Universität Bielefeld

CeBiTec
Center for Biotechnology

Center for Biotechnology
A Decade

Preface

Over the last ten years, the Center for Biotechnology (CeBiTec) at Bielefeld University has become a major international research institution of exceptional quality. It benefits from the tradition of interdisciplinarity that Bielefeld has embraced as a guiding principle since its foundation.

The life sciences have managed to break down traditional barriers more effectively than other interdisciplinary fields. This holds not only for the barriers between the natural sciences themselves, but also for the barriers between natural sciences and technological subjects such as informatics. This is particularly evident in the field of bioinformatics, a combination that was pioneered in Bielefeld. These developments enable the CeBiTec to realize forward-looking ideas: ideas which advance research and solve some of the mysteries of the life sciences. Soon these ideas will be applied in the everyday life of all individuals. Here we witness the development of the scientific basis for the development of new pharmaceuticals, for new technology used in environmental protection such as the application of biofuels, and for new microscope techniques, to name just a few.

The CeBiTec has continually expanded and developed a clear and well-formed structure, that has enabled optimal interdisciplinary exchange between various fields. It consists of the Institutes for Bioinformatics, Genome Research and Systems Biology, Biophysics and Nanoscience, and Biochemistry and Bioengineering. All institutes run different technology platforms providing state-of-the-art technical infrastructure and methods. Additionally, the CeBiTec runs a Graduate Center.

This compact organizational structure is complemented by physical structure. CeBiTec now has its own laboratory building, meeting the highest demands of modern research while being architecturally appealing at the same time. After the official opening of the first part of the CeBiTec building in February 2007, an extension of the building was completed in March 2009. These activities constitute the most extensive building project at Bielefeld University since the construction of the main building about 40 years ago. The University feels extremely fortunate that it was possible to realize this construction project in spite of the current adverse climate in public spending. This demonstrates the importance attached to the research carried out at the CeBiTec, as seen from a political perspective. The University also sees this as a sign of appreciation of its significant efforts in building a profile for seminal scientific fields.

Although the general financial conditions have been unusually difficult, Bielefeld has strengthened the interdisciplinary focus over the past few years through its appointment strategy. The CeBiTec has focussed on being attractive to ambitious researchers, and will definitely continue to do so in the future. There are no less than 300 members of staff, working under optimal conditions and in close mutual proximity.

Prof. Dr.-Ing. Gerhard Sagerer

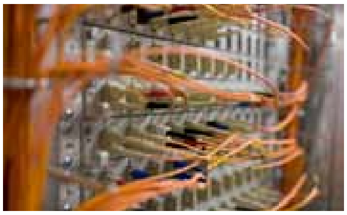
At the same time, the CeBiTec plays an important role in the promotion of the next generation of young scientists by consistently encouraging their development. The Graduate Center offers ideal conditions for post-doctoral research and provides a starting platform for a successful scientific career. Furthermore, the CeBiTec has developed innovative new courses of study which attract gifted students from all over the world.

The CeBiTec not only shines in the international world of science; it also has important cooperation partners in many branches of industry, including multinational companies. With these industry partners, the CeBiTec is able to realize high calibre projects. The CeBiTec also advises companies in the Bielefeld area operating in a nanoscience context.

I am sure that the CeBiTec will initiate important impulses for development in the key areas of the natural sciences and technology, and would like to wish all participating scientists the highest possible amount of success in their work. I hope you will enjoy reading this brochure, which provides insight into the diversity of research carried out by the CeBiTec.

Prof. Dr.-Ing. Gerhard Sagerer
Rector, Bielefeld University, May 2010

Table of Contents



The Center for Biotechnology
A Central Scientific Institution of Bielefeld University

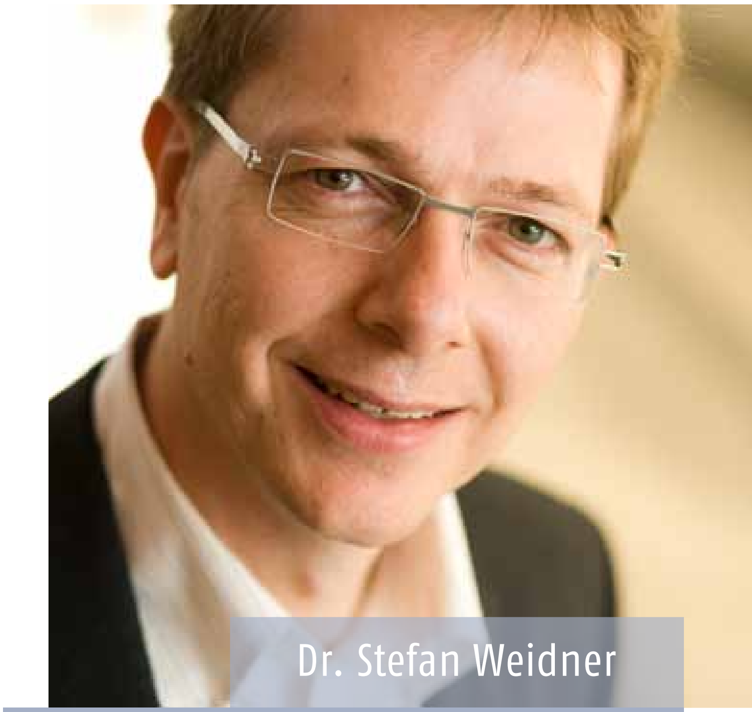
The development of the CeBiTec during the last decade

Biotechnology is a field of great importance and is a key technology requiring the multidisciplinary cooperation between biologists, chemists, physicists, mathematicians, information scientists and engineers. The idea to constitute a scientific center dedicated to biotechnology was born in 1995. This center would bundle the activities and interests of the groups focusing on biotechnology within the Faculties of Biology, Chemistry, and Technology of Bielefeld University and increase the visibility of the field biotechnology at Bielefeld University. In the following years the plan became more concrete and finally in September 1998 the senate of Bielefeld University officially established the scientific institution 'Center for Biotechnology – CeBiTec' (Table 2). In itially its scientific members were heavily engaged in defining joint research projects. Particularly the combination of Genome Research and Bioinformatics turned out to be of highest importance for the further development of the CeBiTec. A grant from the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG) which was part of the 'DFG Initiative Bioinformatics' together with a matching fund from Bielefeld University supported the establishment of two CeBiTec institutes, namely the 'Institute for Bioinformatics' (founded in 2002) and the 'Institute for Genome Research' (founded in 2003). The latter was later renamed the 'Institute for Genome Research and Systems Biology'. In 2001, a further important measure could be established, the 'International NRW Graduate School in Bioinformatics and Genome Research'. This was made possible by a grant from the state North Rhine-Westphalia which secured its finances until 2009. Also in 2001, the CeBiTec was successful with its proposal for the establishment of the competence network 'Genome Research on Bacteria relevant for Agriculture, Environment and Biotechnology – GenoMik', financed by the

Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*, BMBF). In 2006, after the first five years, the competence network was extended for further three years. In 2004 a third CeBiTec institute was founded, namely the 'Bielefeld Institute for Biophysics and Nanosciences – BINAS'. The highlight of the year 2004 was without doubt the beginning of the construction of the CeBiTec laboratory building on the campus of the University of Bielefeld. The first part was officially opened 2007 and the second part was completed 2009. The CeBiTec developed its current structure in 2007, with the establishment of the 'Institute for Biochemistry and Bioengineering'. The interdisciplinary approach of the scientific groups becomes apparent in the different topics of the CeBiTec symposium series. The CeBiTec organizes symposia on current topics annually, starting in 2006 with a symposium on Molecular Systems Biology (Figure 1).

The CeBiTec houses all modern disciplines playing a role in biotechnology

Today the Center for Biotechnology is a central scientific institution at Bielefeld University serving as a focal point of interdisciplinary research and strategic research planning in life sciences. Its mission is to encourage and to support the development of innovative projects crossing discipline boundaries. The close collaboration of scientists from the Faculties of Biology, Chemistry, Physics, and Technology in various research projects is supported by the German Research Foundation (DFG), the Federal Ministry of Education and Research (BMBF), the State North Rhine-Westphalia, the European Union and by industrial cooperation. Today four institutes, namely the 'Institute for Bioinformatics', the 'Institute for Genome Research and Systems Biology', the 'Institute for Biophysics and Nanosciences', and the 'Institute for Biochemistry and Bioengineering' form the main pillars of the CeBiTec. These institutes run different technology platforms particularly for bioinformatics and for genomics providing the technical infrastructure as well as the methodological procedures and applications essential for the research fields of the CeBiTec groups. Furthermore, the graduate center currently offers high-level PhD programs. Initially starting with a program dedicated to bioinformatics and genome research, currently two further programs are offered. One program is dedicated to industrial biotechnology while the second is an international graduate programme in bioinformatics of signaling networks. In total CeBiTec has more than 300 members. The detailed structure of the institution is presented in Figure 2. Due to the shortage of highly equipped laboratory space the different groups of the CeBiTec were dispersed in different parts of the university main building and their available amount of space was insufficient. As already mentioned before, Bielefeld University fortunately acquired funds to erect a new laboratory building. Building work started 2004. The move to the CeBiTec building was possible in June 2007, after completion of the first part. An extension of the building was completed in March 2009. The building now houses the administration of the CeBiTec and several groups of the 'Institute for Genome Research and Systems Biology'. The 'Chair of Genome Research' of Prof. Dr. B. Weisshaar, the senior research group of Prof. Dr. A. Pühler, junior research groups as well as the 'Technology Platform Genomics' reside in the building. Furthermore, the 'Chair of Algae Biotechnology and Bioenergy' of Prof. Dr. O. Kruse of the 'Institute for Biochemistry and Bioengineering' and specially equipped laboratories of the 'Institute for Biophysics and Nanosciences' found their home in the building.



Dr. Stefan Weidner

Stefan Weidner studied biology at Bielefeld University. After completing his PhD in 1996 at the Department of Genetics he worked as a PostDoc in different national and international third-party funded projects, first in the area molecular systematics and later in bacterial genome research. From 2001 he was commissioner for the construction of the CeBiTec laboratory building and co-leader of a coordinating committee for a bioinformatics and genome research initiative at Bielefeld University. Since 2004 he is the Executive Director of the CeBiTec.

Figure 1: Posters announcing the four CeBiTec symposia from 2006 to 2009. Every year an up-to-date topic is elucidated and discussed by international invited speakers.

The CeBiTec is run by an executive committee and supervised by a scientific advisory board

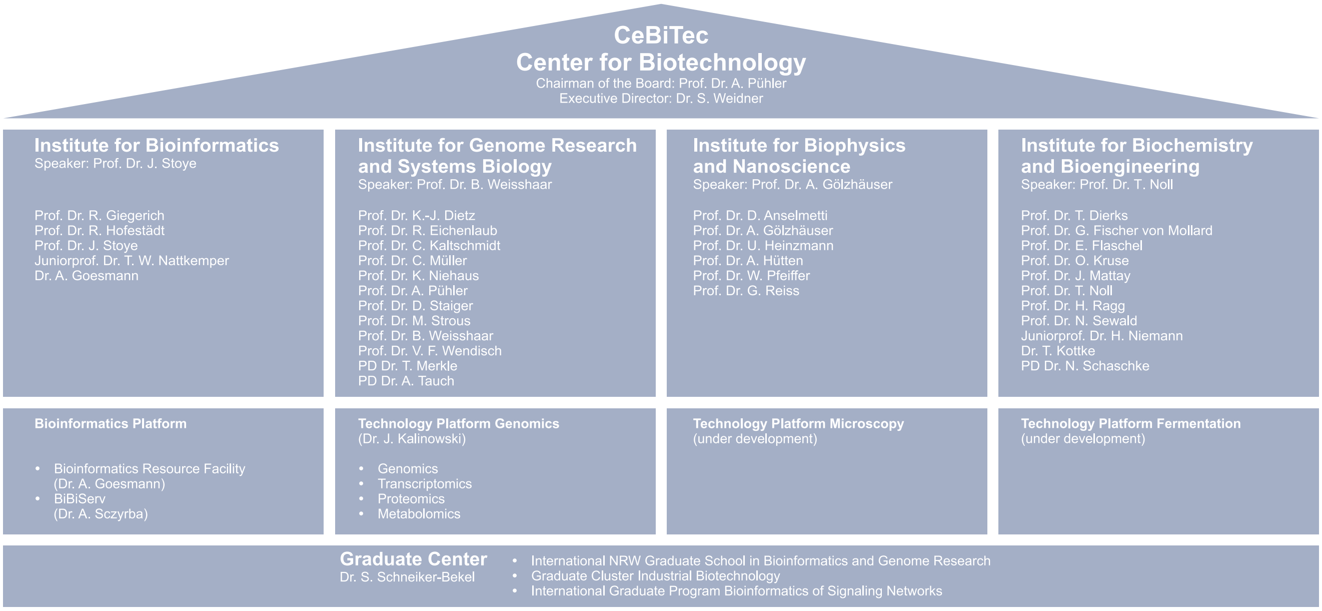
The size of the CeBiTec and the number of participating groups from different faculties of Bielefeld University clearly require an organizational structure capable of managing not only the day-to-day business, but also the coordination of joint research projects and project applications, as well as setting the future direction. The Center for Biotechnology was established by the Senate of Bielefeld University in 1998. Shortly after this official decision the first meeting of the executive committee was held. In particular the 'DFG Initiative Bioinformatics' grant in September 2000 facilitated the expansion of the CeBiTec necessitating a changed management structure. The Executive Committee is now composed by representatives of all status groups of the university. The speakers of the four institutes and two further elected members represent the professorate. The status groups research staff, technical staff, and students are also represented in the executive committee by elected members. The current members of the committee are shown in Figure 3. Since February 2004 Prof. Dr. A. Pühler is the chairmen of the executive committee and Dr. S. Weidner is executive director of the CeBiTec.

Since November 2005 the CeBiTec is supervised by a scientific advisory board now composed of eight members from industry and academia (Table 1) with an excellent expertise in the fields of activity of the CeBiTec. Prof. Dr. R. Amann (*Max-Planck-Institut für Marine Mikrobiologie*, Bremen, Germany) is the elected speaker of the board. The Board meets bi-yearly and advises the executive committee in all matters of the CeBiTec, particularly concerning principles of its scientific work and future development.

Junior research groups benefit from laboratory space in the CeBiTec building

As outlined above and illustrated in Figure 2 four institutes form the pillars of the CeBiTec. In chronological order the first two institutes, namely the 'Institute for Bioinformatics' and the 'Institute for Genome Research and Systems Biology' were funded by the 'DFG Bioinformatics Initiative' grant together with a matching fund from Bielefeld University. Furthermore, this DFG Initiative provided for the establishment of junior research groups in these institutes. Multiple junior research groups were set up in the different institutes with funds from various sources. The CeBiTec building provides highly equipped laboratory space, particularly modern wet laboratories, laboratories for practical student courses, seminar and meeting rooms. The policy of the CeBiTec is to support its junior research groups by supplying laboratory space and allocating the appropriate basic equipment and infrastructure. As a matter of course the groups also benefit from the modern equipment of the technology platforms. Therefore, the stage is set for a successful start of new groups at the CeBiTec.

Over the years the number of junior research groups varies. On the one hand several new groups could be established; on the other hand several scientists very successfully moved into both academia and industry. Currently, the junior research groups of PD Dr. T. Merkle and PD Dr. A. Tauch reside in the building. Furthermore, the former junior research group and now 'Chair of Algae Biotechnology and Bioenergy' of Prof. Dr. O. Kruse found place in the building.



CeBiTec
Center for Biotechnology
Chairman of the Board: Prof. Dr. A. Pühler
Executive Director: Dr. S. Weidner

Institute for Bioinformatics
Speaker: Prof. Dr. J. Stoye

Prof. Dr. R. Giegerich
Prof. Dr. R. Hofestädt
Prof. Dr. J. Stoye
Juniorprof. Dr. T. W. Nattkemper
Dr. A. Goesmann

Institute for Genome Research and Systems Biology
Speaker: Prof. Dr. B. Weisshaar

Prof. Dr. K.-J. Dietz
Prof. Dr. R. Eichenlaub
Prof. Dr. C. Kaltschmidt
Prof. Dr. C. Müller
Prof. Dr. K. Niehaus
Prof. Dr. A. Pühler
Prof. Dr. D. Staiger
Prof. Dr. M. Strous
Prof. Dr. B. Weisshaar
Prof. Dr. V. F. Wendisch
PD Dr. T. Merkle
PD Dr. A. Tauch

Institute for Biophysics and Nanoscience
Speaker: Prof. Dr. A. Gölzhäuser

Prof. Dr. D. Anselmetti
Prof. Dr. A. Gölzhäuser
Prof. Dr. U. Heinzmann
Prof. Dr. A. Hütten
Prof. Dr. W. Pfeiffer
Prof. Dr. G. Reiss

Institute for Biochemistry and Bioengineering
Speaker: Prof. Dr. T. Noll

Prof. Dr. T. Dierks
Prof. Dr. G. Fischer von Mollard
Prof. Dr. E. Flaschel
Prof. Dr. O. Kruse
Prof. Dr. J. Mattay
Prof. Dr. T. Noll
Prof. Dr. H. Ragg
Prof. Dr. N. Sewald
Juniorprof. Dr. H. Niemann
Dr. T. Kottke
PD Dr. N. Schaschke

Bioinformatics Platform

- Bioinformatics Resource Facility (Dr. A. Goesmann)
- BiBiServ (Dr. A. Sczyrba)

Technology Platform Genomics
(Dr. J. Kalinowski)

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics

Technology Platform Microscopy
(under development)

Technology Platform Fermentation
(under development)

Graduate Center
Dr. S. Schneiker-Bekel

- International NRW Graduate School in Bioinformatics and Genome Research
- Graduate Cluster Industrial Biotechnology
- International Graduate Program Bioinformatics of Signaling Networks

Figure 2: Organization chart of the CeBiTec with its four institutes, technology platforms, and Graduate Center.

State-of-the-art scientific infrastructure supports research projects

The scientific members of the CeBiTec undertake projects in a multitude of fields of research. Undoubtedly the main emphasis is on projects in microbial genomics, postgenomics and biotechnology. One major project is the competence network 'Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology – GenoMik' and the follow-up program 'GenoMik-Plus', both financed by the Federal Ministry of Education and Research (BMBF). In recent times projects in the field of genome research on industrial microorganisms gained in importance. The participation in the initiative 'CLIB^{2021} – Cluster Industrial Biotechnology' is a prominent example for this new direction.

Next to microbial genomics, genome research on plants, particularly on *Arabidopsis thaliana* constitute another focus. A prominent example is the participation in the program 'GABI – Genome Analyses in the Biological System Plant', which focuses on future-oriented plant genome research and is financed by the BMBF and by private enterprises. A latter research field concerns solar biofuel production with an emphasis on the use of microalgae.

A comprehensive technological infrastructure for all fields of genome and post genome research is crucial to successfully pursuing these goals. The 'Technology Platform Genomics' possesses the state-of-the-art technical equipment as well as the methodological procedures and applications for genomics, transcriptomics, proteomics, and metabolomics. Likewise important as the availability of genome and postgenome methodologies is the management, processing and analysis of the data obtained. In parallel to the establishment of the genomics platform the

Table 1: Members of the Scientific Advisory Board of the CeBiTec

Prof. Dr. Rudolf Amann	Max Planck Institute for Marine Microbiology, Bremen, Germany, Director, Head of Department of Molecular Ecology
Dr. Rolf Apweiler	EMBL Outstation – Hinxton, European Bioinformatics Institute, Cambridge, United Kingdom, Head of Sequence Database Group
Prof. Dr. Michael Grunze	Heidelberg-University, Chair of Applied Physical Chemistry
Dr. Klaus Huthmacher	Evonik Degussa AG, Hanau, Leader F&E Degussa Feed Additives
Prof. Dr. Reinhard Krämer	Institute of Biochemistry, University of Cologne, Germany, Executive Director
Dr. Eduard Sailer	Miele & Cie. KG, Gütersloh, Germany, Executive Director
Prof. Dr. Martin Vingron	Max Planck Institute for Molecular Genetics, Berlin, Germany, Director, Head of Computational Molecular Biology Department
Prof. Dr. em. Christian Wandrey	Institute of Biotechnology 2, Forschungszentrum Jülich GmbH, Jülich, Germany

Figure 3: Executive committee of the CeBiTec, May 2008. From left to right: Prof. Dr. A. Pühler (chairman), S. Konermann, Dr. J. Kalinowski, Prof. Dr. J. Stoye, Prof. Dr. T. Noll, Prof. Dr. A. Gölzhäuser, Dr. S. Weidner (executive director), E. Schulte-Berndt, T. Wittkop, Prof. Dr. B. Weisshaar, D. Greif.

CeBiTec installed a comprehensive hardware and software infrastructure centralized in the 'Bioinformatics Resource Facility – BRF'. The BRF provides general hardware and software support for all research groups of the CeBiTec within genome and post genome projects. A high-performance compute cluster is available for large scale computations like whole genome annotations or metagenome analysis. Furthermore, a comprehensive software suite to systematically store and analyze all data sets from genomics, transcriptomics, proteomics, and metabolomics has been developed.

Further technology platforms of the CeBiTec are currently under development. Driven by bio- and nano-physicists a platform for a new generation of microscopes especially appropriated for the analysis of biomolecules is currently being planned. Furthermore, the available equipment and expertise in the fermentation of microorganisms and of animal cell culture are planned to be consolidated in a technology platform.

The CeBiTec organizes Symposia, Colloquia and Distinguished Lectures

The CeBiTec organizes various events to foster international contacts and collaborations and exchanges with other scientists. Regularly during the course of the year a colloquium provides a forum for lectures and discussion rounds with project and cooperation partners, and with other guest scientists. In 2009 speakers from Argentina, Canada, Germany, Sweden, and USA presented their topics.

A series of Distinguished Lectures was launched, where excellent and outstanding speakers present on current research to members of the CeBiTec and to the public. The series was opened by Prof. Dr. R. Amann, Director of the Department of Molecular Ecology at the Max Planck Institute for Marine Microbiology (Bremen, Germany) with the topic 'Genome-Enabled Earth System Studies'. In February 2010 Prof. Dr. M. Hecker, Head of the Division of Microbial Physiology and Molecular Biology, Ernst-Moritz-Arndt University of Greifswald (Germany) presented a lecture entitled 'From the blueprint of life to life – Physiological protemics of Gram-positive Bacteria'. Lectures for the upcoming semesters are under way.

