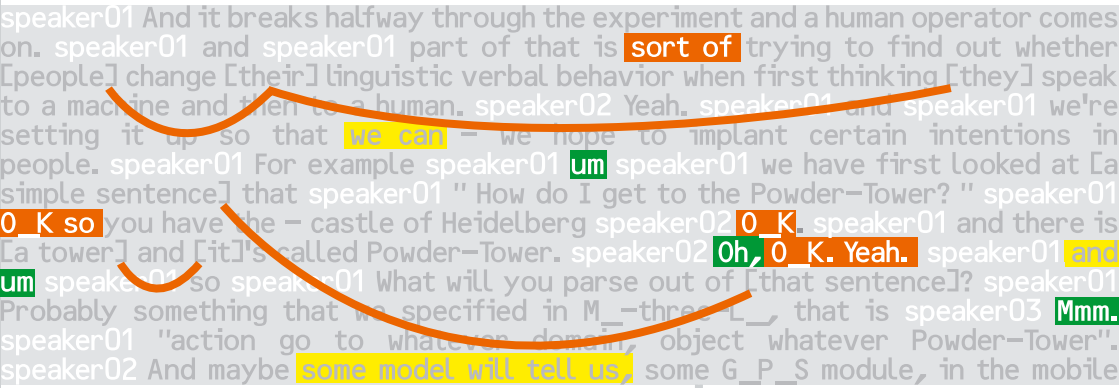
Annual Report



EML

Research

2005

Annual Report

EML Research gGmbH

2005

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)

All brand names and product names used in this report are trade names, service marks, trademarks, or registered trademarks of their respective owners. (In diesem Bericht werden eingetragene Warenzeichen, Handelsnamen und Gebrauchsnamen verwendet. Auch wenn diese nicht speziell als solche ausgezeichnet sind, gelten die entsprechenden Schutzbestimmungen.)

All rights reserved.

ISSN 1438-4159

Table of contents

Andreas Reuter

2005 has proved to be a very successful year for EML Research. One is tempted to characterize it as „unusually successful", but I will not do that for two reasons: First, with little more than two years of existence, our lab has not been around for long enough to allow for any statistical characterizations. And second, we hope that our work will continue to be successful at the same level, so there will, in fact, not be anything unusual about it.

After this display of enthusiasm, the reader will be curious to see the evidence supporting such a bold introduction. It will be provided in considerable detail in the following chapters, so I can restrict myself to briefly summarizing the key aspects of our work in 2005. In principle, there are five criteria that are relevant for assessing the work of a research organization:

- quality of results in the scientific projects;
- success in acquiring third-party funding;
- visibility of the organization and its research staff members in the scientific community;
- activities in the educational domain;
- transfer of results.

The first aspect can hardly be cast into a summary, so for that I have to refer the reader to the chapters describing the results of our research projects.

Regarding third-party funding, EML Research was successful at the European level, at the national level and in the area of contract research. As will be explained in the following, we started two new projects funded by the EU in the context of FP6. A research contract by Astra Zeneca was extended for another year, and the NLP group was awarded their second research grant from DFG. The latter project, however, has moved to the Technical University of Darmstadt in the meantime (together with the principle investigator), but that in itself has to be considered a success with respect to visibility and transfer of results.

The visibility of EML Research and its (senior) scientists is well documented by the rapidly increasing number of conferences, workshops and other events they are involved in in key functions or which are initiated by them in the first place. The same holds for positions in editorial boards, advisory boards, for invited lectures etc. Another indicator for the increasing visibility of our lab is the growing number of high-quality applications we get for scientific positions at all levels. The scholarship program for PhD students that we started jointly with KTS about two years ago is very popular, and because of the large number of very good candidates the research groups in EML have decided to augment the program (thus increasing the number of scholarships) by allocating money from their own budgets to it.

The activities in the educational domain have been continued at the level of the previous years, with lectures at universities, with support for master students, internships and the like. As with the applications from scientists we can observe an increasing demand from the students' side – in many cases much more than we can handle with the resources available.

Transfer of results has been an area of major activity in the reporting period. As is documented in the following chapters, our research groups have completed a number of software packages that are being made available under different licensing schemes, depending on the intended purpose of use and the legal status of the licensee. In some cases, the number of users has quickly grown to a level that future support of such a large community may pose problems in the years to come.

Collaboration with our sister organisation, the European Media Laboratory, has proceeded along the lines that were established in the previous years: We have continued to work on our joint project Morabit, and we still share a large number of formal events, such as the lab meeting, and informal ones as well.

For all these very positive results and perspectives we have to thank many people and organizations; without their encouragement and support we would not be enjoying such a wonderful, stimulating research environment. Of course, we cannot mention them all, but one deserves to be singled

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

out: The Klaus Tschira Foundation, which is the base rock on which all the interesting, crazy, challenging, … projects can be built. Our many friends and colleagues should feel included in this comprehensive „THANK YOU“ – they are the audience we are addressing with our work.
And if this report leaves anything unclear – don't hesitate to ask; we will be more than happy to discuss those issues with you. The year 2005 was dedicated to Albert Einstein, who left us (among many other things) with the sage advice: „Never stop asking.“

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 the autumn of 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 addition to the ongoing projects Simulation of Biochemical Pathways, Copasi (Complex Pathway Simulator), and Sycamore, two EU-funded projects, BioSim and UniNet, were launched in 2005. Therefore, two new members, Femke Mensonides and Iulian Stoleriu, have joined the group. In order to strengthen some aspects of the existing projects and to replace Jürgen Zobeley, who left the group later in the year, Irina Surovtsova and Natalia Simus have also joined the group. Finally, Gerold Baier, a collaborator

Group Leader	Research Associates
Dr. Ursula Kummer Tel.: +49-6221-533225 Fax: +49-6221-533298	Ralph Gauges, Femke Mensonides (since July 2005), Dr. Sven Sahle, Dr. Natalia Simus (since June 2005), Dr. Iulian Stoleriu (since June 2005), Dr. Irina Surovtsova (since April 2005), Dr. Andreas Weidemann, Dr. Jürgen Zobeley (until November 2005)
	Doctoral Students
	Jürgen Pahle, Katja Wegner
	Students
	Sebastian Boland, Lebriz Ersoy, Noemi Hummel, Dr. Anette Kurz, Ralph Voigt
	Visiting Scientist
	Dr. Ibrahim Coumbassa (September 2005), Prof. Gerold Baier (since August 2005)
	Visiting Student
	Jens Christian Brasen (November 2004)
	Interns
	Anh Do, Sarah Lilienthal

on correlation methods from the University of Cuernavaca, Mexico, is spending a sabbatical with us.

In 2005 the Copasi simulation suite made publicly available in 2004 became a widely used software, with ca. 6000 downloads within one year. The positive feedback from the community is encouraging.

In September 2005 the group organized the 4th Workshop on the Computation of Biochemical Pathways and Genetic Networks.

Fig. 2: Solid as a rock: The BCB group in 2005

2.2 The Scientific Databases and Visualization (SDBV)

Fig. 3: The SDBV group in 2005 (from left to right): Ulrike Wittig, Olga Krebs, Renate Kania, Saqib Mir, Stefanie Anstein, Martin Golebiewski, Isabel Rojas, Jasmin Saric, Andreas Weidemann

The main development for the Scientific Databases and Visualization (SDBV) group during the year 2005 was the SABIO-Reaction Kinetics (SABIO-RK) database (see Section 3.5, SABIO-RK). With this database and its web-based user interface (first beta-version released early 2006) the group provides methods and tools supporting researchers in the analysis and management of scientific data, more specifically of biochemical data. This project addresses problems such as the integration of different data types and formats, the storage of these data, and methods for querying and navigating the data. During the coming year the group will be turning to issues connected with the visualization of the data in the database. The various approaches pursued will enhance the understanding of these data, the modeling and storage of experimental data, and the procedures employed in obtaining them.

Another main research path for the group is the extraction of information from biochemical texts. This plays an important role in the process of database population and curation. Besides this, the experience gained in the manual revision and information-extraction from literature sources allows the definition and evaluation of automatic information extraction rules. To provide researchers with the possibility of incorporating SABIO-RK as part of their research workflows, the group is researching and developing semantic web/grid services for biochemical data, mainly related to the kinetics of reactions. More information on information extraction and web/grid service developments in the group during 2005 can be found in section 3.4 (AMBOS).

Group Leader

Dr. Isabel Rojas
Tel.: +49-6221-533231
Fax: +49-6221-533298

Research Assistants

Martin Golebiewski, Renate Kania, Dr. Olga Krebs, Jasmin Saric, Dr. Andreas Weidemann, Dr. Ulrike Wittig

Klaus Tschira Foundation Scholarship Holder

Saqib Mir

Students

Stefanie Anstein, Gerhard Kremer, Xiaon Chen

2.3 Natural Language Processing (NLP)

Lexical Semantic Processing (Iryna Gurevych)

For school graduates the decision about their future profession is absolutely essential. It is also a hard one to make. While it is fairly easy to describe to people what they might like to do using their native language, which professions match the description best? Professional career counseling is time-consuming, both for advice-seekers and advice-givers. The automation potential for this process is immense. Advice-seekers can then obtain an on-the-spot set of suggestions based on their description.

The Natural Language Processing Group is working on minimizing the human labor involved in accessing the relevant information efficiently and conveniently. We are developing algorithms that automatically extract the meaning of a user's information needs from its natural language description. This "meaning" is compared with the meanings of documents available as potential answers to information searches. The most relevant documents are returned to the users immediately. The semantically enhanced information retrieval system has been applied to the task of career counseling. Figure 5 shows the first prototype in action.

Computing document relevance is crucial to the process. We have designed an algorithm that goes beyond conventional keyword searches. Instead, it uses linguistic knowledge, more specifically, the semantic relatedness of words.

Group Leader	Research Associates
Dr. Michael Strube Tel.: +49-6221-533243 Fax: +49-6221-533298	Dr. Iryna Gurevych (until Oct. 2005), Margot Mieskes, Christoph Müller
	Klaus Tschira Foundation Scholarship Holders
	Tomasz Marciniak, Simone Paolo Ponzetto
	Diploma Students
	Vesna Cvoro, Hendrik Niederlich
	Students
	Svetlana Dedova, Kerstin Heß, Elena Loupanova, Violeta Sabutyte, Iryna Schenk
	Interns
	Katja Filippova, Aleksandrs Galickis, Bela Usabaev

Fig. 4: The NLP group 2005: (from left to right) Michael Strube, Grainne Toomey, Violeta Sabutyte, Simone Paolo Ponzetto, Vanessa Doyle, Christoph Müller, Margot Mieskes, Ekaterina Filippova, Matthew White

At the heart of the algorithm is a computational model of semantic relatedness extended to handle different kinds of words, such as “cake – bakers”, “to program – software engineer”.
The semantic distance between two words can be determined by this model. Accordingly, the program would know that “cakes” are more closely related to “bakers” rather than to “software engineers”.

In order to test the effectiveness of our programs, we have created a collection of word pairs. Each word pair is la-

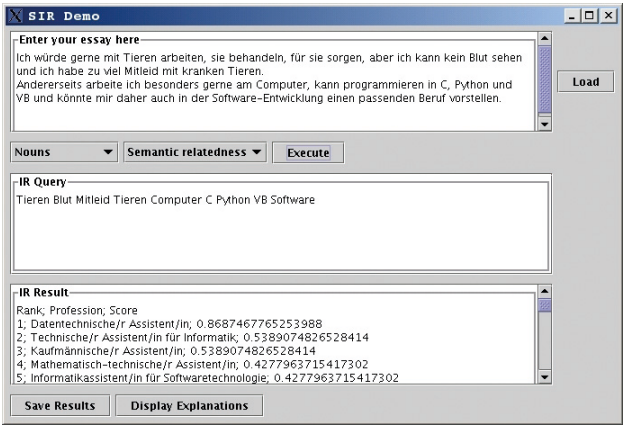


Fig. 5: The first prototype of the semantically enhanced information retrieval system

beled with human judgments of semantic relatedness. Thus we can permanently monitor the advantages and shortcomings of each algorithm. To make the user's search even more effective, the system automatically expands the user's request with semantically related words. Then a search for "cakes" would return documents containing e.g. "sponge cakes" and "fruit cakes."

Alongside our basic-research activities we have also been implementing software for Natural Language Processing and have made this work available to interested researchers in the field. A Java-based API to a German lexical-semantic wordnet, GermaNet, was released in July. This software already enjoys great popularity in the research community working with the German language.

Natural Language Generation (Tomasz Marciniak)

Our work in the area of Natural Language Generation (NLG) spans two areas of research relevant to the field: domain modeling and tactical generation. The first objective is to design an ontological model for the domain of route directions, focusing on those elements of the process of route-following that find direct manifestation in the linguistic description. Accordingly, the primary motivation for the identification of the relevant entities in the route direction domain comes from the analysis of the relevant texts. The ontology needs to serve as a formal specification of the content of route directions, and as such define the input to the tactical generator. While being necessarily domain-specific, the ontology should be designed to anticipate extensions to other instructional domains.

The second goal is the development of a data-driven tactical generator capable of realizing the linguistic form of route directions from the underlying conceptual specification sanctioned by the ontology. Formally, the task of linguistic realization can be defined as one-to-many mapping between the conceptual content and the grammatical form of a linguistic expression. It has long been recognized as a knowledge-intensive process, involving a range of linguistic decisions applying to different levels of the linguistic organization, i.e. discourse and clause and phrase levels. To

handle these decisions an NLG system requires a substantial amount of linguistic knowledge about how both syntactic and lexical constructions can be used to encode the intended meaning.

Two major considerations involved in designing an NLG system are: the architecture of the system and the knowledge source. We propose a generation model that provides a novel perspective on both issues. Firstly, we subscribe to the principle of lexicon and syntax continuity proposed by Langacker in 1988 and draw no dichotomy between lexical and syntactic decisions, regarding them all as forming a single category of generation tasks. Each such task is further considered as constituting a classification problem. The generation process is then modeled as a series of classifications integrated within a discrete optimization model (i.e. in the framework of Integer Linear Programming). Secondly, the linguistic knowledge required for generation is not specified in an explicit way. Instead, in the model of generation proposed here, an annotated corpus is used as a source of linguistic data from which machine-learning algorithms can learn how to perform the individual tasks.

Coreference Resolution (Simone Paolo Ponzetto)

Explaining how people use words to talk about things in the world is a very old problem (dating back at least to the Greek philosophers) that pervades our daily lives. In a nutshell, robust natural language understanding involves being able to recognize what things we are talking about, and how these entities interact with each other.

People tend to refer to the same things with different words. One might for instance refer to one and the same soul-funk musician as “Prince”, “The Minneapolis Genius”, “The artist formerly known as Prince”, “TAFKAP”, or simply “The artist”. Nevertheless it is always the same person we are referring to. This is the task we are engaging with in the framework of coreference resolution. It involves identifying the unique persons or things we are talking about on the basis of different ways of describing them. Another core challenge for a state-of-the-art information extraction system is that of identifying the relations between the

participants involved in the situation or event described by a text. For instance, one would like to be able to extract from a sentence such as "The Minneapolis Genius played tirelessly for almost three hours" the knowledge that Prince is the performer at the event described, namely, a concert. This amounts to performing shallow semantic parsing of unrestricted text in terms of Semantic Role Labeling (SRL), i.e. identifying phrase chunks that are in a given relation to the predicate – e.g. identifying the AGENT, PATIENT, TIME, and PLACE expressions in the sentence. As this implies being able to identify "who" (did) "what" (to) "whom", "when", "where" and "why" in a text, this represents an essential task for Information Extraction, Question Answering, Summarization, and, in general, all those NLP tasks in which natural language understanding is needed.

Building on work in coreference resolution carried out in previous years, we investigated how we could bring these two problems together, as both deal with semantic text mining and are essential for building robust, scalable, real-world NLP applications requiring a level of semantic representation. We explored novel data representation techniques for learning semantic parsing (Ponzetto and Strube, 2005) and merged this level of information into a machine-learning-based coreference-resolution engine. Future work will center on exploiting natural language taxonomies (e.g. WordNet and other repositories of world knowledge) as sources of semantic information for coreference resolution.

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 studying 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.

The research of the MCM group in 2005 is described in this report under four project headings:

• SYCAMORE: SYstems biology's Computational Analysis and MOdeling Research Environment (see section 3.3). The aim of the SYCAMORE project is to develop a system of

Group Leader	Research Associates
Dr. Rebecca Wade Tel.: +49-6221-533247 Fax: +49-6221-533298	Dr. Razif Gabdoulline, Dr. Stefan Henrich, Dr. Matthias Stein, Dr. Ting Wang, Dr. Peter Winn (since February 2005)
	IT Specialist
	Dr. Stefan Richter (since October 2005)
	Klaus Tschira Foundation Scholarship Holders
	Dr. Vlad Cojocaru (since October 2005), Anna Feldman-Salit (since August 2005), Domantas Motiejunas
	Students
	Tim Johann, Frederik Ferner, Matthias Janke, Bruno Besson (March-August 2005), Stefan Ulbrich (until August 2005), Yang Xiang Zhou (July 2005)
	Guest Scientist
	Dr. Renate Griffith (January to March 2005)

software tools and computational methods to support concerted computational and experimental approaches to systems biology problems. The project is a collaboration between the BCB and MCM groups at EML Research. It is part of the Modeling Platform of the Federal Ministry of Education and Research's 'Hepatosys' Systems Biology program. The MCM group is working on the development and application of methods to harness protein structural information for systems biology projects. This year, we have worked on three aspects aimed at the estimation of kinetic parameters from protein structural information: Firstly, the creation, in SYCAMORE, of integrated software tools for querying sequence, structure, and enzyme kinetic databases to assemble the data necessary for deciding whether parameters can be estimated on the basis of structural data, and, if so, what approach can be taken to doing so. Secondly, the development of qPIPSA (quantitative Protein Interaction Property Similarity Analysis), based on our PIPSA methodology, to estimate kinetic parameters from the molecular interaction fields of proteins. Thirdly, the application of ligand docking approaches to investigate both substrate and protein-specific influences on enzyme kinetic parameters.

• Modeling and Simulation: From the Molecule towards the Cell (see section 3.8). The principal aim of this project is the development and application of computer-aided methods for predicting and simulating the interactions of proteins. The methods are primarily designed to exploit the three-dimensional structures of macromolecules. Software tools are made available to users and can be found from the web addresses below. PIPSA2 was released this year for public use in studying protein interaction properties. Improvements were made in the NPSA implicit solvent model and the RAMD method for molecular-dynamic simulation of proteins, and their software implementations were made available to other users. The group carries out applications in collaborative projects. These include modeling the structures of protein-ligand and protein-protein complexes, predicting protein interaction properties and energetics, studying how protein dynamics affects binding processes, and investigating enzyme mechanisms.

• TASSFUN: Target-Specific Scoring Functions (see section 3.9). In this project, COMBINE (COMparative BINding Energy) analysis is being applied and developed to address the problem of designing drugs that are selective between related protein targets. This project is being carried out in collaboration with AstraZeneca, Sweden.

• Modeling Macromolecular Motions in the Cell by Brownian Dynamics Simulations (see section 3.10) The aim of this project is to develop the methodology to simulate the diffusional motion of many proteins or of large proteins made up of relatively rigid domains connected by flexible linkers via Brownian dynamics simulation based on atomic detail protein structures. Our SDA (Simulation of Diffusional Association) software is being developed for this purpose. The main focus this year has been on introducing hydrophobic interactions into intermolecular forces and suitably calibrating and testing these. The project started this year and is being carried out by Razif Gabdoulline with the support of a postdoctoral fellowship from the BIOMS Center for Modeling and Simulation in the Biosciences, Heidelberg.

For more information on the group's projects, see:

- www.eml-research.de/english/research/mcm
- http://projects.villa-bosch.de/mcm.

Fig. 6: The MCM group 2005 (from left to right): Rebecca Wade, Vlad Cojocaru, Peter Winn, Mai Zahran, Stefan Henrich, Stefan Richter, Anna Feldman-Salit, Matthias Stein, Razif Gebdoulline, Domantas Motiejunas

2.5 Information Technology in the Health Sector: Dr. Feelgood

The Dr. Feelgood project was successfully completed in 2005. As mentioned in the previous report, the focus of the work was on integrating additional data sources into the original DigiCoach, which is a set of accelerometers. Those sources include video recordings of training laps and special techniques for analyzing the fluid dynamic processes in the water as a result of the swimmer's motions.

The outcome of the project has generally been considered a great success, notably by the Institute of Sport and Sport Science of the University of Heidelberg and the Olympic Training Center, both of which had been project partners for the development and evaluation of DigiCoach from the outset. As a matter of fact, the Olympic Training Center plans to make routine use of the DigiCoach technology in the training of their junior athletes.

EML Research has decided not to continue its activities in this area. Clearly, the reason is not lack of success. It has, however, become evident that in order to stay competitive in this field, which in the meantime has become very interesting for different sectors of industry, we would have to allocate substantially greater resources to it – especially in personnel terms. But for a small company like EML Research this would have severely affected other project areas, so we have no choice but to discontinue our research activities in an area that has brought us very considerable success.

