*	22	81
Windows 2000TS	1	0
Windows TabletPC	0	8
Windows CE	0	10
SGI IRIX	1	1
Linux Debian	14	38
MAC OS	0	3

* Windows (div) includes: Windows 2000/2003 and XP (clients only).