Finally, a series of symposia about current research topics are organized annually. In June 2006 an international workshop on 'Molecular Systems Biology' took place at the Center for Interdisciplinary Research of Bielefeld University. This workshop brought together leading scientists from the field of molecular biology, mathematics, bioinformatics, chemistry and physics with the aim to foster the interdisciplinary collaboration and to establish the University of Bielefeld in the international systems biology community. The workshop was the starting point of a series of international CeBiTec symposia. A second CeBiTec symposium on 'The Future of Genome Research in the Light of Ultrafast Sequencing Technologies' took place in July 2007. It was dedicated to the delineation of concepts in the field of ultrafast sequencing methods and the the requirements on bioinformatics concerning the analysis of the emerging huge data sets. The third symposium on 'Solar Bio-Fuels' in August 2008 addressed the topics: general bioenergy aspects, biomass for bioenergy, and sun light to storable fuels. In August 2009 the series was continued with a symposium on Bioimaging, presenting interdisciplinary topics from the fields Biophysics, Systems Biology and Bioinformatics. In May 2010 the fifth Symposium entitled 'New Frontiers in Microbial Genome Research' will take place. ■

Table 2: Developmental steps concerning the CeBiTec

1998-09-25	Establishment of the CeBiTec through the Senate of the Bielefeld University
2000-09-14	Grant from the German Research Foundation (DFG) for the establishment of Institutes for Bioinformatics and Genome Research at Bielefeld University
2000-10-01	Grant from the German Research Foundation (DFG) for the establishment of a Research Training Group Bioinformatics
2001-06-21	Grant from the Ministry of Education and Research (MSWF) of the State North Rhine Westphalia for the establishment of an International Graduate School in Bioinformatics and Genome Research
2001-08-15	Grant from the Federal Ministry of Education and Research (BMBF) for the Competence Network Genome Research on Bacteria relevant for Agriculture, Environment and Biotechnology – GenoMik
2002-12-05	Inauguration of the Institute for Bioinformatics
2003-02-13	Inauguration of the Institute for Genome Research
2004-04-22	Inauguration of the Bielefeld Institute for Biophysics and Nanoscience (BINAS)
2004-10-19	Laying of the cornerstone for the CeBiTec building at the Bielefeld University
2005-10-20	Topping-out ceremony of the first part of the CeBiTec building
2005-11-18	Constitutive meeting of the Scientific Advisory Board at the Bielefeld University
2006-06-06	1st CeBiTec-Symposium Molecular Systems Biology
2006-07-24	Grant from the Federal Ministry of Education and Research (BMBF) for the Competence Network Genome Research on Bacteria relevant for Agriculture, Environment and Biotechnology – GenoMik-Plus
2007-01-30	Inauguration of the Institute of Biochemistry and Bioengineering (BioChemTech)
2007-02-28	Official inauguration of the CeBiTec building
2007-07-04	2nd CeBiTec Symposium The Future of Genome Research in the Light of Ultrafast Sequencing Technologies
2008-04-28	5. General meeting of the CeBiTec and election of the current Executive Board
2008-08-12	3rd CeBiTec Symposium Solar Bio-Fuels
2009-03-01	Completion of the second part of the CeBiTec building
2009-08-25	4th CeBiTec Symposium bioIMAGING

The DFG Bioinformatics Initiative at the CeBiTec

In 2000, Bielefeld University was granted one of six 'DFG Bioinformatik-Zentren' (BIZ-CeBiTec), funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG) Bioinformatics Initiative. This program, running up to 2007, has substantially shaped the early years of the CeBiTec. This contribution gives a short overview of the activities and the impact of the DFG Bioinformatics Initiative as implemented at the CeBiTec.

History, structure and funding of BIZ-CeBiTec

The starting point
When the DFG Bioinformatics Initiative was announced in 1999, the Bielefeld situation was as follows:

1. Contacts between the groups of A. Pühler and R. Giegerich had led to a series of seminars, lectures, diploma theses, and to students graduating in bioinformatics as a track of the diploma curriculum *Naturwissenschaftliche Informatik*.
2. The university's 'Center on Structure Formation', chaired by A. Dress, had already built up a reputation in biomathematics.
3. The Bielefeld 'Center for Biotechnology' (CeBiTec) had been established by the university senate – without staff and budget – as a forum of interdisciplinary discourse and strategic planning.
4. A group of CeBiTec researchers from biology, computer science and mathematics had written a proposal for a DFG-Research Training Group (*Graduiertenkolleg*) in Bioinformatics.
5. The university had announced an open full professor position in bioinformatics, to our knowledge the first one in Germany.

6. In refutation to an outside call, A. Pühler had acquired substantial funding to establish genome research at Bielefeld University.
7. When the arising bottleneck of bioinformatics expertise was discussed in the life sciences community in the late 1990s, thanks to reports in Nature, Bielefeld had already gained reputation as the first university providing dedicated bioinformatics education.

The DFG Initiative Bioinformatics had resulted from the experience of this very same bottleneck in Germany. Quite unusual for a DFG movement, it called for bioinformatics research closely integrated with educational efforts.

The CeBiTec Proposal

The DFG Initiative allowed us to connect all these ongoing endeavors. The goals of the BIZ-CeBiTec as spelled out in the original proposal of 2000 and the intermediate proposal of 2003, were such:

> 'The general goal of CeBiTec is to foster interdisciplinary research in the Life Sciences. Being a University site exclusively, without external research institutes, Bielefeld takes a special responsibility to anticipate changes of research paradigms and to react early by providing innovative educational and scientific programs. With respect to the scientific alliance of genome research and computer science, CeBiTec strives to provide both concrete bioinformatics support to ongoing research projects in genomics, and fundamental algorithms and tools applicable in the worldwide bioinformatics community.'

Our proposal was to create the two institutes of 'Genome Research' and 'Bioinformatics' under the – already existing – roof of the CeBiTec, and complement the already existing educational bioinformatics track by full B.Sc. and M.Sc. curricula in 'Bioinformatics and Genome Research'. Two new C4-groups were to be created, together with several junior research groups, computational resources and technical staff. Figure 1 depicts the structure of the 'Bioinformatik-Zentrum CeBiTec' (BIZ-CeBiTec) as seen in the intermediate proposal in 2003.
All these goals were implemented without deviations, and in retrospect, without major complications.

Prof. Dr. Robert Giegerich

Robert Giegerich studied Computer Science and Mathematics at the TU Munich and at Stanford University, USA. He received his PhD from TU Munich in 1981 with a thesis in the area of compiler generation systems. In 1989, he acquired a chair for Practical Computer Science in the newly founded Faculty of Technology at Bielefeld University. R. Giegerich created the Bielefeld Curriculum *Naturwissenschaftliche Informatik*, and starting 1992, he shifted his research interests to bioinformatics. Prof. Giegerich's most recent research work has centered around algorithmic methods to explore the relation between sequence, structure and function of RNA.

Implementing BIZ-CeBiTec

Recruiting in the times of the DFG Bioinformatics Initiative was extremely difficult. Centers in Berlin, Bielefeld, Leipzig, Munich, and Tübingen simultaneously announced their openings, competing for a rather small number of qualified candidates. The situation was further aggravated by a subsequent BMBF initiative with similar goals and dimensions. Even our open C4 Bioinformatics professorship, although announced one year ahead, got entangled in the arising, nation-wide mesh of negotiations. Eventually, however, we were quite successful in our recruiting, bringing Ralf Hofestädt, Jens Stoye, Bernd Weisshaar, and Ellen Baake to Bielefeld, and many more – see Figure 2.
In supporting the national establishment of the new discipline, the CeBiTec researchers created the German Bioinformatics Summerschool. First taught at the Bielefeld ZiF in 2000, it went

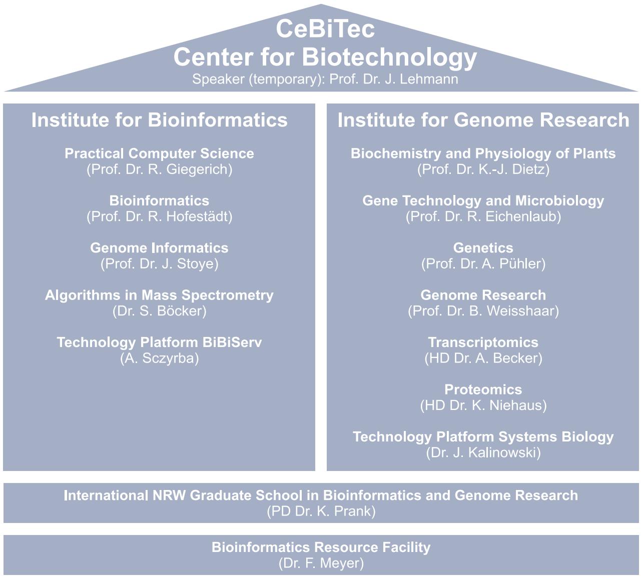

CeBiTec
Center for Biotechnology
Speaker (temporary): Prof. Dr. J. Lehmann
Institute for Bioinformatics
Practical Computer Science
(Prof. Dr. R. Giegerich)
Bioinformatics
(Prof. Dr. R. Hofestädt)
Genome Informatics
(Prof. Dr. J. Stoye)
Algorithms in Mass Spectrometry
(Dr. S. Böcker)
Technology Platform BiBiServ
(A. Sczyrba)
Institute for Genome Research
Biochemistry and Physiology of Plants
(Prof. Dr. K.-J. Dietz)
Gene Technology and Microbiology
(Prof. Dr. R. Eichenlaub)
Genetics
(Prof. Dr. A. Pühler)
Genome Research
(Prof. Dr. B. Weisshaar)
Transcriptomics
(HD Dr. A. Becker)
Proteomics
(HD Dr. K. Niehaus)
Technology Platform Systems Biology
(Dr. J. Kalinowski)
International NRW Graduate School in Bioinformatics and Genome Research
(PD Dr. K. Prank)
Bioinformatics Resource Facility
(Dr. F. Meyer)

Figure 1: Structure of the CeBiTec around 2003.

on for two further issues in Göttingen and Tübingen in subsequent years. Bielefeld was also the site of the German Bioinformatics Conference in 2004.

Indirect impacts of the DFG funding

Once awarded the funding for both the BIZ-CeBiTec and the DFG-*Graduiertenkolleg* 'Bioinformatics', we were able to attract substantial additional resources, which allowed us to go beyond our original goals in manifold respects. It was agreed from the beginning that Bielefeld University would permanently establish the infrastructure and curricula created within the BIZ-CeBiTec project. This provided a solid basis for medium-term planning and recruiting.

The major additional resources were the following:

1. The university acquired funds of 19 Mio. Euro to erect a new building – and has actually built it. Officially named *Laborgebäude*, it is better known as the CeBiTec.
2. The NRW state government provided 1 Mio. DM for enhancement of the computational infrastructure for bioinformatics education.
3. Furthermore, we were awarded one of the six 'International NRW Graduate Schools' with an annual budget of 1 Mio. Euro.
4. In program calls for genome research by the federal government, Bielefeld won large-scale projects of about 18 Mio. Euro, and in addition to this, the Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*, BMBF) supported the CeBiTec as a Bioinformatics Service Provider with approximately 2.8 Mio. Euro.

Research topics within BIZ-CeBiTec

Having said so much about funding, let us sketch some of the main research topics of the CeBiTec groups in the years of the DFG Initiative – leaving it to the reader to judge if all the money was well spent.

Practical Computer Science – Prof. Dr. R. Giegerich

While the original denomination of the group relates to pure computer science topics such as programming languages and compilers, our bioinformatics activities started in the mid-nineties. Originally, a prime concern was the introduction of index structures for large-scale sequence matching, such as suffix trees and suffix arrays. These techniques are commonly used today, and still an area of active research in other groups, both in Bielefeld and worldwide. A new focus was placed on RNA, where several tools were developed that today mark the state of the art in RNA bioinformatics, such as *pknotsRG*, *RNAforester*, *RNAshapes*, *RNAhybrid*, *RNAcast*, *GUUGle*, *Locomotif*, and more.

The other major line of work has been the development of the algebraic dynamic programming technique. It makes dynamic programming fun, which used to be a tedious and error prone task. Gradually, our algebraic technique is entering bioinformatics education in other places.

Genome Informatics – Prof. Dr. J. Stoye

The group was newly established in March 2002 as an immediate consequence of the funding by the DFG Initiative. The main focus of the group was on sequence analysis at a genomic scale, where rapid DNA database search techniques employing suffix trees, suffix arrays, and similar index structures were developed and tested.

Another research focus of the group were algorithms and computer programs for the comparison of genomes at a higher level, where a genome is modeled as a linear (or circular) order of its genes. In a series of publications written in collaboration with Anne Bergeron (Montreal), simpler methods for distance computation and rearrangement studies among genomes were developed. This resulted in corrected and faster algorithms and, last not least, improved teaching material.

The Stoye group was joined by three (temporary) junior research groups, funded from different sources:

- Informatics for Mass Spectrometry in Genomics and Proteomics (2003–2006)
  Head: Dr. Sebastian Böcker
  Funding: DFG *Aktionsplan Informatik*
- Computational Methods for Emerging Technologies (2004–2007)
  Head: Dr. Sven Rahmann
  Funding: Bielefeld University
- Combinatorial Search Algorithms in Bioinformatics (2005–2008)

Head: Dr. Ferdinando Cicalese
Funding: Sofja Kovalevskaja Award, Alexander-von-Humboldt-Stiftung

Bioinformatics and Medical Informatics – Prof. Dr. R. Hofestädt

The Bioinformatics/Medical Informatics group headed by Prof. Dr. Hofestädt was established in 2001 at the Faculty of Technology. The research concentrates on biomedical data management, modelling and simulation of metabolic processes, parallel computing and multimedia implementation of virtual scenarios. Within the scope of these topics, different national and international co-operations and research projects have been created and successfully finished (RAMEDIS, CELLECT).

The group has worked on the development of integrative methods for the modelling and simulation of metabolic processes. For example, they developed a complex information system which represents three levels: tools for the data integration of molecular databases, tools for the automatic implementation of user-specific databases and a rule-based method for the simulation of metabolic processes.

Biomathematics and Theoretical Bioinformatics – Prof. Dr. E. Baake

The group's main area of research is the mathematical theory of biological evolution, in particular, population genetics. Here, the main research results concern:

- a variational approach to mutation-selection models: For a large class of such models, the balance between mutation and selection could be proved to result from a competition for a maximal long-term growth rate, as given by the difference between the current mean reproduction rate, and a long-term loss related to the mutation process.
- a stochastic approach to recombination: The group investigates the Moran model with recombination. Surprisingly, in the case of single crossovers, the type frequencies follow the well-known solution of the corresponding deterministic model in expectation – quite an unusual behavior.
- statistical recognition: T cells recognise foreign antigens against a selfbackground. The group has started to investigate the mathematical principle behind the foreign-self distinction, and can, so far, summarize it as 'T-cells use large deviations to recognize foreign antigens'.

Bioinformatics of Regulation – Dr. M. Rehmsmeier

The groups' interests are RNA Bioinformatics and Regulatory DNA Elements. In RNA Bioinformatics, the focus is on the analysis of microRNAs (miRNAs) and their targets. The group has developed a program for miRNA target prediction, RNAhybrid, which considers binding energies of miRNA/target duplexes, statistical significance of individual and multiple binding sites, and evolutionary conservation of miRNA/target relationships. Further contributions are in the field of RNA secondary structure analysis.

The focus in the area of Regulatory DNA Elements is the analysis of Polycomb/Trithorax Response Elements (PRE/TREs or PREs for short). PREs are epigenetic switch elements that maintain previously determined transcription states of their associated genes over many generations of cell divisions, thus establishing a memory of transcriptional history. An important result of this work is the prediction of PREs in *Drosophila melanogaster*.

Bielefeld Bioinformatics Server – Dr. A. Sczyrba

The BiBiServ group supports Internet-based collaborative research and education in bioinformatics. More than 30 software tools and various educational media are available online. These include tools from different areas such as RNA structure analysis, genome comparison, classical methods of sequence analysis, and PCR primer design. The BiBiServ makes tools developed at different CeBiTec groups available to the bioinformatics community. The group supports authors in integrating their tools into a stable server environment and designing both HTML-based interfaces as well as Webservices. Reliable service and user support are provided, beyond the point where the author of the tool may have left Bielefeld University.

The lack of persistence and usability of bioinformatics software is often lamented on in the community. The BiBiServ model, created within BIZ-CeBiTec and continuing today, shows how a mid-size research institute like the CeBiTec can organize itself to overcome this problem.

Genetics – Prof. Dr. A. Pühler

The research carried out by the Chair of Genetics concentrated on genomics and postgenomics of prokaryotes. This is demonstrated by a series of larger research projects which were performed in recent years. In this respect, it is of special interest that the Chair of Genetics was selected to become the coordinator of a BMBF research network entitled 'Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology'. The structure of the network is composed of more than 20 groups from universities, research institutes and companies. As a consequence, a technology platform for genome sequencing, microarray technology, proteomics and metabolomics could be established. In addition, a computer cluster could be installed which supports the storage and analysis of the huge amount of data produced by the different 'omics' technologies.

Genome Research – Prof. Dr. B. Weisshaar

Building on expertise and long standing interests in the analysis of transcription factor networks and functional genomics, several projects integrating molecular biology, high-throughput technology and applied bioinformatics were newly launched at Bielefeld University.

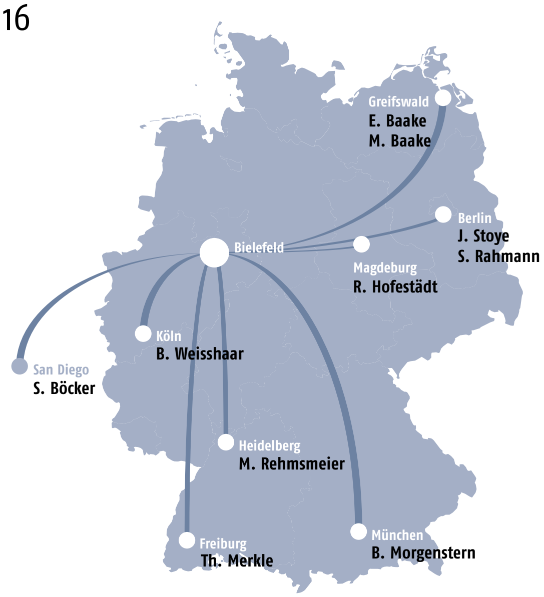


Figure 2: Recruiting for BIZ-CeBiTec

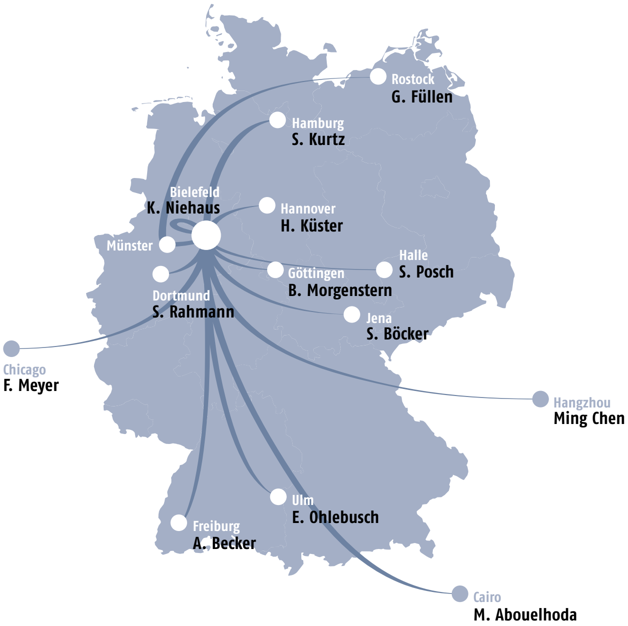


Figure 3: Recruiting away from BIZ-CeBiTec

- The project GABI-Kat aimed at generating about 90,000 T-DNA transformed lines with sequence-indexed insertion sites.
- A platform for TILLING (Targeting Induced Local Lesions IN Genomes) was established at the CeBiTec. TILLING is a 'reverse genetics' approach relying on the detection of small mismatches in dsDNA. The focus is on setting up a web-based LIMS (lab information management system) specifically adopted to TILLING for documentation of TILLING results.
- The Weisshaar group is also involved in generating a physical, BAC-based map of the sugar-beet (*Beta vulgaris*) genome. This map is of central strategic importance for marker assisted breeding, for straight-forward positional cloning of genes, and for the integration of molecular resources that have already been generated by sugar-beet breeders.

Proteome and Metabolome Research – Prof. Dr. K. Niehaus

The group of K. Niehaus started out as a junior research group during BIZ-CeBiTec. The group, today permanently established, developed a strong focus on interdisciplinary approaches in the field of bacterial metabolism and molecular plant microbe interactions.

The group was responsible for the setup of a proteome and metabolome unit that operates two MALDI-TOF-MS (Bruker biflex III, ultraflex) and two GC-MS instruments. A third, state of the art technique, allowing to couple two dimensional GC (GCxGC) to a time of flight MS instrument was installed.

Transcriptomics – Prof. Dr. A. Becker

The junior research group was involved in establishing transcriptomics approaches and tools for 9 prokaryotic organisms and supported transcriptome studies in the model legume plant *Medicago truncatula*, in pea and in poplar.

Moreover, a portal for the symbiotic soil bacterium *Sinorhizobium meliloti* was started. It gives access to *S. meliloti* genome, transcriptome and mutant data and is used by an international consortium comprising 8 external research groups that collaborate on regular updates of genome information. The group of A. Becker has extensively collaborated with the Bioinformatics Resource Facility of the CeBiTec and the groups of Robert Giegerich and Tim Nattkemper (Faculty of Technology). This contributed to platform tools like *GenDB*, *OligoDesigner*, *ArrayLIMS*, and *EMMA*.

RNA-Based Regulations – Dr. Th. Merkle

The junior research group RNA-based Regulation established a pipeline for in silico prediction of novel miRNA targets in *Arabidopsis thaliana*. In co-operation with Marc Rehmsmeier, the program *RNAhybrid* was used as starting point for in silico prediction. Based on the analysis of miRNA-mRNA duplex structures of validated miRNA targets, additional criteria were defined and applied to refine the number of predictions and to reduce the signal-to-noise ratio considerably. 296 targets were predicted, about half of them were novel. Experimental validation of selected novel targets was performed, with special emphasis on MYB transcription factors.

The Bioinformatics Resource Facility – Dr. A. Goesmann

There was one demand that had not been anticipated, resulting the rapid expansion of bioinformatics and genome research beyond the DFG funded activities. A central bioinformatics resource facility became quickly inevitable. It was created based on a series of, sometimes dramatic, searches for funding. Headed initially by F. Meyer, it is today directed by A. Goesmann. The group has not only built up a strong international reputation for dependable service and reliable software, it also takes a major load in the education of our students in the form of programming projects, B.Sc. and M.Sc. theses.

Long-term impacts of BIZ-CeBiTec

How to measure the long-term impact of multi-facette activities such as BIZ-CeBiTec? While the permanent research groups involved are still engaged in the two CeBiTec institutes today (and the CeBiTec has grown to accommodate four Institutes), many others – junior researchers, PhD and M.Sc. graduates, have left Bielefeld to do good research elsewhere. Let us conclude our little historical view with backpage of recruiting, as successful recruitment is only one side of the medal. Figure 3 gives an account of researchers who left Bielefeld to continue their career in other places.

The DFG Bioinformatics Initiative ended with an evaluation meeting in Berlin in Autumn 2007. BIZ-CeBiTec is ready for the next decade. ■

International NRW Graduate School in Bioinformatics and Genome Research

International NRW Graduate School in Bioinformatics and Genome Research

Bioinformatics and Genome Research at Bielefeld University have a long tradition in education and research. In 1989 one of the first curricula for Bioinformatics world-wide was established with the study program Applied Informatics in the Natural Sciences and has been the basis for many Master of Sciences programs. To educate PhD students in this new field of research, the 'International NRW Graduate School in Bioinformatics and Genome Research' was endowed at Bielefeld University in 2001 by a grant from the Ministry of Innovation, Science, Research and Technology (Ministerium für Innovation, Wissenschaft, Forschung und Technologie, MIWFT) of the state of North Rhine-Westphalia (NRW).

Altogether there exist seven Graduate Schools in NRW funded by the MIWFT located in Bielefeld, Bochum, Dortmund, Essen, Köln, Münster and Paderborn. The NRW Graduate Schools serve as a place for interdisciplinary collaboration of distinct faculties. These schools are internationally oriented and deliver an important contribution to the internationalization of the universities. Specially conceived post-graduate study programs build the basis of the structured PhD education, which should be completed within three years. During this time, the PhD students get full scholarships to center all their effort in their PhD studies.

In accordance to this the International NRW Graduate School in Bioinformatics and Genome Research, located as a cross-section department at the Center for Biotechnology (CeBiTec), is ideally placed to set up PhD studies in interdisciplinary fields. With respect to the international orientation of this program, the Graduate School acquired from 2004 to 2007 additional funds from the IPP PHD program by the German Academic Exchange Service (DAAD). The DFG Research and Training Group GK 635 Bioinformatics were integrated into the Graduate School in 2003.

The Graduate School accomodates a faculty formed from distinguished researchers in the fields of computer science, biology, mathematics, chemistry, and physics. In addition to this an international faculty of seven renowned scientists from excellent international research institutions of the USA, Canada and China was established. They visit Bielefeld University on a regular basis to give lectures, conduct workshops on cutting edge science, and advise the PhD students in their projects. The Graduate School is represented by the two Speakers: Prof. Dr. Robert Giegerich and Prof. Dr. Alfred Pühler. The Office of the Graduate School is managed by a Director, Dr. Susanne Schneiker-Bekel, and technical staff.