Group Leader (Provisional)	Doctoral Student
Prof. Dr. Andreas Reuter Phone: +49-6221-533201 Fax: +49-6221-533298	Markus Buchner
	Student Worker
	Alexander Folz

3.1 Simulating Biochemical Pathways (BCB)

The development of computational methods for the analysis and simulation of biochemical networks requires the application of those methods to real-world problems. It is very gratifying to witness 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 of our own.

Simulating the Metabolism of Neutrophilic Leukocytes

Project Manager

Dr. Ursula Kummer

Project Members

Dr. Ibrahim Coumbassa, Anh Do, Noemi Hummel, Jürgen Pahle, Dr. Irina Surovtsova, Dr. Jürgen Zobeley

Neutrophilic leukocytes are white blood cells 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. In this context, we focused in previous years on the computational investigation of two key modules, reactions featuring myeloperoxidase and the central metabolism via glycolysis. In 2005 we started to integrate these modules with new modules on the hexose monophosphate shunt and calcium signal transduction to obtain a more universal picture of the biochemical processes underlying neutrophil activation.

After the experimental verification of our computational predictions that glycolysis is mainly responsible for the frequency of the observed metabolic oscillations, whereas myeloperoxidase is crucial for a change in amplitude, we are now able to see how the different factors involved interconnect.
For example, our model suggests that calcium signal transduction triggered upon stimulation of the leukocytes is able to activate glucose import, which in turn results in a switch of the metabolism from glycolysis to the hexose monophosphate shunt supplying NADPH for the production of reactive oxygen species.

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)

Sponsors

Klaus Tschira Foundation

European Science Foundation

3.2 COPASI (BCB)

Group Leader

Dr. Ursula Kummer

Project Members

Ralph Gauges, Sarah Lilienthal, Jürgen Pahle, Dr. Sven Sahle, Katja Wegner, Dr. Natalia Simus, Ralph Voigt

Cell metabolism is a very complex dynamic system, consisting of thousands of different chemical substances and reactions. One key method for understanding the behavior of such a system is modeling and simulation. The basic idea is that a model reproducing the behavior of the system is a very valuable tool both for understanding the system and for practical applications. In this context the modeling consists of identifying the chemical substances that are relevant in the metabolism and specifying all the reactions that consume or produce those substances. Furthermore, a mathematical description of the speed at which these reactions take place has to be given. A computer can then be used to simulate the model. To support this approach it is important to have software tools that are not only powerful but also easy to use for scientists who may not have expert knowledge of the numerical algorithms involved.
This is the aim of the COPASI project. COPASI is a software package that brings together various modeling, simulation, analysis, and visualization methods for biochemical pathways. It is designed 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.

The user interface of COPASI consists of a tree view on the left and a dialog field on the right (see e.g. Annual Report 2004). From the tree view the user can choose different dialogs for entering a model and can then simulate, analyze, or visualize it. The model can be simulated either via numerical integration or stochastical simulation. The numerical integration algorithm we use is the well-established LSODA, while the stochastical simulation is done with Gibson's modified Gillespie algorithm. Analysis methods that can be used on a model include stability analysis of steady states, metabolic control analysis (MCA), and detection of elementary modes or mass conservation relations. In 2005 different methods for parameter fitting and optimization were implemented. In addition, methods for scanning the parameter space have been made available (Figure 7). Output of results can be achieved with a simple integrated plotting tool.

To gain acceptance in the scientific community it is important for a software tool like COPASI to be able to exchange data with other tools. Therefore COPASI supports SBML (Systems Biology Markup Language) as a way of reading and writing model information. Also, COPASI can now export a mathematical description of the model, either as a differential equation or as a source code in the C programming language. This makes the model available for software that cannot import SBML files.

COPASI is closely linked to the SYCAMORE project, which involves the construction of models from different information sources. One of the problems in combining information

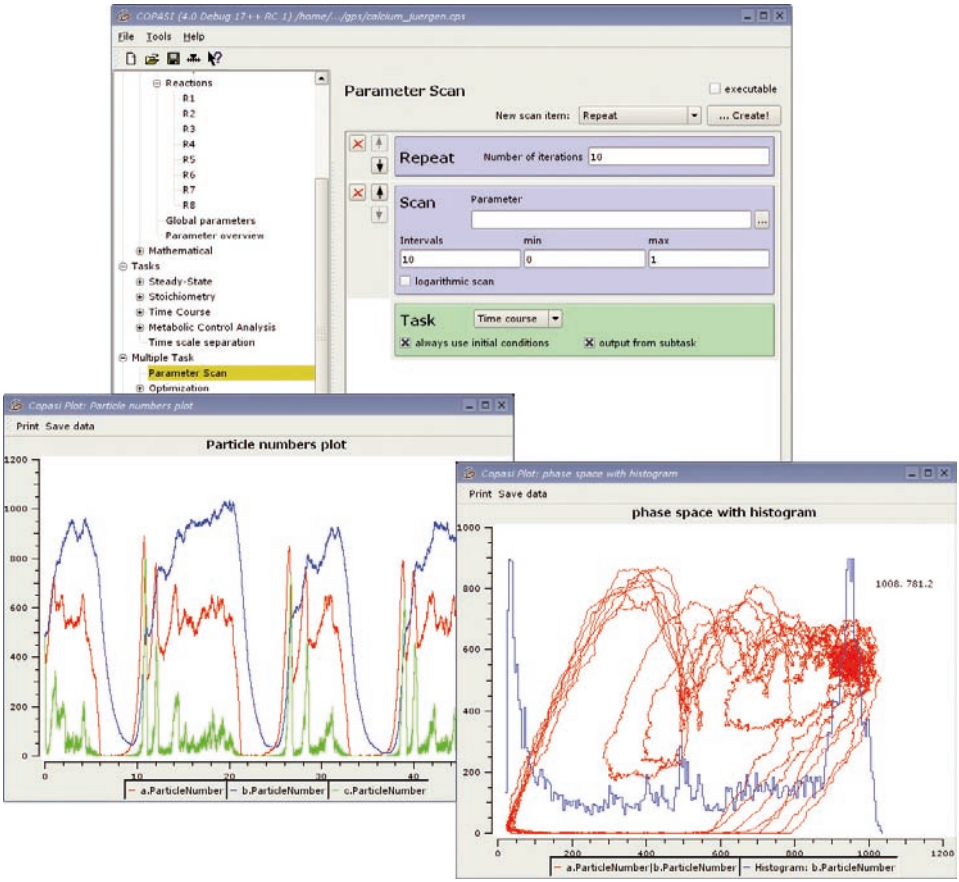


Fig. 7: Screenshot of the Copasi GUI showing both the window for the parameter scans and output windows

from different sources into one mathematical model is that the software must be able to tell whether two mathematical expressions have the same meaning even if they look different. In 2005 we employed symbolic algebra techniques to develop algorithms for solving this problem.

COPASI is a joint project between our group and Dr. Pedro Mendes' research group at the Virginia Bioinformatics Institute in Blacksburg, USA. Test versions of COPASI have been published since 2004 and have been 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

SBML Layout Extension

Work on the SBML Layout Extension continued in 2005. Although only minor changes were made to the specification itself, a lot of work has gone into implementation. One implementation is based on libsbml (www.sbml.org/libsbml.html), a library for reading, writing, and checking SBML files. Like libsbml, the layout extension is written in C++, but language bindings for C, Java, and Python also exist. This implementation is no longer available as a separate patch-set for application to the libsbml sources. Instead, the whole implementation has been integrated into the source-code tree of libsbml. As the implementation is now distributed together with libsbml, users do not have to invest additional time getting the layout extension to work. To make sure that changes to the core libsbml source code do not break the layout implementation, an extensive set of several hundred unit tests have also been implemented, thus allowing for fast, automatic code checking prior to a release.

An additional implementation in the form of an XSLT style sheet has also been written..This stylesheet can be used to convert an SBML file with layout information into a Scalable Vector Graphic (SVG). As the name suggests, SVG is a vec-

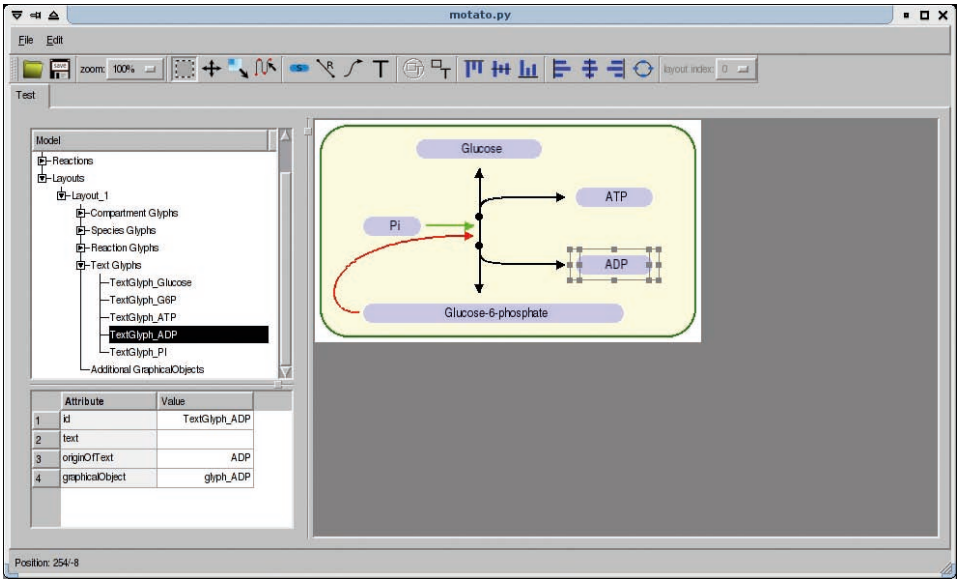


Fig. 8: New GUI of the MOTATO tool

tor-based graphics format, therefore the image can be scaled to any resolution without sacrificing image quality, as would be the case with pixel-based images. Accordingly, the SVG images generated from layout information are ideal for printing, e.g. in publications, or for presentations and posters.

Work on MOTATO, a tool for the generation and editing of reaction network layouts, has also been progressing, and a lot of new functionality has been implemented. Currently, MOTATO is being translated from Python to C++ using the new Qt4 toolkit to generate the user interface. Once this translation is finished, MOTATO will even be able to handle large reaction layouts, which, due to some performance bottlenecks, is not the case with the current Python implementation.

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 to be incorporated into Copasi include:

Graphical Visualization of Biochemical Pathways and their Simulation Results

Modeling and simulation techniques are becoming increasingly important for the understanding of biochemical processes in living cells. These processes consist of single reactions that build up reaction pathways. In a computational model these pathways can be visualized as graphs. Graph nodes represent reactants, while edges represent reactions between reactants. During the last few years we have developed a graph layout algorithm [Wegner 2005a] that automatically places such graph nodes with due consideration for biological conventions.

Researchers create mathematical models of such pathways and use the graphical representations to find visual relationships between reactants. These models describe how the concentration values of reactants change over time and are simulated with the aid of computer programs, e.g. COPASI. These simulations result in a large amount of numerical data. Since numerical data can be better analyzed and interpreted visually, in 2004 we started to integrate simulation results into a graph representation.

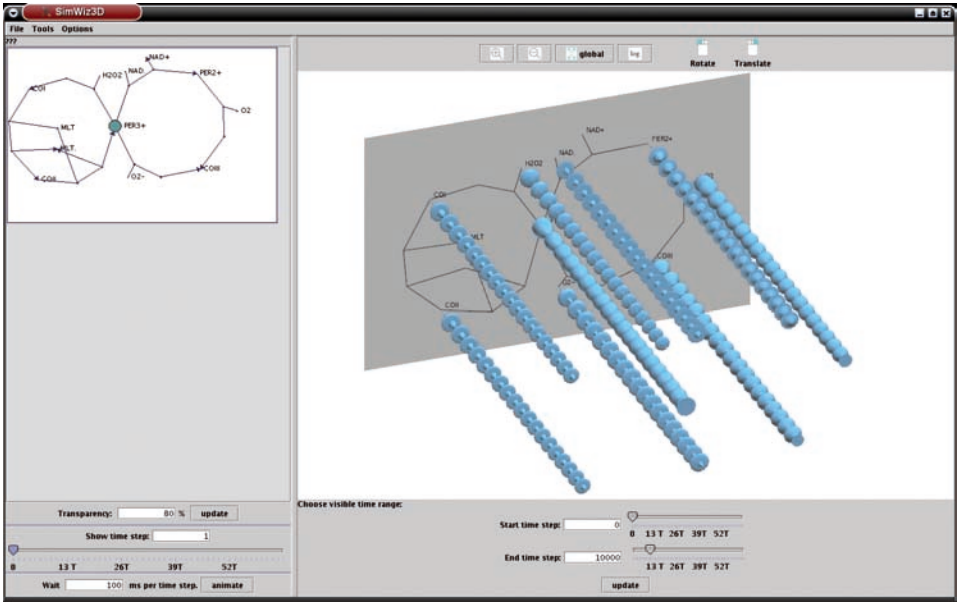


Fig. 9: GUI of SimWizD: The metabolic network is situated in a plane, the third dimension depicting time. The concentration of each metabolite is coded as the size of its corresponding node

In 2005 we developed SimWiz3D [Wegner 2005b] using the x,y plane to display the graph representation and the z-axis as time axis. Tubes with increasing and decreasing diameters represent concentration changes (Figure 9). SimWiz3D contains a lot of user interactions for filtering, resampling, or highlighting data. In addition, we have intregrated correlation analysis methods assisting reseachers in interpreting the data. We have also developed a new correlation method specially suited for spatio-temporal data [Muller 2005c].

Collaboration Partners

Prof. Dr. Jürgen Hesser, University of Mannheim, Mannheim, Germany

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 (BCB, MCM)

Systems biology aims at a better understanding of the biochemical network in the living cell. In order to achieve this aim, experimental and theoretical investigations have to go far beyond studying isolated genes, proteins, and reactions to study whole systems of biochemical reactions and their emergent properties.
To support this research new methods have to be established to integrate experimental and theoretical methods in a more efficient manner.

In this project, we are developing a system of tools and methods to support the combination of diverse computational and experimental approaches. Sycamore is a joint project of the Bioinformatics and Computational Biochemistry Group and the Molecular and Cellular Modeling Group at EML Research. It is supported by the Federal Ministry of Education and Research. Sycamore is closely linked to Copasi (3.2), with Copasi providing some of the methodology used in Sycamore.

Apart from the implementation of Sycamore, new computational methods have to be developed for the ambitious goals the project has set itself:

Complexity Reduction and Sensitivities in Large Biochemical Reaction Networks

Focussing on increasingly large and complex cellular systems, the aim of „understanding“ consists of two key aspects: first, imitating essential features (both qualitatively and quantitatively), and second, analyzing their sensitivity to biochemical conditions. “Precise” simulation requires a quantitative description of the dynamics (usually in the form of nonlinear differential equations) and implementing its numerical solution. Due to the complexity of biochemical reaction networks, however, so-called complexity-reduction algorithms play a crucial role in making simulations realizable „in silico“ (i.e. on a computer). Furthermore, they can act as tools for gaining insight into the relevance of metabolites. With respect to later comparisons with experimental findings, these algorithms also provide criteria for the required precision of biochemical parameters.
Each complexity-reduction method is based on a criterion for evaluating the computational relevance of subsystems.

Our first approach, initiated in 2003, refers to the different time scales of biochemical kinetics and captures the distinction between „fast“ and „slow“ modes detected adaptively. This modified ILDM method („intrinsic low-dimensional manifold“) has come up with good results for the peroxidase-oxidase reaction (see Annual Report 2003/04). Meanwhile it has also produced interesting results when applied to glycolysis in liver cells. These numerical results also prove that a more detailed mathematical model is required here. Accordingly, we intend to focus on this modeling challenge.

A further complexity-reduction design combined with sensitivities has been started recently. The key criterion for minimizing numerical errors directly is to replace the distinction between „fast“ and „slow“ modes by studying the aspects of error estimates and analysis. We are hoping to come up with an effective new method soon.

Project Managers

Dr. Ursula Kummer (BCB)
Dr. Rebecca Wade (MCM)

Project Members

Dr. Razif Gabdoulline (MCM)
Dr. Matthias Stein (MCM)
Bruno Besson (MCM)
Ralph Gauges (BCB)
Jürgen Pahle (BCB)
Sven Sahle (BCB)
Jürgen Zobeley (BCB)
Sebastian Boland (BCB)
Lebriz Ersoy (BCB)
Dr. Anette Kurz (BCB)
Dr. Irina Surovtsova (BCB)
Dr. Andreas Weidemann (BCB)

Collaboration Partners

Dr. Dirk Lebiedz, Julia Kammerer, IWR, University of Heidelberg, Heidelberg, Germany

Approximate Stochastic and Hybrid Simulation Methods in Biochemistry

Random fluctuations of molecule numbers inside a cell can significantly change its dynamic behavior. For this reason, stochastic simulation methods have been developed that are able to reproduce these fluctuations correctly. In contrast to the conventional deterministic approach, which utilizes differential equations (ODEs) and does not take random effects into account, the stochastic methods are based on Monte Carlo algorithms for computing the probabilistic system behavior. However, the stochastic simulation of biochemical networks is computationally demanding, especially when high particle numbers are present in the system.

A variety of different methods have been proposed to mitigate this problem. Among them are the approximation of the discrete particle numbers by continuous variables (Stochastic Differential Equations), or the grouping of reaction events (PW-DMC, τ-Leap Method). Hybrid methods

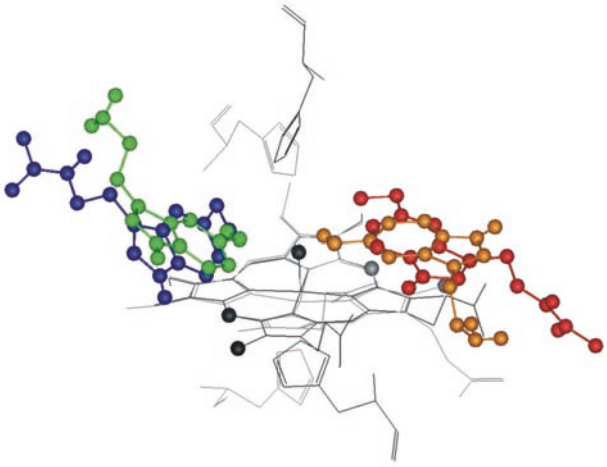


Fig. 10: Superposition of the active sites of the compound I states of horseradish peroxidase (HRP) (black) and myeloperoxidase (MPO) (grey) showing differences in selected docking modes of the indole substrates melatonin (blue, red) and serotonin (green, orange). The hemes and proximal and distal histidines of the active sites are shown. C18 methyl, C20, and ferryl oxygen are shown as balls (black: HRP; grey: MPO). The docking modes shown orient their indole substituent inwards. The docking modes to HRP are on the left side, those to MPO on the right side of the figures (corresponding to the channel locations in the two proteins). See [Hallingbäck 2006]

are an attempt to combine the advantages of the deterministic and stochastic approaches. They divide the whole biochemical system into sub-systems and make parallel use of appropriate deterministic or stochastic methods on those sub-systems.

We have integrated the τ-Leap Method, published by Gillespie in 2001(„Approximate accelerated stochastic simulation of chemically reacting systems“, J. Chem. Phys., 115/4, pp.1716-33, 2001), as a module into our COPASI software system. This method performs approximate stochastic simulations of biochemical networks and thus represents an

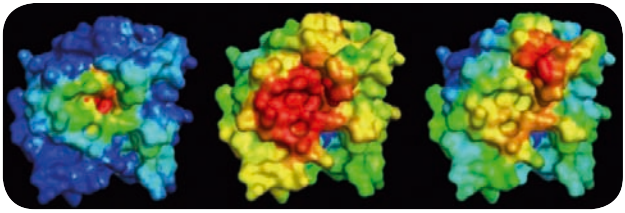


Fig. 11: The surface in the region of the active site of a glycolytic enzyme colored to show the regions responsible for different kinetic properties. (Left) Conservation of electrostatic potential (increasing from blue to red). The active site displays the greatest conservation and is colored red. The correlation of the electrostatic potential with the Kcat/Km (middle) and) Km kinetic parameters (right) is shown with increasing correlation from blue to red. The red region is centered on the active site for Kcat/Km and on a flexible loop forming a lid over the active site when a substrate is bound

intermediate approach between pure stochastic and deterministic simulation methodologies. It completes our suite of stochastic, deterministic, and hybrid simulation modules already implemented in COPASI and thus made available through Sycamore.