PhD Education in Bioinformatics and Genome Research

This interdisciplinary PhD program combines the areas of bioinformatics and experimental genome research, allowing students to acquire a PhD in either biology, computer science, chemistry or biotechnology. The biological focus of the program lies in the fields of genomics, transcriptomics, RNomics, proteomics, metabolomics. High throughput experimental techniques in genome research present manifold challenges for algorithmic data analysis and data management. For the experimental work in this field and the analysis of the experimental data, knowledge and expertise in molecular biology, biomathematics, algorithmics, modeling and simulation has to be combined.

The three-year PhD project is defined by the faculty upon acceptance to the Graduate School. The PhD student is supported by two supervisors. A project plan for the first year is laid out by the supervisors and the PhD student, and is formally approved by the faculty. In addition to this, highest standards of PhD education are ensured through yearly written progress reports, scientific retreats, control of the educational study program, soft skill trainings and language courses. Funds for travelling and publications foster the international visibility of the research work. The following paragraphs will give you a detailed description of the above mentioned instruments of structured PhD education:

Instruments of structured PhD Education

Progress reports have to be handed in by the PhD students every year to continue their fellowships. They consist of two parts, a project report and a scientific essay. The project report contains the list of lasts years project plan with comments on what was achieved, what failed, and where the plan was changed together with a plan specifying the project details for the next year. The scientific essay consists of a summary of the scientific results obtained in the PhD project so far. It may be replaced by a publication or technical report.

Dr. Susanne Schneiker-Bekel

Susanne Schneiker-Bekel studied horticultural sciences at the Leibniz Universität Hannover. After completing her PhD in Genetics at Bielefeld university in 2001, she worked as a scientist at the sequencing department of Qiagen. In 2003 Susanne Schneiker-Bekel went back to Bielefeld University and worked in the field of bacterial genome sequencing at the competence centre and since 2004 in the NRW Graduate School in Bioinformatics and Genome Research, first as a mentor and then as a managing director. Currently, she is heading the Graduate Center of the CeBiTec.

Twice a year, the students and the faculty of the Graduate School spend two days at a conference center or hotel near Bielefeld. Students give presentations of their work after the first and the second project year. This a good opportunity for everyone to get an overview of all scientific activities at the Graduate School and to discuss findings and progress of projects. Last but not least, social activities in the evening allow students and faculty to form personal relationships.

In addition to the scientific research work, which can be done in the wet-lab and/or on the computer, the PhD students have to complete the educational part of the PhD program consisting of activities in seven different areas: Advanced courses in bioinformatics and genome research; working group seminars;

organisation and participation in workshops; assistance in giving lectures and courses; publications on conferences or in journals; support of students during their theses (bachelor, master or diploma degree); additional qualifications and soft skills.

Additional qualifications

The Graduate School offers soft skills training courses on demand by inviting professional lecturers. In 2008, we invited international lecturers for 'Academic Writing in English' and 'Presentations on International Conferences'. Moreover, the University offers very useful soft skills courses on various levels and topics such as 'Time Management' or 'Application Training', open to all students.

Although the language of the Graduate School is English and knowledge of German is not required for the enrollment in this PhD program, we encourage our foreign students to take part in language courses in German offered by the University's PunktUm program. Tailored to the students' profiles, we provide funding for English or German language courses.

The Graduate School welcomes the willingness of PhD students to present their results at scientific conferences. The costs to travel to international scientific conferences or the costs of publishing in international scientific journals given the approval of the supervisors are funded. Moreover, research stays abroad were financed.

Special Events of the Graduate School

In 2002 the BREW series of workshops was initiated to give young researchers an introduction to scientific conferences, including submission, peer review and presentation of scientific papers. The BREW workshop is jointly organised by some of the most distinguished european PhD programmes in bioinformatics: ComBi Programme (University of Helsinki, Finland) EBI PhD Programme (European Bioinformatics Institute, England), Molecular and Computational Biology Research School (University of Bergen, Norway), International Graduate School in Bioinformatics and Genome Research (University of Bielefeld), Max Planck Research School for Computational Biology and Scientific Computing (Berlin, Germany). Special aims of the BREW are not only to introduce PhD students at an early stage in their PhD work with the work modes of international conferences (Table 1). In addition to this, the BREWs bring together PhD students and experienced researchers in an atmosphere of cooperation and inspiration, and help to establish research contacts across Europe, which could be utilized in the student's subsequent research.

Since 2004 each of the seven International NRW Graduate Schools award prizes of 1,500 Euros to young scientist for outstanding publications in their respective fields of research (Table 1). The awards in the respective fields of research are thought to foster the early development of research experience and practice during undergraduate studies. Applicants were authors or co-authors of outstanding publications in high-ranking scientific journals or proceedings of peer reviewed international conferences. The awardees of the International NRW Graduate School in Bioinformatics and Genome Research were Benjamin Schuster-Böckler from the Freie Universität Berlin (2004), Martin Bader from the Universities of Magdeburg and Hannover (2005) and Nikolaos Sgourakis from the University of Athens (2006/2007). The winner of the Young Scientist Award in 2009 was Martin Simonsen from the University of Aarhus.

Table 1: Special Events of the CeBiTec NRW Graduate School

Event	Organizer
BREW workshop April 2003 Bielefeld	DFG Research Training Group 'Bioinformatics'
Undergraduate Science Award April 2005 Academy of Sciences Düsseldorf	NRW Graduate Schools
Undergraduate Science Award May 2006 Academy of Sciences Düsseldorf	NRW Graduate Schools
Undergraduate Science Award May 2007 Academy of Sciences Düsseldorf	NRW Graduate Schools
BREW workshop April 2008 Bielefeld	DFG Research Training Group 'Bioinformatics'
Visit of the Dalai Lama Sept. 2008 Münster	NRW Graduate Schools
Young Scientist Award Nov. 2009 Academy of Sciences Düsseldorf	NRW Graduate Schools

Ongoing associations and future perspectives

Since 2001 there have been 92 students admitted to the study program of which a third are female. The students were funded by different sources: 58 with scholarships from the NRW Graduate School, 21 with scholarships from the DFG, 3 with scholarships from the DAAD and the rest from other sources. About one third of the students of the Graduate School comes from foreign countries all over the world: Asia, South America, European countries, GUS, USA and New Zealand. Until now, 47 students have successfully finished their PhD education. About one third of the NRW

Figure 1: Students and faculty of the Graduate Cluster Industrial Biotechnology

Graduate School's alumni works in German academia in renowned universities and Max-Planck Institutes, another third works in distinguished German companies, the last third has left Germany and works in industry and academia of different European countries, the Switzerland, the USA and Canada. Two previous directors of the NRW Graduate School, Dr. K. Prank and Dr. D. Evers, have acquired highly ranked positions in industry and one of our Junior Group leaders, Dr. B. Morgenstern, has become Professor for Bioinformatics at Göttingen University.

To steady and stabilize structured PhD education of interdisciplinary natural sciences the CeBiTec is building up a Graduate Center. The CeBiTec Graduate Center should coordinate and administrate the structured doctoral programs of all institutions of the CeBiTec. At present, there exist three structured doctoral programs which are jointly coordinated by the Graduate Center: the newly established international graduate program Bioinformatics of Signaling Networks, the 'CLIB Graduate Cluster Industrial Biotechnology' with research focus on Polyomics, and the International NRW Graduate School in Bioinformatics and Genome Research and associated students of the 'DFG Research Training Group Bioinformatics' which take part in the study program of the Graduate School. ■

International Graduate Program
Bioinformatics of Signaling Networks
CLIB
Graduate Cluster
Industrial Biotechnology
International Graduate School
Bioinformatics & Genome Research

Figure 2: Logos of the three current doctoral programs coordinated by the Graduate Center

Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology

History

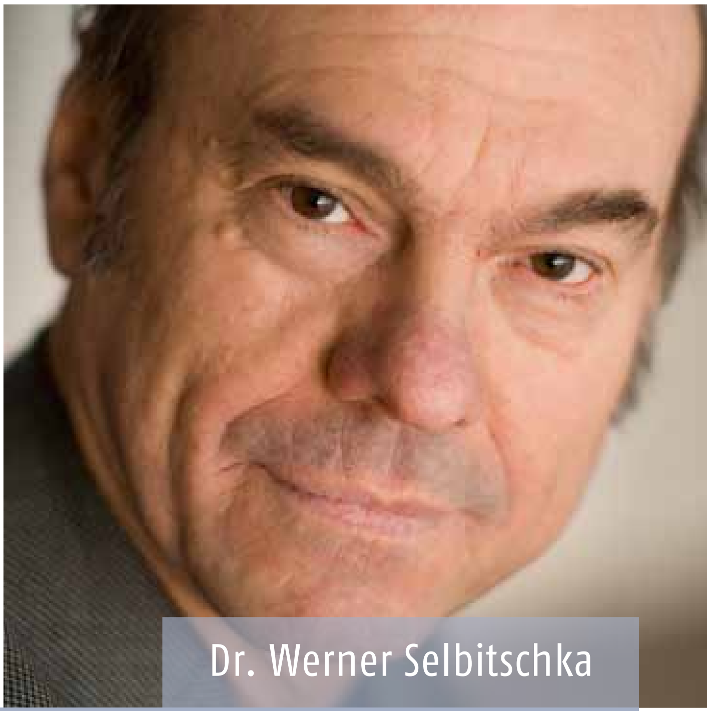
As early as 1996, several renowned German scientists prepared a memorandum entitled 'Research and Funding Initiative: Integrated Genome Research on Bacteria'. The authors emphasized that bacterial genome research in Germany was largely neglected in the past and that there was an urgent need for financial support of this research area. The memorandum influenced two consecutive papers, a paper of the German Research Foundation (DFG) entitled 'Perspectives of Genome Research' as well as a paper of the Federal Ministry of Education and Research (BMBF) designated 'Genome Research in Germany – Current Status and Future Perspectives'. The memorandum as well as both papers contributed strongly to the publication of the BMBF call for proposals 'Genome Research on Microorganisms – GenoMik' in October 2000. In the framework of the GenoMik program, three German-wide competence networks were selected for funding. Würzburg University was chosen as the coordinating centre of the competence network 'PathoGenoMik' which focussed on genome research on human pathogens. Göttingen University was chosen as the coordinating unit of the BiotechGenoMik network which concentrated on genome research on bacteria for the analysis of biodiversity and their use for production processes. Finally, Bielefeld University was awarded as the coordinating centre of the competence network 'Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology'. Thus, initiation of bacterial genome research in a more systematic way occurred in Germany just in time. This can be concluded from the enormous increase of bacterial genome projects finished from the year 1996 until today (Figure 1).

Structure of the Bielefeld network

The Bielefeld competence network was launched in June 2001 and is working in its overall structure to date, albeit with some minor modifications. During the GenoMik program (2001–2006) the Bielefeld network pooled the excellence of 21 research groups, and thereafter, during the GenoMik-Plus program (2006–2009) of 26 German research groups which are affiliated with universities, research institutes and companies. Figure 2 shows the German-wide geographic distribution of the partners of the current Bielefeld GenoMik-Plus network.

The network is co-ordinated by a competence centre which is located at the Center for Biotechnology (CeBiTec) of Bielefeld University. The competence centre consists of two core facilities, the network management and a technical facility designated technology node. The network management executes the scientific and administrative co-ordination of the network whereas the technology node provides the essential techniques for genome research to its network partners, i.e. bioinformatics, transcriptomics as well as proteomics. The Bielefeld network consists of three research clusters, namely 'Plant-Associated Bacteria', 'Primary Metabolite Producers' and 'Secondary Metabolite Producers', respectively. Moreover, the Bielefeld network center houses the Technology Platform for Microbial Genome research (TPMG)-Bioinformatics.

The TPMG-Bioinformatics provides services for all German GenoMik-Plus partners by offering its hardware infrastructure as well as the in house developed open-source software modules for bacterial genome annotation (GenDB) and for the interpretation of transcriptomics data (EMMA), respectively. The overall scientific structure of the Bielefeld network funded under the GenoMik-Plus guideline is shown in Figure 3.


Dr. Werner Selbitschka

Werner Selbitschka studied Biology at the Julius-Maximilians Universität Würzburg and the Friedrich-Alexander Universität Erlangen-Nürnberg. After completing his PhD in Microbiology, in 1988 he moved to the Department of Genetics at Universität Bielefeld and since then works in the area of microbial ecology, initially on biological safety in the frame of the deliberate release of genetically modified organisms (GMO). He is currently heading the competence centre of the national network 'Functional genome research on bacteria relevant for agriculture, environment and biotechnology' which is located at Bielefeld University.

Scientific focus of the Bielefeld network

Since its start in 2001, genome as well as postgenome analyses of bacterial species possessing outstanding metabolic capabilities for use in agricultural, environmental and biotechnological applications are in the focus of the network's research. During the GenoMik program, the genome sequences of six bacterial isolates were established. For the area 'Agriculture', the genome sequence of the nitrogen-fixing *Azoarcus* sp. was decoded. *Azoarcus* sp. is of considerable agricultural relevance since this bacterium exerts a growth stimulating effect on rice plants. Moreover, the genomes of the phytopathogenic bacteria *Xanthomonas campestris* pathovars campestris and vesicatoria as well as *Clavibacter michiganensis* subsp. michiganensis which cause great economic losses in agriculture worldwide were sequenced. Within the area 'Environment' the genome sequence of the marine bacterium *Alcanivorax borkumensis* was established. *A. borkumensis* is able to degrade oil and thus, may be useful for the cleansing of oil-polluted environments. In the area 'Biotechnology' the genome of the myxobacterium *Sorangium cellulosum* was deciphered. *S. cellulosum* is well known for its ability for secondary metabolite production. Based on the sequence information, genome-wide microarrays were constructed for all of these bacterial strains. Last but not least, selected cosmids of the *Streptomycetes* were sequenced which contain biosynthetic gene clusters encoding secondary metabolite production. Further postgenome analyses in the areas 'Agriculture' and 'Biotechnology' addressed the symbiotic nitrogen fixing soil bacteria *Sinorhizobium meliloti* and *Bradyrhizobium japonicum* as well as the amino acid producer *Corynebacterium glutamicum*, respectively. For these organisms genome-wide microarrays were established to study gene expression under specific conditions. Some of the scientific results achieved by the Bielefeld network partners during the initial

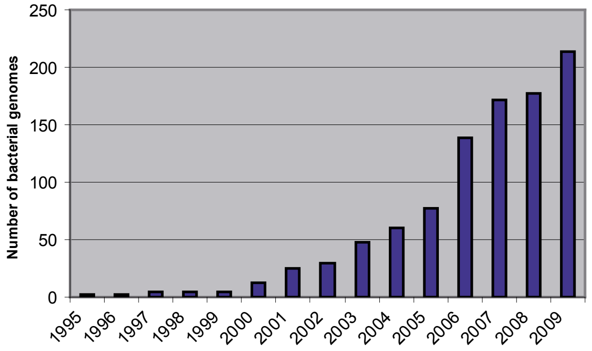


Figure 1: Survey of the number of finished bacterial genome projects from 1995 to 2009

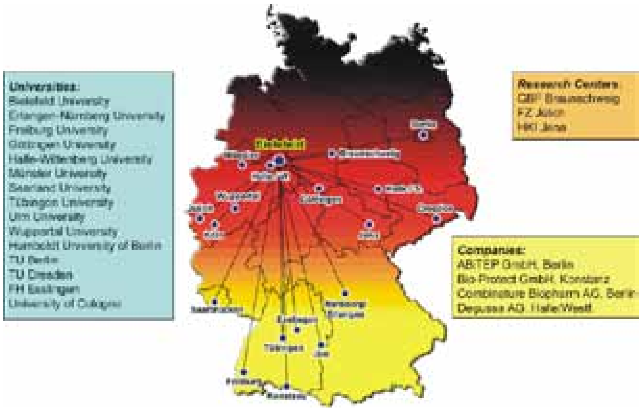


Figure 2: Geographic distribution of partners of the German-wide network

stage of the GenoMik program were published in a special issue of the Journal of Biotechnology (2003, Vol. 106, Issues 2–3; eds. A. Pühler and W. Selbitschka).

Based on the scientific progress achieved during the GenoMik funding phase, i.e. the establishment of the genome sequences and the subsequent construction of genome-wide microarrays, the focus of the network's research changed to functional studies with the start of the GenoMik-Plus program. This included the use of the -omics technologies in the comparative analyses of wild type and mutant strains. Moreover, the Bielefeld network underwent some structural changes during the transition from the GenoMik to the GenoMik-Plus program. A new project focussing on *Bacillus amyloliquefaciens* was included in the network's research. *B. amyloliquefaciens* is of biotechnological relevance due to the synthesis of enzymes which can be used in biotechnological production processes. Table 1 gives an overview on the various genome and postgenome projects of the Bielefeld network. Several of the scientific results achieved by the Bielefeld network partners during the GenoMik-Plus program were published in a second special issue of the Journal of Biotechnology (2009, Vol. 140, Issues 1–2; eds. A. Pühler and W. Selbitschka). Below, some selected highlights of the network's research will be briefly presented.

Selected highlights of genome research on bacteria relevant for agriculture and environment

Two projects representative for the areas 'Agriculture' and Environment' are briefly described. *Clavibacter michiganensis* subsp. michiganensis is a plant-pathogenic actinomycete that causes bacterial wilt and canker of tomato (Figure 4A). Under EU legislation, *C. michiganensis* subsp. michiganensis is classified as a quarantine organism. The *C. michiganensis* subsp. michiganensis genome project aimes at contributing to the development of diagnostic tools and, in the long-run, providing clues for the breeding of tolerant or resistant tomato varieties. The bioinformatics interpretation of the *C. michiganensis* subsp. michiganensis circular, 3.298 Mb genome sequence showed no similarities to genes of Gram-negative phytopathogenic organisms known to be involved in the host plant infection process. This result indicates that the infection process of the Gram-positive model organism is completely different from that of Gram-negative bacterial phytopathogens. The genome sequence also provided some clues for the poor survival of *C. michiganensis* subsp. michiganensis in the soil environment. The apparent lack of a sulfate reduction pathway could well account for this observartion. As a consequence, *C. michiganensis* subsp. michiganensis is dependent on reduced sulfur compounds for growth.

Alcanivorax borkumensis is a marine bacterium that uses oil hydrocarbons as single source of carbon and energy (Figure 4B). The bioinformatics interpretation of the *A. borkumensis* SK2 genome data revealed that *A. borkumensis* is perfectly adapted to the challenges of its habitat. *A. borkumensis* has a multitude of genes mediating a wide hydrocarbon substrate range and efficient oil-degradation capabilities. Furthermore, the genome specifies a numer of systems relevant for oil-degradation such as biofilm formation at the oil-water interface or biosurfactant production. The availability of the *A. borkumensis* genome sequence provides a sound basis for the future design of bioremediation strategies to fight oil pollution in the marine environment.

Selected highlights of genome research on bacteria relevant for biotechnology

Based to their biotechnological relevance, two projects of the area 'Biotechnology' addressing primary and secondary metabolite producing bacteria will be briefly discussed. *Corynebacterium glutamicum* is a Gram-positive soil bacterium well known for its ability to produce amino acids such as L-lysin (Figure 4C). L-lysine is an essential amino acid which must be obtained by humans or

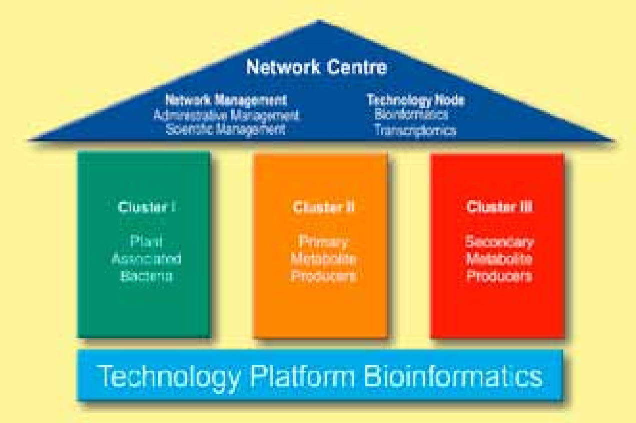

Network Centre
Network Management
Administrative Management
Scientific Management
Technology Node
Bioinformatics
Transcriptomics
Cluster I
Plant Associated Bacteria
Cluster II
Primary Metabolite Producers
Cluster III
Secondary Metabolite Producers
Technology Platform Bioinformatics

Figure 3: Scientific structure of the network during the GenoMik-Plus funding phase (2006–2009)

animals from external sources such as food or feed. Consequently, this amino acid is used as a feed additive for lifestock production in order to prevent yield losses. Genome-wide transcriptomics analyses were performed to elucidate the complex regulatory network of sulfonate utilization. These analyses led to the identification of the repressor McbR and the activator SsuR, which negatively and positively regulate the corresponding genes.

Bacteria of the species *S. cellulosum* belong to the myxobacteria, a group of soil bacteria which show gliding motility and undergo complex morphogenetic changes resulting in fruiting body

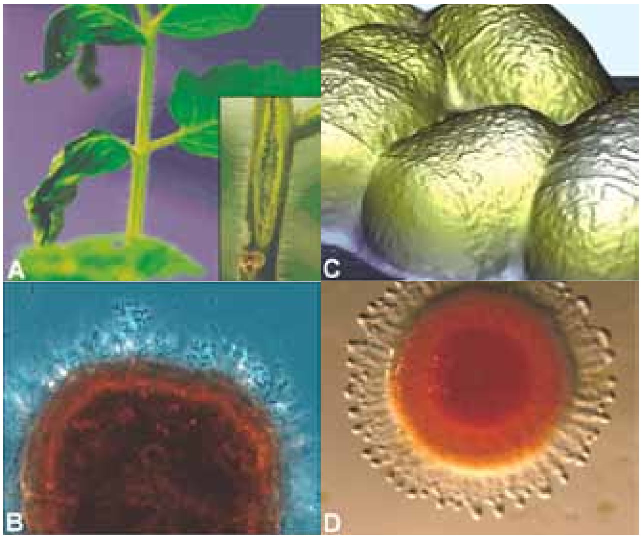


Figure 4: Phenotypic characteristics of selected bacterial strains of agricultural, environmental and biotechnolgical relevance. Disease symptoms on leaves and the shoot of a tomato plant infected with the phyto-pathogen *C. michiganensis* subsp. michiganensis (A). Micrograph of an oil-droplet surrounded by *A. borkumensis* cells (B). Atomic-force microscopic image of whole cells of the amino acid producer *C. glutamicum* (C). Colony morphology of the myxobacterium *Sorangium cellulosum* strain Soce 56 (D).