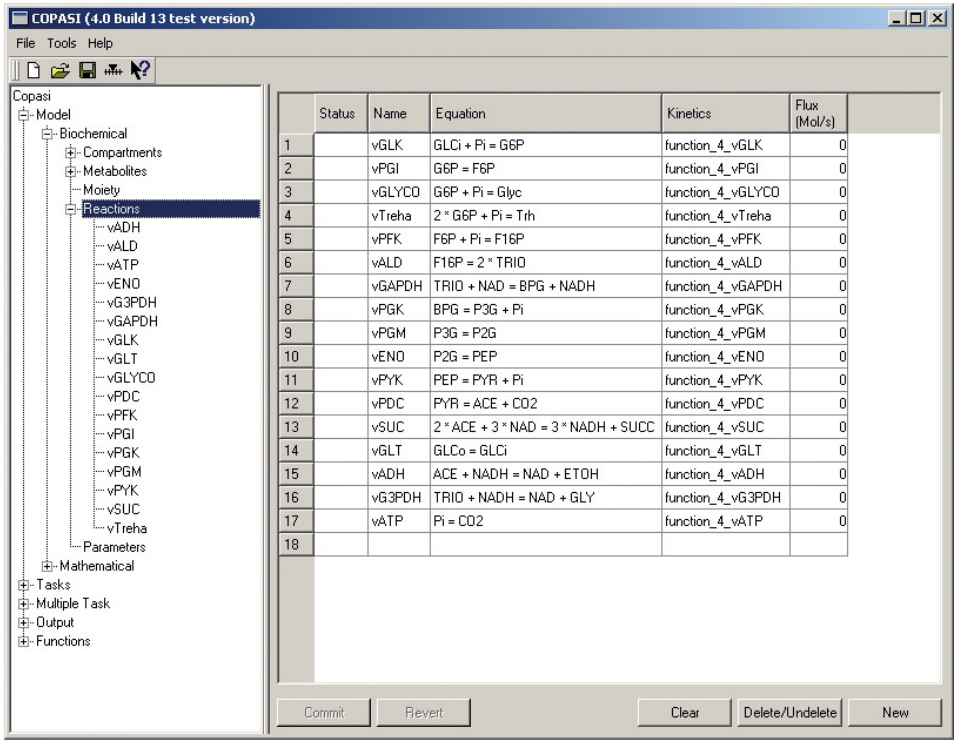
We have also extended our collection and analysis of the different published simulation methods to include new approaches like the two different hybrid methods proposed by Salis and Alfonsi (both published in 2005). In addition, we have set up simple criteria for the classification of hybrid simulation methods, including their partitioning policy or their ability to account for variable probabilities during one computation step, and have classified all known hybrid methods according to that scheme.

Using Protein Structures in Systems Biology Projects

In the SYCAMORE project, the MCM group is working on the development and application of methods to harness 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 the development of methodology and software for the protein structure-based estimation of kinetic parameters. Firstly, we have created integrated software tools for querying sequence, structure, and enzyme-kinetic databases from a web browser to assemble the data necessary for deciding whether parameters can be estimated on the basis of structural data, and, if so, what approach should be taken in doing so [Besson 2005]. Secondly, we are developing qPIPSA (quantitative Protein Interaction Property Similarity Analysis), based on our PIPSA methodology, as a cost-effective approach to estimating kinetic parameters from molecular interaction fields [Wade 2005a]. PIPSA2 was made publicly available on our website this year and permits the classification of proteins by quantifying the similarity of their 3D molecular interaction fields, for example, their electrostatic potentials. qPIPSA is now undergoing validation tests. It has been applied to enzymatic reactions

Fig. 12: Screenshot of the SYCAMORE user interface



	Status	Name	Equation	Kinetics	Flux (Mol/s)
1		vGLK	GLCi + Pi = G6P	function_4_vGLK	0
2		vPGI	G6P = F6P	function_4_vPGI	0
3		vGLYCO	G6P + Pi = Glyc	function_4_vGLYCO	0
4		vTreha	2 * G6P + Pi = Trh	function_4_vTreha	0
5		vPFK	F6P + Pi = F16P	function_4_vPFK	0
6		vALD	F16P = 2 * TRIO	function_4_vALD	0
7		vGAPDH	TRIO + NAD = BPG + NADH	function_4_vGAPDH	0
8		vPGK	BPG = P3G + Pi	function_4_vPGK	0
9		vPGM	P3G = P2G	function_4_vPGM	0
10		vENO	P2G = PEP	function_4_vENO	0
11		vPYK	PEP = PYR + Pi	function_4_vPYK	0
12		vPDC	PYR = ACE + CO2	function_4_vPDC	0
13		vSUC	2 * ACE + 3 * NAD = 3 * NADH + SUCC	function_4_vSUC	0
14		vGLT	GLCo = GLCi	function_4_vGLT	0
15		vADH	ACE + NADH = NAD + ETOH	function_4_vADH	0
16		vG3PDH	TRIO + NADH = NAD + GLY	function_4_vG3PDH	0
17		vATP	Pi = CO2	function_4_vATP	0
18					

in glycolysis for a diverse range of organisms. The method is helpful in identifying mechanistically important regions in enzymes, where the computed interaction fields are correlated with kinetic constants (see Figure 11). The PIPSA approach is also being used to investigate nucleotide exchange factor proteins involved in the endocytosis pathway studied in Dresden (A. Deutsch, TU, M. Zerial, MPI-CBG) and cytochrome P450s studied in Stuttgart (M. Reuss, R. Schmidt, J. Pleiss, University of Stuttgart), both in the context of the Ministry of Education and Research's HepatoSys network. Thirdly, we have applied ligand docking approaches to investigate both substrate and protein-specific influences on kinetic parameters in an investigation of the catalysis of indole substrates by two peroxidases, myeloperoxidase and horseradish peroxidase [Hallingbäck 2006] (see Figure 10).

Implementation of Sycamore

In 2005, the user interface for Sycamore was implemented, including some additional functionalities. Both SABIO-RK, structure-based methods and Copasi are already directly accessible through the interface. SBML-files can be loaded, and an editor for the files has also been implemented.

Figure 12 shows a screen shot of the user interface. As with Copasi, users can navigate through the system via a tree in the left panel.

Collaboration Partners

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

Sponsors

Federal Ministry of Education and Research (BMBF)

Klaus Tschira Foundation

3.4. UniNet (BCB)

Project Manager

Dr. Ursula Kummer

Project Member

Dr. Iulian Stoleriu

This project, which started in summer 2005, is part of the Unifying Networks for Science & Society (UniNet) Consortium, a European research initiative centering on the identification and investigation of common mathematical structures in various complex networks, such as genetic, metabolic, neuronal, economic, and ecological networks. Within the project we are aiming at the development of methods for using the topological information w.r.t. metabolic networks without knowledge of the specific parameters to an extent where in some cases it is possible to e.g. predict the stability of the network in question. For this purpose, stoichiometric network analysis (SNA) is one of the methods employed.

SNA provides a convenient tool for the decomposition of the entire network operating at a stationary state into subnetworks. It also indicates those subnetworks that can cause changes in the network's steady-state stability. SNA works by splitting up the whole network into subnetworks generated by the elements of a basis in the space of steady states. Each element in this basis is called extreme current (or, a similar notion, elementary flux mode), and the subnetwork generated by it usually has a smaller order than the initial network, thus making it easier to handle. From the physiological point of view, one can think of these elementary flux modes as feasible biochemical pathways operating at a steady state in a metabolic network. One of our goals is the implementation of the stability analysis of these modes in COPASI. At present we are studying the effects that changes in the kinetics type have on the extreme currents to see what happens to the stability of the currents when the system of equations is rescaled.

Collaboration Partners

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

Sponsors

European Union

Klaus Tschira Foundation

3.5 BioSim (BCB)

Project Manager
Dr. Ursula Kummer
Project Member
Femke Mensonides

In this project, which started in 2005, we attempt to develop methods for computationally supporting the prediction of the biochemical fate of drugs in an organism. For drug development it is important to determine these possible metabolic fates of pharmaceutically active compounds (PAC) inside living cells. Moreover, the stereo-geometry of PACs and the intermediates of their metabolic fates essentially determine the physiological effects of a chiral drug, as the thalidomide (contergan) tragedy showed in the 1950s. The analysis of the metabolic fates of PACs, or xenobiotics in general, demands a systems approach, since individual components of the metabolic networks are closely linked both by common substrates or cofactors and by allosteric enzyme regulation. Here, we use a well-characterized model system, the yeast Saccharomyces cerevisiae. As model PACs we consider the three ketones ethyl acetoacetate, ethyl 4-chloro-acetoacetate, and ethyl 4,4,4-trifluoro-acetoacetate. These ketones enter the cell and are reduced by at least two alternative enzymes to a mixture of the two enantiomeric forms of the corresponding carbinol.

The metabolism of the three ketones in yeast has been characterized experimentally by our collaborator, Martin Bertau. Furthermore, a preliminary kinetic model of the reduction of the xenobiotics (PharmaBiosim) has been developed by Lutz Brusch. Analysis of the model suggests a key role for the cofactor NADH in the flux distribution of the system.

Recent experimental results show the activation of an extensive biochemical network of stress response due to the presence of the ketone, which, in turn, has an effect on the metabolism of the compound. These findings call for an extension of the existing model. However, since the new network is very large, it is hard to translate into a kinetic framework at once. Therefore we have commenced with the parallel development of a stoichiometric model of the ketone metabolism. In addition, both experimental and theoretical work has been initiated to extend the model to account for aerobic conditions, which will, for instance, change the availability of the important cofactor NADH.

Collaboration Partners

The BioSim consortium, especially Dr. Martin Bertau; Dr. Lutz Brusch (TU Dresden)

Sponsors

European Union

Klaus Tschira Foundation

3.6 AMBOS - Algorithms and Methods for the Development of Biochemical Ontology-Based Database Systems (SDBV)

The main objective of the AMBOS project is the development of methods and applications to support the integration and analysis of data related to networks of biochemical reactions. Its underlying database (SABIO System for the Analysis of Biochemical Pathways) contains information about biochemical reactions, their compounds, their kinetics, and the pathways in which they participate. All this is related to information on proteins, genes, enzymes, and organisms, as well as to experimental results, simulation models, and simulation runs.
During this last year the main developments in the project have centered on providing support for the storage, querying, and export of information related to the kinetics of metabolic reactions. Work in the development of biochemical ontologies has continued, alongside the development of methods for the semi-automatic integration and curation of data coming into the database from other sources available on the Internet.

Building Protein Interaction Networks from PubMed

During this last year we have continued with the development of a rule-based biological information extraction system (BIEST) for extracting structured information on gene expression in yeast from Pubmed abstracts. This has been a carried out in collaboration with the Bork group at the European Molecular Biology Laboratory (EMBL). The resulting gene-regulation networks have been used to automatically populate databases. However, the biomedical scientific literature provides other equally important information, like the de-/phosphorylation of proteins. The main development during the last year was achieving the capability for

capturing some of this other important information, as well as linguistic constructions previously overlooked.
We have been working on four model organisms: the human being, yeast, E.coli, and the mouse.
From almost one million PubMed abstracts related to our four model organisms we have managed to extract regulatory networks and binary phosphorylations comprising 3319 relation chunks. Depending on the organism involved, accuracy is between 83 and 90%, and between 86 and 95% for gene expression and de-, phosphorylation relations, respectively. It should be noted that the initial extraction system has been developed for gene expression in yeast, but it has also proved to be applicable on modifications for other organisms as well. The details of the system are described in [Saric 2005].

To illustrate how the rule-based IE system operates and how the extracted information bits can be combined to form a network, we present here a series of automatically annotated examples with a corresponding graph representation (see Figure 13). The first example shows a phosphorylation relation in the active voice. The participating proteins are expressed in bold letters. The relation word is underlined, and the selective negation is marked by a negation bracket. From the following example we extract the information that Lyn phosphorylates CrkL:

[phosphorylation_active
Lyn, *[negation* but not **Jak2***]*
phosphorylated **CrkL***]*

Applying the same principles as before, we extract the information that Lyn phosphorylates syk from:

[*phosphorylation_active*
Lyn
also participates in
[*phosphorylation* the tyrosine phosphorylation
and activation of **syk**]]

The following two examples show that we can extract phosphorylation relations from nominalizations as well. The arguments (i.e. the participating proteins) are identified through the attached prepositional phrases (of- and by-).

Project Manager

Dr. Isabel Rojas

Project Members

Dr. Olga Krebs
Jasmin Saric
Dr. Ulrike Wittig

Students

Xiaoyan Chen
Saqib Mir
(KTF Scholarship)

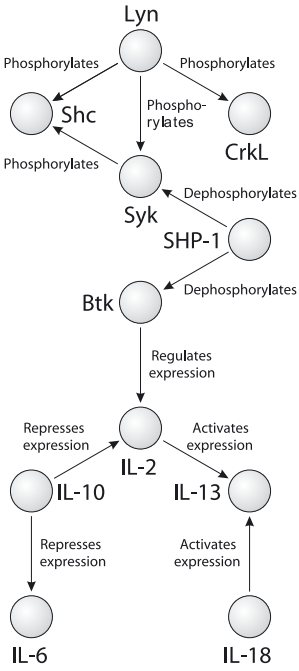


Fig.13: An example network extracted for the mouse organism. The picture exemplifies the various types of relation extracted by our rule-based approach

[*phosphorylation_nominal*
the phosphorylation of the *adapter protein* **SHC**
by the Src-related kinase **Lyn***]*

[*phosphorylation_nominal*
phosphorylation of **Shc** by the
hematopoietic cell-specific tyrosine *kinase* **Syk**]

Our system also recognizes dephosphorylations :

[*dephosphorylation_nominal*
Dephosphorylation of
Syk and **Btk**
mediated by
SHP-1]

The last three examples illustrate the extraction of gene expression relations. In the first example we identify *activation*. The second exemplifies *repression*, while the third demonstrates neutral *regulation*:

[*expression_repression_active*
Btk
regulates
the **IL-2** ***gene***]

[*expression_repression_active*
IL-10
also decreased
[*expression* mRNA expression of
IL-2 and **IL-6** cytokine receptors]

[*expression_activation_passive*
[*expression* **IL-13** *expression*]
induced by
IL-2 + **IL-18**]

Bringing more Ontology to GENIA

The GENIA ontology developed by the Tsujii Lab (www.tsujii.is.s.u-tokyo.ac.jp) at Tokyo University is a formal model of cell signaling reactions in humans. It contains 46 nominal concepts related to gene expression and its regulation, including cell signaling reactions, proteins, DNA, and RNA. However, the GENIA hierarchy contains no relations between these concepts. It is used as a basis for the annotation of PubMed abstracts, which resulted in the GENIA corpus. The overall purpose of this annotation was to support the development of natural language processing applications comprising, among other things, information retrieval, information filtering, and information extraction.

The goal of our work was to set up a system for unsupervised learning of arbitrary relations between GENIA concepts (the details are described in [Ciaramita 2005]) for the purpose of supporting manual ontology-building. We started with the GENIA corpus, which has named entities (i.e. concepts) annotated according to the formal model. In a second step we parsed the corpus with a statistical parser, drawing upon the entities already annotated. For each syntactic tree we generated a dependency graph (see Figure 14 below). By using relational words as a governor it

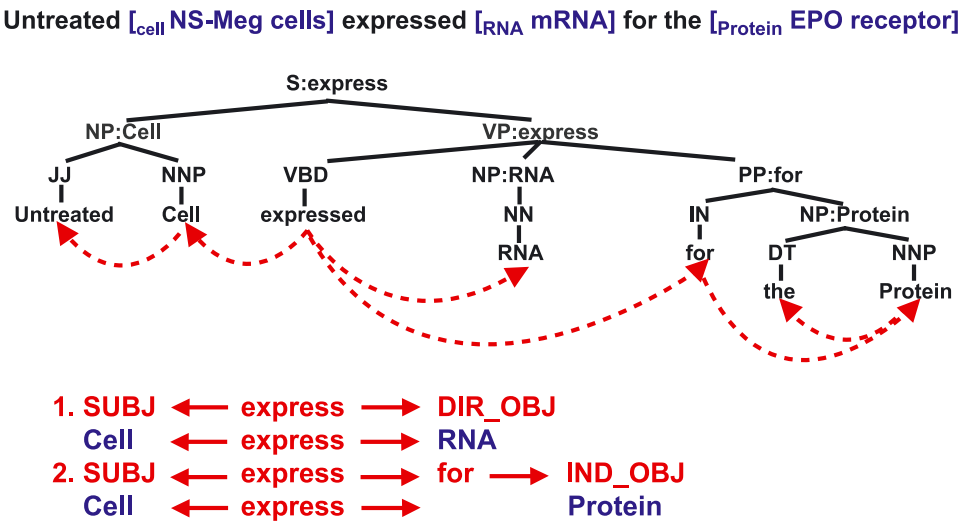


Fig.14: Example of a syntactic tree with its corresponding dependency graph

was possible to formalize semantic relations between the named entities or concepts. Relations between named entities are thus learned from the GENIA corpus by means of several standard natural language processing techniques. An in-depth analysis of the output of the system shows that the model is accurate and has good potential for text-mining and ontology-building applications. This work was carried out in cooperation with the Laboratory for Applied Ontology, CNR, in Rome, Italy, and Esther Ratsch from the University of Würzburg.

We used the outcome of this experiment for the manual ontology-building process, which consists of manual curation of the learned relations and an introduction of new concepts described in what follows.

The 78 learned relations were initially a flat list of relations without any hierarchical ordering. Thus the first step in the manual curation of these relations consisted in defining an appropriate hierarchical structure for them. Initial ontological ordering was undertaken in accordance with the domain and range of the relations, yielding 5 new top-level relations:

- BiologicalToChemicalRelation: Domain: biological object (super class of "source") Range: chemical object (super class of "substance")
- ChemicalToBiologicalRelation: Domain: chemical object Range: biological object
- intraBiologicalRelation: Domain and range: biological object
- intraChemicalRelation: Domain and range: chemical object
- hybridRelation: Domain and range can be subclasses both of biological object and chemical object.

In the next step, the learned relations were ordered hierarchically according to their (biological) meaning, e.g.:

- "ActivatedIn" isA "RegulatedIn" isA ("RegulationRelation" and "LocationRelation");
- "Inhibits" isA "Regulates" isA "RegulationRelation".

This also leads to the addition of some more top-level relations, like RegulationRelation, LocationRelation, etc.
In addition to incorporating new relations into the GENIA ontology, many of the ontological classes were refined. Most of our work was done on the refinement of the so-called "other_name" class, which was widely used in many relationships and thus interfered with its understanding.
The extension of the GENIA ontology defines the basis for a re-annotation of the data to improve the quality of information extraction from biological texts.

Semantic Web/Grid Services for SABIO

Major ongoing work in the Systems Biology community is dedicated to defining standards for data exchange and annotation. The SBML (Systems Biology Mark-Up Language) is an XML-based standard defined for exchanging biochemical models. So far, however, there have been no common standards or directories for the definition of web services providing either data or processes.

During the last quarter of 2005 we started with the definition of semantic web-services for the data on biochemical reactions and their kinetics. We have taken SBML as a starting point for the definition of the ontology describing the information that can be provided by using a service. However, this format is very limited, and we are thus working on an extension designed to offer more possibilities and better semantics for the sharing of data amongst applications. The advantage of using SBML is that there are many tools that support it (see http://sbml.org for more information).

We are currently reviewing existing semantic grid middleware, such as myGrid for the definition of workflows. Other platforms will also be revised in an attempt to re-use as much existing technology as possible.

Our aim is to facilitate the incorporation of the applications and platforms we have developed into the workflow of systems biology processes. These include SABIO-RK, COPASI, and Sycamore.

Collaboration Partners

Peer Bork Group, EMBL, Heidelberg, Germany

Laboratory for Applied Ontology, CNR (National Research Center), Rome, Italy

Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany

Sponsor

Klaus Tschira Foundation

3.7 SABIO-RK (SDBV)

The biosciences have undergone some dramatic changes in the last few years. Novel lab approaches like high-throughput methods enable scientists to rapidly produce an enormous amount of data. For researchers this poses problems connected with retaining an overview of this data and accessing it. Thus one of the biggest challenges in biological science at present is to achieve data comparability and ease of access for the scientific community. To attain this goal, the published data from different sources has to be standardized, and databases integrating this data have to be developed, including tools for data mining.

Project Manager

Dr. Isabel Rojas

Project Members

Martin Golebiewski
Renate Kania
Dr. Olga Krebs
Dr. Andreas Weidemann
Dr. Ulrike Wittig

Students

Stefanie Anstein
Gerhard Kremer

SABIO-RK (http://sabio.villa-bosch.de/SABIORK) is a web-accessible, relational database system designed to support researchers with information about biochemical reactions and their kinetics. It contains and merges information about biochemical reactions: their reactants and effectors (e.g. cofactors, activators or inhibitors), details about the catalyzing enzyme (or enzyme complex), the organism, tissue, and cellular location where the reaction takes place, and equations describing the kinetic laws for reaction rates. The latter are shown with reference both to their parameters, including their values, and to the experimental conditions (e.g. pH, temperature, buffer) under which they were obtained. Links to other databases enable the user to gather further information about reactions and enzymes or to refer to the original publication.