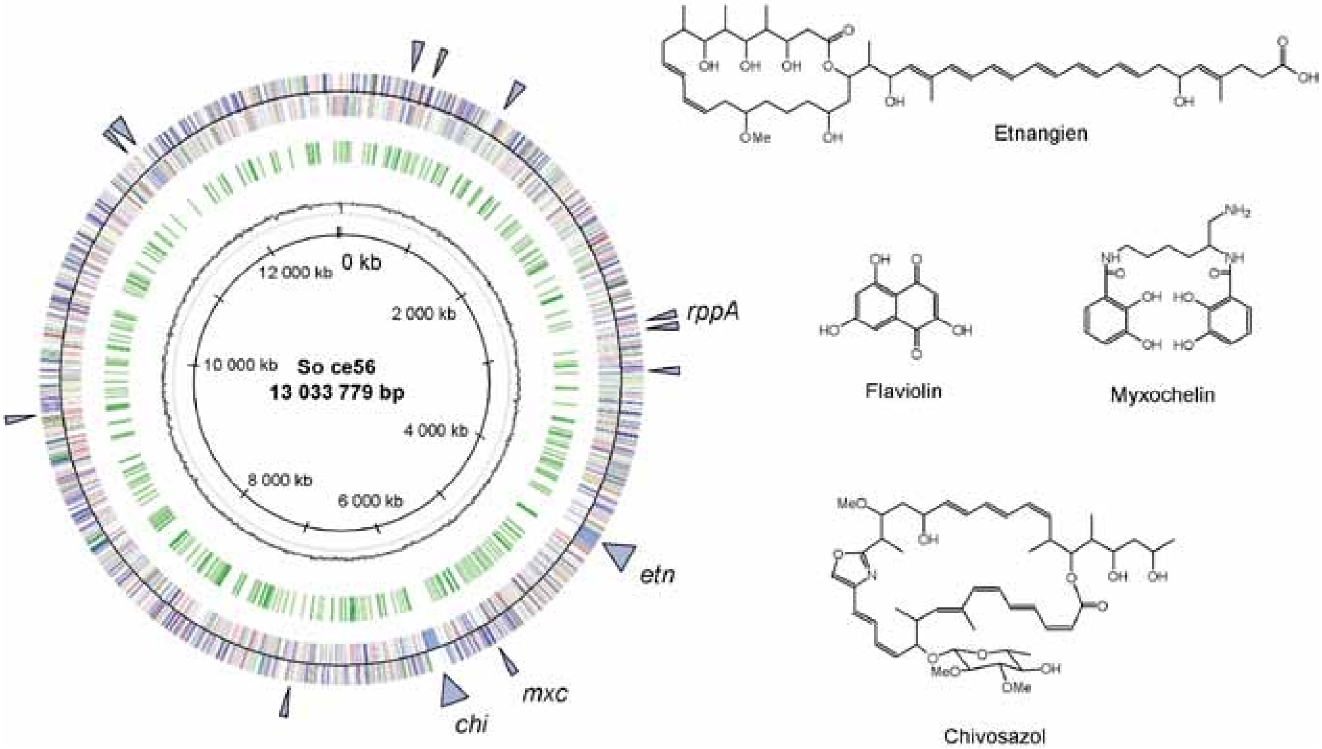


Figure 5: Circular map of the *Sorangium cellulosum* Soce 56 genome. The locations of gene clusters for the biosynthesis of the four different secondary metabolites chivosazol (chi), myxochelin (mxc), etnangien (etn) as well as flaviolin (rppA) are indicated as well as the chemical structures of these natural products.

Table 1: Genome and post-genome projects of the Bielefeld network during the BMBF-funded GenoMik and GenoMik-Plus programs, respectively.

Bacterial Strain	Genome Size	Number of Putative Genes	Principal Investigator	References
Azoarcus sp.	4,37	3.992	B. Reinhold-Hurek (Bremen University)	Nat. Biotechnol. (2006) 24:1385
Alcanivorax borkumensis	3,12	2.755	K. Timmis (HZI Braunschweig)	Nat. Biotechnol. (2006) 24:997
Bacillus amyloliquefaciens	3,92	3.693	R. Borriss (HU Berlin)	Nat. Biotechnol. (2007) 25:1007
Clavibacter michiganensis subsp. michiganensis	3,40	3.080	R. Eichenlaub (Bielefeld University)	J. Bacteriol. (2008) 190:2138
Corynebacterium glutamicum	3,28	3.002	J. Kalinowski (Bielefeld University)	J. Biotechnol. (2003) 104:5
Sinorhizobium meliloti	6,69	6.204	A. Becker (Freiburg University)	Science (2001) 293:668
Sorangium cellulosum	13,03	9.367	R. Müller (Saarland University)	Nat. Biotechnol. (2007) 25:1281
Xanthomonas campestris pv. campestris	5,08	4.471	K. Niehaus (Bielefeld University)	J. Biotechnol. (2008) 134:33
Xanthomonas campestris pv. vesicatoria	5,42	4.726	U. Bonas (Halle-Wittenberg University)	J. Bacteriol. (2005) 187:7254

Table 2: Bacterial genome projects currently under way in the Bielefeld network.

Biotechnological Relevance	Bacterial Strain	Principal Investigator – Affiliation
Plant Symbionts	*Sinorhizobium fredii*	M. Göttfert – TU Dresden / A. Becker - Freiburg University
	Sinorhizobium meliloti 41	A. Becker – Freiburg University
Plant Pathogens	*Xanthomonas campestris* pv. vesicatoria 75-3	U. Bonas – Halle-Wittenberg University
	Xanthomonas campestris pv. translucens	K. Niehaus – Bielefeld University
	Clavibacter michiganensis subsp. nebraskensis	R. Eichenlaub – Bielefeld University
Primary Metabolite Producer	*Corynebacterium* spec. ATCC 21341 'C. glycinophilum'	J. Kalinowski – Bielefeld University
Producers of Antibiotics	*Actinoplanes friuliensis*	D. Schwartz/R. Biener – Esslingen Univ. of Applied Sciences
	Streptomyces collinus	W. Wohlleben – Tübingen University
	Saccharothrix espanensis	A. Bechthold – Freiburg University
Producers of Natural Products	*Sorangium cellulosum* So ce1525	R. Müller – Saarland University
	Sorangium cellulosum So ceGT47	R. Müller – Saarland University
	Sorangium cellulosum So ce38	R. Müller – Saarland University
	Bacillus amyloliquefaciens F	R. Borriss – Humboldt University of Berlin
Human Pathogens	*Pseudomonas aeruginosa* WS136	J. Heesemann – Max-von-Pettenkofer Institute, Munich
	Pseudomonas aeruginosa WS394	J. Heesemann – Max-von-Pettenkofer Insttitute, Munich
	Pseudomonas aeruginosa MH27	D. Jahn – TU Braunschweig
	Pseudomonas aeruginosa MH38	D. Jahn – TU Braunschweig
	Mycobacterium tuberculosis Beijing6	S. Kaufmann – Max-Planck-Insitute Berlin
	Mycobacterium tuberculosis 7199/99	S. Niemann – FZ Borstel
	Mycobacterium tuberculosis 4546/04	S. Niemann – FZ Borstel
	Neisseria meningitidis α522	M. Frosch – Würzburg University
	Neisseria meningitidis α704	M. Frosch – Würzburg University

formation. Figure 4D shows a colony of the myxobacterium *S. cellulosum* Soce56. Myxobacteria are known for a long time as extremely talented producers of bioactive secondary metabolites and are therefore, the subject of intensive research. Genome analyses revealed that *S. cellulosum* Soce 56 has a circular genome of 13,03 megabases encoding nearly 10.000 genes (Figure 5). This is by far the largest genome of a bacterial organism known to date, i.e. the Soce 56 genome holds the world record in genome size. In this context it is worth noting, that the human genome contains some 25.000 annotated genes. This corresponds roughly to 2.5-fold more genetic information of humans compared to this soil bacterium. The bioinformatics interpretation of the sequence data revealed that the Soce 56 genome encodes several natural products such as Etnangien or Chivosazol, which are of potential commercial value (Figure 5). Parallel to the publication of the *S. cellulosum* Soce56 genome sequence, a bioactive secondary metabolite synthesized by strain Soce 90 of *S. cellulosum* was approved by the Food and Drug Administration (FDA) of the United States as an anticancer drug. The epothilone B derivat Ixabepilone was commercialized under the label IXEMPRA by Bristol-Myers Squibb in 2007. This example amply underpins the enormous biotechnological potential of the microorganisms under study in the Bielefeld network.

Ongoing and future projects

With the establishment of the state-of-the-art next generation sequencing technology, a new era of this research field has started at the CeBiTec of Bielefeld University. The use of the Genome Sequencer (GS)-FLX machine of Roche Applied Science which is in operation since 2008 greatly accelerates the *de novo* sequencing of bacterial genomes. In the framework of additional grant money allocated to the Bielefeld GenoMik-Plus network, the establishment of the genome sequences of more than 20 bacterial isolates with medical, agricultural and biotechnological relevance is currently under way, further strengthening this exciting research field at Bielefeld University (Table 2). In addition, metagenome analyses addressing the microbial communities of relevant habitats are performed. This includes analyses of the microbial communities of biogas plants as well as the rumen of cattle. ■

Insertion mutants from GABI-Kat
An invaluable tool in plant functional genomics research

Background

After completion of the genome sequence of the model plant *Arabidopsis thaliana* in the year 2000, research emphasis focused on the assignment of biological function to genes. The genome annotation initially predicted about 25,500 genes, but experimental evidence for a biological function was available for only about ten percent of the genes. These functions were deduced or at least confirmed in almost all cases by using mutants that were selected from phenotypic screens. This classical genetic approach is now referred to as 'forward genetics'. Obviously, the availability of a genome sequence reverses the experimental design. In reverse genetic approaches, the first goal is to identify a mutant for a given (predicted) gene, while classically a gene was 'cloned' from a locus defined by a mutant.

Since higher plants do not allow targeted gene replacement, reverse genetics in plants is based on large collections of random mutants from which the relevant line with a mutation in the gene of interest needs to be selected. The mutagen of choice should generate null alleles with a good frequency, and the mutations should be easily detectable in the genome. A mutagenesis system that meets these requirements makes use of a process that naturally occurs during infection of plants with *Agrobacterium tumefaciens*. This soil bacterium is able to transfer a part of its DNA, called T-DNA, into plant cells where it stably integrates at random position into the plant genome. The T-DNA can be modified to carry a selectable marker that allows efficient mutant recovery, its sequence is big enough to result in gene knock outs, and serves as a tag for the insertion locus. Due to these advantages a number of T-DNA insertion mutant collections were generated. The mutagenized populations were initially screened for insertion alleles of the gene of interest by using DNA pools and PCR. A big improvement was the generation of flanking sequence tags (FSTs) from all individual mutants of the populations. The FSTs contain sequences from the genome region immediately adjacent to the inserted T-DNA. For such sequence-indexed populations the FST data are stored in a database and users can apply BLAST or key word searches to easily select lines with knockout alleles according to their special interest.

Based on the assumption that T-DNA insertion into the *A. thaliana* genome is random, it has been estimated that 180,000 independent T-DNA lines are required to find a single knockout allele of a specific gene of 2.1 kb in length with 95% probability. Unfortunately, the selected insertion alleles are not randomly distributed in the genome, but cluster shortly before and after the transcribed region of genes. In addition, not all insertions are resulting in a null allele even if the insertion is located within the transcribed region. Quite some genes are smaller than the average value of 2.1 kb which results in significantly higher numbers of insertion mutants to 'hit' these genes with high probability. And finally, some of the insertions that are predicted to exist from FST data are not found in the offspring of the initially analysed mutagenized plant.

The GABI-Kat collection, which consists of about 90,000 T-DNA insertion mutants, provides roughly 65,000 lines with at least one FST that allows the prediction of a genomic location of the T-DNA. It is the largest European collection and the second largest worldwide. In total, there are about 300,000 insertion alleles predicted from FST data available to the arabidopsis scientific community.

History of GABI-Kat

GABI-Kat was initiated at the Max-Planck-Institute for Plant Breeding Research (MPIZ; Cologne, Germany) in the year 1999, before the *A. thaliana* genome sequence was finished. The project was and still is funded in the context of the German Plant Genome Research Program GABI (Genomanalyse im biologischen System Pflanze), which is a private-public partnership that is financed by the BMBF (German Federal Ministry for Education and Research) and the WPG (Business Platform promoting GABI Plant Genome Research e.V.). The initial PIs of GABI-Kat were Bernd Weisshaar, Koen Dekker, Bernd Reiss and Heinz Saedler. After the Weisshaar group moved from the MPIZ to Bielefeld University in 2003, the project was gradually transferred from MPIZ to the CeBiTec (Center for Biotechnology) of Bielefeld University. Since the beginning of 2007, GABI-Kat is running completely at the Institute for Genome Research and Systems Biology (IGS), which is one of the institutes of the CeBiTec.

The main goal of the GABI-Kat project is to provide sequence-indexed T-DNA insertion mutants of *Arabidopsis thaliana* to the scientific community. The project initially aimed at generating at least 70,000 T-DNA transformed lines for the sequenced accession Columbia (Col-0). During the extensions of the project, the number of mutants was increased to more than 90,000. After transformation by *A. tumefaciens*, the first generation of transgenic plants (called T1) was identified due to their resistance to the antibiotic sulfadiazine provided by the T-DNA insertion. Genomic DNA from T1 plants was extracted, and seeds of single lines (T2 seeds from T1 plants) were harvested. FST production was based on the T1 DNA and followed a linker-ligation-PCR approach to access genomic DNA sequence directly adjacent to the individual insertion sites. The FST data was then mapped to the genome and the results were made available via the Internet (*http://www.gabi-kat.de*, see screenshot in Figure 4). Insertion lines can be ordered via a web form and seeds (usually T2) of confirmed insertion lines are then delivered to users.

Within the first phase of GABI-Kat, selection, growth and leaf harvest for DNA extraction from more than 90,000 single T-DNA-transformed lines was completed until February 2006. Up to now, FST-sequences have been generated for 77,612 lines of which 65,376 match the arabidopsis genome. Upon user request

Prof. Dr. Bernd Weisshaar

Bernd Weisshaar studied biology and biochemistry at the University of Cologne. After completing his PhD in Animal Virology in 1988, he worked initially as postdoctoral researcher and since 1991 as a group leader at the Max-Planck-Institute of Plant Breeding (MPIZ) in Cologne. The research topic was regulation of secondary metabolite accumulation in plants. Between 1995 and 2003, he established in addition to running a research group focussing on transcriptional control phenylpropanoid biosynthesis and large transcription factor gene families the DNA core facility at MPIZ. Since 2003 he is professor for genome research at the Faculty of Biology at Bielefeld University.

Figure 1: Catalogued GABI-Kat seed collection of T-DNA mutagenized *A. thaliana* plants. Seeds are packed in paper bags and these are subsequently stored in cabinets at a temperature of 9°C and a relative humidity of 20–25 %.

the insertions in these lines are confirmed by PCR and sequence analysis of an amplicon that spans the insertion site. The template DNA for this confirmation PCR is prepared from the offspring generation (T2); this confirmation is performed at GABI-Kat prior to delivery. In a second phase of GABI-Kat, which in fact started in 2005, T3 seed that are harvested from individual T2 plants are donated to a biological resource centre serving the scientific community (see below). In addition, great efforts are made to further improve the quality of the mutant population (Figure 1) and the database, and to continuously serve the arabidopsis community by providing confirmed insertion lines (Figure 2).

A problem that becomes increasingly relevant is the decreasing germination rate of the roughly ten-year-old T1 seeds in the GABI-Kat collection. Therefore, propagation to the T2 generation and production of T3 seeds is obligatory to preserve the lines and the insertion mutants.

Resistant seedlings are grown in the green house descending from T2 seeds, which usually segregate for the T-DNA and therefore also for the sulfadiazine resistance in this generation. The putative insertion is confirmed by sequencing the PCR product obtained from the T2 DNA using a T-DNA- and a gene-specific primer. T2 seeds of successfully confirmed lines are sent out to the user. The confirmation process is repeated two times if necessary. If this does not yield success, the respective insertion is given up and the user is notified. Overall, the confirmation rate in GABI-Kat is 80%.

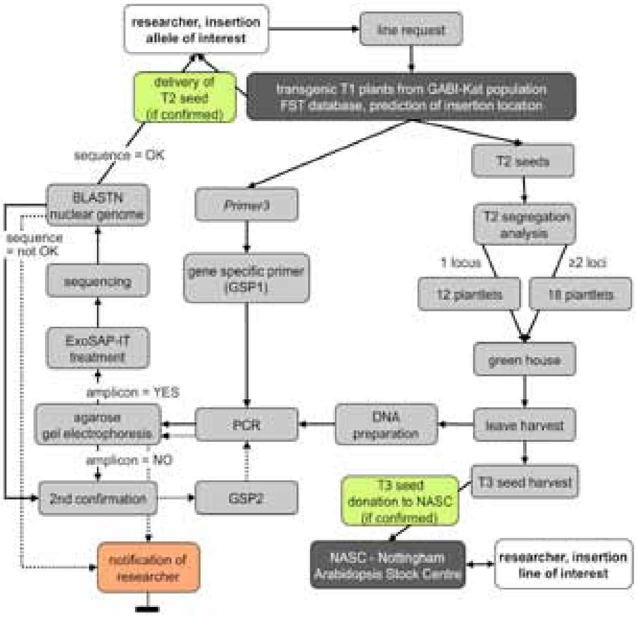

researcher, insertion allele of interest
line request
delivery of T2 seed (if confirmed)
transgenic T1 plants from GABI-Kat population FST database, prediction of insertion location
sequence = OK
T2 seeds
BLASTN nuclear genome
Primer3
T2 segregation analysis
sequence = not OK
1 locus
≥2 loci
sequencing
gene specific primer (GSP1)
12 plantlets
18 plantlets
ExoSAP-IT treatment
green house
amplicon = YES
agarose gel electrophoresis
PCR
DNA preparation
leave harvest
amplicon = NO
T3 seed harvest
2nd confirmation
GSP2
T3 seed donation to NASC (if confirmed)
notification of researcher
NASC - Nottingham Arabidopsis Stock Centre
researcher, insertion line of interest

Figure 2: Flow chart of workflow in the confirmation process after a line request has been placed.

New Tasks for GABI-Kat in Bielefeld

The main focus of the recent work at GABI-Kat is the donation of confirmed lines to the 'Nottingham Arabidopsis Stock Centre' (NASC). After harvesting the T3 seeds of lines selected for confirmation by users the complete T3 set is donated to NASC. In addition, an important goal of GABI-Kat is the confirmation of all lines that contain insertions in genes that are not hit in other populations. These lines are identified through the evaluation of the predicted gene hits in relation to other sequence-indexed insertion mutant populations (in Col-0). Moreover, it is often desired to confirm mutant phenotypes from two independent insertions. Therefore '2nd alleles' are identified as well. The selected GABI-Kat lines, which are designated 'GK only lines', enter the confirmation process with exactly the same protocol as lines requested by users. Upon confirmation, sets of T3 seeds with separate seed lots from individual T2 plants are donated to the NASC. Subsequently, researchers can order the sets of T3 seeds of confirmed GABI-Kat insertion lines at the NASC directly. The access to T3 seeds from a segregating family provides the advantage that homozygous seeds for the desired insertion are included with a high probability, which can save the time of a plant generation. Until May 2009, user requests for 5,998 insertions were responded positively and a total of 8,684 insertions were confirmed by PCR. T3 sets of 6808 lines have been donated to the NASC.

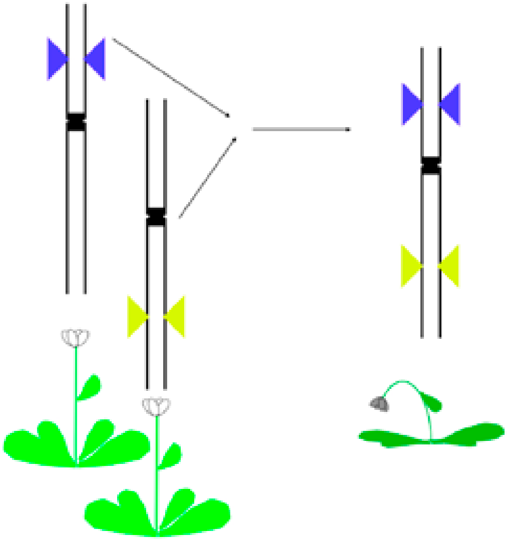


Figure 3: Basic principle of GABI-DUPLO.
Homozygous double mutants for paralogous genes are generated by crossing homozygous single mutants (T-DNA insertion lines) and genotyping the F2 generation. Phenotypical effects, which may not be visible in the homozygous single mutants, might be revealed in the double mutants.

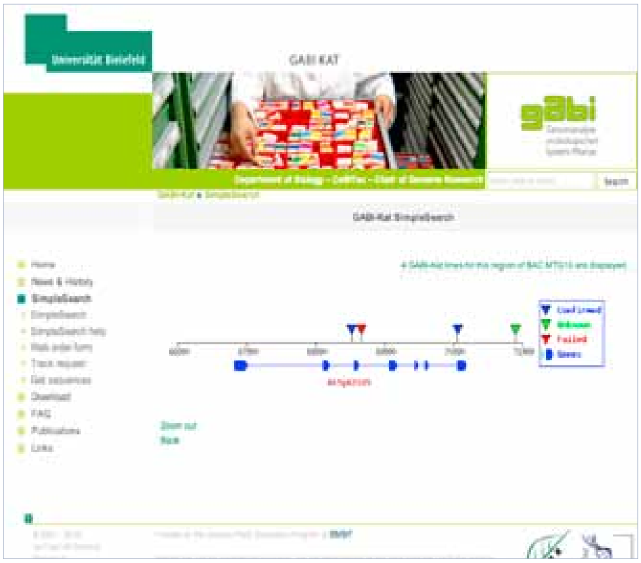


Figure 4: Screen shot from the GABI-Kat website. The picture from the SimpleSeach database shows the gene structure of the *A. thaliana* gene At5g62165, and the positions of four GABI-Kat T-DNA insertions relative to this gene.

Perspective

Besides the usefulness of GABI-Kat lines for individual researchers, the collection is used as the main resource for a new 'large resource project', which is part of the GABI-FUTURE program of BMBF: GABI-DUPLO (Figure 3). Because a significant proportion of the genome of *A. thaliana* consists of paralogous genes that resulted from segmental genome duplications, the analysis of such genes is impeded since a single gene knock out might not lead to phenotypical effects. The project GABI-DUPLO aims to provide double homozygous lines for insertions in unlinked duplicated genes. GABI-DUPLO is a joint effort of different German groups. Besides the GABI-Kat group, the Helmholtz Zentrum München (HZM; Klaus Mayer) and the Ludwig-Maximilians-Universität München (LMU; Dario Leister, Cordelia Bolle) participate. The Bielefeld-part is to provide homozygous single mutants of GABI-Kat lines and also SALK lines (which are T-DNA insertion lines in Col-0 background from the population of Joe Ecker located at the SALK Insitute in the US). Moreover, the workflow database support for the coordination of the project is provided by the Bielefeld group. The prediction of paralogy is performed by the group of Klaus Mayer and yielded so far a list of 865 candidate gene pairs. Crossing of the homozygous single mutants and identification of the double mutants is performed at the LMU, which received 196 lines corresponding to 101 gene pairs until May 2009. The double mutants generated in the project will be made available to the scientific community by donation to NASC.

New applications for the GABI-Kat collection (like DUPLO) as well as the constantly high number of user requests for GABI-Kat lines at GABI-Kat and at NASC prove the usefulness of and the need for the collection, even ten years after the beginning of the project. ■

Table 1: Some statistics: Numbers concerning NASC donations and other seed deliveries.

Total Number of Lines donated to NASC	6,808
Number of T3 seed bags donated to NASC	96,294
Total number of insertions delivered to individual users	5,998
Number of insertions confirmed for 'GABI-Kat only' or phenotypic reasons	2,686

CeBiTec Contributions to the Cluster of Industrial Biotechnology CLIB^{2021}

General features of the CLIB^{2021} project

The starting point for CLIB^{2021} was the BioIndustry2021 competition of the Federal Ministry of Education and Research (BMBF) in 2007. This competition was to figure out the best cluster concept for establishing industrial biotechnology. The proposal developed in North Rhine-Westphalia designated CLIB^{2021} got the highest award which comprises graduation funds of 20 million Euros from the BMBF for cooperative projects in research and development between industrial companies including start-ups and academia. The focus of CLIB^{2021} concerns the utilization of renewable resources and the development of novel materials and active substances for all markets and areas of life. One principal goal of CLIB^{2021} is to contribute to climate protection with production processes that reduce greenhouse gases. CLIB^{2021} will give decisive impulses for innovation in many fields such as the chemical industry, but also household applications and medicine. CLIB^{2021} is composed of three sections, namely projects in research and development organized by industrial companies, as well as a Technology Platform and a Graduate Cluster both run by universities located at Bielefeld, Dortmund and Düsseldorf. In this article, the contributions of the CeBiTec to the Technology Platform and the Graduate Cluster of CLIB^{2021} will be presented.

Research at the Technology Platform PolyOmics at the CeBiTec

An important section of CLIB^{2021} concerns the Technology Platform which concentrates on PolyOmics, Biocatalysis, Expression and Downstream Processing. The Technology Platform located at Bielefeld University is financed by the federal state NRW in the frame of the competition Bio.NRW. The PolyOmics part of the Platform is integrated into the CeBiTec at Bielefeld University and makes use of the sophisticated equipment in the fields of genomics, transcriptomics, proteomics, metabolomics and bioinformatics.

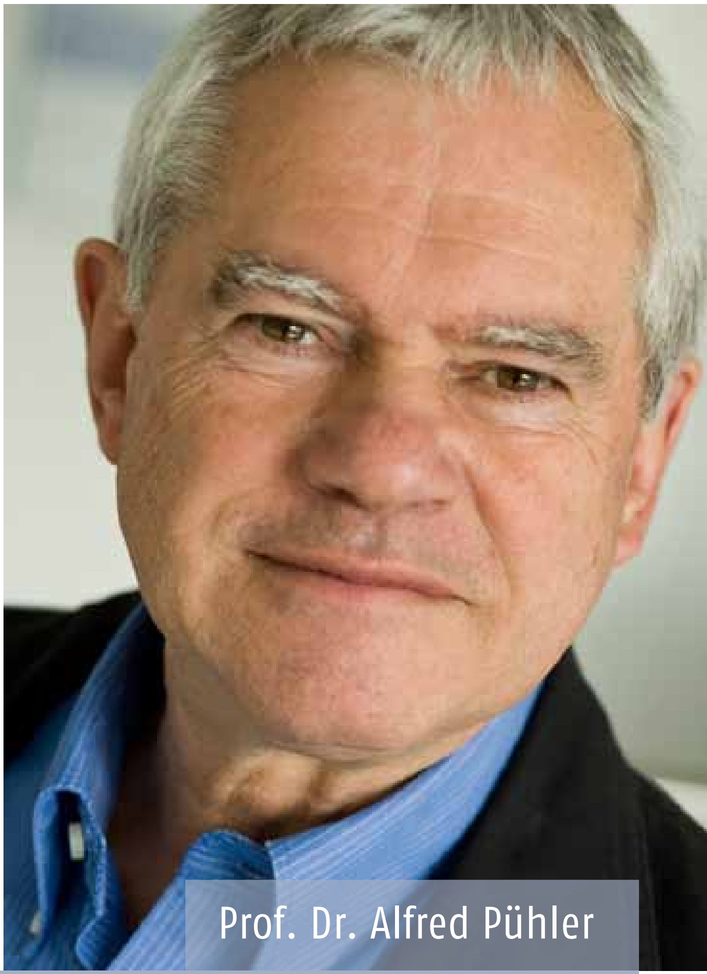
Goals of the PolyOmics Platform are the genome-based development of industrial production strains, as well as the optimization of biotechnological production processes. To reach these goals a team consisting of three scientists and two technicians was set up. This scientific team will be supported by PhD students of the CLIB Graduate Cluster 'Industrial Biotechnlogy'.

In the field of genome-based development of industrial production strains, specific research topics were selected. One topic concerns the improvement of ultrafast sequencing techniques applied for the establishment of prokaryotic and eukaryotic genome sequences. Furthermore, it is planned to analyze messenger RNAs by sequencing appropriate cDNA libraries. The sequenced cDNA libraries of eukaryotic microorganisms will play a role in the process of gene identification, since a certain percentage of eukaryotic genes have a mosaic structure and are therefore difficult to predict with bioinformatic tools. A further research topic that belongs to the field of synthetic biology, deals with the development of techniques reducing the genome size of microorganisms important for biotechnology. Those tools can be used to remove insertion elements and prophage genomes which are known to contribute to genome instability. In addition, it is planned to develop regulatory systems for the targeted initiation or termination of gene transcription. These systems should be easy to handle under fermentation conditions.

The second goal of the PolyOmics Technology Platform concerns the optimization of production processes by analyzing the corresponding production strains. One research project concentrates on the use of a new type of microarray which is produced by the synthesis of oligonucleotides directly on the chip. Another project deals with the role of small RNAs modulating gene expression in microorganisms relevant for industrial production. Such small RNAs can be studied by the above-mentioned microarrays, but also by ultrafast sequencing techniques. A further research project supports the metabolome analysis of production strains under fermentation conditions. In this respect it is planned to extend the existing metabolite database by analyzing metabolic mutants of selected microorganisms. Last but not least, an additional research project will contribute to the metabolomics field by developing new harvesting methods for cells which should guarantee that the measured concentrations of cellular metabolites reflect the actual situation.

Education in the CLIB Graduate Cluster 'Industrial Biotechnology'

To effectively enlarge the distinctive competencies of the CLIB^{2021} Cluster of Industrial Biotechnology and to advance the interdisciplinary work of the technology platforms, a Graduate Cluster 'Industrial Biotechnology' was founded and integrated into the



Prof. Dr. Alfred Pühler

Alfred Pühler studied Physics at the Friedrich-Alexander University Erlangen Nürnberg, got his PhD degree in Microbiology in 1971 and habilitated in 1976. In 1980, he became head of the Chair of Genetics at Bielefeld University. Since 2004, he is chairman of the Executive Board of the Center for Biotechnology (CeBiTec) of Bielefeld University. The CeBiTec is equipped with all omics technologies and also with a Bioinformatics Resource Facility. From 1999 to 2005, Alfred Pühler was member of the Science Council installed by the Federal President of Germany. He also is member of the North Rhine-Westphalian Academy of Sciences, of the German Academy of Sciences Leopoldina and of the German Academy of Science and Engineering. Since 2008, he works as a Foreign Secretary for the Union of the German Academies of Sciences and Humanities. His research interests are focused on genome research of industrially relevant microorganisms such as coryneform bacteria.

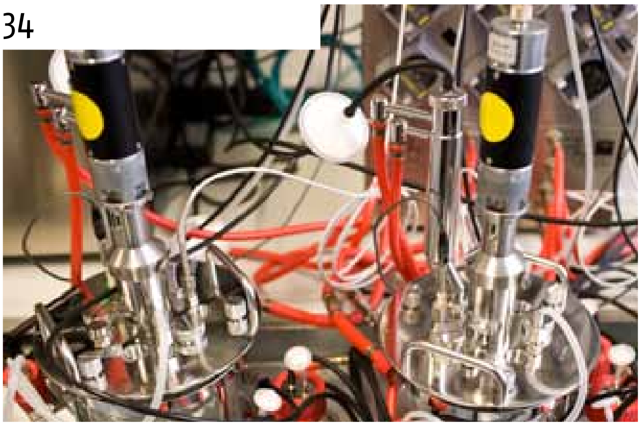


Figure 1: Bioreactor system for parallel operation. The fermenter system is capable of controlling temperature, pH, pO_2, agitation speed, gas composition etc. in different culture vessels independently.

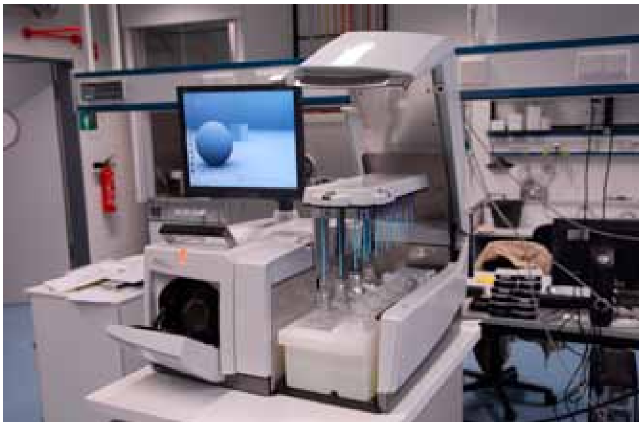


Figure 2: Roche Genome Sequencer FLX system. With different sequencing chemistries the system is suited for *de novo* sequencing of genomes and for sequencing studies of cDNA libraries.

CLIB2021 cluster. On April 1st 2009 the Graduate Cluster 'Industrial Biotechnology' started officially as a joined initiative of the three member universities: Bielefeld University, Heinrich Heine University Düsseldorf and TU Dortmund University. During the first three years the programme offers 84 PhD positions to excellent young scientists of which 28 are located at Bielefeld University. The North Rhine-Westphalian Innovation Ministry supports the 7.2 million Euro project with 4.1 million Euro. The remaining 3.1 million Euro are covered by the universities and by industrial companies organized in CLIB2021.

For the PhD programme highly-qualified students from the fields biology, bioinformatics, biotechnology, chemistry, process engineering or medicine are selected by means of international tendering and are invited to apply. The CLIB Graduate Cluster 'Industrial Biotechnology' represents a structured PhD programme that offers fast track PhD studies in English language. The programme is characterized by two key aspects that distinguish this programme from other well known graduate schools: First, the education of the scholars takes place in close agreement to research projects supported by CLIB2021. Second, the CLIB Graduate Cluster allows a joint education of the scholars across locations, supervised by the three universities and industrial companies. This structure enables a direct transfer of knowledge between university and industry. The competitive advantage of the Graduate Cluster 'Industrial Biotechnology' is therefore its industrial orientation and its close relationship to member companies of CLIB2021.

Thematically, the Graduate Cluster 'Industrial Biotechnology' is based on the four CLIB-associated technology platforms located at three universities: PolyOmics at Bielefeld University, Expression at Düsseldorf University, Biocatalysis at TU Dortmund University and Downstream Processing at TU Dortmund University.

At the CeBiTec in Bielefeld the scientific branch PolyOmics is realized. This newly designed term comprises the methods of genome and post-genome research and harbours projects dealing with research in the fields genomics, transcriptomics, proteomics, metabolomics and bioinformatics. In April 2009 the first four PhD students received their scholarships. In the meantime, altogether sixteen PhD students were accepted in the Bielefeld branch of the Graduate Cluster.

During their PhD studies, the participants get the opportunity of 3-months practical industrial training which will take place at a company (CLIB2021 member). The selection of the company is subject to the approval of the scientific supervisor and CLIB2021. The studies are planned for at least 36 months, during which 30 credits have to be collected by each participant. The lectures and practical training in all four thematic areas of the technology platforms are offered as 3-day blocks during the semester break. The students have to take part in at least one block, the topic of which has to be different from the subject of their own PhD project. Apart from scientific excellence, also key competences in the fields of soft skills and innovation management will be trained. Every year a retreat will take place in which students will have the opportunity to present their results. It is obvious that in addition to the supervision of the scientific work the CLIB Graduate Cluster 'Industrial Biotechnology' follows the intention to prepare the PhD students also for a future position in industry.

Future Contribution to the CLIB2021 project

The research at the Technology Platform 'PolyOmics' and the education in the Graduate Cluster 'Industrial Biotechnology' were both successfully started in the year 2009. For the following years, it is now of highest importance that the Technology Platform and the Graduate Cluster cooperate in a tight manner. This can be achieved by selecting PhD topics which are close to the research programme of the Technology Platform.

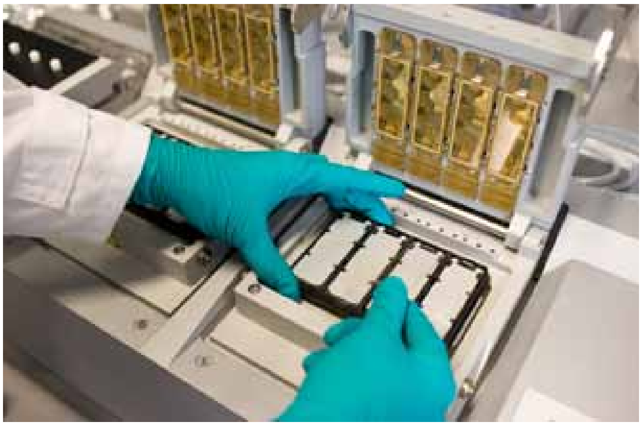


Figure 3: Microarray hybridization station for automated and reproducible microarray processing in transcriptomics analysis.

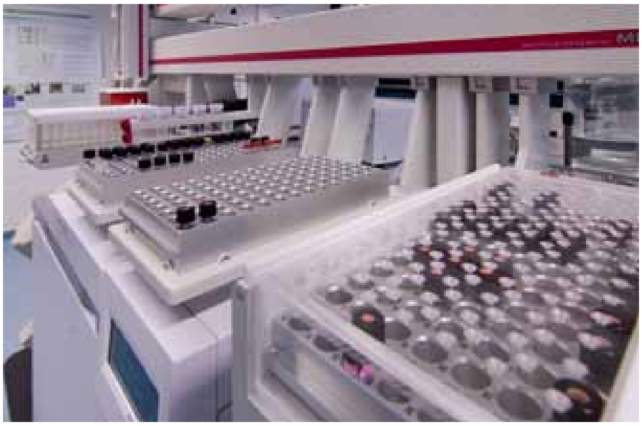


Figure 4: Two-dimensional gas chromatography with time-of-flight mass spectrometer (GCxGC-TOFMS) equipped with a multifunctional autosampler and sample preparation robot. The GCxGC-TOFMS system allows high resolution metabolomics analysis.

As mentioned in the first section describing the general features of CLIB^{2021}, one aspect concerns industrial research projects which are carried out between industry and academia. The CeBiTec also contributes to this aspect with a project entitled 'FerDi: Fermentative production of 1,3-dihydroxy-2-amino-octadecen (Sphingosine)'. Sphingosine and sphingosine-containing ceramides are special ingredients of cosmetics and are currently produced by Evonik Industries in a cost-intensive chemosynthesis process. However, the yeast strain *Pichia ciferrii* naturally secretes sphingolipids into the growth medium. The main task of the FerDi project therefore is to develop a cost-efficient fermentative process to produce sphingosine starting from biorenewables. In this context, knowledge of the whole-genome sequence of *P. ciferrii* is a basic step to investigate the sphingosine metabolism and to illuminate bottlenecks within the metabolic pathways. The CeBiTec is therefore engaged to determine the genomic sequence and to establish a pipeline for gene finding and annotation in eukaryotic organisms. In the meantime, this project has been successfully completed. We are now looking for further industrial partners running biotechnological research projects which can make use of the technologies provided by the CeBiTec Technology Platform 'PolyOmics'.

To summarize: It is obvious, that the CeBiTec represents an important partner in the CLIB^{2021} consortium. It carries out industrially guided research projects, runs the Technology Platform 'PolyOmics' and contributes to the Graduate Cluster 'Industrial Biotechnology'. ■

CLIB
Graduate Cluster
Industrial Biotechnology
tu technische universität dortmund
Universität Bielefeld
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF
CLIB
2 0 2 1
CLUSTER INDUSTRIELLE BIOTECHNOLOGIE

Figure 5: The CLIB^{2021} Graduate Cluster. The logos of the CLIB^{2021} Graduate Cluster and participating institutions are presented.

Genome Informatics Research Group

Research in the Genome Informatics group spans a broad spectrum of activities, starting from the low level of DNA sequence comparison and going up to the higher levels of comparative genomics, metagenomics and phylogenetics.

The primary material analyzed by the methods of genome informatics are genomic sequences, i.e. textual representations of DNA molecules that encode an organism's function. Thanks to modern technologies, large-scale DNA sequencing is becoming easier and cheaper, and for many organisms even the whole genomes are nowadays completely sequenced. Moreover, in the recently emerged field of metagenomics, the output of a sequencing project does not any more originate from a single organism, but rather from a large microbial community that is studied in its natural environment.

However, acquiring the genomic sequences is only the first step towards getting deeper biological insights. The next challenge is to interpret these sequences, i.e. locate their interesting regions, extract the higher-level information encoded in them and, based on this, compare different organisms. Moreover, such an analysis must be done in quantitative terms, and this poses the need for sound mathematical models, efficient algorithms and user-friendly software.

Efficient comparison of DNA sequences

The primary data sources in genome analysis are DNA sequences which, from a formal viewpoint, are strings written in the four-letter DNA alphabet {A,C,G,T}. A fundamental task is then the comparison of two such strings, with the aim of finding regions that match to each other exactly or approximately. This task arises naturally when searching in large sequence databases, in order to retrieve sequences that are similar to a query sequence that we want to characterize. The constantly increasing size of

Dr. Epameinondas Fritzilas and Prof. Dr. Jens Stoye

modern datasets raises the need for time-efficient search algorithms that detect exact or approximate matches. A common approach in this direction is indexing of the sequence database, i.e. offline preprocessing and storage in an appropriate data structure, which then allows query search to be performed very quickly. In the Genome Informatics group we have focused on data structures, such as suffix trees, suffix arrays and q-gram indices, that are specialized for full-text indexing and fast string matching.

In particular, we have developed a practical algorithm for the construction of suffix arrays, that is less complex than other construction algorithms and fast for all kinds of strings (*http://bibiserv.techfak.uni-bielefeld.de/bpr*). On the theoretical side, we have also investigated questions related to the combinatorial properties of suffix arrays. In particular, for fixed alphabet size and string length, we counted the number of strings sharing the same suffix array and the number of such suffix arrays. These results have applications to succinct suffix arrays and also build the foundation for the efficient generation of appropriate datasets for testing suffix array-based algorithms.

We have also developed *Swift*, a tool for fast local alignment (*http://bibiserv.techfak.uni-bielefeld.de/swift*). Contrary to other popular alignment heuristics, *Swift* guarantees that all alignments matching its query parameters will be found. More specifically, it guarantees to find all epsilon-matches, where an epsilon-match is a local alignment longer than a given threshold, with an error ratio of at most epsilon. The algorithm behind *Swift* uses a q-gram index and a sliding window to quickly and efficiently identify parallelograms in the implied dynamic programming matrix with which epsilon-matches may overlap. In computational experiments, we found out that *Swift* can in general attain sensitivity levels similar to that of Blast, while being more than 25 times faster. Currently, we are using the *Swift* algorithm also in the context of genome assembly, in order to quickly map the contigs that are produced in a genome project onto one or more reference genomes.

Epameinondas Fritzilas received a diploma in electrical and computer engineering from the National Technical University of Athens in 2003 and an M.Sc. degree in bioinformatics from the University of Athens in 2005. He completed his PhD studies at Bielefeld University in 2009 with a scholarship from the International Graduate School in Bioinformatics and Genome Research.

Jens Stoye studied applied computer science in the natural sciences at Bielefeld University, where he received the PhD degree in 1997 on a topic related to multiple sequence alignment. After postdoctoral positions at the University of California at Davis and the German Cancer Research Center in Heidelberg, he became head of the Algorithmic Bioinformatics Group at the Max Planck Institute of Molecular Genetics in Berlin. Since 2002, Dr. Stoye has been a professor of genome informatics at Bielefeld University.

Detection of repeated regions

Genomes often contain several repetitive regions that do not encode any proteins. However, these repetitive elements are interesting to study for two reasons. Firstly, they may interfere with other protein-encoding regions, thereby resulting in the development of genetic disorders. Secondly, they may be a source of difficulties in sequencing projects, because they make it harder to design specific PCR primers and assemble the short sequence reads.

We have investigated the problem of identifying families of repetitive sequences based on data from partially sequenced genomes. This approach uses the data that are already available in the databases and does not require the sequencing of the whole

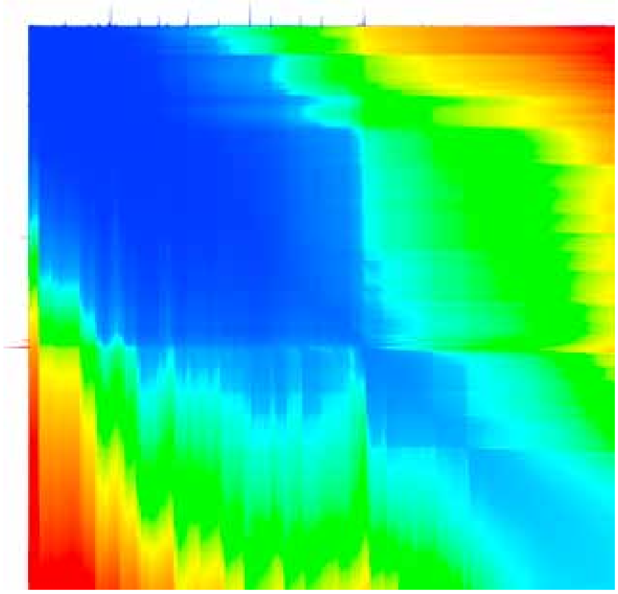


Figure 1: A visualization of the cumulative distance matrix between two chromatograms of an experiment measuring 14 differet natural and synthetic amino acids. Each colored pixel represents the optimal value of the alignment at the respective position. Colors range from blue (low distance) via green to red (high distance). Visible at the top and at the left side are the Total Ion Currents of the two aligned chromatograms.

genome, which is the traditional prerequisite for repeat detection. Our approach is based on the so-called de Bruijn graphs, whose highly regular structure results in several appealing combinatorial properties. In computational biology, de Bruijn graphs have been traditionally used in the context of sequence assembly, but their use for other applications has not been explored. We investigated the use of de Bruijn graphs for the detection of repeats and developed algorithms to efficiently represent and handle the sparse graphs that arise in practice. This project has been carried out in collaboration with the Genome Research group of the Institute for Genome Research and Systems Biology.

Computational comparative genomics

In computational comparative genomics, the basic task is to compare the genomes of two or more organisms, not any more at the level of DNA sequence, but at the higher level of gene content and gene order. A typical approach usually assumes that, across the genomes to be compared, several genes have already been identified and classified in families. Then, from an abstract viewpoint, each gene can be modeled as a signed integer, where the absolute value encodes the gene family and the sign encodes the DNA strand orientation.

Starting from the representation of two genomes as strings of signed integers, we set out to devise a quantitative measure for the similarity (or equivalently distance) between these genomes. Such a pairwise distance measure can, in turn, be used in phylogenetic studies.

As a reasonable distance measure we can use the minimum number of rearrangement operations that are required in order to transform one genome into another. How exactly to compute the distance and a corresponding shortest rearrangement scenario depends on which operations we consider as biologically plausible and allow in our model. In uni-chromosomal genomes the most common rearrangements are inversions, while in multi-chromosomal genomes we must be more flexible and also consider translocations, fusions and fissions.

In a series of papers we have contributed considerably to the generalization and simplification of genome rearrangement studies, including the first linear-time algorithm for computing the so-called Hannenhalli-Pevzner distance between two multichromosomal linear genomes. Moreover, we have formally defined the Double-Cut-and-Join (DCJ) distance, a very general measure that can model all classical genome rearrangement operations, also for circular chromosomes (*http://bibiserv.techfak.uni-bielefeld.de/dcj*).

Another important task in whole genome comparisons is the detection of gene clusters, i.e. sets of genes that occur co-localized in several genomes. This is motivated by the simple, but biologically verified assumption that interacting proteins are often encoded by genes that are located in close genomic proximity. Therefore, if groups of genes are observed to be co-located in several, not too closely related species, this may hint at functional association of the encoded proteins.

An active area of research in our group is the development of efficient algorithms for the detection of gene clusters across a given set of genomes. The computational complexity of this task depends crucially on the number and sizes of the considered genomes and on the flexibility that we allow in our definition of a gene cluster.

The most constrained model requires that a gene cluster must appear with exactly the same gene composition in the considered genomes. For this case, we have developed a quadratic time algorithm for finding all gene clusters of two genomes and also extended it for finding gene clusters in more than two genomes. However, for most applications it is more realistic to relax the definition of a gene cluster, so that it allows for small deviations in the gene content of the cluster occurrences. On the other hand, introducing this flexibility increases the computational complexity of cluster detection drastically. For the detection of such approximate gene clusters, we have introduced a concept called median gene cluster that improves over existing models and presented efficient algorithms for its computation.

Our algorithms for the efficient detection of gene clusters, coupled with several interactive visualization features, are implemented in the software tool Gecko (*http://bibiserv.techfak.uni-bielefeld.de/gecko*).