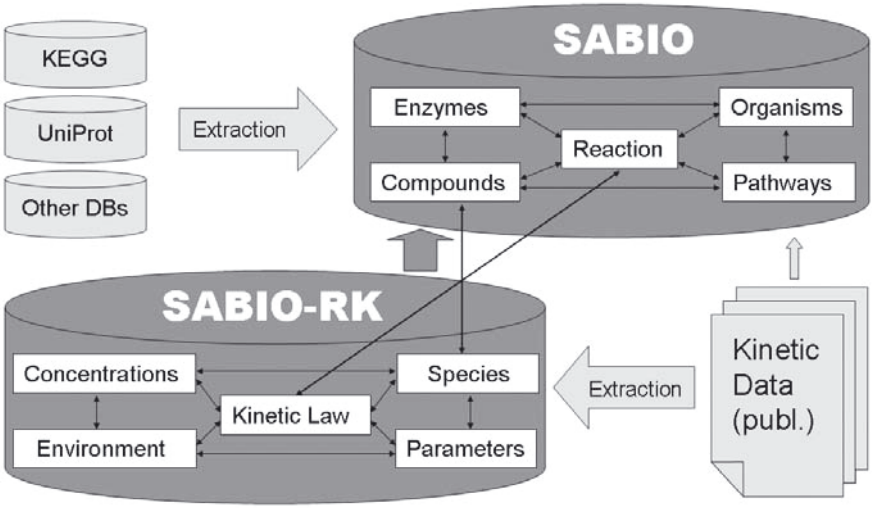


Fig.15: Population, content and schematic data relation of SABIO and SABIO-RK. SABIO contains general information about biochemical pathways and reactions in different organisms, including details about corresponding enzymes and reactants. Most of these data are collected from other databases like KEGG or UniProt. SABIO-RK extends SABIO by storing information about the reactions' kinetic properties, such as the kinetic laws with their corresponding parameters and environmental conditions under which they were determined. These kinetic data are manually extracted from literature and inserted into SABIO-RK. All data are highly interrelated within and in-between the two databases

The database has been conceived to serve the Systems Biology community as its main user. It aims to support modelers with high quality data in setting up in-silico models describing biochemical reaction networks. Users can select kinetic data exported by the system in SBML (Systems Biology Mark-up Language) format, thus allowing use of the data as the basis for the definition of biochemical network models. These models then can be used to simulate complex biochemical processes. However, SABIO-RK also bundles information for lab experimenters interested in comparing reaction-kinetic data originating from different sources.

We integrate data from different sources with a view to establishing a broad information basis. Most of the reactions,

their associations with biochemical pathways, and their enzymatic classifications (EC classifications) are extracted from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database. By contrast, the kinetic data contained in the database is manually extracted from published scientific articles and then verified by curators. At the moment, it is very difficult to extract this information automatically since most of the data is stored in tables, formulas, or graphs.

The extraction work is done by students using a web-based interface to enter the extracted information into a temporary database. The students are provided with a list of publications presumed to contain kinetic data and obtained by complex keyword searches in NCBI's Pubmed database. Identifying articles containing the information required has proved to be a very tedious and time-consuming task. We hope to make it less laborious in the future by the application of document-retrieval and information-extraction techniques. Once they have finished the processing of an article, the students mark the resulting entries as completed. Before the data is finally transferred to SABIO-RK, it is checked, complemented, and verified by a team of biological experts so as to detect possible errors and inconsistencies. Subsequently, the data is transmitted to a second temporary database from which it is finally integrated into SABIO-RK. As of December 2005, 543 publications were inserted into the intermediate database, of which about 60% have already been curated and inserted into the SABIO-RK database. From these publications about 5,200 single database entries, 270 organisms, and 1,451 reactions now figure in the database.

During the curation process, data are unified and structured consistently in order to facilitate the comparison of kinetic data. As there are no existing standards for publishing kinetic data, students and curators are faced with problems like synonymous or aberrant notations of compounds and enzymes, multiplicity of parameter units, or the absence of information about assay proceedings and experimental conditions. The description of a buffer can be very complex, containing for example information about coupled enzyme reactions or synthetic derivatives of physiological compounds. Chemical compounds often have various alter-

native names, organisms can be described by their common or systematic name, and units of kinetic parameters and concentrations can be written in different ways. For consistency, and to avoid duplicate entries, lists of compounds, reactions, organisms, tissues, compartments, and parameter units already existing in the SABIO-RK database are given to the students for selection. These lists also contain synonyms referring to the same content so as to enable the students to search for alternative names of compounds, tissues etc. Furthermore we are faced by the problem of missing or partial information in the literature. For example, a reaction definition can be incomplete, which means that only substrates of reactions are named without a definition of the reaction products. If the chemical mechanism of the enzymatic reaction is known, the reaction equation can be completed, but in most cases this work is very time-consuming, and the result may also be imprecise. Work on the development of new NLP (Natural Language Processing) methods to semi-automatically support the curators will start in early 2006.

Natural Language Processing Support for the Curation of Data

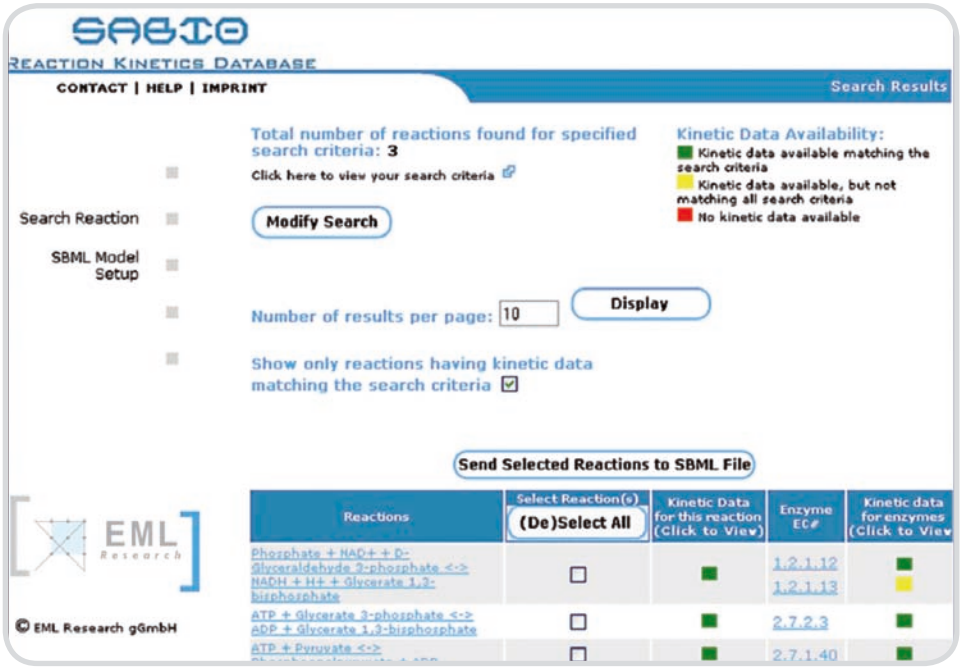
As part of their diploma work, Stefanie Anstein and Gerhard Kremer of the Universiy of Stuttgart are working on the linguistic analysis of chemical terminology - more precisely the names of organic compounds - as a key to biochemical text processing and semi-automatic database curation.
They have developed a system (CHEMorph) that analyzes systematic and semi-systematic names, class terms, and otherwise underspecified names by using a morpho-syntactic 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 structure and also classifies the terms analyzed. It has been implemented in Prolog as a prototype and a basis for further development to support research in the life sciences.

The Web-Based User Interface

The web-based user interface enables the user to search for reactions and their kinetics by specifying characteris-

tics of the reaction. These characteristics may include the pathways in which the reaction participates, reactants of the reactions (substrates and products), organisms, tissues and cellular locations in which the reaction takes place, as well as enzyme classifications of the enzymes catalyzing the reaction. The experimental conditions (currently only pH and temperature) are considered solely for the retrieval of kinetic data. The system retrieves all entries satisfying the given criteria and indicates whether there is kinetic information for the associated reaction under the search criteria specified (organism, tissue, cellular location, and experimental conditions). Apart from this, the system will also indicate whether there is kinetic data available for the enzymes catalyzing each reaction in the result set (see Figure 16). This approach has been selected to support the variations in the definition of the reactions composing a pathway, e.g. where a reaction can be substituted for by a very similar reaction with a slight change in the reactants. The next version of the interface will also enable the user to search for networks or paths of reactions between two compounds or

Fig.16: Search results in SABIO-RK



enzymes. The kinetic data can then be viewed and selected for export in SBML (Systems Biology Mark-Up Language) format. Reactions with no kinetic data can also be included in the SBML file. The SBML file is generated using the LibSBML library (http://sbml.org/software/libsbml). Given the restrictions of the SBML format, the system includes the kinetic data in the SBML file by making certain unilateral decisions. For example, if a parameter value is defined as a range, the system takes the middle value of this range as the parameter's value in SBML, given that the SBML file does not support the definition of ranges for parameter values. We plan to include annotations to reaction parameters in order to provide the user with information about the data from which the parameter value has originated, as well as information about the experimental conditions under which they were determined and the concentration of compounds.

Federal Ministry of Education and Research (BMBF)

Klaus Tschira Foundation

3.8 Modeling: From the Molecule towards the Cell (MCM)

Project Manager	Project Members
Dr. Rebecca Wade	Bruno Besson, Dr. Vlad Cojocaru, Anna Feldman-Salit, Frederik Ferner, Dr. Razif Gabdoulline, Matthias Janke, Tim Johann, Domantas Motiejunas, Dr. Stefan Richter, Dr. Matthias Stein, Dr. Ting Wang, Dr. Peter Winn
Guest Scientists and Visitors	
Dr. Renate Griffith, Stefan Ulbrich, Yanxiang Zhou	

Klaus Tschira Foundation (and additional agencies as specified for the subprojects)

3.8.1 Bioinformatics Resources on the Web

During this year, the following web-based resources have been 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

We are working on a major extension of the functionality of ProSAT, ProSAT2, to manage custom annotations and permit more versatile usage [Ulbrich 2005]. This will facilitate the mapping to protein structures of position-specific data from mutation and sequence-variant databases. ProSAT2 is also being developed to map results from text mining for protein mutation data to protein structures for visualization (in collaboration with Christopher Baker, Concordia University, Canada, and René Witte, University of Karlsruhe, Germany).
A new release of the PIPSA software for Protein Interaction Property Similarity Analysis, PIPSA version 2, was made available on our website this year. PIPSA can be used to compute and analyze the pairwise similarity of the 3D interaction property fields [Wade 2005a] of a set of structurally related proteins. Changes resulting from the major rewriting of the code in PIPSA2 include much-improved computational efficiency, additional analysis tools, e.g. for making epograms (trees showing electrostatic similarity relationships in a protein family), more transparent usage, and greater flexibility in the way PIPSA can be applied.

3.8.2 Methodological Developments

3.8.2.1 Incorporating Sequence and Experimental Information into Protein-Protein Docking Procedures

Protein-protein interactions are crucial in many biological processes. Examples include interactions involved in signal transduction, control of biochemical pathways, inhibition and activation of enzymes, and gene expression. A good understanding of the structures of protein-protein complexes and their association mechanisms can help to explain biological processes and can also be of direct value in applications to medicine and biotechnology. The number of protein structures solved experimentally by X-ray crystallography and NMR is constantly increasing. However, the experimental determination of the structure of a pro-

tein-protein complex is a difficult and time-consuming task. Therefore computational docking methods represent an attractive approach to predicting possible protein-protein complexes from experimental or modeled structures. Computational methods can also reveal key regions responsible for specific complex formation and provide insights into the association process itself.

We are developing procedures consisting of a rigid-body docking phase followed by a flexible-body docking phase (see next section) for protein-protein docking. These procedures are designed to make use of biochemical data, e.g. from mutagenesis experiments or sequence conservation information, that may indicate which amino acid residues play an important role in binding during docking. We have modified our SDA (Simulation of Diffusional Association) program for rigid-body docking. The structures of the unbound proteins are used, their diffusional association simulated, and the structures of encounter complexes that satisfy the constraints of the biochemical data are collected, clustered, and ranked. Representative highly-ranked structures of complexes are then subjected to flexible refinement. This year, we have introduced an implicit treatment of limited protein flexibility in the rigid-body docking phase and developed clustering procedures and empirical functions for scoring encounter complexes. These have been validated against a number of structurally and functionally diverse protein complexes, and we are now working on achieving an automated parameterization and implementation of the rigid-body docking procedures. The docking procedures are being applied to several protein-protein complexes in collaboration with experimentalists (see section 3.8.3).

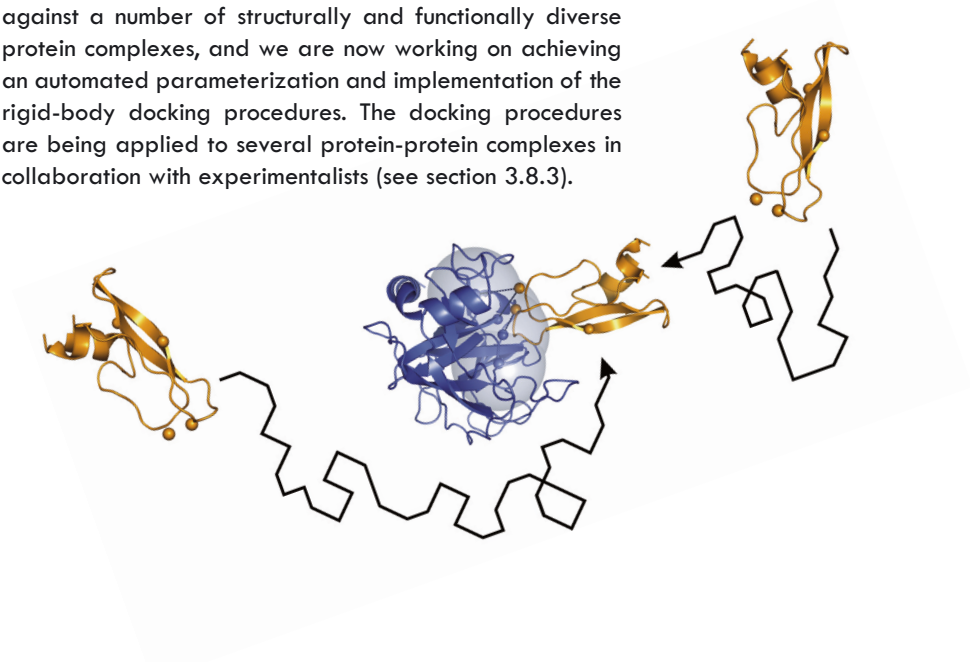


Fig.17: Schematic illustration of two trajectories from Brownian Dynamics simulations of the diffusional association of two proteins to form an encounter complex satisfying constraints from biochemical data. Two atoms in the ligand protein (orange spheres) expected to be at the interface, on the basis of biochemical data, are within the required distance (large blue spheres) from expected interface atoms of the target protein (small blue spheres) in the docked encounter complex

3.8.2.2 Treating Protein Flexibility in Protein-Protein Docking

The experimentally determined structure of an unbound protein generally differs from that of the protein when it is bound to another molecule, and the extent and type of the difference can vary widely. We are currently working on employing molecular dynamics simulations to optimize structures obtained from rigid-body protein-protein docking. Standard molecular dynamics procedures alone are inadequate for this purpose, as they do not permit sufficiently efficient sampling. We are thus developing enhanced sampling methods aimed at improving the speed and accuracy with which docked structures can be obtained. One of the methods we have investigated this year is the Random Accelerated Molecular Dynamics (RAMD) method, which we originally developed to study ligand egress from buried cavities in proteins. We have extended RAMD to assist protein-protein docking. A biasing function has been introduced, and protocols combining RAMD and standard MD for docking purposes have been developed [Johann 2005]. We have implemented the RAMD method in the AMBER8 software suite, and it is available to other users as a patch. Initial tests show that the RAMD approach can improve sampling towards the correct docked complex. Refinements of the methodology are being undertaken against a diverse set of protein-protein complexes.

3.8.2.3 Implicit Solvent Models for Protein Simulations

Molecular dynamics simulations with an explicit representation of solvent water molecules are computationally demanding due to the large number of atoms present. Treating the water implicitly in the simulation is one way of reducing computational demands. We have updated our NPSA (Neutralized, Polarized ionizable side chains with a solvent-accessible Surface Area-dependent term) implicit solvent model for use with the ff03 AMBER force field for proteins [Wang 2006]. We have implemented the NPSA method in the AMBER8 software suite, and it is available to other users as a patch. We have tested the NPSA implicit solvent models in simulations of protein unfolding and protein docking.

We found that simulations of the temperature-induced unfolding of beta-sheet structure protein domains (WW domains) provide a sensitive test for force field parameterization. The secondary structure propensities observed in protein simulations depend heavily on the force field parameters used. The existing empirical force fields often have difficulty in balancing the relative stabilities of helical and extended conformations. The resultant secondary structure bias may not be apparent in short simulations at room temperature starting from the native folded states. However, it can manifest itself dramatically at high temperatures and lead to large deviations from experimentally observed secondary structure propensities. We have investigated several AMBER force fields, as well as the parameterization of the NPSA solvent model, in high-temperature simulations of a WW domain. Older force fields resulted in overstabilization of either helical or beta-sheet structures. The newer ff03 force field was able to reproduce the beta-sheet-coil transition and experimentally observed unfolding pathways with both an explicit water solvent and the NPSA implicit solvent model at relatively low temperatures. However, the protein domain became predominantly helical after unfolding. Modification of the solvation parameter

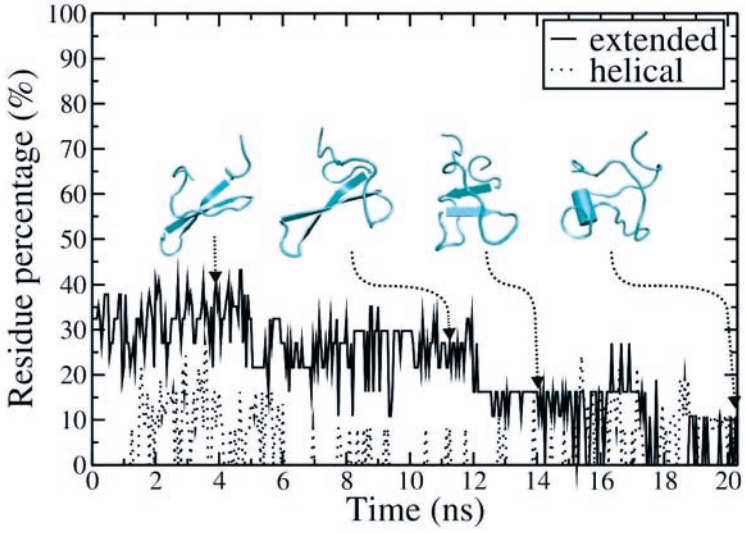


Fig.18: Time development of the percentage of residues in extended (solid line) and helical (dash line) conformations in molecular dynamics simulations of a WW protein domain with the NPSA model and the AMBER ff03 force field at 430K. Although notable helical conformations are present in the trajectory, they are discontinuous until unfolding is completed, losing all strands, and a clear unfolding pathway can be observed. The early loss of the third strand was followed by the loss of the first two strands. See [Wang 2006]

in the NPSA model was not sufficient to remedy this problem. The results imply that the intrinsic secondary structure bias in a force field cannot easily be solved by modifying a single parameter such as backbone torsion potential or a solvation parameter of a solvent model. Nevertheless, the results show that the AMBER ff03 force field, together with either an explicit solvent model or the NPSA implicit solvent model, is a useful tool for studying the unfolding of both alpha- and beta-sheet structure protein domains. We have applied these simulation procedures to several mutant WW domains in an investigation of the stability determinants and the unfolding mechanisms in collaboration with Maria Macias and Ximena Ramirez (IRBB-PCB, Barcelona, Spain), who have performed experiments by NMR.

3.8.2.4 Simulation of Biomacromolecular Diffusion

Improvement of the methodology in our SDA (Simulation of Diffusional Association) software is being carried out both in this project and in the 'Modeling Macromolecular Motions in the Cell by Brownian Dynamics Simulations' project (see section 3.10). Using the model of non-polar forces that we have developed, simulations of the diffusional association of electron transfer proteins have been carried out. Inter-protein electron transfer is characterized by protein interactions on the millisecond time-scale. Simulations have been applied to compute electron transfer rates between plastocyanin and cytochrome f for cyanobacterial and plant species. We have found that both electrostatic and non-polar interactions are relevant in determining observed electron transfer rates, and that these are dependent on both diffusional and electron transfer activation barriers.

Volkhard Helms and Alex Spaar (Center for Bioinformatics, Saarland University, Saarbrücken) recently devised a way of computing free energy landscapes from Brownian dynamics trajectories and introduced this into the SDA software. In a collaborative project, the free energy landscape for the diffusional association of two well-characterized proteins, barnase and barstar, was investigated. This study showed how single point mutations may drastically change the free energy landscape and significantly alter the popu-

lation of the two free energy minima for barstar's encounter with barnase. This implies that certain protein-protein pairs may require careful adaptation of the encounter positions and transition states when interpreting the effects of mutation on kinetic rates of association and/or dissociation [Spaar et al 2006].

3.8.3 Applications

3.8.3.1 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 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 for optimization of their properties in biotechnological applications. Our role is to employ modeling and simulation techniques to guide the engineering of haloalkane dehalogenases to alter their substrate specificity and optimize their activity. Mutants have been designed on the basis of COMBINE analysis to affect substrate specificity [Kmunicek 2005], and the RAMD simulation method is being used to investigate the mechanism of product release in haloalkane dehalogenases.

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

3.8.3.2 Modeling and Simulation of Cytochrome P450 Dynamics and Interactions

Cytochrome P450 enzymes contribute to the disposal by the human body of about 80% of all medicines. These enzymes are important for detoxification processes and for the synthesis of many important molecules, such as the sex hormones, progesterone, and testosterone. An understanding of how cytochrome P450s function is important for their medical and biotechnological application. We are investigating how cytochrome P450s interact with low molecular weight compounds (substrates, products, and inhibitors) and with the proteins that transfer electrons to cytochrome P450s for catalysis [Wade 2005b].

Cytochrome P450: Substrate/Product Interactions

The active site of the cytochrome P450s is buried in their center, so understanding how substrates access and products leave is critical to understanding the enzyme's function. Using the RAMD simulation technique (see section 3.8.2.2), we have studied ways into and out of the center of CYP2C5, a cytochrome P450 from rabbit liver, for both substrates, including progesterone, and products. These simulations indicate a channel that is different from those we have seen so far in simulations of cytochrome P450s found in bacteria [Schleinkofer 2005]. Based on our simulations, we have proposed two mechanisms by which compounds can

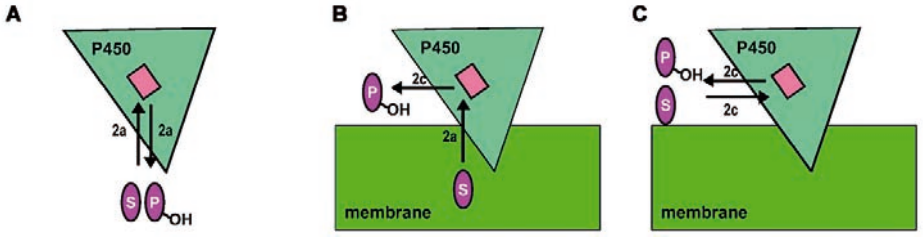


Fig.19: Schematic representation of putative routes for substrates and products in bacterial and mammalian P450s identified by molecular dynamics simulations. (A) Pathway 2a serves as a substrate-access and product-egress channel in soluble P450s. (B) One-way route through mammalian membrane-bound P450: lipophilic substrates access the active site via pathway 2a and products exit via pathway 2c. (C) Soluble substrates access and products exit the active site via pathway 2c. Substrate (S) and product (P) are depicted in pink, the membrane in green, the heme cofactor (active site) in light red, and cytochrome P450 in light blue. Mammalian P450s are anchored to the membrane via the N-terminal transmembrane helix (not shown) and by interactions with the region around the FG loop of the protein

travel to and from the active site of cytochrome P450s: (1), a ‚one-way' route whereby fat-soluble (lipophilic) substrates enter the enzyme from the membrane and products leave the active site, via the newly discovered channel, directly into solvent; and (2) a ‚two-way' route for access and egress of water-soluble compounds solely via the new channel. The proposed differences in substrate access and product egress routes between the mammalian and bacterial cytochromes P450 highlight the adaptability of the P450 family to the requirements of different cellular locations and substrate specificity profiles. We are currently extending our study to human cytochrome P450s, for which crystal structures have recently been solved.

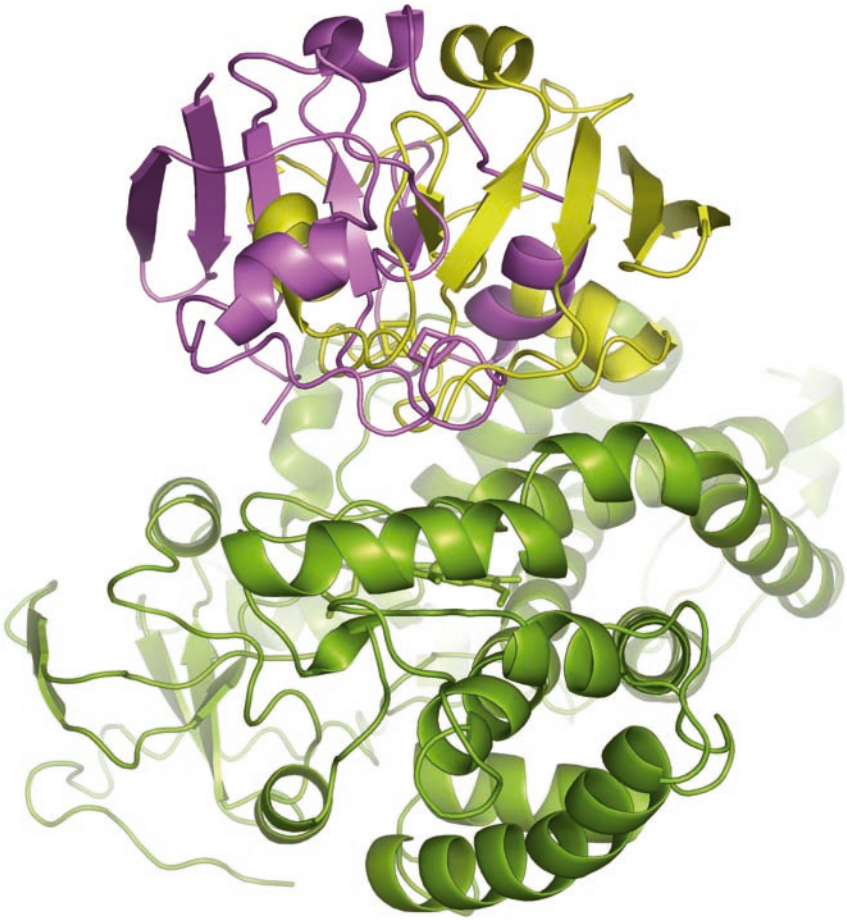
Cytochrome P450: Protein Interactions

CYP101 (P450cam) from the soil bacterium Pseudomonas putida is a well-characterized P450 protein. It catalyzes the hydroxylation of camphor. An important component in this reaction is another protein, putidaredoxin (Pdx), which serves as an electron shuttle between the NADH-dependent flavoprotein, putidaredoxin reductase and CYP101. The structures of both CYP101 and Pdx are available, but the structure of the CYP101/Pdx complex is unknown. Previous models made of the complex were inconsistent with recent data from FTIR experiments by Christiane Jung and Andrei Kariakin (MDC, Berlin). Therefore we used rigid-body protein-protein docking with the SDA program, followed by flexible refinement with MD and the NPSA implicit solvent model, to dock the most recent crystallographic structures of CYP101 and Pdx. Clustering and energetic ranking of the docked complexes yielded two prominent docking positions of Pdx. These models of the complex are consistent with all the available experimental data and suggest that D38 residue of Pdx is a "hot-spot" residue in CYP101/Pdx complex formation [Wade 2005b].

In mammalian liver, there are a number of different cytochrome P450s that aid the process of excretion of toxins and drugs. For these cytochrome P450s, the electron transfer protein is not Pdx but cytochrome P450 reductase. The different cytochrome P450s compete with each other for electrons from this protein. This interaction may affect the metabolism of different drugs, and we are thus apply-

ing computational tools, including those developed in the SYCAMORE project (section 3.3), to investigate these protein-protein interactions.

Fig. 20: The two most prominent binding modes (yellow, mauve) of putidaredoxin to cytochrome P450cam (CYP101) (green) as identified by computational docking (See [Wade 2005b])



3.8.3.3 Mechanisms of Cellular Regulation by Ubiquitin

Ubiquitin is a small protein essential for the regulation of cellular function. Malfunctions in ubiquitin-controlled cellular regulation contribute to many forms of cancer, as well as to neurodegenerative diseases such as Alzheimer's and Parkinson's. In 2004, the Nobel Prize for chemistry was awarded 'for the discovery of ubiquitin-mediated protein degradation'. The E2 ubiquitin-conjugating enzymes are key enzymes in the ubiquitin and ubiquitin-like protein ligation pathways, but little is known about their biochemical mechanisms. To understand the functionality of the different E2 enzymes, we have taken a two-pronged approach. Firstly, we have carried out an in-depth study to investigate the catalytic mechanism of one important E2 enzyme by combining computational studies with experimental studies of mutants by Amit Banerjee (Wayne State University, Detroit). Insights gained in this work will assist the development of chemical inhibitors for use both as experimental tools and as possible leads for therapeutic applications. Secondly, we have developed an automated computational pipeline to model the three-dimensional structures of all known Ubc proteins [Winn 2005]. This provides a basis for comparative studies of the different E2 enzymes. This pipeline will be integrated into our 'ubiquitin and ubiquitin-like protein web resource' (www.ubiquitin-resource.org) in early 2006.

Prof. A. Banerjee, Wayne State University Detroit, USA — Collaboration partner

National Institute of General Medical Science (NIGMS), USA — Sponsor

3.8.3.4 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 thereby make sulfur available to animals and humans where it 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", composed of the enzymes Serine Acetyl-Transferase (SAT) and O-Acetyl-Serine-(Thiol)-Lyase (OAS-TL). The biological function of this complex and the mechanism of reciprocal regulation of the constituent enzymes are still poorly understood. In a collaboration with Rüdiger Hell and Markus Wirtz (Faculty of Plant Science, University of Heidelberg) started this year, 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 3.8.2) to model their complex.

3.8.3.5 Signal Recognition Particle-Receptor Interactions

The Signal Recognition Particle (SRP) and its receptor (SR) target nascent proteins destined for secretion or membrane integration to the endoplasmic reticulum in eukaryotes or the plasma membrane in prokaryotes. The principal components of SRP and SR are conserved across all kingdoms. We have used molecular modeling methods to investigate the structures of SRP and SR proteins and their complexes, which were recently determined via crystallography by Irmi Sinning's research group (Biochemistry Center, University of Heidelberg). Protein-protein docking techniques (see section 3.8.2) and analysis of electrostatics were applied to study the process of SRP/SR association. Our calculations suggest an important role for the N domain of SRP and SR in the association process, and we have designed point mutations to investigate the binding mechanism. We have also identified energetically favorable regions for ions and water molecules in the unbound and bound structures of SRP and SR, and these suggest a possible mechanism of GTP hydrolysis and SRP/SR dissociation (see Figure 21).

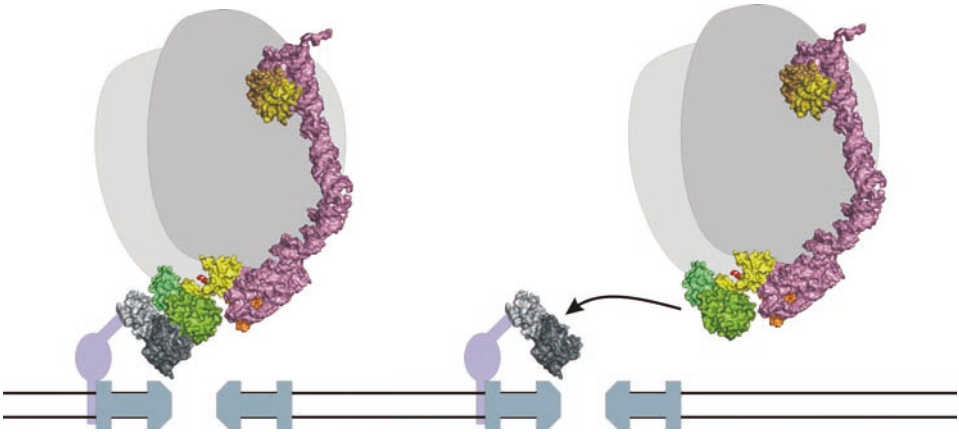


Fig. 21: Schematic figure showing how the SRP (green) (a) binds to ribosome-nascent chain complex and (b) subsequently associates with the SR (gray) on the endoplasmic reticulum membrane

3.9 TASSFUN: Target-Specific Scoring Functions (MCM)

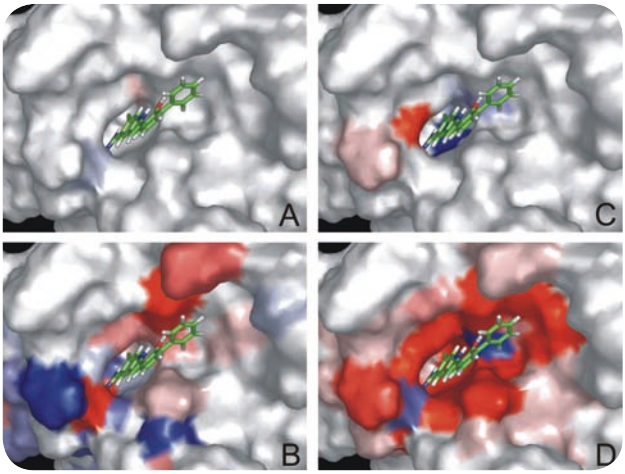
Project Manager
Dr. Rebecca Wade
Project Member
Dr. Stefan Henrich
Collaboration Partner
Niklas Blomberg, AstraZeneca R&D, Mölndal, Sweden
Sponsor
AstraZeneca

Thromboembolic diseases are common in humans. They are often affected by enzymes of the blood coagulation cascade, so the specific inhibition of these enzymes is a major goal in drug design. In this project, Comparative Binding Energy (COMBINE) analysis is being developed to generate target-specific scoring functions (TASSFUN) for blood coagulation cascade enzymes. It will be applied in virtual ligand screening and for the prediction of the bioactivity of new inhibitors.

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 to bioactivity or binding free energy values by Partial Least Squares (PLS). Subsequently, the target-specific scoring function thus derived is used to

predict the bioactivity of inhibitors for which no experimental values are available.
On the methodological side we have worked on streamlining the structural modeling of complexes and the energetic calculations necessary for COMBINE analysis so that these steps can be automated for large data-sets. We have also addressed difficult issues in the prediction of the structures of protein-ligand complexes posed by the enzymes studied, such as the presence of alternative protein side-chain and ligand functional group positions, water molecules in the binding site, and very short hydrogen bonds. We are assembling suitable data-sets of diverse enzyme-inhibitor complexes with structural and binding-affinity information and are in the process of building and testing COMBINE models for different protein targets.

Fig. 22: Visualization of interaction energy terms. Electrostatic (A, B) and van der Waals (C, D) interaction values between a receptor model and one of the compounds in the training set were mapped onto the receptor-binding site surface (stabilizing and destabilizing regions for complex formation in red and blue, respectively). The upper panel (A, C) shows the unweighted interaction energy terms and the lower panel (B, D) the corresponding terms weighted by the COMBINE model. It is apparent that not all the interaction energies have the same impact on binding affinity



3.10 Modeling Macromolecular Motions in the Cell by Brownian Dynamics Simulations (MCM)

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 permits 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 for the computation of association rate constants and the study of encounter complex formation. The goal of the project is to develop this software and the methodology so that the diffusional motion of many macromolecules or large proteins

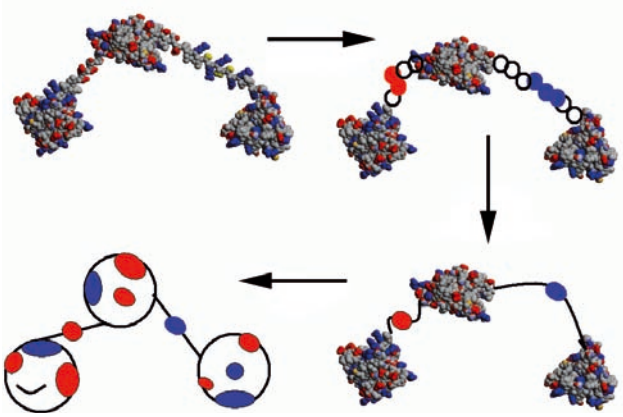


Fig. 23: Simplified models to simulate large proteins consisting of rigid domains connected by flexible linkers. From upper-left, clockwise: (1) A protein shown in atomic detail, (2) the domains are treated as rigid bodies and the flexible linkers as spherical particles, (3) the linkers are treated as restraints, (4) the domains are treated as spheres with interaction patches

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 the forces due to hydrophobic interactions. This year, an empirical model for these non-polar interactions has been constructed and calibrated, and is now undergoing detailed testing. To simulate large systems over very long times it is also necessary to develop reduced models. We are consequently developing models in which, e.g., one protein domain is represented by one sphere, while its interaction properties are derived from the (non-spherical) atomic-detail structure of the domain. These models are being applied to simulation of the interaction of the large modular insect muscle protein kettin with filamental actin, an interaction that is important for muscle elasticity and which is being studied experimentally by C. Mühle-Goll (EMBL).

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

Sponsor BIOMS Center for Modeling and Simulation in the Biosciences, Heidelberg
www.bioms.de

3.11 DIANA-Summ (NLP)

Deutsche Forschungsgemeinschaft

DFG

Project Manager
Dr. Michael Strube
Project Members
Margot Mieskes Christoph Müller

DIANA-Summ is EML Research's first DFG-funded project (DFG = German Research Foundation). Its aim is to generate meeting minutes, a task that we consider to be a special application of automatic speech summarization. Since we do not intend to develop a new summarization system, but rather to apply one that already exists, we focus on making transcriptions of multiparty dialog accessible to an existing summarizer. Hence we have been working mainly on preprocessing to label the dialogs from the ICSI meeting corpus with parts of speech, to detect and mark disfluencies, and to recognize and mark topic boundaries, i.e., to divide a dialog into segments containing parts of the meeting dealing with the same topic.
The annotation of these linguistic phenomena has proved to be extremely difficult in some cases. This is also true of the annotation of anaphoric relations in the ICSI meeting corpus, which serves as the empirical basis for the development of a component for resolving anaphoric expressions in spoken multi-party dialog.

Part-of-Speech (POS) Tagging

Assigning part-of-speech (POS) tags to words in a document (either manually or automatically) is called POS tagging. POS tags are abbreviations like CC (coordinating conjunction), NN (noun), and VBZ (verb, 3rd person singular, present tense), as shown in Figure 24. Each element in a sentence is assigned a POS tag. The process is shown in Figure 24 and the result in Figure 25.

Adding POS information to the data is a prerequisite for many natural language processing tasks, including summarization. Assigning POS tags manually requires a lot of time and human work. We set out to make the task less demanding for human annotators - and less expensive for us - by using available POS taggers to automatically as-

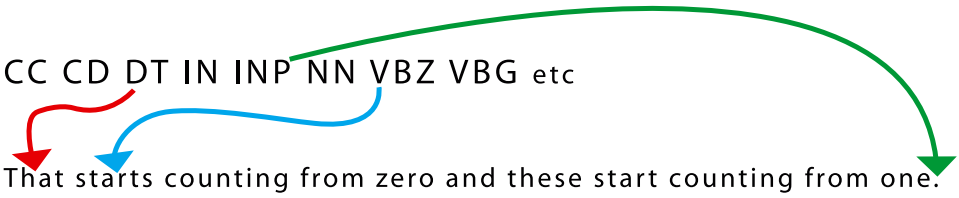


Fig. 24: Some POS tags, an example sentence, and the assignment of POS tags to elements of a sentence (tokens)

sign POS tags to words. The four taggers we used were originally trained on text collections like the Wall Street Journal. Spoken language is different from written text, but we still hoped to get reasonable results by using a majority decision on the results from the taggers.

Evaluation and Retraining

The automatically annotated POS tags were then presented to human annotators for manual correction. 12 meetings from the whole collection were corrected by three annotators. We did another majority decision on the manually corrected POS tags. This second majority decision was manually corrected for those cases where the annotators did not reach a majority. These 12 meetings form a "gold standard" for POS tagging.

Based on the gold standard we evaluated the original automatic annotation. The resulting error rates were between 10.5% for the majority decision and 14.3% for the TnT tagger. These error rates are much higher than the error rates reported for the POS taggers on written text, which are between 3% and 4%. Subsequently, nine meetings from the gold standard were used for training. These nine meetings are spread across the whole corpus and contain a large variety of speakers and speech styles.

All the taggers achieve error rates between 3.2% and 4.7%. More training data did not result in remarkably lower error rates. The improvement over and against the original automatic tagging is 8-10% for each tagger. These individual results are very close to those that have been reported for the taggers on written text. The majority decision on the retrained taggers achieves 2.9% on the best setup, which is close to the best tagger on written text (2.8%). Figure 25 below shows the data with POS tags.