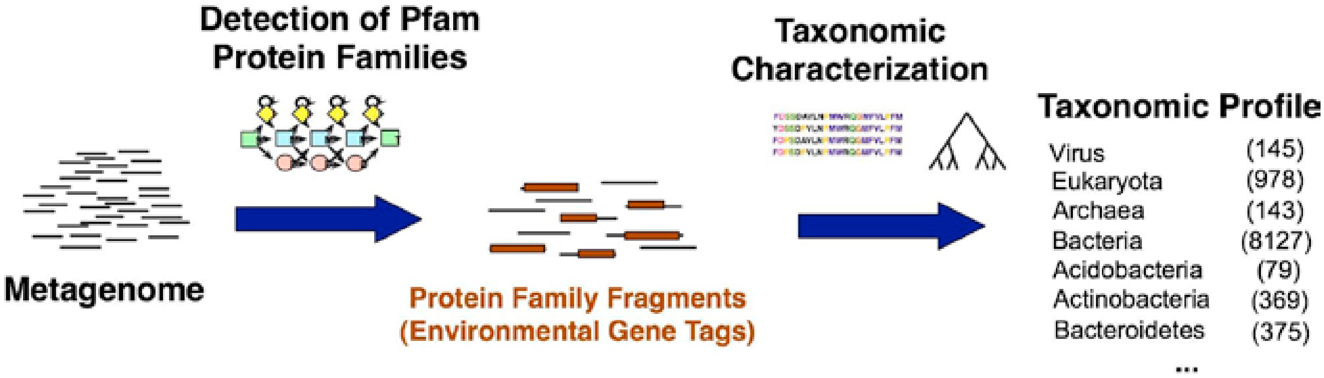


Figure 2: The metagenome is searched for reads that encode for known proteins. Those reads are called EGTs (Environmental Gene Tags) and are used to create a taxonomic profile by means of phylogenetic trees.

In another research trail we aim at putting the concept of gene clusters in the framework of phylogenetics. Here, the focus is not any more on the detection of gene clusters, but on their reconstruction for ancestral genomes.

More specifically, we have developed an algorithm that, given the topology of a phylogenetic tree and the gene orders of the leaf nodes, calculates optimal sets of gene clusters for the internal nodes. The optimization criterion combines two properties: parsimony, i.e. the number of gains and losses of gene clusters has to be minimal, and consistency, i.e. for each ancestral node there must exist at least one gene order that contains all the reconstructed clusters. This algorithm has been implemented in the Rococo software (*http://bibiserv.techfak.uni-bielefeld.de/rococo*).

Computational metagenomics

Metagenomics is a new field of research that studies natural microbial communities. The new sequencing techniques, such as 454 or Solexa-Illumina, can produce huge amounts of data in much shorter time and with less efforts and costs than the traditional Sanger sequencing. However, the produced data comes in short reads (35-50 base pairs with Solexa-Illumina, 100-300 basepairs with 454 sequencing).

Carma (*http://www.cebitec.uni-bielefeld.de/brf/carma/carma.html*) is a new pipeline for the characterization of the species composition and the genetic potential of microbial samples using 454-sequenced reads. The species composition can be described by classifying the reads into the taxonomic groups of organisms they most likely stem from. By assigning the taxonomic origins to the reads, a profile is constructed which characterizes the taxonomic composition of the corresponding community. Carma has already been successfully applied to 454-sequenced communities. In particular, it has been also used to characterize a plasmid sample isolated from a wastewater treatment plant, sequenced and studied by members of the Institute for Genome Research and Systems Biology.

Using samples from a biogas plant we examined the applicability of this approach for the ultra-short Solexa-Illumina reads by comparing the results with those obtained by the 454-sequenced sample. Our results using 77 million 50 bp-reads revealed that this approach indeed produces consistent results. Most differences we have found are in the taxa of higher order, e.g. in the species level, and in general for species with a very low abundance.

In order to apply Carma to high-throughput sequencing data, we had to improve the accuracy and speed of our method in various ways: A preprocessing assembly phase using an adapted q-gram index; adaptation of the pipeline to take the information of mated reads into account in order to ‘increase’ the read length; modification of the amino acid sequence distance function for the construction of the phylogenetic tree; and implementation of a protein q-gram index over a multiple alignment for the read-against-Pfam protein family matching.

Metabolomics and mass spectometry

Modern analytical methods in biology and chemistry use separation techniques coupled to sensitive detectors, such as Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS). These hyphenated methods provide high-dimensional data whose manual comparison in order to find corresponding signals is a tedious task because each experiment usually consists of thousands of individual scans, each one containing hundreds or even thousands of distinct signals. Therefore an accurate automatic alignment and matching of corresponding features between two or more experiments is required. Such a matching algorithm should capture fluctuations in the chromatographic system which lead to non-linear distortions on the time axis.

We have developed the *ChromA* software (*http://bibiserv.techfak.uni-bielefeld.de/chroma*) that performs a retention time alignment of multiple chromatograms of mass spectra. The alignment is calculated with a dynamic programming algorithm known as

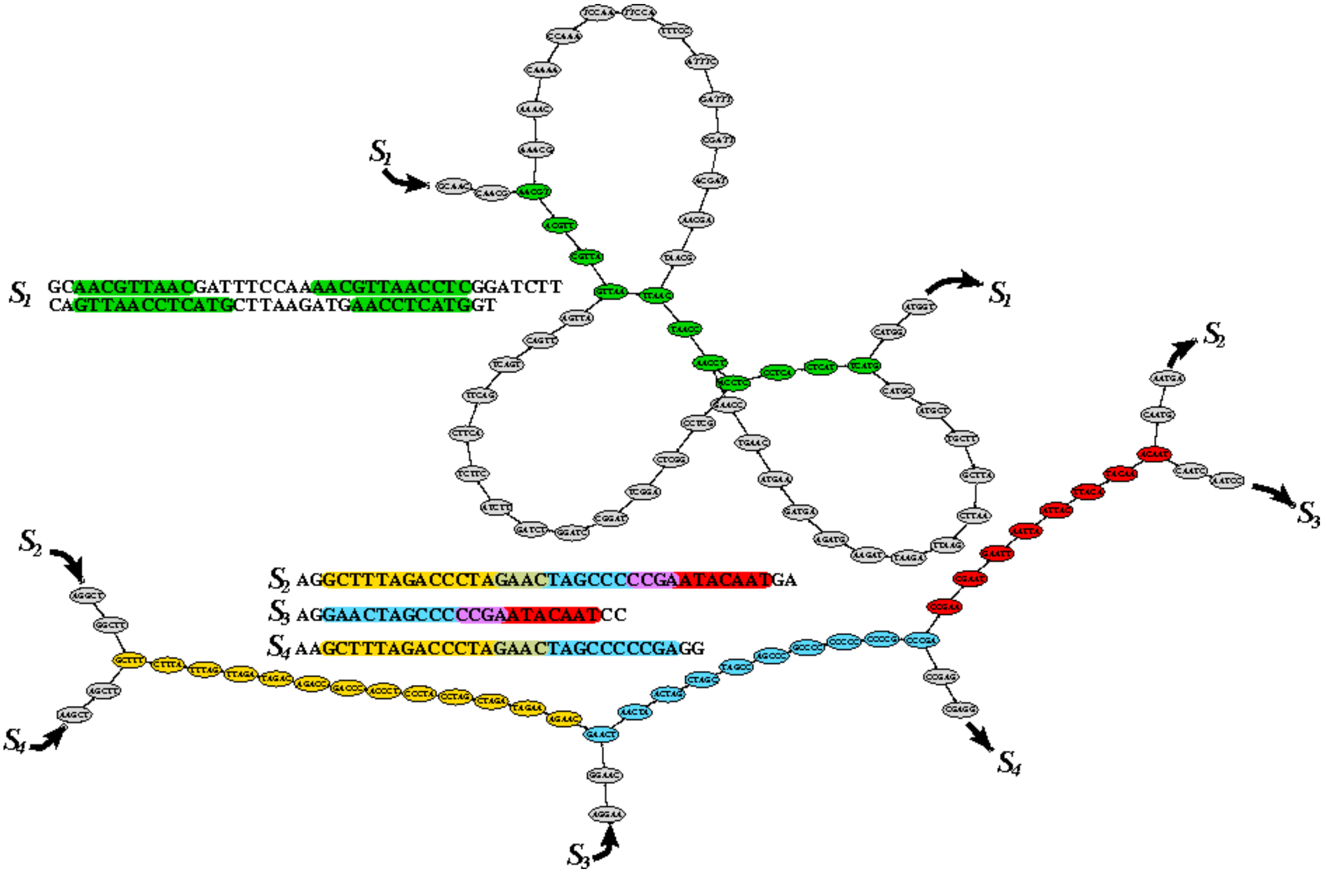

S1
S1 GCAACGTTAACGATTTCCAAAACGTTAACCTCGGATCTT
CAGTTAACCTCATGCTTAAGATGAACCTCATGGT
S1
S2
S2
S2 AGGCTTTAGACCCTAGAACTAGCCCCCGAATACAATGA
S3 AGGAACTAGCCCCCGAATACAATCC
S4 AAGCTTTAGACCCTAGAACTAGCCCCCGAGG
S3
S4
S4
S3

Figure 3: De Bruijn subgraphs for a single sequence (*S1*) containing two repetitive regions (highlighted in green) and for three sequences (*S2*, *S3*, *S4*) that share common subsequences.

dynamic time warping, which is a generalization of classical sequence alignment for continuous data. *ChromA* allows users to experiment with different similarity measures for nominal mass precision mass spectra. Moreover, they can take advantage of any a-priori knowledge about certain substances, which give rise to peaks that have to be aligned together across the chromatograms. The use of these so-called anchors can speed up the computation of the alignment significantly.

The web-based *MeltDB* software platform (*http://meltdb.cebitec.uni-bielefeld.de*) provides a complete solution for the analysis and integration of metabolomics experiment data. *MeltDB* allows easy integration of external programs via its extensible tool concept, for example *ChromA* or other available software like *XCMS*. It is coupled to existing CeBiTec tools such as *GenDB* and *Emma*, combining different -omics techniques to simplify multi-level analysis of samples of biological origin. Statistical analysis of experiments is based on the open-source software *R*.

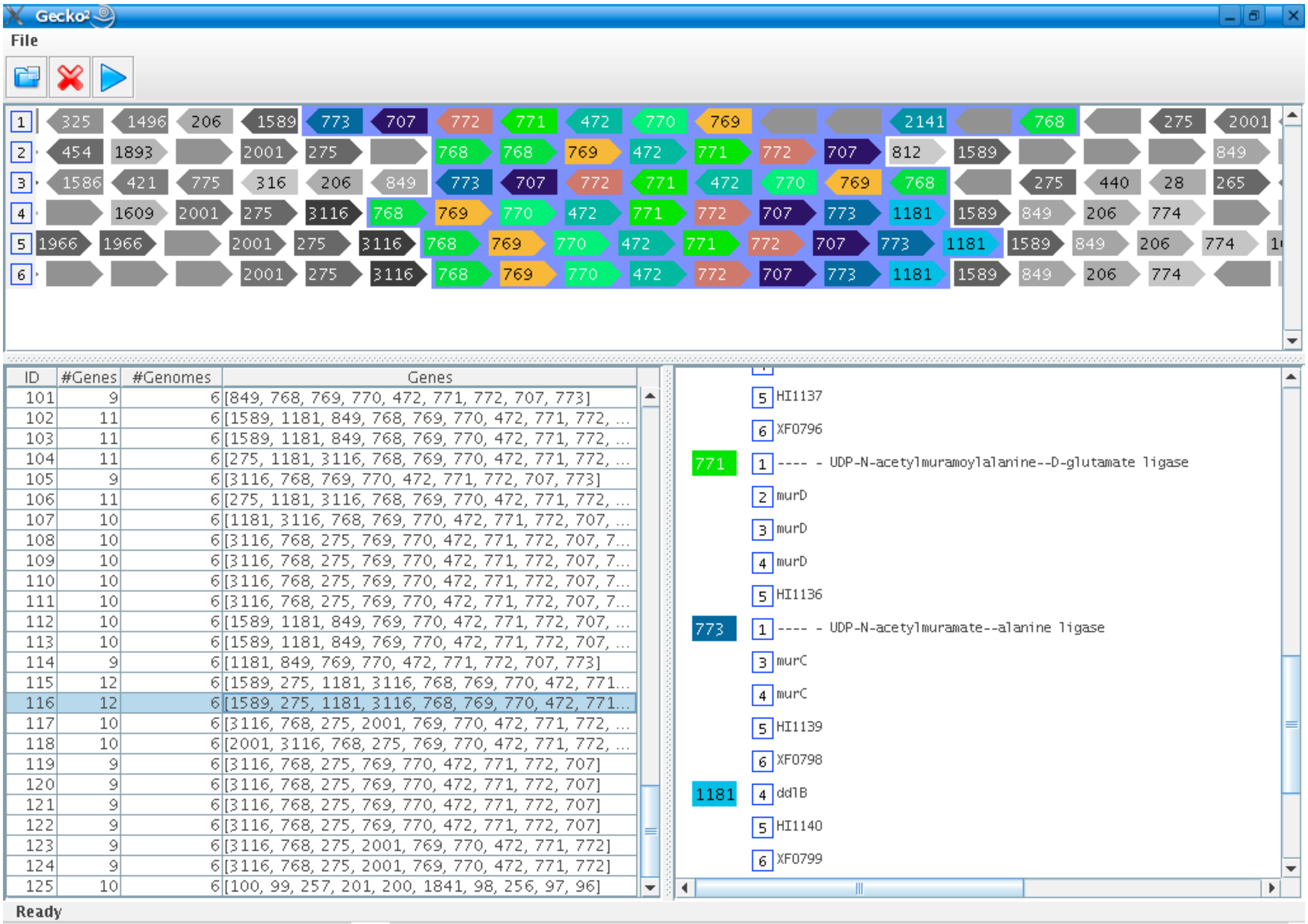
Other research activities

Clustering a set of objects based on a pairwise distance measure is a classical machine learning problem that arises in many different fields. Bioinformatics, in particular, asks for clustering algorithms that can operate on large datasets. For protein datasets, we advocate that a graph-theoretical formulation of the clustering problem, the so-called transitive graph projection, is well-suited.

For its solution we have developed the *Force* heuristic (*http://gi.cebitec.uni-bielefeld.de/comet/force*). The main idea is to find an arrangement of the vertices in the two-dimensional plane, such that vertices from subgraphs with high intra-connecting edge weights are placed close to each other. This layout is then used to define the clusters by Euclidean single-linkage clustering of the vertices' positions on the plane.

With extensive evaluation within the TransClust framework (*http://transclust.cebitec.uni-bielefeld.de*) we have shown that Force outperforms the most popular existing clustering tools, when it comes to clustering large protein datasets.

In a more theoretical trail, we investigate the properties of Probabilistic Arithmetic Automata (PAA) and their applications in bioinformatics. PAA are probabilistic models that combine the features of Deterministic Finite Automata with those of Hidden Markov Models. In our investigations, PAA are used to compute the statistics of random DNA or protein texts that are generated by a first-order Markov process. The knowledge of these statistics



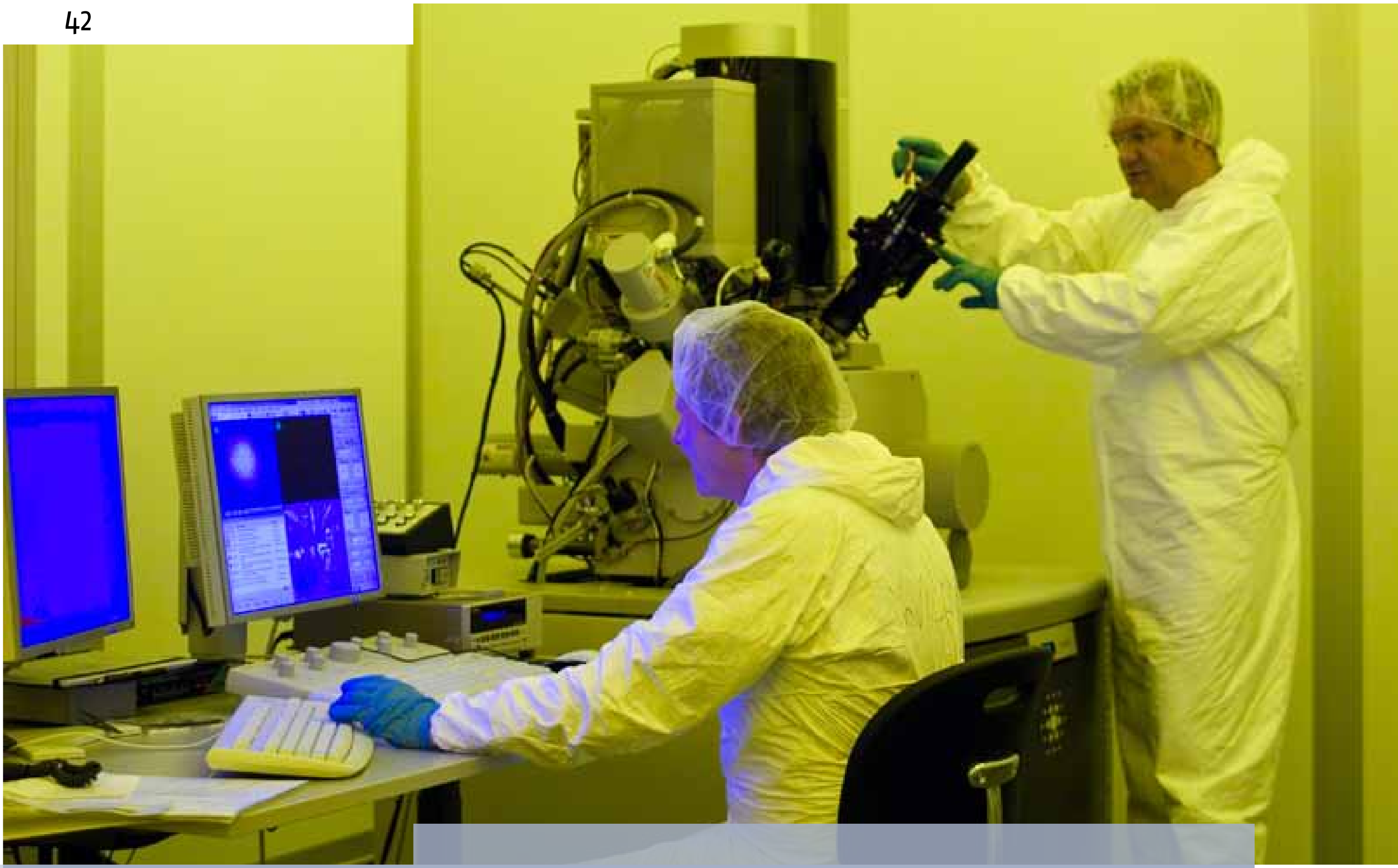
ID	#Genes	#Genomes	Genes
101	9	6	[849, 768, 769, 770, 472, 771, 772, 707, 773]
102	11	6	[1589, 1181, 849, 768, 769, 770, 472, 771, 772, ...
103	11	6	[1589, 1181, 849, 768, 769, 770, 472, 771, 772, ...
104	11	6	[275, 1181, 3116, 768, 769, 770, 472, 771, 772, ...
105	9	6	[3116, 768, 769, 770, 472, 771, 772, 707, 773]
106	11	6	[275, 1181, 3116, 768, 769, 770, 472, 771, 772, ...
107	10	6	[1181, 3116, 768, 769, 770, 472, 771, 772, 707, ...
108	10	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707, 7...
109	10	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707, 7...
110	10	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707, 7...
111	10	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707, 7...
112	10	6	[1589, 1181, 849, 769, 770, 472, 771, 772, 707, ...
113	10	6	[1589, 1181, 849, 769, 770, 472, 771, 772, 707, ...
114	9	6	[1181, 849, 769, 770, 472, 771, 772, 707, 773]
115	12	6	[1589, 275, 1181, 3116, 768, 769, 770, 472, 771...
116	12	6	[1589, 275, 1181, 3116, 768, 769, 770, 472, 771...
117	10	6	[3116, 768, 275, 2001, 769, 770, 472, 771, 772, ...
118	10	6	[2001, 3116, 768, 275, 769, 770, 472, 771, 772, ...
119	9	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707]
120	9	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707]
121	9	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707]
122	9	6	[3116, 768, 275, 769, 770, 472, 771, 772, 707]
123	9	6	[3116, 768, 275, 2001, 769, 770, 472, 771, 772]
124	9	6	[3116, 768, 275, 2001, 769, 770, 472, 771, 772]
125	10	6	[100, 99, 257, 201, 200, 1841, 98, 256, 97, 96]

	5	HI1137
	6	XF0796
771	1	---- - UDP-N-acetylmuramoylalanine--D-glutamate ligase
	2	murD
	3	murD
	4	murD
	5	HI1136
773	1	---- - UDP-N-acetylmuramate--alanine ligase
	3	murC
	4	murC
	5	HI1139
	6	XF0798
1181	4	ddlB
	5	HI1140
	6	XF0799

Figure 4: Screenshot of Gecko, a software tool for the detection of approximate gene clusters. The upper window shows the gene cluster currently selected from the list of detected clusters in the lower-left window. Gene annotations for this cluster are listed in the lower-right window.

is, in turn, useful for several applications. These include, for example, finding overrepresented motifs in protein sequences and optimizing the sensitivity of string matching algorithms.

In another theoretical project we have studied some combinatorial aspects of sparse matrix factorization and identified several connections to classical concepts from graph theory. Our results are useful for any signal processing application that involves a set of signal sources emitting signals and a set of sensors measuring these signals over time. Among other fields of engineering, this general model also arises in two applications in bioinformatics: Firstly, in the analysis of microarray measurements under the presence of cross-hybridizing probes and, secondly, in the quantification of transcription factor activities in simple regulatory networks. ■



From the Nobel Price in Physics towards a unique Microscope for Biotechnology

History

In 2007 Albert Fert and Peter Grünberg have been awarded the Nobel Prize in physics for their discovery of the Giant magnetoresistance (GMR) which is of quantum mechanical origin and was observed in thin film structures composed of alternating ferromagnetic and well conducting nonferromagnetic spacer layers. GMR relies on the orientation of the electron spin with respect to the magnetization direction of a ferromagnetic layer. When adjacent ferromagnetic layers in these thin film structures have the same orientation, electrons of a single spin type parallel to this orientation can move easily between them whereas electrons of the other spin type are being scattered. As a consequence the resulting electrical resistance is low. Are their magnetizations opposed, electrons of both types are scattered, causing a high electrical resistance. The GMR effect manifests itself as a significant decrease of typically 5% to 80% in electrical resistance by changing from an antiparallel orientation of the magnetization in adjacent ferromagnetic layers to a parallel one with increasing external magnetic field. In the absence of this field, the magnetization direction of adjacent ferromagnetic layers can be fixed in antiparallel as a function of the layer thickness of the spacer layers utilizing a so called spacer layer coupling.

This pioneering work of Fert and Grünberg was based on two different stacking sequences. While Grünberg was investigating a sandwich consisting of only three layers Fe/Cr/Fe named as spin-valve, Fert was looking into the characteristics of $\{Fe/Cr\}_N$ multilayers. Nevertheless, both of these GMR device are nothing else than very sensitive magnetic field detectors which were driving the development of a new generation of read-heads (GMR spin-valves) and a new generation of sensors for automotive applications (GMR multilayers). Only about ten years after this discovery the potential of GMR-sensors for the detection of magnetic beads was realized and led to another technological avenue, the development of biosensors for life science applications.

Magnetoresistive Biosensors

Currently, magnetoresistive biosensors use a new detection method for molecular recognition reactions based on a combination of magnetic markers and XMR-sensors. Besides GMR-sensors also Tunneling Magnetoresistance- (TMR) sensors are of great interest. Replacing the spacer layer in GMR spin-valves by a thin insulator such as Al_2O_3 or MgO will lead to a TMR sensor. If this insulating layer is thin enough, e.g. about 2 nm, electrons can tunnel from one ferromagnetic layer into the other – again a strictly quantum mechanical phenomenon. The tunneling probability is associated with the relative orientation of the magnetizations of the two adjacent ferromagnetic layers. A parallel orientation yields a high tunnel current or an electrical state of low resistance whereas an antiparallel orientation is characterized by a low tunnel current or a state of high resistance. Like for GMR devices the TMR sensor can be switched between these two states of electrical resistance employing an external magnetic field.

As is shown in Figure 1, the new detection method consists of superparamagnetic nanoparticles or beads which are specifically attached to a target molecule. The superparamagnetic nature of the nanoparticles or beads enables to switch on their magnetic stray fields by using an external magnetic field. Hence, the localization of the magnetic stray field by an embedded XMR-sensor allows to identify the target molecule on or in close vicinity to the XMR-sensor indicated by a drop in the electrical resistance.