Disfluency Annotation

Disfluencies are a major speech processing problem. We thus decided to develop a component that detects and marks disfluencies. In order to evaluate this component we manually annotated the disfluencies in a portion of the corpus. The categories marked are

• **nlfp**
non-lexicalized filled pauses are elements like *uh*, *um*

• **lfp**
lexicalized filled pauses are elements like *well*, *you know*

• **repet**
single words or phrases that are literally repeated, like *today*, *today*

• **repai**
areas where items in an utterance are corrected/repaired, like *today*, *yesterday*

• **abw**
words that have not been completed, but abandoned at some point , like *tha-*

• **abutt**
utterances that have not been completed, but abandoned at some point , maybe because another speaker interrupted, like *yesterday I went to*

• **disfluent**
marks general disfluent areas that cannot be characterized as one of the categories above

The manually annotated data serve as the gold standard for evaluating the automatic disfluency annotation. In addition to the pattern-matching methods already developed, we are currently exploring machine-learning methods (e.g. Decision Trees) for classifying words or phrases as disfluent or not. In future work, the binary classification will be enhanced to cover the individual categories.
In Figure 25 we have marked the different types of disfluencies with different colors. Dark blue items are nlfp, light blue items are lfp, magenta items are repai, grey items are either abutt or abw, and green items are repet.

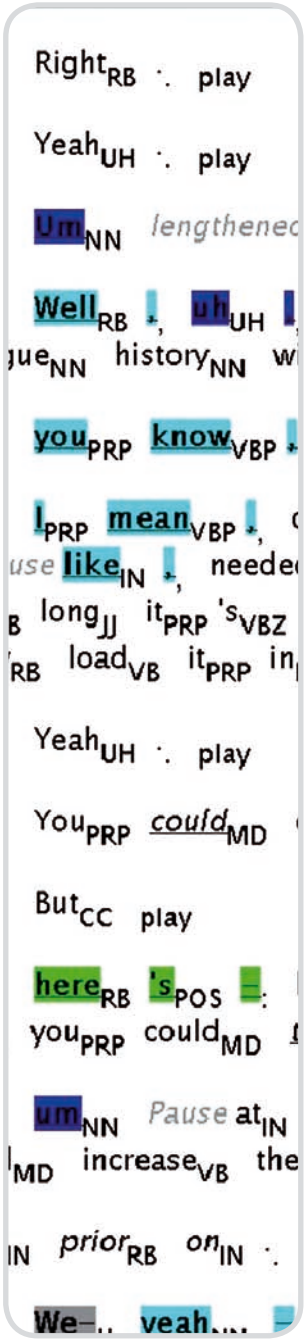


Fig. 25: Screenshot of one of the meetings with POS tags and disfluencies annotated

Topic Boundary Annotation

Topic boundary annotation is one of the first major steps towards summarizing the meetings. Since we aim at producing summaries that look as similar to real minutes as possible (table-like, formal meeting minutes), identifying topic changes within the meetings is crucial. The topic changes provide information about how many topics have been discussed during the meeting. Additionally, they divide the whole meetings into smaller portions, which are easier to summarize than the whole meeting.

The basic unit we use for extracting all these measures is the so-called "spurt". Spurts are sections of continuous speech from one speaker. In the recordings these sections can be identified via pause information. Sections that have shorter pauses than 500ms between them form one spurt.

The procedure currently employed for detecting topic changes takes three information sources into account. First, dotplots, plotting the positions in a text at which a certain word occurs. The positions of the word "house" (e.g. spurt 1 and spurt 5) are marked in a two-dimensional plot, where areas with many dots very close together indicate a cluster of words belonging to the same topic. Areas with fewer dots indicate a topic break.

Second, vocabulary introduction, which counts the amount of new words in each spurt. In each spurt it determines whether the words in it have already been used before or not. Those words that have already been used are ignored. The others are counted. The idea behind this is that sentences or spurts with many new words can indicate a topic change.

Finally lexical chains without external information. Originally, lexical chains relied on exterior knowledge sources like WordNet or a regular thesaurus. Here, lexical chains are computed using word information only. Words can form chains if they occur within a certain number of spurts (e.g. 10 spurts, which is equivalent to a time span anywhere between 50 sec and several minutes, depending on the length of the individual spurts). If the word does not occur again

within the 10 spurts, the chain stops. For example, if the word *house* occurs at the beginning (spurt1) and then again in spurt5 and then again in spurt22, the chain reaches from spurt1 to spurt5 and stops afterwards. In spurt22 a new chain could start. This indicates that there could be a topic change between spurt5 and spurt22.

Together, these three information sources provide a function for finding local maxima indicating a topic break. Using local maxima above a certain threshold and comparing them with the manually annotated topic changes has produced reasonable results that will be improved on in the future.

Coreference Resolution

Apart from the actual automatic summarization, another integral part of the Diana-SUMM project is coreference resolution. So far, coreference-related annotation has been performed for a subset of the ICSI Meeting Corpus. In order to prevent our preconceptions from influencing the annotation, all annotation tasks were performed by two students external to the project.
The first task consisted in the classification of instances of *it* into one of five classes. The pronoun *it* is important because it is both very frequent and at the same time ambiguous. The ambiguity arises from the fact that not all instances of *it* are normal, referential pronouns like in the following example. (Unless noted otherwise, all following examples are from dialog Bed017):

ME003: Eva's got [a laptop], she's trying to show [**it**] off.

Other, non-referential uses of *it* can be found in extraposition constructions like the following:

FN050: And then, after you group them together this - no- the dependencies would - of the queries would be reduced to this. And so, you know, **it**'s easier to specify the C_P_T and all.

or in the form of so-called *prop-it* as in

FE004: So **it** seems like a lot of - some of the issues are the same.

Finally, *it* does also quite often appear in abandoned utterances, like the second but last one in the following example

ME010: Whether the inference gets any faster or not I don't know. Uh, it wouldn't surprise me if **it** - if it doesn't.

What all but the first example of *it* have in common is that they are **not** referential pronouns, i.e. they cannot be anaphoric. Identifying non-referential instances of *it* is an important preprocessing step for automatic coreference resolution: By excluding these from consideration as anaphors, the precision of the coreference resolution module can be improved because less spurious anaphor-antecedent links will be retrieved.
Using more than one annotator for the task allowed us to determine the reliability with which humans can distinguish the different types of *it*. The reliability figure is important because it allows a conclusion as to what the upper bound for the automatic solution of a given task is. In other words, a poor reliability can often be taken to indicate that the task is in itself ill-defined or vague, and that it therefore cannot be expected to be solved by a computer. Indeed, the initial reliability figures on the five-fold (we also included a class *vague* which is not discussed here) classification were below the threshold that is normally required. However, by giving up the subclassification among non-referential *it* and breaking the task down to a binary classification, the reliability could be improved. The data was then used as training and test data for a machine learning classifier which could be used to automatically identify non-referential instances of *it* in the ICSI Meeting Corpus. First experiments indicate a level of performance for the classifier that make it already practically usable as a filtering component for a coreference resolution module.

Sponsor Deutsche Forschungsgemeinschaft (DFG)

3.12 MMAX2 (NLP)

Project Manager
Dr. Michael Strube
Project Member
Christoph Müller

MMAX2 is the multi-level annotation tool that has been developed at EML Research. It is a customizable annotation tool for creating, browsing, visualizing, and querying linguistic annotations at multiple levels. MMAX2 uses so-called „stand-off" annotation to allow for trouble-free co-existence of an arbitrary number of annotation levels in a single document. While this makes MMAX2 one of the very few available tools for practical multi-level annotation, the special requirements for representing multiple levels of annotation do somewhat complicate access to the data. This is true especially for annotation querying. Therefore the main development activities for MMAX2 in the last year centered around improving the expressive power and the usability of the MMAX2 Query Language MMAXQL. The result is simplified MMAXQL. ([Müller 2005a], [Müller 2005b]).

Simplified MMAXQL provides an easier and more concise way of formulating certain types of queries for multi-level annotated corpora. Queries are automatically converted into the underlying query language and then executed. A query in simplified MMAXQL consists of a sequence of query tokens combined by means of relation operators. Each query token queries exactly one base-data element (i.e. word) or one markable (i.e. element of annotation).

Base-data elements can be queried by regular expression matching. Each base-data query token consists of a regular expression in single quotes, which must exactly match one base-data element. The query

‚[Tt]he'

matches all definite articles, but not e.g. ether or there. For the latter two words to also match, wildcards have to be used:

‚.*[Tt]he.*'

Sequences of base-data elements can be queried by simply concatenating several space-separated tokens. The query

‚[Tt]he [A-Z].+'

will match sequences consisting of a definite article and a word beginning with a capital letter.

Markables are the carriers of the actual annotation information. They can be queried by means of string matching and by means of attribute-value combinations. A markable query token has the form

string/conditions

where string is an optional regular expression and c*onditions* specifies which attribute(s) the markable is to match. The most simple ‚condition' is just the name of a markable level, which will match all markables on that level. If a regular expression is also supplied, the query will return only the matching markables. The query

[Aa]n?\s.+/ref_exp

will return all markables from the ref_exp level beginning with an indefinite article (i.e. a or an).
The conditions part of a markable query token can indeed be much more complex. A major feature of simplified MMAXQL is the option of omitting redundant parts of conditions, thus making queries very concise. For example, the markable level name can be left out if the name of the attribute accessed by the query is unique across all active markable levels. Thus the query

/!coref_class=empty

can be used to query markables from the ref_exp level that have a non-empty value in the coref_class attribute, provided that only one attribute of this name exists. The same applies to the names of nominal attributes if the value specified in the query unambiguously points to this attribute. Thus the query

/pn

is sufficient to query markables from the pos level which

have the value *pn*, provided that there is exactly one nominal attribute with the possible value *pn*. Several conditions can be combined into one query token. Thus the query

/{poss_det,pron},!coref_class=empty

returns all markables from the ref_exp level that are either possessive determiners or pronouns and that are part of some co-reference set. The curly braces notation is used to specify several OR-connected values for a single attribute, while a comma outside curly braces is used to AND-connect several conditions relating to different attributes.

The whole point of querying corpora with multi-level annotation is to relate markables from different levels to each other. The reference system with respect to which the relation between different markables is established is the sequence of base-data elements, which is the same for all markables on all levels. Since this bears some resemblance to different events occurring in several temporal relations to each other, we adopt this as a metaphor for expressing the sequential and hierarchical relations between markables, and we use a set of relation operators inspired by James Allen 1991 (www.cs.rochester.edu/~james). This set includes (among others) the operators before, meets (default), starts, during/in, contains/dom, equals, ends, and some inverse relations.
The following example gives an idea of how individual query tokens can be combined to form complex queries by means of relation operators. The example uses the ICSI meeting corpus of spoken multi-party dialog, the data also used in the DIANA-Summ project (see 3.11). This corpus contains, among other things, a segment level with markables roughly corresponding to speaker turns, and a meta-level containing markables representing e.g. pauses, emphases, or sounds like breathing or mike noise. These two levels and the base-data level can be combined to retrieve instances of ‚you know' that occur in segments spoken by female speakers (the first letter of the participant value encodes the speaker's gender) that also contain a pause or an emphasis:

```
,[Yy]ou know‘ in (/participant={f.*}
dom /{pause,emphasis})
```

For comparison, the following equivalent (but much more verbose and complicated) MMAXQL query is automatically generated from the above:

```
let $10=segment (*participant={f.*});
let $11=meta (type={pause,emphasis});
let $22=contains($10, $11);
let $20=basedata (*basedata_text={[Yy]ou});
let $21=basedata (*basedata_text={know});
let $2=during(meets($20, $21), $22);
display $2;
```

Further work on simplified MMAXQL will mainly consist of adding support for wildcards and proximity operators, and doing some general optimization.

Sponsor Klaus Tschira Foundation

3.13 Information Technology in the Health Sector: Dr. Feelgood

DigiCoach: Analyzing the Swimming Technique of Competitive Swimmers

Project Manager
Prof. Dr. Andreas Reuter
Project Member
Markus Buchner
Student Worker
Alexander Folz

In competitive swimming, not only the swimmer's technique itself but also other phases like the start and the turn play an important role in the final showing. Analysis of races at an international level has revealed that start times account for more than 10 % of the total time for a 50 m race and are still as high as 5% in 100 m races. With short-course races in 25 m lanes becoming more and more popular, the importance of the start phase is even more dramatic. Far more than a quarter of the total time taken to reach the other side of the pool is spent on the starting phase. The same applies to turns, which also play a significant role, particularly when lanes are short.

Accordingly, we have set about improving our DigiCoach device with a view to investigating those phases and improving swimmer performance. Our aim is to expand the set-up into a measurement system for the diagnosis and training of these specific phases.

The first stage of this plan was to set up a new diagnosis system with four digital high-speed cameras capable of recording the athletes' performance with up to 100 fps simultaneously (see Figure 27). By means of contact plates and video metric measurements, timestamps are taken to calculate velocities and time spans between certain body positions. The analysis of these values provides important information on biomechanical performance and deficiencies in technique. A frame sequence of a female swimmer at the start is shown in Figure 26. The device was used during the last mass trials at the Olympic Training Base in Heidelberg. We are currently extending the set-up with the DigiCoach system to include acceleration measurements in the analysis (decisions). The continuation of this project will take place in conjunction with the Rhine-Neckar Olympic Training Base and a new start-up company.

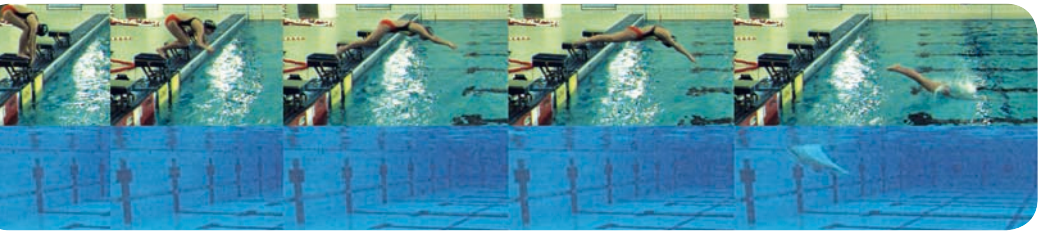


Fig. 26: Frame sequence of a start. Two digital high-speed cameras with frame rates of up to 100 fps record the swimmer simultaneously. A contact plate on the block measures reaction time and block time

Fig. 27: New analysis set-up for starts and turns. Four digital high-speed cameras monitor the athletes during their performance. One of the two underwater cameras can be seen in the background

Collaboration Partners

The project is conducted in collaboration with Dr. Klaus Reischle of the Institute of Sport Sciences at the University of Heidelberg and Wolfgang Döttling at the Rhine-Neckar Olympic Training Base. The facilities and technical equipment available at these institutions are used for the measurements.

Deutscher Schwimm-Verband e.V. (German Swimming Association), Kassel, Germany

Sponsors

The project is supported by the Deutscher Sportbund (German Sport Federation) and by research funds from the Federal Institute of Sport Science (Bundesinstitut für Sportwissenschaft) (Rec.No. VF0407/06/15/2002).

Klaus Tschira Foundation

3.14 MORABIT and MORABIT-FT

Project Manager
Prof. Dr. Andreas Reuter
Acting Project Manager
Dr. Rainer Malaka, European Media Laboratory
Project Member
Matthias Merdes

In the MORABIT project we are investigating new methods for testing component systems at run-time. Special attention is given to properties found in mobile systems, namely the inherent scarceness of resources, such as computing power and network bandwith, and the dynamic nature of such systems. In order to gain new insights in this area it is necessary to develop 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 needed is essentially a run-time infrastructure for software components that enables the 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 provide services to other components (server role), and 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 need test definitions and metadata describing the relationship between the core components and the tests. Test definitions normally take the form of well-known programmed test cases , e.g., in the popular “JUnit” testing framework. These tests check whether a potential service-providing component has some property or behavior important to the component requiring the test. Traditional defect tests, such as JUnit tests, are performed at development 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 execu-

tion of the program. As the test cases are bundled with the core components, they are often called “built-in tests“. The necessary component meta-data 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.

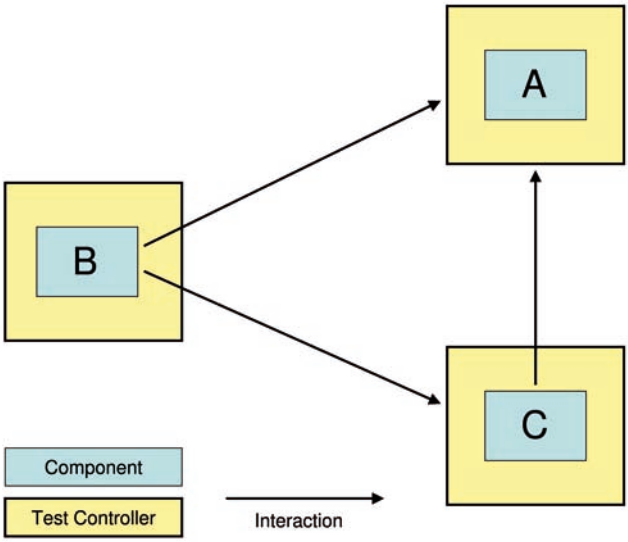


Fig. 28: Schematic interaction of three Morabit components wrapped with test controllers

We have implemented an initial prototype for such an infrastructure that is able to execute programs consisting of MORABIT components. The infrastructure uses the meta-data of the component to execute tests at certain well-defined points during the execution of the software. These logical points in time include the first contact between collaborating components and the invocation of a service from another component. The prototype also includes a facility for measuring the current state of resources, such as computing power and memory. The test execution management of the infrastructure can contact the resource measurement facility before executing a test in order to determine if there are sufficient resources to actually perform the test. Beyond this basic resource-aware test strategy, more sophisticated resource-aware test strategies are being developed. Dur-

ing the execution of the program the ‚normal' components can be thought of as being enhanced with testing capabilities: So-called test controllers are "wrapped around" the original components. A simplified illustration can be seen in Figure 28. The infrastructure allows the developer to turn off all testing, thus allowing for the development of a ‚normal' application first. The current prototype for the infrastructure includes rich support for logging static configuration data, as well as dynamics test request, execution, and result histories, and other important diagnostic information. A screenshot of part of the user interface of the prototype can be seen in Figure 29.

The MORABIT infrastructure is implemented at a meta-level with respect to the deployed component application, as its test execution behavior requires reflection about the execution of this application. Hence a sound understand-

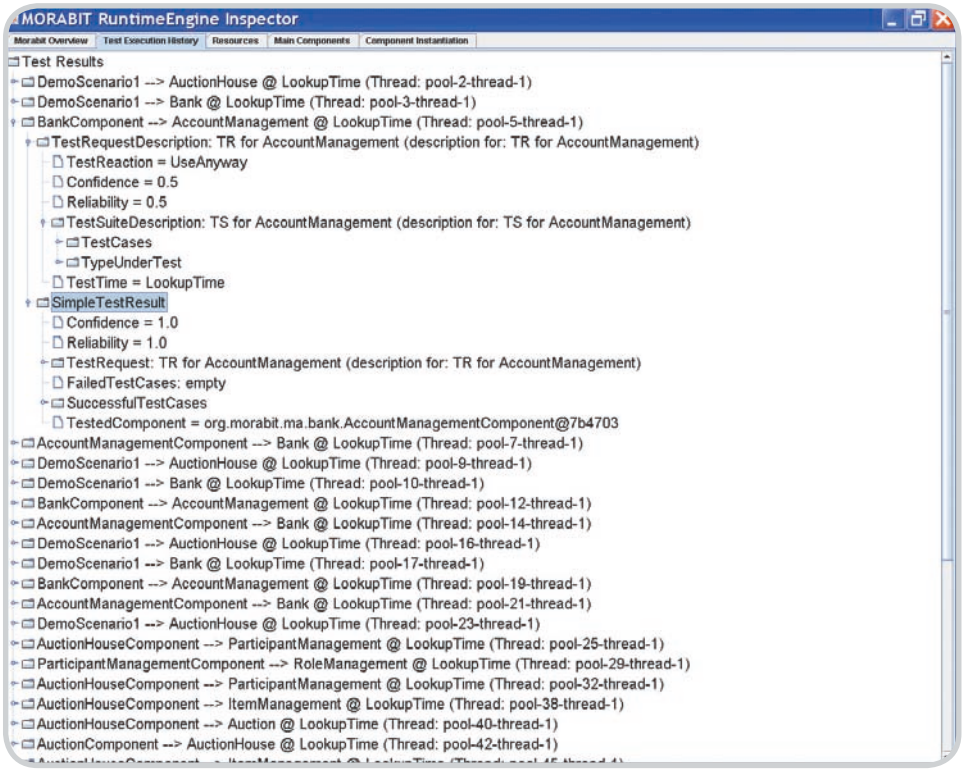


Fig. 29: Screenshot of the test execution history view in the Morabit Runtime Inspector

ing of the actual sequence of interactions at run-time is of paramount importance. During the specification of software systems, such behavioral sequences are typically modeled with UML sequence diagrams. As these diagrams are also very useful for understanding the actual behavior of a running application, we have developed a UML sequence diagram reengineering tool in a subproject. This tool is called 'Sequence4J' and is able to noninvasively observe the behavior of a running Java application and generate a sequence diagram while the program is running. With Sequence4J it is possible to visualize both the behavior of the core component application itself and the (reflective) behavior of the infrastructure.

In addition to generating diagrams during the execution of a program, Sequence4J is also able to store and post-process diagrams interactively after the recording. Generally, the fine-grained structure of object-orented software will often lead to very large diagrams. This makes it important to limit the size of a diagram, both via specifying filters before and by post-processing and editing diagrams after generation. The pre-processing is realized through detailed filter-mechanisms allowing the definition of include and exclude filters from a coarse-grained level like package-definitions up to the fine-grained level of a single method signature. The post-processing facility supports a multi-dimensional navigation on, and a selection of, the different elements of the graphical representation of the sequence diagram, such as participants in the interaction and exchanged messages. An enhancement of the diagram's readability has been achieved via hiding/deletion of a selection and automated generation of diagram fragments that can be exported from, and linked in, the original.

Before information on interaction in a program can be visualized, it has to be recorded in some way. This involves both the possibility of restricting the recording with filters as described above, and an efficient mechanism for extracting the necessary information from a running program. Ideally, the running program will be disturbed as little as possible, and the program itself (in source-code or execut-

able form) will remain unchanged. These requirements can be satisfied by using aspect-oriented technologies such as AspectJ, which enables users to insert the recording functionality of Sequence4J while loading the actual program under inspection. Diagram and model information can be stored permanently and edited later. Export is possible for diagrams in a vector graphics format and for model information for import into commercial UML modeling tools. The intention is to maintain Sequence4J in two forms, one closely integrated with the MORABIT infrastructure and the other as a stand-alone general-purpose tool.

Project Partners

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

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

Sponsor

Landesstiftung Baden-Württemberg

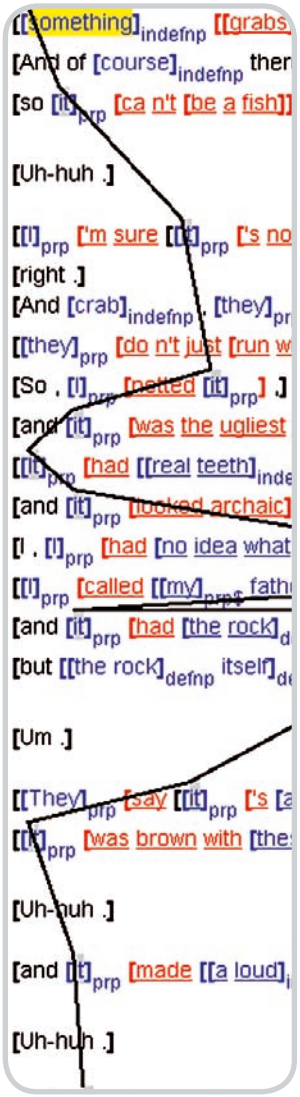


Fig. 30: MMAX2

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
- Downloads (2005): about 300
- Academic and commercial research licenses:
  15 single, 4 multiple
- Users in Germany, France, UK, the Netherlands, USA, Brazil, Singapore, and Japan

GermaNet Application Programming Interface (API) (NLP Group, Iryna Gurevych)
API developed to access GermaNet, a lexical semantic database for German represented in XML.

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.

- Downloads: 21
- Licences: 17

(April-December 2005)

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.
- Downloads: 62
- Licences: 59
(April-December 2005)

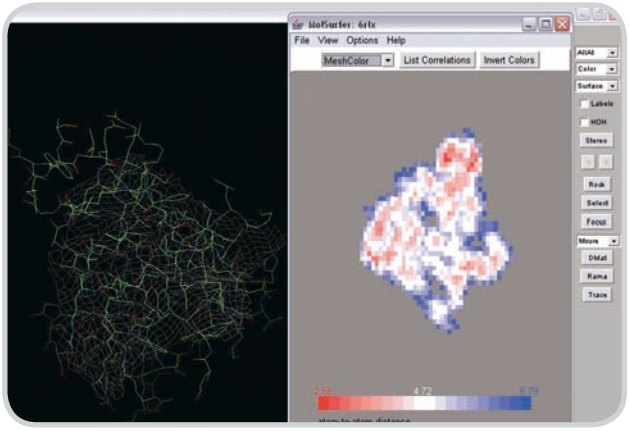


Fig. 31: MolSurfer

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.
Downloads: 25
Licences: 25
(April-December 2005)
(*All MCM licences distributed in 2005 were free*)

5.1 Workshops

5.1.1 First South Eastern European Workshop on Practical Approaches to Computational Biology

Opatija, 1-4 September 2005

The main aims of this event were the education of young scientists in South Eastern Europe in the field of computational biology and the establishment of a dialog both among scientists in the region and with scientists from Germany. While some areas of the computational sciences, such as quantum mechanical studies of small molecules, are well represented in South Eastern Europe, research on computational biology in that area is still in its infancy. Accordingly, the scientific program was designed to provide an introduction to new, cutting-edge areas of computational biology, such as systems biology and metagenomics, as well as to present state-of-the-art, top-quality work in topics ranging from bioinformatics to molecular simulation and structure-based drug design.

The scientific program of the workshop consisted of 23 lectures and a poster session. The lectures were divided into five sections: Bioinformatics, Structure-Based Drug Design, Systems Biology, Molecular Simulations, and Towards Quantum Chemical Study of Bio(macro)molecules. Eight lectures were given by prominent invited speakers from Germany, and four lectures by young scientists from South Eastern Europe invited to talk about their current work in laboratories abroad. The remaining 11 lectures were delivered by scientists working in South Eastern Europe.

57 scientists (19 women, 38 men) participated in the workshop. They came from Croatia (21), Germany and Hungary (10 each), Slovenia and Romania (5 each), Serbia and Montenegro (3 scientists), Bulgaria (2 scientists), and Austria (1 scientist). The participants ranged from diploma students (1), doctoral students (22), and postdoctoral fellows (7) to senior scientists, with two participants from industry.

The workshop was very successful both scientifically and socially, so that there was strong support for the idea that this

workshop should be the first in a series. More details can be found at http://projects.villa-bosch.de/mcm/conferences.

Organizers

Sanja Tomic (Ruder Boskovic Institute, Zagreb) and Rebecca Wade (EML Research, Heidelberg)

Program Committee

Simona Funar-Timofei (Romania), Dusanka Janezic (Slovenia), Hugo Kubinyi (Germany), Zvonimir Maksic (Croatia), Ilza Pajeva (Bulgaria), Michael Ramek (Austria), Sanja Tomic (Croatia), Rebecca Wade (Germany)

Sponsor

Alexander von Humboldt Foundation, Special Program for Academic Recovery in South Eastern Europe through the Stability Pact for South Eastern Europe

5.1.2 4th Workshop on Computation of Biochemical Pathways and Genetic Networks

September 12-13, 2005

Villa Bosch Studio, Heidelberg

Computational approaches are an integral part of systems biology. They are needed to support and complete the experimental investigation of the highly complex biochemical network in the living cell. Among these computational approaches modeling, simulation, and network analysis play a major role. Like its predecessors, the 4th Workshop on Computation of Biochemical Pathways and Genetic Networks focused on the current state of computational methodology in this research field at different levels of abstraction.

We had the following sessions:

- Network Analysis
- Stochastic Methods
- Reconstruction of Genetic Networks
- Tools

The following speakers were invited:

- Gregory Batt, INRIA, Rhone-Alpes, Saint Ismier, France
- Søren Brunak, DTU, Copenhagen, Denmark
- Roland Eils, DKFZ, Heidelberg, Germany
- Didier Gonze, ULB, Brussels, Belgium
- Marcelle Kaufman, ULB, Brussels, Belgium
- Edda Klipp, MPI Molecular Genetics, Berlin, Germany
- Yuri Kuznetsov, Universiteit Utrecht, Netherlands
- Reinhard Laubenbacher, VBI, Blacksburg, USA
- Nicolas Le Novere, EBI, Hinxton, UK
- Linda Petzold, UCSB, Santa Barbara, USA
- Isabel Rojas, EML Research, Heidelberg, Germany

Fig. 32: Participants of the 3rd Workshop on Computation of Biochemical Pathways and Genetic Networks in the Studio of the Villa Bosch

5.2 Colloquium Presentations

Dr. Monika Henzinger
Director of Research at Google Inc.

January 31, 2005:
The Past, Present, and Future of Web Information Retrieval

Prof. Dr. Jens Reich
Max Delbrück Center, Berlin

March, 8, 2005:
The moral status of the human being in vitro

Prof. Dr. Manfred Euler
Leibniz-Institut für die Pädagogik
der Naturwissenschaften (IPN), Kiel

May 17, 2005:
Lernen durch Experimentieren: Herausforderungen
an den naturwissenschaftlichen Unterricht (in German)

Dr. Toby Gibson
Team Leader at European Molecular Biology Laboratory

June 28, 2005:
Unstructure Determines Function

Prof. Dr. Dirk Helbing
Managing Director Institute for Transport & Economics, University of Technology, Dresden

September 19, 2005:
Dynamics of production, supply and traffic networks: From the slower-is-faster effect to signal control and business cycles

Dr. Roland Kaehlbrandt
CEO, gemeinnützige Hertie-Stiftung

October 25, 2005:
Deutsch – eine Sprache im Niedergang? (in German)

Prof. Dr. Hinrich Schütze
Chair of Theoretical Computational Linguistics, Institute for Natural Language Processing, University of Stuttgart

November 28, 2005:
The future of natural language processing: Machine learning or cognitive science?

Prof. Dr. Hans Schöler
Director, Max Planck Institute for Molecular Biomedicine, Münster

December 19, 2005:
Gewinnung und Einsatz Patienten-spezifischer pluripotenter Stammzellen (in German)

5.3 Miscellaneous

5.3.1 Talking about Careers in Science

Fostering public understanding of science has always been one of the priority goals of the Klaus Tschira Foundation. Accordingly, EML Research scientists seldom miss an opportunity to tell the public what they are doing.

In an ongoing series of events called "Mutmacher" ("Going Places") and designed to encourage personal enterprise, the organizers invited Dr. Iryna Gurevych from the NLP group to participate in an event called "Mut zur Karriere" ("The Way to a Career") on 30 January at the Kulturbrauerei pub in Heidelberg.

Before a packed audience she talked to compères Roland Ackermann and Rolf Kienle about the way she has been pursuing her career. "In my opinion," she said, "a career is not a stairway to mount, but more like a path I have chosen for myself and that I simply follow." Iryna also gave an outline of what computational linguists actually get up to and told the audience about current projects going on at EML Research. The other distinguished guests discussing different aspects of carving out a career for oneself in business, the arts, and science were the new chief conductor of the Heidelberg Philharmonic Orchestra, 24-year-old Cornelius Meister, and Adriana Nuneva, global marketing CEO of Heidelberger Druckmaschinen AG (Heidelberg Printing Machines).

Fig.33: Promoting science: Iryna Gurevych

Fig.34: On stage at the Kulturbrauerei: Cornelius Meister, Roland Ackermann, Irina Gurevych, Adriana Nuneva, and Rolf Kienle (from left to right)

5.3.2 The Run Goes On: Heidelberg Half-Marathon 2005

For the second time, EML Research members participated in the Heidelberg Half-Marathon. On 24 April 2005, Renate Kania, Michael Strube, and EML Research director Andreas Reuter took part once again in the 21-kilometer slog through some of Heidelberg´s most romantic spots. This year they were joined by KTF scholarship-holder Tomasz Marciniak and Christian Elting from the sister organization European Media Laboratory.

After an interval of several years, an EML team also entered the half-marathon competition. The result was respectable indeed, with the EML runners coming in 28th out of 255 teams. Once again, Michael Strube was the fastest runner in the team, crossing the finishing line after 1:28:17 hrs. and ranking 41st out of 3,000 starters.
Tomasz Marciniak clocked in after 1:38:16 hrs., which put him 226th in the overall ranking. Christian Elting breasted the (non-existent) tape after 1:53:45 hrs., thus ranking 984th.
Andreas Reuter finished the course in 2599th place (2:27:07 hrs., two minutes faster than the year before). Renate Kania, finally, took 2:37:07 hrs., ranking 2790th.

5.3.3 Bioinformatics for Schoolkids ("Kinder-Uni")

Children are extremely curious to know what is going on in the world around them. When it comes to natural phenomena and science, they often ask more pertinent questions than adults. This "researcher´s approach" is well-suited for interviews with experts. Accordingly, the local newspaper "Rhein-Neckar-Zeitung" started a project several years back called "Kinderuni im Netz" (Kids' University on the Web) to give schoolchildren between 9 and 13 the opportunity to interview scientists and experts about different topics, from the Stone Age to the age of computers, from astronomy to whales.

After interviewing the respective scientist, they would then write an article for publication both on the internet and in the paper.

One topic they chose was bioinformatics. For this purpose the young journalists visited EML Research to listen to BCB group leader Ursula Kummer. She first told them how we are able to transform a sugar cube or a cookie into sheer energy by simply swallowing it. "It´s a burning process, done not by a flame, but by a lot of proteins that are in every cell of our bodies."

To help them understand the complexity of reaction processes, Ursula Kummer showed the kids a map of biochemical pathways – just a small part of all the reactions taking place in a cell. She then described how computing can help to explore the biochemical processes in cells, and how EML Research scientists co-operate with their colleagues in the "wet labs."

The children asked a lot of questions, made a lot of notes, and apparently had a good time, well-supplied as they were with orange juice and cookies. Of course the cookies were only consumed in an experiment designed to illustrate reaction processes!

Fig. 35: The sugar cube lesson - Ursula Kummer talking about protein reactions

Fig. 36: A map of biochemical pathways – just a small part of all the reactions taking place in a cell

Fig. 37: Only for experimental purposes: Orange juice and cookies

6.1 Publications

[Ciaramita 2005] Massimiliano Ciaramita, Aldo Gangemi, Esther Ratsch, Jasmin Saric and Isabel Rojas: Learning of Semantic Relations between Concepts of a Molecular Biology Ontology. In: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence Edinburgh, UK, July 30, 2005 - August 5, 2005

[Ehrlich 2005] Lutz P. Ehrlich, Michael Nilges and Rebecca C. Wade: The Impact of Protein Flexibility on Protein-protein Docking. Proteins: Structure, Function and Bioinformatics, 2005, 58, pp. 126-133

[Gurevych 2005a] Iryna Gurevych: Anwendungen des semantischen Wissens über Konzepte im Information Retrieval. In: Proceedings of „Knowledge eXtended: Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten". Jülich, Germany, November 2 - 4, 2005 (to appear)

[Gurevych 2005b] Iryna Gurevych and Thade Nahnsen: Adapting Lexical Chaining to Summarize Conversational Dialogues. In: Proceedings of Recent Advances in Natural Language Processing Conference (RANLP'2005), Borovetz, Bulgaria, September 21 - 23, 2005 , pp. 212-218

[Gurevych 2005c] Iryna Gurevych and Hendrik Niederlich: Computing Semantic Relatedness in German with Revised Information Content Metrics. In: Proceedings of „OntoLex 2005 - Ontologies and Lexical Resources". IJCNLP'05 Workshop Jeju Island, Republic of Korea, October 15, 2005, pp. 28-33

[Gurevych 2005d] Iryna Gurevych Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005) Jeju Island, Republic of Korea , October 11, 2005 - October 13, 2005 , 767-778.

[Gurevych 2005e] Iryna Gurevych and Hendrik Niederlich: Accessing GermaNet data and Computing Semantic Relatedness. In: Companion Volume of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'2005), Michigan, Ann Arbor, USA, June 25 - 30, 2005 , pp. 5-8

[Gurevych 2005f] Iryna Gurevych and Hendrik Niederlich: Measuring Semantic Relatedness of GermaNet Concepts. In: Bernhard Fisseni, Hans-Christian Schmitz, Berhard Schröder and Petra Wagner (editors): Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: „Applications of GermaNet II" Frankfurt am Main: Peter Lang, pp. 462-474

[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. Intl. J. Quant. Chem. 2006, 106, in press

[Kmuníček 2005] Jan Kmuníček, Kamila Hynková, Tomáš Jedlička, Yuji Nagata, Ana Negri, Federico Gago, Rebecca C. Wade and Jiří Damborský: Quantitative Analysis of Substrate Specificity of Haloalkane Dehalogenase LinB from Sphingomonas paucimobilis UT26, Biochemistry, 2005, 44, pp. 3390-3401

[Kummer 2005a] Ursula Kummer and Lars Folke Olsen: No music without melody: How to understand biochemical systems by understanding their dynamics. Topics in Current Genetics, Vol., 13, pp.18-93

[Kummer 2005b] Ursula Kummer, Borut Krajnc, Jürgen Pahle and Marko Marhl: Transition from Stochastic to Deterministic Behaviour in Calcium Oscillations . Biophysical Journal, Vol. 89, pp.1603-1611

[Marciniak 2005a] Tomasz Marciniak and Michael Strube: Beyond the pipeline: Discrete optimization in NLP. In: Proceedings of the 9th Conference on Natural Language Learning (CONLL ,05) Ann Arbor, Mich., USA , June 29 - 30, 2005, pp. 136-143

[Marciniak 2005b] Tomasz Marciniak and Michael Strube: Discrete optimization as an alternative to sequential processing in NLG. In: Proceedings of the 10th European Workshop on Natural Language Generation (ENLG ,05), Aberdeen, Scotland, August 8 - 10, 2005 , pp.101-108

[Marciniak 2005c] Tomasz Marciniak and Michael Strube: Using an annotated corpus as a knowledge source for language generation. In: Proceedings of the Workshop on Using Corpora for Natural Language Generation (UCNLG ,05), Birmingham, UK, July 14, 2005, pp. 19-24

[Marciniak 2005d] Tomasz Marciniak and Michael Strube: Modeling and annotating the semantics of route directions. In: Proceedings of the 6th International Workshop on Computational Semantics (IWCS ,05) Tilburg, The Netherlands , January 12-14, 2005 , pp.151-162

[McCammon 2005] J. Andrew McCammon and Rebecca C. Wade. Pushing the limits: Editorial overview. Curr. Opin. Struct. Biol. 2005, 15, pp.135-136

[Merdes 2005] Matthias Merdes, Jochen Häußler and Alexander Zipf: GML2GML: generic and interoperable round-trip geodata editing - concepts and example. 8th AGILE Conference on GI-Science, Association of Geographic Information Laboratories in Europe. Estoril, Portugal, May 26 - 28, 2005

[Müller 2005a] Christoph Müller: Simplified MMAXQL: An Intuitive Query Language for Corpora with Annotations on Multiple Levels. In: Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialog (DiaLor'2005) Nancy, France, June 09 - 11, 2005 , pp.151-153

[Müller 2005b] Christoph Müller: A Flexible Stand-Off Data Model with Query Language for Multi-Level Annotation. In: Proceedings of the Interactive Posters/Demonstrations Session at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'2005), Ann Arbor, Michigan, USA, June 25- 30, 2005, pp.109-112

[Müller, M. 2005] Markus Müller, Katja Wegner, Ursula Kummer and Gerold Baier: The quantification of cross-correlations in stochastic spatio-temporal systems. Phys. Rev. E. (accepted)

[Ponzetto 2005] Simone Paolo Ponzetto and Michael Strube: Semantic role labeling using lexical statistical information. In: Proceedings of the 9th Conference on Natural Language Learning (CONLL '05), Ann Arbor, Mich., USA, June 29- 30, 2005, pp.213-216

[Saric 2005] Jasmin Saric, Lars J. Jensen, Rossita Ouzounova, Isabel Rojas and Peer Bork: Extraction of regulatory gene/protein networks from Medline. Bioinformatics Advance Access published online on July 26, 2005

[Schleinkofer 2005] Karin Schleinkofer, Sudarko, Peter J. Winn, Susanna K. Luedemann and Rebecca C. Wade: Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? EMBO Reports, 2005, 6, pp.584-589

[Spaar 2006] Alexander Spaar, Christian Dammer, Razif R. Gabdoulline, Rebecca C. Wade and Volkhard Helms: Diffusional encounter of barnase and barstar. Biophys. J. 2006, in press

[Strube 2005] Michael Strube: Anaphora and Coreference Resolution, Statistical. In: Brown, Keith (Editor in Chief). The Encyclopedia of Language and Linguistics, 2nd. Edition. Vol.1 Oxford, UK: Elsevier, pp. 216-223

[Wade 2005a] Rebecca C. Wade: Calculation and Application of Molecular Interaction Fields. In 'Molecular Interaction Fields. Applications in Drug Discovery and ADME Prediction', Ed. Cruciani, G., Wiley-VCH, Weinheim. (ISBN 3-527-31087-8) 2005, Ch. 2, pp.27-42

[Wade 2005b] Rebecca C. Wade, Domantas Motiejunas, Karin Schleinkofer, Sudarko, Peter J. Winn, Amit Banerjee, Andrei Kariakin and Christiane Jung: Multiple molecular recognition mechanisms. Cytochrome P450—A case study. Biochimica et Biophysica Acta. 2005, 1754, pp.239-244