The challenges of the development of such a combined tool for single molecule detection is fourfold: (1) the magnetic core of magnetic nanoparticles has to be stabilized by organic ligands so as to define their size distribution and simultaneously to preserve their magnetic property by preventing them from oxidation, (2) to functionalize the tail groups of the ligands such that biomolecules can easily be marked by these magnetic nanoparticles, (3) to design and realize XMR-sensors which are capable of detecting the magnetic stray field of magnetic nanoparticles enabling to count the number of magnetically labeled biomolecules covering the sensors surface and (4) to incorporate the sensors into a fluidic environment so as to ensure that all magnetically labeled biomolecules will pass by in low heights so as to ensure their binding onto the sensors surfaces in a static mode resulting in an interaction between their magnetic stray fields and the XMR-sensors or allowing an interaction between the magnetic stray fields and the XMR-sensors while passing by in a dynamic mode of analysis.

Prof. Dr. Andreas Hütten

Andreas Hütten studied physics at the Georg-August-Universität Göttingen and received his PhD in 1989. As a postdoctoral researcher he was working in the US at UC Berkeley and the Lawrence Berkeley National Laboratory and in Germany at the Leibniz Institute for Solid State and Materials Research Dresden. In 1997 he was appointed as an Assistant Professor (C1) and in 2001 as an Associate Professor (C2) at the Department of Physics at Bielefeld University and received his Habilitation in 2001. In 2005 he took a position as a group leader for Magnetic Nanostructures at the Institute of Nanotechnology of the Forschungszentrum Karlsruhe. Since 2007 he is Professor of Physics of Nanostructures at the Department of Physics at Bielefeld University. At present, his research activities are focused on magnetic nanoparticles, magnetoresistive biosensors and materials development for magnetoelectronic devices.

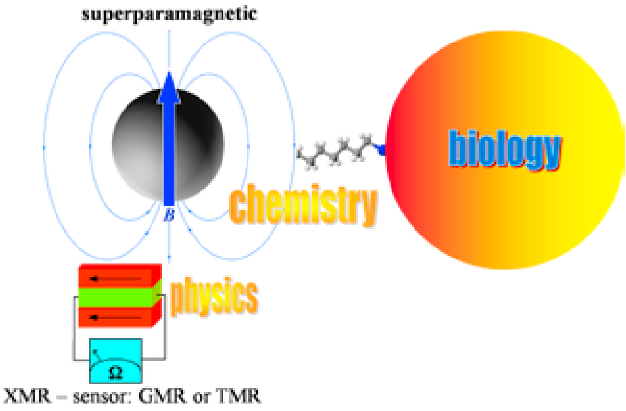


Figure 2: Schematic presentation of a combined tool for single molecule detection.

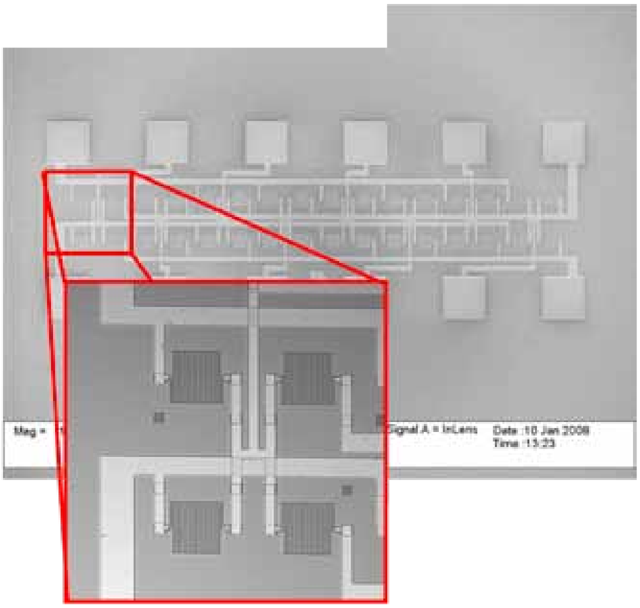


Figure 1: SEM image of a 32 GMR-sensor array.

The XMR – sensor design

Ideally, XMR-sensors are built in form of sensor arrays as can be seen in Figure 2. This provides a lot of options in magnetic as well as and electrical direction. Magnetically, the GMR-characteristics of the individual sensors can be tuned for different magnetic beads allowing biotests in parallel. Electrically, all sensors within such an array can individually be read out. That way it is possible to allocate different tasks to individual sensors. To enhance the statistical significance of a biotest, for example, to determine the mean coverage of biomolecules on the sensor surface, one can compare the results of all individual sensors assuming a homogeneous coverage of all sensors. Functionalizing the surface of individual sensors with different biolayers of antibodies the detection of different antibody parameters can also be made in parallel.

XMR-sensors based on the TMR-effect have several advantages when compared to current-in-plane GMR-sensors: (1) their very large TMR-amplitude can be translated into a large magnetic field range allowing to detect beads with different magnetic properties, (2) laterally, they can be made very small which translates into the detection of several nanoparticles only, (3) they can be extended to a sensor array which enables to detect magnetic beads with high lateral resolution, (4) furthermore, they allow to determine the orientation of the bead`s magnetic moment relative to magnetization orientation of the sensing magnetic layer of the sensor and (5) assuming that the stray field of one magnetic bead is detected at the same time by several neighboring sensors, then the trajectory of that bead can be dynamically measured.

Figure 3 is showing an experimental realization of two sensor arrays with 20 TMR-sensors, each, together with all current leads. In addition, two current lines integrated into the TMR chip are shown so as to magnetostatically attract the beads towards the individual sensors.

When testing the performance of the individual sensors integrated into the array very small magnetic Co-nanoparticles can be used as is demonstrated in Figure 4. A self organized monolayer of 14 nm Co-nanoparticles reveals a hysteresis characterized by a coercivity of 47.4 Oe. Hence, the TMR characteristic of an individual TMR sensor should reflect the same behavior when it is covered by a monolayer composed of these 14 nm Co-nanoparticles. This is also seen in Figure 4 where the TMR characteristic of an individual TMR-sensor displays a coercivity upon covering with a monolayer of 14 nm Co-nanoparticles. The resulting coercivity is about 47 Oe and hence in very good agreement with that of the monolayer of 14 nm Co-nanoparticles.

The microfluidic environment

The fluidic environment which ensures an enhanced probability of binding labeled biomolecules onto XMR-sensor surfaces is addressed in the following section. One approach that can be used to transport biomolecules attached to magnetic beads is the on-off ratchet, a transportation phenomenon in the presence of diffusion and some perturbation that drives the system out of equilibrium without introducing a priori an obvious bias into one or the other direction of motion. The on-off ratchet mechanism employs a periodic switching between the on- and off-state of a potential as is illustrated in Figure 5. The first state is the on-state, where beads move to their potential minimum. The second state is the off-state, where beads diffuse freely. Due to the asymmetry of the potential, which can experimentally be realized by a superposition of an assembly of spatially periodic conducting lines with a homogeneous magnetic field perpendicular

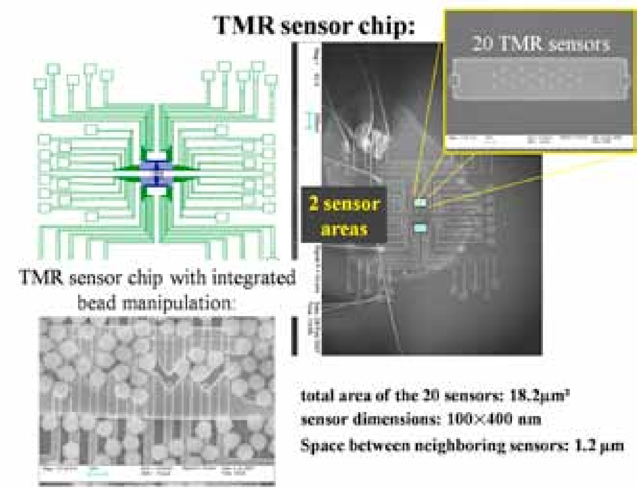

TMR sensor chip:
20 TMR sensors
2 sensor areas
TMR sensor chip with integrated bead manipulation:
total area of the 20 sensors: 18.2µm²
sensor dimensions: 100×400 nm
Space between neighboring sensors: 1.2 µm

Figure 3: Design and realization of two TMR-sensor arrays with 20 individual sensors within each array is displayed. In the lower left corner additional current lines are integrated into the TMR chip so as to magnetostatically attract magnetic beads towards the individual sensors.

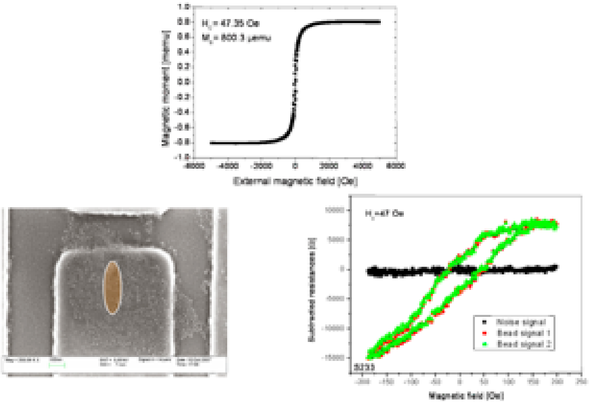


Figure 4: Testing individual TMR sensors within the array of 20 sensors by covering with a monolayer of 14 nm Co-nanoparticles. The coercivity of the Co monolayer is also reflected in the characteristic of the TMR sensor which proofs the proper performance.

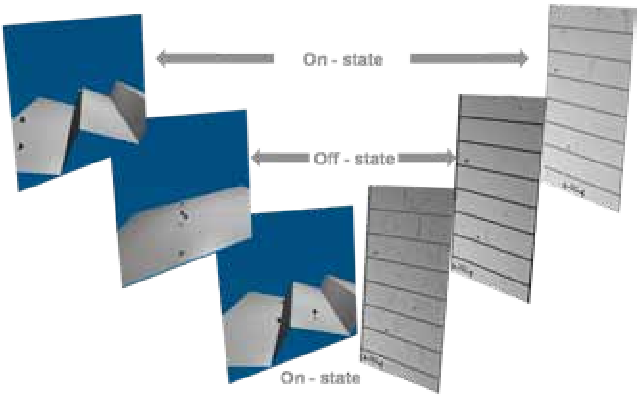


Figure 5: Schematic drawing of the principle of the magnetic on-off ratchet on the left side. The magnetic potential acting in the off-state is picture as an undulating landscape. In comparison, the experimental realization of the magnetic on-off ratchet on the right side is showing the corre-sponding bead movement. Clearly visible is the net flux of beads imaged as trace pattern. It arises from the asymmetric geometry of the magnetic potential visualized on the left hand side.

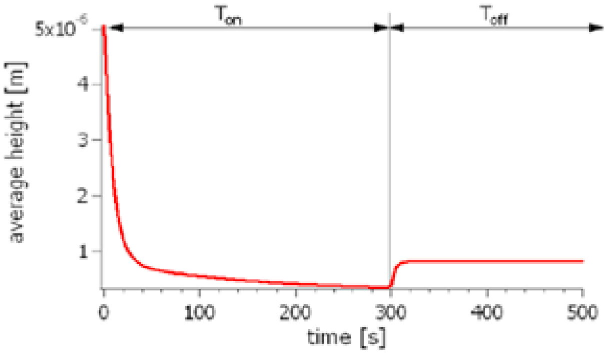


Figure 6: Average height of a 1 µm bead trajectory in an on-off ratchet. These beads passing the positions of potential minima in the off-state within their magnetic stray field range.

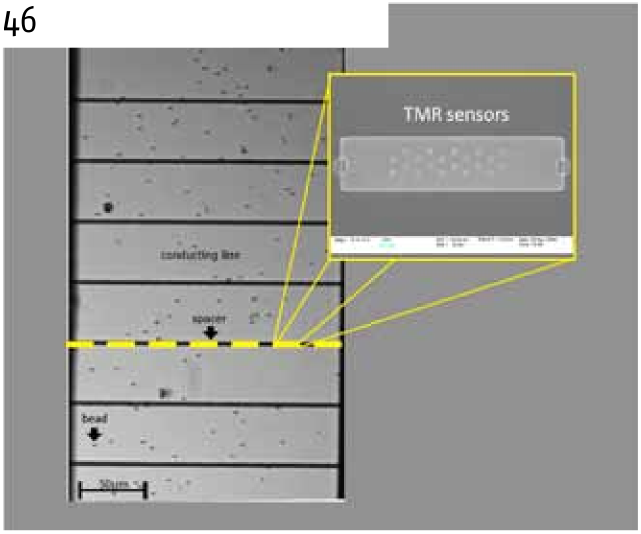


Figure 7: The setup of a magnetoresistive microscope (MRM) integrating TMR-sensor array at the potential minima of a magnetic on-off ratchet.

to the conduction lines, the distance to the maximum point of the potential on the steeper slope side is shorter than that on the gently inclined side. Thus the chance for beads to pass the potential barrier during the off-state on the steeper side is larger than that on the gently inclined side and hence a net flux of beads arises as is shown in Figure 5 as well.

One benefit of such a ratchet used to control the motion of the beads can be seen in Figure 6. This ratchet concept keeps the trajectories of biomolecules attached to magnetic beads passing an array of TMR-sensors in stray field range allowing for binding to the sensors surfaces and for detection without employing micropumps.

The magnetoresistive microscope

The idea of a magnetoresistive microscope is taking shape by integrating TMR-sensor arrays into the magnetic on-off ratchet structure at positions characterized by potential minima. Such a combination is sketched in Figure 7 and the underlying concept becomes visible. In the off-state the biomolecules attached to magnetic beads can freely move driven by Brownian motion about the positions of potential minima. Since their trajectories are always in stray field range of the TMR-sensors located there as well it is possible to monitor their movement by simultaneously reading out the resistance drop of all sensors. The exact location of the center of one biomolecules can be calculate from the fraction of magnetic stray field interaction with all sensors involved. This is similar of finding the position of a receiver with help of several satellites employing GPS. In analogy to an optical microscope which enables to track the biomolecules movement by imaging this technique allows to monitor this movement by following a chain of electrical signals initiated by the stray field interaction of magnetic beads attached to the biomolecules. Due to (1) the incredible sensitivity of the TMR-sensors, (2) the possibility to realize TMR-sensor arrays with a high lateral sensor integration density and (3) employing high moment magnetic nanoparticles with a very sharp particle size distribution for magnetic labeling this proposed microscope can work far below an optical resolution and hence enables to locate not only biomolecules en bloc but also certain sequences of their molecular structure when magnetically labeled. Furthermore, the concept of the magnetic on-off ratchet intrinsically separates size distributions of biomolecules by transferring their size differences in time differences when passing by. In addition, this allows for molecular recognition in parallel.

Currently, this MRM is being tested against optical microscopy in terms of spatial and time resolution aiming for:

- a quantitatively detection of Lawsonia bacteria,
- determining the on- and off-rate of the binding process between biotin and streptavidin,
- determining diffusion constants of different molecule in parallel.

It is dedicated to the results of these experiments whether the MRM will entering the laboratory for biotechnology as a new generation of microscopes. ■

Research on Biofuels at the CeBiTec

Introduction

A number of recent reports have emphasized that the development of CO_2-neutral fuels is one of the most urgent challenges facing our society to prevent negative effects of climate change. These reports and the increasing public awareness regarding global energy supply and the environmental problems associated have led European policy makers to the decision to introduce targets to decrease carbon emission and to incorporate carbon neutral fuels like biodiesel, bioethanol, biomass-to-liquid (BTL) fuels, or biohydrogen into their energy portfolios.

According to the Energy Information Administration (USA), the global energy demand will increase in the future from 488 exajoule in the year 2005 to around 650 exajoule in 2020 as a consequence of the growing industry in China, India and other countries (Figure 1).

Only one third of the global energy demand is consumed as electrical power, whereas two thirds accounts for fuels. Fossil fuel production from oil cannot be increased significantly in the future anymore, because all easy accessible major oil fields are already being exploited. Therefore, the potential market for biofuels is great.

The capacity of photosynthetic organisms to capture solar energy and convert it into chemical energy and biopolymers can be used to provide bioenergy and biofuels. These so called ‘solar biofuels’ have got a number of advantages compared to fossil fuels (oil, natural gas, coal). One major benefit is that carbon emissions from solar biofuels are low or even zero. Furthermore, photosynthetic organisms are able to grow and proliferate autonomously without the need for expensive chemical ingredients. Despite the obvious importance, CO_2-neutral fuel production systems are technically far less developed than fossil fuel production and electricity generation (e.g. by nuclear power plants).

For these reasons solar biofuel production has been identified as an important research field within the Center for Biotechnology (CeBiTec) at Bielefeld University. By far the largest proportion of biofuels is produced from higher plants today. In nature, this chemical energy is stored in a diverse range of molecules (e.g. lignin, cellulose, starch, oils). Lignin and cellulose, the most dominant carbohydrates in plant biomaterial, can also be converted into biofuels by lignocellulosic processes, which are currently being developed as second generation biofuel systems. Similarly, starch (e.g. from corn) and sugar (e.g. from sugarcane) are already being converted into bioethanol by fermentation, while oils (e.g. from canola, soy and oil palm) are being used as a feedstock for the production of biodiesel. However, solar biofuels which rely on higher plants have got the disadvantage that they are produced on arable land, therefore directly competing with the cultivation of food or feed plants. A major research stream at the CeBiTec is focused on the use of microalgae, which can be cultivated on non-arable land. Microalgae are able to efficiently produce carbon hydrates, starch and oils in large amounts. Some microalgae and cyanobacteria can also produce biohydrogen which could be an interesting alternative biofuel in the future.

The Consortium Bioenergy OWL

The Consortium Bioenergy OWL was founded in November 2007 and is based at the CeBiTec (Figure 2). The founding members Bielefeld University, the University of Applied Sciences Bielefeld, Stadtwerke Bielefeld GmbH and Biogas Nord GmbH signed a memorandum of understanding to cooperate in the field of bioenergy.

Bioenergy research activities at the CeBiTec and related public relations are bundled and coordinated within the Consortium Bioenergy OWL.

Since 2008, research projects of the Consortium have focused on solar biohydrogen production with green algae, light to biomass conversion and anaerobic fermentation of plant biomass to biomethane (see below).

An international CeBiTec symposium on solar biofuels was organized and held in August 2008. More than 100 scientists from 16 countries participated and discussed progress and perspectives in utilization of solar biofuels.

The Consortium Bioenergy OWL also actively participated in the science festival *Geniale* in October 2008. Presentations and guided tours were offered for the interested public.

Prof. Dr. Olaf Kruse

Olaf Kruse studied biology at Bielefeld University. After completing his PhD in Plant Cell Physiology in 1994, he worked as a postdoctoral fellow at Imperial College London with research topics on plant biochemistry and molecular biology. In 1997 he returned to Bielefeld University where he took over a group leader position at the Department of Biology in the field of Molecular Biology on microalgae. In 2006 he was appointed as the Head of 'Life, Earth & Environmental Sciences' at the 'European Science Foundation' (ESF) in Strasbourg. In 2007 he became a Research Fellow at the University of Queensland before he returned to Bielefeld where he built up the 'Algae Biotechnology & Bioenergy Group 'at the CeBiTec focusing on systems biology with microalgae and on molecular aspects of light to biomass conversion. In 2009 he was appointed as a full professor at the Faculty of Biology.

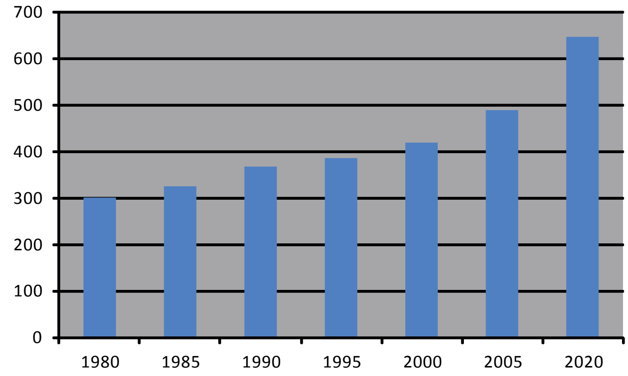
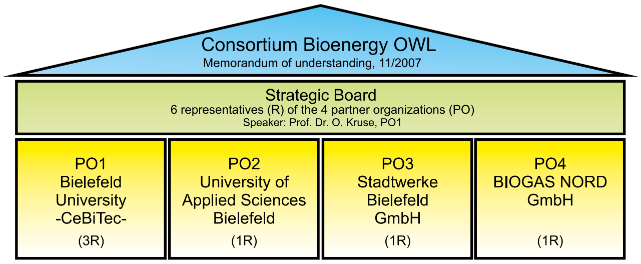


Figure 1: Global primary energy demand in the past and prognosis in Exajoule. From *Energy Information Administration* (USA), Report 2005



Consortium Bioenergy OWL
Memorandum of understanding, 11/2007

Strategic Board
6 representatives (R) of the 4 partner organizations (PO)
Speaker: Prof. Dr. O. Kruse, PO1

PO1	PO2	PO3	PO4
Bielefeld University -CeBiTec- (3R)	University of Applied Sciences Bielefeld (1R)	Stadtwerke Bielefeld GmbH (1R)	BIOGAS NORD GmbH (1R)

Figure 2: Organization scheme of the Consortium Bioenergy OWL

Research highlights and future projects

Photosynthesis research has a long standing-history at Bielefeld University. In recent years, new aspects such as solar energy to biomass conversion and solar biofuel production have increasingly become major research focuses. A number of approaches are investigated in the group of Prof. Dr. O. Kruse at the department for Algae Biotechnology and Bioenergy. The first step and basis for the production of any kind of solar biofuel is the capturing of solar irradiation. Understanding and optimization of light capturing in green algae has therefore been one of the main research topics. In higher plants and green algae, light is captured by specialized light harvesting complex proteins (LHCs). These are encoded by a large gene family that exhibits a high degree of homology and their expression is dependent on the prevailing environmental condition (e.g. light intensity). These proteins bind the bulk of the chlorophyll and carotenoids in the plant and play a role both in light capture and in the dissipation of excess energy which would otherwise inhibit the photosynthetic reaction centers. Excitation energy used to drive the photosynthetic reactions is funneled to the photosynthetic reaction centers. In nature, LHC proteins are produced in excess. This causes problems when microalgae are used in biotechnological application in high concentrations. Here, the outer cell layers absorb the light energy efficiently and dissipate excess light irradiation as heat or fluorescence. This loss of light irradiation leads to shading of cells deeper inside of the bioreactor and overall bad process efficiency. Research has been carried out to optimize light harvesting. RNAi technology was applied successfully to reduce the amount of light harvesting proteins in the model organism *Chlamydomonas* (Figure 3). The resulting strain was characterized by a light green phenotype as a result of less chlorophyll molecules (Figure 3A, B), reduced photoinhibition (Figure 3C) and increased growth rates under high light conditions (Figure 3D).

Within the project it could be demonstrated that optimization of the light harvesting system of microalgae can be achieved by genetic manipulation and results in improved growth parameters in high light (i.e. most of the day in moderate climate zones).

Another major research project carried out in the lab of Prof. Kruse is the analysis and optimization of photosynthetic biohydrogen production with microalgae.
The green algae *Chlamydomonas reinhardtii* is used as a model organism. It has the remarkable ability to capture solar energy, split water and recombine protons and electrons under anaerobic conditions to generate molecular hydrogen gas. Hydrogen is a very promising alternative fuel, as the combustions does not release harmful green house gases. Molecular biotechnology has been applied which resulted in a successful stepwise improvement of hydrogen production efficiency in *C. reinhardtii* cells. To further improve this biological approach of light energy to hydrogen conversion in the future, a systems biology approach has been established which aims at better understanding the metabolic pathways of hydrogen production (Figure 4).
Time points have been selected which represent the characteristic phases of the hydrogen production cycle (Figure 4A). Samples were taken for coordinated transcriptomics (Figure 4B), proteomics (Figure 4C) and metabolomics (Figure 4D) studies. The combined data will be used to obtain a complete picture of cellular processes during hydrogen production in *Chlamydomonas*.

Bioenergy production currently is not competitive to fossil fuels without subsidization, because production costs are still too high and the technology still is in its infancy. This will change in the future because easy accessible oil and natural gas reserves will deplete, whereas the energy demand will increase. Today bioenergy concepts with microalgae can only be commercially viable in a combined biorefinery system (Figure 5A). Here, high-value product (HVP) synthesis is the first production step. A number of potential HVPs from microalgae have been identified, e.g.