[Wade 2005c] Rebecca C. Wade: Recognition Highlights and Commentaries J. Mol. Recognit. 2005, 18, 191

[Wang 2005] Ting Wang and Rebecca C. Wade: Force Field Effects on a ß-sheet Protein Domain Structure in Thermal Unfolding Simulations. J. Chem. Theory Comput, 2005, in press

[Wegner 2005a] Katja Wegner and Ursula Kummer: A new dynamical layout algorithm for complex biochemical reaction networks. BMC Bioinformatics (http://www.biomedcentral.com/1471-2105/6/212), 2005, 6:212

[Wegner 2005b] Katja Wegner: SimWiz3D - Visualising biochemical simulation results. In: Proceedings of the MediViz 2005 London, UK, July 6-7, 2005, pp. 77-82

[Winn 2005] Peter J. Winn, James N. D. Battey, Karin Schleinkofer, Amit Banerjee and Rebecca C. Wade: Issues in High-Throughput Comparative Modelling: A Case Study Using the Ubiquitin E2 Conjugating Enzymes. Proteins: Structure, Function and Bioinformatics, 2005, 58, pp.367-375

[Zobeley 2005] Jürgen Zobeley, Dirk Lebiedz, Julia Kammerer, Anton Ishmurzin and Ursula Kummer: A New Time-Dependent Complexity Reduction Method for Biochemical Systems. Transactions in Computational Systems Biology, Vol.1, pp.90-110

6.2 Guest Speaker Activities

Markus Buchner

Messplätze im Schwimmen, SpoTech, Magdeburg, June 30, 2005

Diagnostik im Schwimmen, Symposium „Medienkompetenzen in Sportunterricht und Training“ in Heidelberg, October 12 – 14, 2005

Ursula Kummer

Graduate Seminar „Nonlinearities and Disorder“. University of Magdeburg, Germany, January 31, 2005

Tomasz Marciniak
Michael Strube

Generating Natural Language Route Directions Using Classifiers and Linear Programming. Colloquium of the SFB/TR8 Spatial Cognition at the University of Bremen, Bremen, Germany, May 27, 2005

Jürgen Pahle

Biochemical Simulations: Which method to use? Berlin Center for Genome Based Bioinformatics, Berlin, Germany, November 30, 2005

Andreas Reuter

Keynote at the International Symposium on Applications and the Internet SAINT05, Trento/Italy, February 3, 2005

Invited Talk Joint DAIS/FMOODS Conference, Athens, June 17, 2005

Welcome Address (stand in for Dr. h. c. Klaus Tschira) at the 19th International Supercomputer Conference ISC2004, Heidelberg, June 22, 2005

11th International Workshop on High Performance Transaction Systems (HPTS), Asilomar, USA, September 25 – 28, 2005

Keynote at the ER2005 24th /International Conference on Conceptual Modeling, Klagenfurt, October 27, 2005

Introduction to Ontology Web Language (OWL), FH Heidelberg, Germany, April 5, 2005 — **Isabel Rojas**

"From Hydrogen Activation to Systems Biology". University Milano-Bicocca: October 5-7, 2005 — **Matthias Stein**

„Meeting minutes at the push of a button": Automatically summarizing spoken multi-party dialogue. Computerlinguistik, Universität Tübingen, Tübingen, Germany, November 21, 2005 — **Michael Strube**

"Estimation of protein-ligand binding affinity and specificity". 5th European Workshop on Drug Design, Certosa di Pontignano, Siena, Italy, May 31, 2005 — **Rebecca Wade**

"Cytochrome P450s: Multiple recognition mechanisms". Institute of Technical Biochemistry, University of Stuttgart, June 15, 2005

"Multiple molecular recognition mechanisms. Cytochrome P450 – a case study". 4th International Conference on Inhibitors of Protein Kinases and Workshop on Molecular Recognition, University of Warsaw, Poland, June 29, 2005

"Simulation of Protein-Ligand Interactions". 1st South Eastern European Workshop on Practical Approaches to Computational Biology', Opatija, Croatia, September 2, 2005

"Exploring Biomolecular Recognition Mechanisms by Modeling and Simulation". Molecular Graphics and Modelling Society International Meeting on 'Biomolecular Simulation', Trinity College Dublin, Ireland, September 12, 2005

"Brownian Dynamics Simulations". FEBS Course / 1st International NorStruct Workshop in Structural Biology on "Theoretical modeling of ligand binding and enzyme catalysis", University of Tromso, Norway, October 19th, 2005

"From protein structures to cellular biochemical networks". M2Cell, Royal Abbey of Fontevraud, France, Dec 3-7, 2005

6.3 Presentations

Talks

Markus Buchner

Technikdiagnostikmessplatz „DigiCoach“. KLD-Workshop, Heidelberg, 5.-6. April 2005

Razif Gabdoulline

“Deriving enzyme kinetic properties from structural similarities: application case of PIPSA (Protein Interaction property similarity Analysis)”. BMBF Hepatosys Systems Biology Programme Platform Meeting, Modeling & Bioinformatics, Berlin, January 27-28, 2005

”Towards modeling macromolecular motions in the cell by Brownian dynamics simulations”. BIOMS meeting, Heidelberg, February 23, 2005

”Simulation of diffusional association of proteins with electrostatic and hydrophobic interactions”. Workshop on Computer Simulations and Theory of Macromolecules, Hünfeld, April 22-24, 2005

Iryna Gurevych

Measuring Semantic Relatedness of GermaNet Concepts. GLDV Conference 2005, Bonn, March 31, 2005

Using the Structure of a Conceptual Network in Computing Semantic Relatedness. 2nd International Joint Conference on Natural Language Processing (IJCNLP‘2005), Jeju Island, Republic of Korea, October 13, 2005

Computing Semantic Relatedness in German with Revised Information Content Metrics. IJCNLP‘05 Workshop „OntoLex 2005 - Ontologies and Lexical Resources“, Jeju Island, Republic of Korea, October 15, 2005

Renate Kania
Ulrike Wittig
Martin Golebiewski
Andreas Weidemann
Olga Krebs
Isabel Rojas

SABIO-RK: a reaction kinetics database. First International Biocurators Meeting, Asilomar Conference Center, Pacific Grove, CA, USA, December 8 - 11, 2005

Ursula Kummer

Mathematical modeling: Chosing the right simulation method. FEBS Advanced Course Systems Biology, Gosau, Austria, March 12-17, 2005

BMBF 'HepatoSys' Systems Biology Programme Meeting, Heidelberg, April 28-29, 2005

Software tools for parameter identification and their limitations in applications to large systems. ECMTB 2005, Dresden, Germany, July 18 – 22, 2005

Do protein buffers influence the stochasticity of calcium oscillations? ECMTB 2005, Dresden, Germany, July 18 -22, 2005

Biochemical Pathways and Kinetics. MMS 2005, Heidelberg, Germany, July 25-27, 2005

A systems biology approach to the mechanisms of leukocyte activation. Swedish Bioinformatics Workshop, Goteborg, Sweden, November 25-26, 2005

Multi-Level Annotation with MMAX2. Joint TALK/AMI Workshop on Standards for Multimodal Dialogue Context, Edinburgh, Scotland, December 12, 2005 **Christoph Müller**

Multi-level annotation of linguistic data with MMAX2. Learntec 2005, 13. Europäischer Kongress und Fachmesse für Bildungs- und Informationstechnologie, Kongresszentrum Karlsruhe, Karlsruhe, Germany, February 15 – 17, 2005 **Michael Strube**

"Using protein structures in systems biology". BMBF 'Hepatosys' Systems Biology Programme Modeling and Simulation Platform Meeting, Charite, Berlin, January 27, 2005 **Rebecca Wade**

"Using protein structures in systems biology". BMBF 'Hepatosys' Systems Biology Programme Meeting, DKFZ, Heidelberg, April 29, 2005

"Force Field Effects on a Beta-sheet Structure Protein Domain." Methods of Molecular Simulation 2005 Workshop, Heidelberg, Germany, July 25-27, 2005 **Ting Wang**

Posters

Anna Feldman-Salit

"Protein-Protein Docking Guided by Biochemical Data". Motiejunas, D., Wang, T., Feldman-Salit, A., Johann, T., Gabdoulline, R.R. and Wade R.C. Katzir Conference on Molecular Perspectives on Protein-protein Interactions, Eilat, Israel, November 6-10, 2005

Iryna Gurevych

Anwendungen des semantischen Wissens über Konzepte im Information Retrieval Knowledge eXtended: Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten, Jülich, Germany, November 3, 2005

Stefan Henrich

"The Application of COMBINE Analysis to Generate Target-Specific Scoring Functions", Henrich, S., Wang, T., Blomberg, N., Wade, R.C., 19. Darmstädter Molecular Modelling Workshop, Computer-Chemie-Centrum, Erlangen, Germany, May 3-4, 2005

"The Application of COMBINE Analysis to Generate Target-Specific Scoring Functions", Henrich, S., Wang, T., Blomberg, N., Wade, R.C., 1. German Conference on Chemoinformatics, Goslar, Germany, November 13-15, 2005

Domantas Motiejunas

"Rigid Body Protein-Protein Docking Using Biochemical Data", Motiejunas, D., Gabdoulline, R.R., Wade, R.C., Molecular Graphics Modeling Society (MGMS) International Meeting 2005, Trinity College Dublin, "Biomolecular Simulations – from Prediction to Practice", September 11th-14th, 2005

"Rigid Body Protein-Protein Docking Using Biochemical Data", Motiejunas, D., Gabdoulline, R.R., Wade, R.C., 4th European Conference on Computational Biology 2005, Madrid, September 28th – October 1st, 2005

Matthias Stein

"Integrating Structural and Kinetic Enzymatic Information in Systems Biology", Stein, M., Gabdoulline, R.R., Winn, P.J., and Wade , R.C. 4th Workshop on Computation of Biochemical Pathways and Genetic Networks, Villa Bosch, Heidelberg, Germany. 12-13 September 2005

"Integrating Structural and Kinetic Enzymatic Information in Systems Biology", Stein, M., Gabdoulline, R.R., Winn, P.J., and Wade , R.C. BMBF Evaluation Workshop Hepatosys, Berlin, Germany, 29-30 November 2005

"Integrating Structural and Kinetic Enzymatic Information in Systems Biology", Stein, M., Gabdoulline, R.R., Winn, P.J. and Wade, R.C. M2Cell, Royal Abbey of Fontevraud, France, Dec 3-7, 2005	
"Simulating Protein-Protein Docking by Molecular Dynamics." Wang, T., Johann, T., Motiejunas, D., Gabdoulline, R.R. and Wade , R.C. 6th European Symposium of the Protein Society, Barcelona, Spain, April 30-May 4, 2005	**Ting Wang**
SABIO-RK - a reaction kinetics database. Sixth International Conference on Systems Biology (ICSB 2005). Boston, MA, USA, October 19-24, 2005	**Andreas Weidemann** **Renate Kania** **Ulrike Wittig** **Martin Golebiewski** **Olga Krebs** **Isabel Rojas**
"High-Throughput Analysis of Conjugating Enzymes (E2s) of Ubiquitin and Ubiquitin-like Proteins". Winn, P.J., Battey, J.N.D, Religa,T. L., Banerjee, A. and Wade, R.C. New Approaches in Drug Design and Discovery, Schloss Rauschholzhausen, Marburg, March, 2005	**Peter Winn**
"High-Throughput Analysis of Conjugating Enzymes (E2s) of Ubiquitin and Ubiquitin-like Proteins" Winn, P.J., Battey, J.N.D, Religa, T.L., Banerjee, A. and Wade, R.C. German Conference on Bioinformatics GCB05 conference, Hamburg, October 5-7 2005	

Demos

A User Interface for Estimating Semantic Relatedness of Words. GLDV Conference 2005, Bonn, Germany, March 31, 2005	**Iryna Gurevych** **Hendrik Niederlich**
Accessing GermaNet Data and Computing Semantic Relatedness. 43rd Annual Meeting of the Association for Computational Linguistics (ACL'2005), Ann Arbor, Michigan, USA, June 26, 2005	

6.4 Memberships

Ursula Kummer	Coordination and Member of the Scientific Committee of BIOMS
	Member of the Steering Committee of MTBIO
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
	Core member of the German-Japanese Forum on Information Technology
	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 Technology Council of T-Systems debis Systemhaus Solutions for Research GmbH (SfR)
	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
	Member of the Board of Trustees of Fraunhofer Gesellschaft for Integrated Publication and Information Systems
	Mitglied Fachgremium "Landesstiftung Baden-Württemberg"
	Member of the Board of Trustees of Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
	Member of the Editorial Board, Distributed and Parallel Databases, an International Journal, Springer-Verlag

Mitglied des Forschungsverbundes Wissenschaftliches Rechnen Baden-Württemberg „WiR“

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

Section Editor (with J. A. McCammon): Current Opinion in Structural Biology, Theory and Simulation issue, vol 15, April, 2005. **Rebecca Wade**

Editor: Journal of Molecular Recognition

Editorial Board: Journal of Computer-aided Molecular Design; 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

European Conference on Mathematical and Theoretical Biology (ECMTB) 2005, Dresden, Germany, July 18-22, 2005 **Ursula Kummer**

43rd Annual Meeting of the Association for Computational Linguistics (ACL ,05). Area Chair for the Area Discourse, Dialogue, and Multimodality. Ann Arbor, Michigan, USA, June 25-30, 2005 **Michael Strube**

Human Language Technology Conference/10th Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP ,05). PC Member. Vancouver, B.C., Canada October 6-8, 2005

Editorial Work

Proceedings of the 4th Workshop of Computation of Biochemical Pathways and Genetic Networks Heidelberg, Germany, September.12-13, 2005 **Ursula Kummer, Jürgen Pahle, Irina Surovtsova, Jürgen Zobeley**

ACM Transactions on Asian Language Information Processing (ACM TALIP) Computational Linguistics Journal Speech Communication **Michael Strube**

Workshop Organization

Ursula Kummer 4th Workshop on Computation of Biochemical Pathways and Genetic Networks. Heidelberg, Germany, September 12 – 13, 2005

Rebecca Wade Co-organizer (with Sanja Tomic) of the "1st South Eastern European Workshop on Practical Approaches to Computational Biology", Opatija, Croatia, September 1-4, 2005 (funded by the Alexander von Humboldt Foundation)

6.6 Patent

Iryna Gurevych
Michael Strube U.S. Provisional Patent No. 60/676,652 „Method for Generating Definitions of Words and Concepts Automatically and Its Application to the Task of Computing Semantic Relatedness of Words and Concepts". April 29, 2005
(*Patent rights belong to EML Research gGmbH*)

Science Degrees

[Anstein 2005] Anstein, S.: Analysing Names of Organic Chemical Compounds- From Morpho-Semantics to SMILES Strings and Classes, Diploma thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, apl. Prof. Dr. Uwe Reyle, 2005, and EML Research: Isabel Rojas

[Besson 2005] Besson, B.: Development of a software to aid the search of protein structural information to estimate kinetic parameters in systems biology applications. Internship report, Bioinformatics Programme, INSA Lyon, France, and EML Research: Matthias Stein and Rebecca Wade, 2005

[Cvoro 2005] Cvoro, V.: Recognition of Emotional Preferences for Professional Features, Master Thesis, Computational Linguistics Department at the Ruprechts-Karls University of Heidelberg (Prof. Dr. Peter Hellwig), and EML Research: Iryna Gurevych, 2005

[Jäger 2005] Jäger, J.: Word Sense Disambiguation in a Spoken Dialogue Summarization System (orig.title: „Wortlesartendisambiguierung in einem Zusammenfassungssystem für spontansprachliche englische Dialoge"), Master Thesis, Computational Linguistics Department at the Ruprechts-Karls University of Heidelberg (Dr. Anke Holler), and EML Research: Iryna Gurevych, 2005

[Johann 2005] Johann, T.: Perturbing Methods in Molecular Dynamics Simulations with Implicit Solvent Models. Diploma Thesis, Faculty of Physics, Ruprecht-Karls-University, Heidelberg (Prof. Jeremy Smith), and EML Research: Rebecca Wade and Ting Wang, 2005

[Kremer 2005] Kremer, G.: Analysing Names of Organic Chemical Compounds - From Morpho-Semantics to SMILES Strings and Classes, Diploma thesis
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, apl. Prof. Dr. Uwe Reyle, 2005, and EML Research: Isabel Rojas

[Niederlich 2005] Niederlich, H.: Document-Retrieval Models in a Language Based Recommender System (orig. title: „Document-Retrieval-Modelle in einem natürlichsprachlichen Beratungssystem"), Diploma Thesis, Computer Science Department at the University of Applied Sciences of Lübeck (Prof. Dr. Ralf Schiffer), and EML Research: Iryna Gurevych, 2005

[Schenk 2005] Schenk, I.: Resolution of Pronouns with Non-NP Antecedents in Spontaneous English Dialogues (orig. title: „Auflösung der Pronomen mit Nicht-NP-Antezedenten in spontansprachlichen englischen Dialogen"), Master Thesis, Computational Linguistics Department at the Ruprechts-Karls University of Heidelberg (Prof. Dr. Peter Hellwig), and EML Research: Iryna Gurevych, 2005

[Ulbrich 2005] Ulbrich, S.: The Protein Structure Annotation Tool ProSAT2. Studienarbeit, Faculty of Computer Science, University of Karlsruhe (Dr. Feldbusch) and EML Research: Razif Gabdoulline and Rebecca Wade, 2005

Lectures

Iryna Gurevych

Introduction to Natural Language Processing, School of Information Technology, International University in Bruchsal, January - April 2005, Germany

Processing Large Corpora Using Methods of Lexical Semantics (orig. title: Erschließung großer Korpora mit Verfahren der lexikalischen Semantik). Department of Computational Linguistics, Ruprechts-Karls University of Heidelberg, April - July 2005, Germany

Lexical Semantic Processing in NLP. BA/MA Program „International Studies in Computational Linguistics (ISCL)", Seminar für Sprachwissenschaft, University of Tübingen, October - December 2005, Germany

Renate Kania

Was Daten in der Biologie anrichten können, Course Bio-NLP, Summer 2006, Institute for Natural Language Processing. University of Stuttgart, Germany

Ursula Kummer

Simulation in der Biochemie, University of Heidelberg, WS 2005/06

Matthias Stein

Seminar on Structural Biology, University Heidelberg, with Prof. Jeremy Smith, July, 2005

Rebecca C. Wade

"Brownian Dynamics Simulations", FEBS Course / 1st International NorStruct Workshop in Structural Biology on "Theoretical modeling of ligand binding and enzyme catalysis", University of Tromso, Norway, October 20th, 2005

Einführung in die objektorientierte Programmierung mit Java. Fachhochschule Furtwangen - Informatica Feminale (September 2005), Furtwangen, Germany — **Katja Wegner**

Biologische Ressourcen, Course Bio-NLP, Summer 2006, Institute for Natural Language Processing. University of Stuttgart, Germany — **Ulrike Wittig**

Practicals

Student Projects in Natural Language Processing. School of Information Technology, International University in Bruchsal, January - April 2005, Germany — **Iryna Gurevych**

An Independent Study Project: Information Retrieval by Using Semantic Information. School of Information Technology, International University in Bruchsal, May - August 2005, Germany

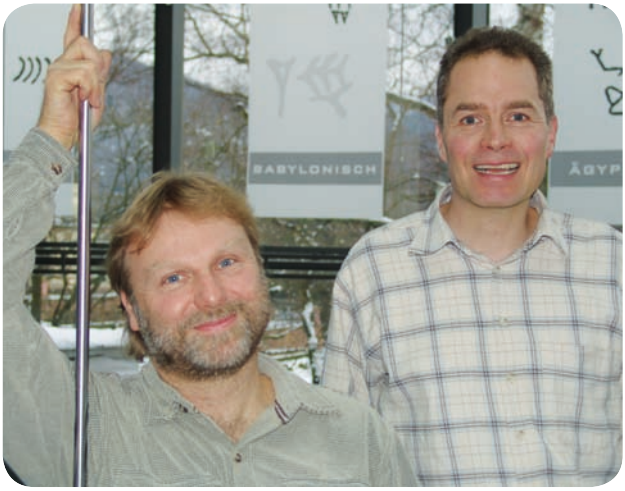


Fig. 38: Norbert Rabes (left), Achim Beck

The computer network of EML Research is currently based on Microsoft Windows 2000 and makes up 85% of the IT backbone. In realizing the active directory structure, particular attention has been paid to security, simple and flexible administration, scaling, upgradability, and support of 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 was expanded to GigaBit. Optical (LWL)-GigaBit lines connecting the two buildings, while within the buildings ordinary Cu-GigaBit is used for the network.

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

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.

EML Research has 22 Windows 2000/2003 servers (MS Exchange2000, MS SQL 2000, Oracle 9), fourteen 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 email filtering server ("spam server") was put into service running on a FreeBSD system and open source's ‚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. Libero ibrary management softwarewith 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. All Linux servers are backed-up via an intermediate RAID storage system with a capacity of 4TB, from which the data are saved to tape.

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).

ISSN 1438-4159

www.eml-research.de