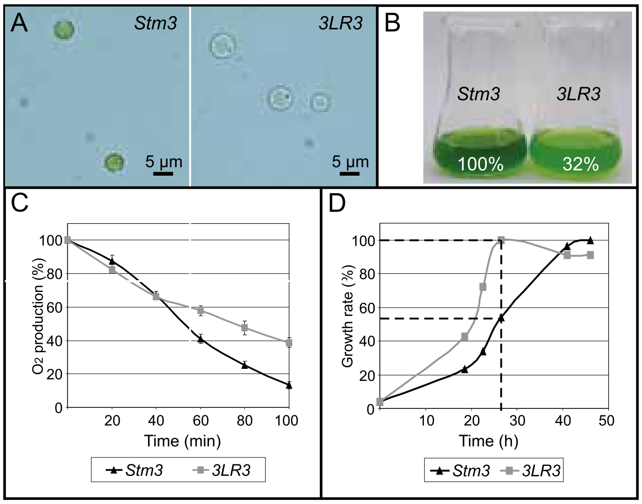


Figure 3: LHC downregulation by RNAi technology.
A) Microscopic cell images of parental strain Stm3 and LHC-RNAi strain 3LR3.
B) Cultures of Stm3 and 3LR3 at the same cell density. Relative chlorophyll concentrations are indicated.
C) O2 production capacity of Stm3 and 3LR3 during high light treatment
D) Mixotrophic growth rates of Stm3 and 3LR3 cultures under high light conditions

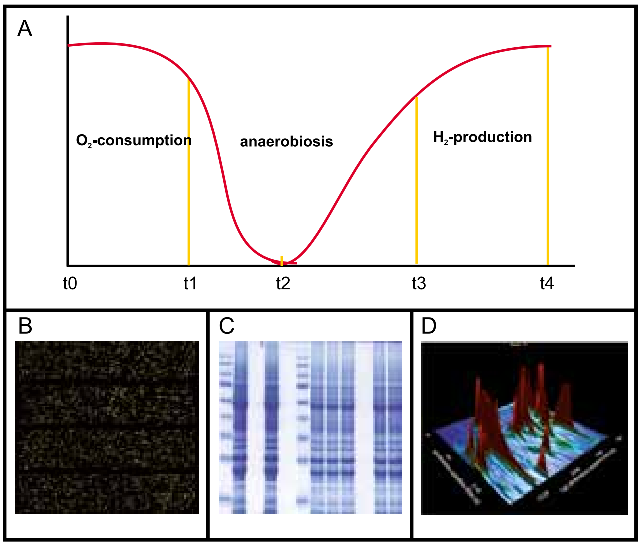


Figure 4: Systems biology for hydrogen production.
A) Time points selected for sample extraction during the hydrogen production cycle.
B) Transcriptome analysis with microarrays.
C) Proteome analyses via SDS PAGE and subsequent mass spectrometry.
D) Metabolome analyses via two-dimensional GC/GC TOF MS.

biohydrogen, sulfated polysaccharides, lipids and fatty acids, pigments or other bioactive molecules. After HVP extraction, the residual biomass is then used to generate bioenergy, e.g. biomethane. These biorefinery concepts are being tested at the CeBiTec. Screening for HVPs is conducted via different bioactivity tests and analytical procedures (HPLC, mass spectrometry). The potential of microalgae as a renewable plant substrate for biomethane production is investigated in 100ml batch tests (Figure 5B) and 20L fermenters (Figure 5C).

The experiments are carried out in close collaboration and with financial support from the industrial partners Biogas Nord GmbH and Stadtwerke Bielefeld GmbH and are planned to be intensified in the future. ■

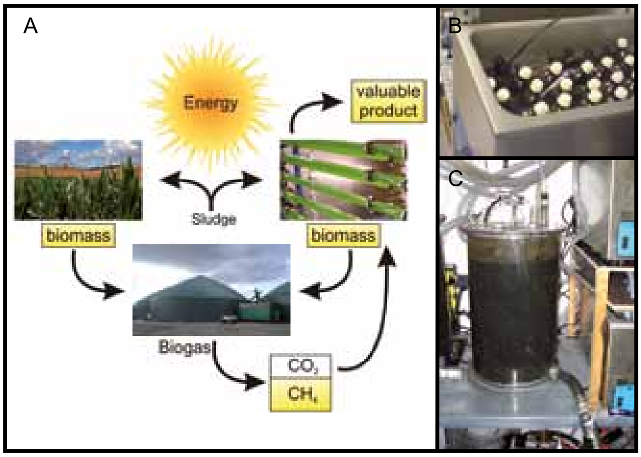


Figure 5: Biorefinery concept (A) and biogas fermentation in 100ml batch tests (B) and 20L fermenters (C).

From the Genome Sequence to the Transcriptional Regulatory Network

Junior Research Group 'Systems Biology of Regulatory Networks'

A Historical Conclusion

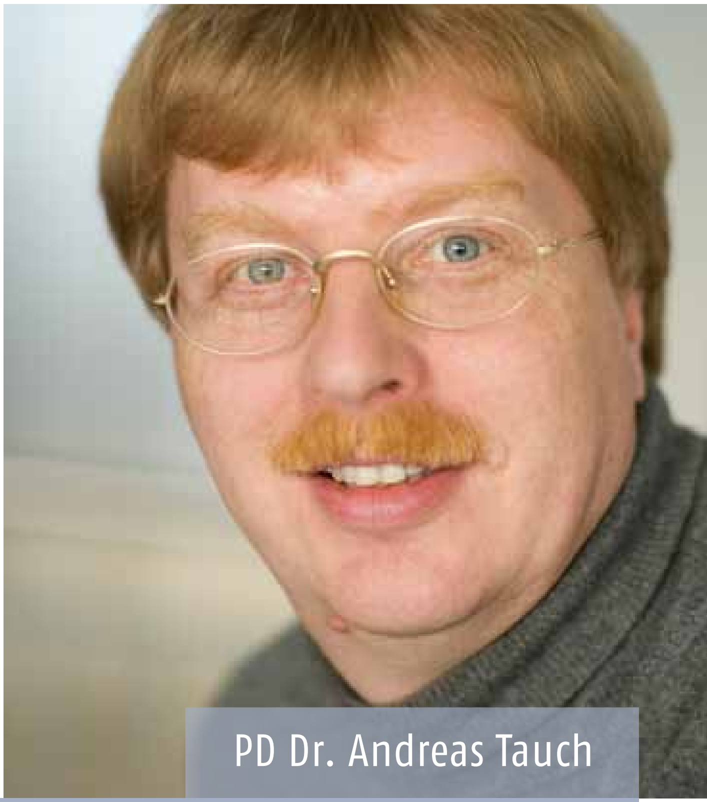
In 1965 the Nobel Prize in Physiology or Medicine was awarded to François Jacob, André Lwoff and Jacques Monod 'for their discoveries concerning genetic control of enzyme and virus synthesis'. In his Noble Lecture entitled 'Genetics of the Bacterial Cell', François Jacob has drawn the famous conclusion that 'the message inscribed in the genetic material contains not only the plans for the architecture of the cell, but also a program to coordinate the synthetic processes, as well as the means of insuring its execution.' Forty-five years later and in the era of 'omics' technologies, we are able to conceptualize this idea of a coordinated cellular program in bacteria as a network that is termed the transcriptional regulatory network.

The Concept of the Transcriptional Regulatory Network

The transcriptional regulatory network (TRN) of bacteria is a fundamental biological system controlling the flow of information from the internal and external environment to the gene level and thus to specific cellular functions. The regulation of gene expression at the transcriptional level is typically mediated by DNA-binding proteins (termed transcription regulators) that sense diverse internal or external stimuli, recognize and bind to specific DNA sequences (termed operators) and, either alone or in combination with other factors, control and modulate the expression of one or more target genes (collectively termed regulon). This basic genetic principle allows the reconstruction of the flow of environmental and genetic information in form of a directed graph. In such graphs, nodes represent transcription regulators and their target genes, and directed edges represent the regulatory interaction between them, exerted via the operator. The sum total of such transcriptional regulatory interactions in a bacterial cell can be conceptualized as the TRN. Within the TRN, transcription regulators and target genes form regulatory units cross-linked to network motifs that are characteristic topological elements. These motifs generally are not isolated but connected to complex structures forming the architectural backbone of the TRN.

Corynebacterium glutamicum A Model Organism for TRN Analysis in Gram-Positive Bacteria

The Junior Research Group 'Systems Biology of Regulatory Networks' is using *C. glutamicum* as a model organism to investigate a bacterial TRN. *C. glutamicum* is used by the biotechnological industry for the large-scale fermentative production of numerous metabolites such as L-amino acids and organic acids. The type strain of the species *C. glutamicum*, nowadays generally referred to as ATCC 13032, was originally isolated from a soil sample and turned out to be a natural producer of L-glutamic acid. Due to the remarkable feature of *C. glutamicum* to secrete large amounts of metabolites under suitable culture conditions, it has been used for more than 50 years in the industrial production of L-amino acids, for example L-glutamic acid and L-lysine. These amino acids are applied in human and animal nutrition respectively and are produced on a large scale by so-called high-performance strains of *C. glutamicum*. To further improve the efficiency of the biotechnological amino acid production by *C. glutamicum*, it is important not only to know in detail the metabolic pathways leading to these industrially important products but also to understand their transcriptional regulation. The availability of the complete genome sequence of the wild-type strain *C. glutamicum* ATCC 13032 had an enormous impact on gene expression profiling during the last years, since it opens the way to establish genome-wide transcriptional analysis techniques with DNA microarrays. Moreover, rapid advances in the development of ultra-fast sequencing technologies and associated bioinformatics approaches enabled new strategies for deciphering the structure of the TRN in corynebacteria by comparative genomics. Besides the genome of *C. glutamicum* ATCC 13032, four additional corynebacterial genomes were completely sequenced at the CeBiTec (Table 1) and many others are in progress.



PD Dr. Andreas Tauch

Andreas Tauch studied biology at Bielefeld University. After completing his doctoral thesis in genetics in 1996, he moved to the R&D department of the Degussa AG, Halle (Westf.), Germany. In 2000 he returned to the Technology Platform Genomics at the Center for Biotechnology (CeBiTec), and since then worked in the area of genome research, in particular in the fields of corynebacterial genomics and transcriptomics. In 2008 he received the habilitation in genetics at the Faculty of Biology of Bielefeld University. He is currently heading the Junior Research Group 'Systems Biology of Regulatory Networks' at the CeBiTec.

Deciphering Nodes and Edges in the TRN of *C. glutamicum*

A TRN encompasses sets of genes, whose expression levels are altered in response to an internal or external signal that is mediated by a transcription regulator interacting with its cognate operators. The detection of these components in TRNs is an essential step towards the creation of a framework for the systems-based analysis of transcriptional regulatory processes in a bacterial cell. We applied for example profile-based methods for the detection of operators to predict regulon members for a particular transcription regulator in corynebacteria and to infer regulons in a collection of closely related genomes (Figure 1). Transcription regulators distributed across bacterial lineages can be linked to multiple lineage-specific regulons with conserved

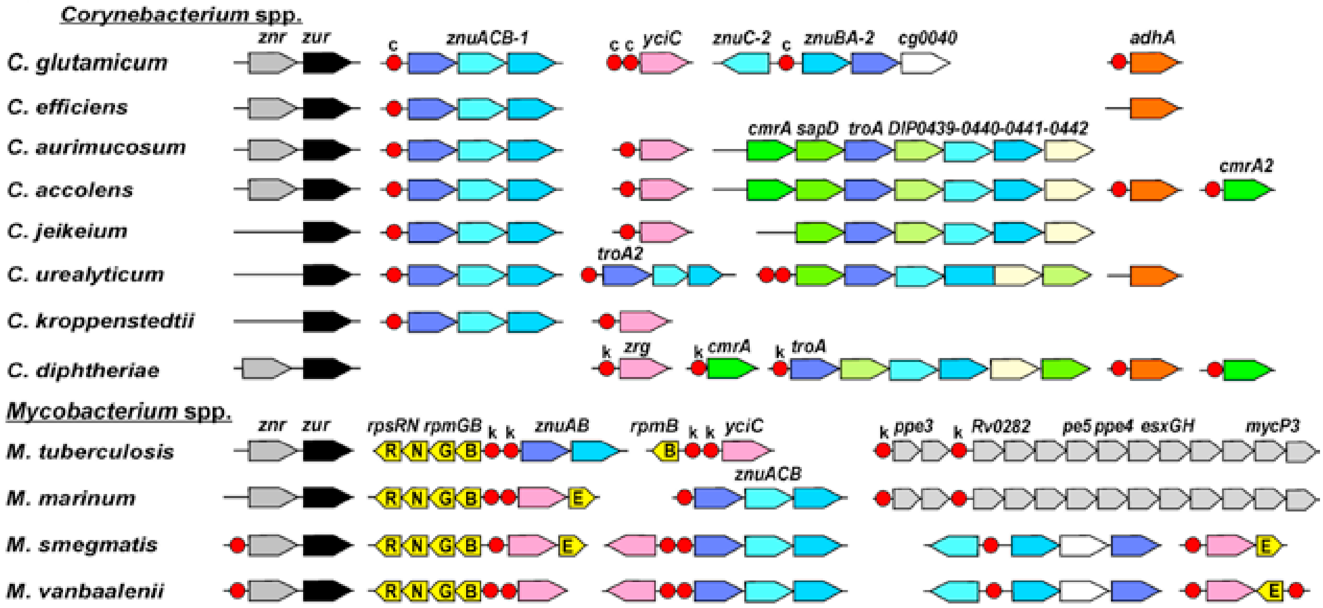


Figure 1: The predicted regulon of the zinc uptake regulator Zur in *C. glutamicum* ATCC 13032 and the corresponding regulog in a selected set of corynebacteria and mycobacteria. The conserved operator sequences of Zur proteins are marked by red circles and represented by the consensus sequence TAATGA(4)TNAT(4) CATTA [BMC Genomics 2010, 11(1):12].

operator motifs. This collection of lineage-specific regulons (termed regulog) allows the analysis of conservation of various regulon contents across a group of genomes. The regulog represents the main outcome of the application of comparative genomics for regulon reconstruction in a group of genomes. This type of computational prediction can be verified by measuring the differential expression of candidate regulon members under certain genetic or environmental conditions and by demonstrating *in vitro* the sequence-specific DNA-binding of the transcription regulator under investigation.

From the Genome Sequences *via* Regulons to the TRN of *C. glutamicum*

Over the last years, considerable information has been accumulated on transcriptional regulation in *C. glutamicum* and stored in the online reference database CoryneRegNet (*http://www.coryneregnet.de*). The combination of several computational methods revealed the transcriptional regulatory repertoire of *C. glutamicum*, consisting of at least 159 regulatory proteins. About 80 transcription regulators were characterized experimentally during the last years, leading to a comprehensive data set of almost 1000 interactions between regulatory proteins and their target genes. The reconstruction of the TRN from *C. glutamicum* with this data is facilitated by the graph visualization feature of CoryneRegNet. The reconstruction leads to a highly connected network of regulatory interactions that displays a modular and hierarchical structure without feedback regulation at the transcriptional level (Figure 2). The modularity of the TRN accounts for the robustness of the entire system in that way that damage is kept in a certain part of the network. The hierarchy of the TRN provides several executive levels to control gene expression and to link target genes of different functional context to respond jointly to an environmental signal. The TRN of *C. glutamicum* is currently among the most detailed reconstructions of regulatory interactions in bacteria, providing valuable insights into its global topological organization. Moreover, the regulatory repertoire of *C. glutamicum* serves for comparative analysis on gene regulation with other sequenced corynebacteria by genome-scale network transfer methods. By this means, a small core set of 24 regulatory genes present in all available corynebacterial genome sequences has been detected.

Further Studies: Transcriptomics, Comparative Genomics and TRN Evolution

Current projects in the Junior Research Group 'Systems Biology of Regulatory Networks' are focused on the sequencing of corynebacterial genomes and the subsequent utilization of the decoded genetic information for genome-wide transcription studies with DNA microarrays. Both industrially and medically relevant species of the genus *Corynebacterium* are under investigation. The research projects comprise (i) the aforementioned systematic reconstruction of the TRN from *C. glutamicum* to improve the industrial production of amino acids, (ii) the global transcription profiling of the skin inhabitant *C. jeikeium* to analyze its role in

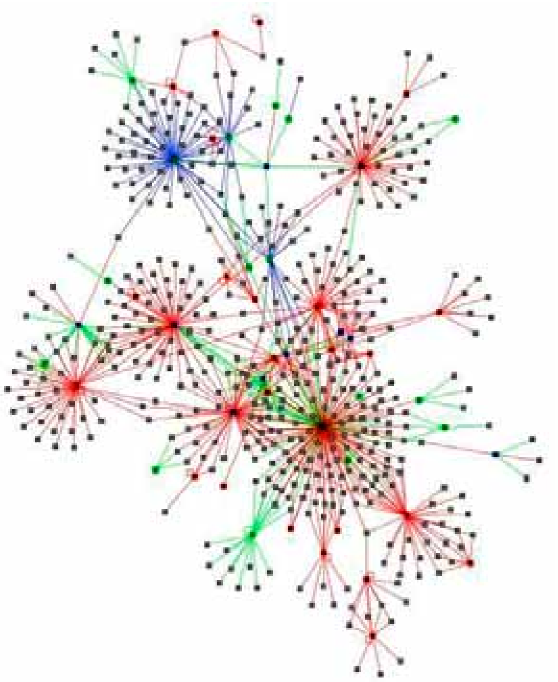


Figure 2: The reconstructed TRN from *C. glutamicum* ATCC 13032.

human body odor formation and (iii) comparative genomics of C. *diphtheriae*, the causative agent of diphtheria, to detect potential virulence factors. In the framework of the international 'Corynebacterium Genome Initiative' additional high-quality draft sequences of corynebacterial genomes are continuously generated by pyrosequencing at the CeBiTec, i.e. the genomes of 27 corynebacteria from various parts of the phylogenetic tree of the genus *Corynebacterium* have been decoded so far. The genome data provide comprehensive insights into the gene composition and metabolic capabilities of the specific corynebacteria and offer diverse possibilities for comparative genomics approaches. Moreover, this data can be used to unravel the (common) architecture and the evolution of the corynebacterial TRN. ■

Table 1: Corynebacterial genomes sequenced at the CeBiTec

Species	Strain	Genome Size	No. of Genes	GenBank No.
C. glutamicum	ATCC 13032	3,282,708 bp	3138	NC_006958
C. jeikeium	K411	2,462,499 bp	2165	NC_007164
C. urealyticum	DSM 7109	2,369,219 bp	2082	NC_010545
C. kroppenstedtii	DSM 44385	2,446,804 bp	2083	NC_012704
C. aurimucosum	ATCC 700975	2,790,189 bp	2602	NC_012590

The Technology Platform Genomics

Supporting Genome and Post-Genome Research by High-Throughput Technologies

The Technology Platform Genomics (TPG) was founded in the late 1990ies as an infrastructure to support the first genome sequencing projects carried out at the Department of Genetics, Faculty of Biology. Today, the TPG is part of the Institute of Genome Research and Systems Biology at the CeBiTec. The TPG today comprises four sections: genomics, transcriptomics, proteomics, and metabolomics, representing the most important technologies in genome and post-genome research.

Genomics

Genomics at the TPG mainly comprises genome sequencing. Two of the most recent high-throughput sequencing systems, the Roche *Genome Sequencer flx* and the Illumina *Genome Analyzer GA IIx*, are currently installed and running. These sequencing machines have been used to sequence the entire genomes of bacteria and yeasts and are currently applied to even larger genomes and complex DNA samples like meta-genomes.

On the first hand, these machines deliver huge sets of short, overlapping sequences (36 to 450 bases), that have to be assembled with the help of bioinformatics tools to larger, continuous stretches (contigs). At this stage, a 'draft' genome is formed that can be analyzed and annotated with the help of genome annotation software. Later, classical methods using PCR or large-insert clone libraries are used to fill the sequence gaps between the contigs, leading to a finished genome.

In the last years, the speed and capacities of sequencing machines have increased dramatically. Today, the *Genome Sequencer flx* delivers around 500.000.000 bases (500 megabases) per 10 hour run. This means that a medium bacterial genome is sequenced with more than 100fold coverage. Typically, the capacity that is sequenced in a single run is allocated to sequence up to 8 small, 4 medium-sized or 2 large bacterial genomes. Alternatively, a yeast genome might also be sequenced in one run. An advantage of this machine is the long read length (around 450 bases), that is very helpful in assembling genomes *de novo*, without the help of a reference sequence.

The *Genome Analyzer GA IIx* on the other hand delivers shorter reads, from 36 bases in a 48 hours run to two times 100 bases in a twelve days run. Although delivering shorter reads, this machine has the advantage to produce hundreds of millions, summing up to 55.000.000.000 bases (55 gigabases) in the long run. This data is normally used for re-sequencing, the assembly of reads with the help of a closely related reference sequence.

Transcriptomics

Transcriptomics at the TPG comprises all techniques used to analyze transcripts, with a focus on mRNA. This includes genome-wide determination of gene expression by microarrays as well as gene-specific, quantitative approaches by real-time *reverse-transcriptase* (RT)-PCR. The TPG has the equipment for printing series of microarrays with up to 20.000 features (gene probes) on a single slide. Probes are usually long, single-stranded oligonucleotides, each representing a single gene. The slides are then hybridized using fluorescent cDNA samples obtained from different strains (e.g. wildtype and mutant) or under different conditions (e.g. stressed and unstressed) and laser-scanned. Differential analyses by bioinformatics software deliver genes showing more or less transcript in individual strains or under specific environmental conditions.

Equipment to hybridize and scan self-printed microarrays and commercial high-density microarrays with up to a million features per slide (Agilent Microarray Scanner) is also available at the institute. With these machines, human genome microarrays are hybridized in projects with a medical focus.

Whereas genome-wide microarrays are an excellent tool for analyzing all genes of an organism, they usually have shortcomings with respect to sensitivity and to dynamic range. These obstacles are normally overcome by using the microarrays as a screening instrument and later testing the candidates derived from microarray analyses with the more sensitive and more specific real-time RT-PCR. This method is gene-specific, very sensitive and can be used in an absolute quantitative way. On the other hand, only smaller sets of genes can be analysed simultaneously.

Current transcriptomics methods are either comprehensive or quantitative. Therefore, the sequencing of complete transcriptomes by the novel high-throughput sequencers has the perspective to replace microarray analysis in the near future. Although being more expensive at the moment, transcriptome sequencing (RNA-Seq) delivers exact numbers for each gene and additionally gives information on transcription start sites, RNA processing and transcriptional organisation (operons).

Dr. Jörn Kalinowski

Jörn Kalinowski studied Biology at Bielefeld University. After completing his PhD at the Department of Genetics at Bielefeld University in 1990, he was appointed group leader for *Corynebacterium glutamicum* research and obtained a permanent position in 1993. From 1999 on, he was in charge for building up a technology platform for genome and post-genome research. Today this platform is the Technology Platform Genomics at the Institute for Genome Research and Systems Biology of the CeBiTec. Beside his function as head of the Technology Platform, he is member of the Executive Board of the CeBiTec.

Figure 1: Loading sequencing samples on an Illumina Genome Analyzer GA IIx.

Figure 2: Setting up a robot for microarray printing in the clean room of the Cebitec building.

Proteomics

Proteomics at the TPG comprises gel-based and non-gel-based separation of proteomes as well as identification and quantification of proteins by tryptic fingerprints and MALDI-TOF mass spectrometry or analysis of protein modification by MALDI-TOF MS/MS or by LC-ESI mass spectrometry.

In contrast to DNA or RNA, proteins represent a more heterogeneous population of molecules and there is no single method to separate all proteins. Generally, a first step in proteomics is to fractionate the cellular proteomes. In bacteria, the extracellular proteome can be separated from the outer membrane proteome, the inner membrane proteome and the cytosolic proteome fraction. In higher organisms, more subcellular compartments and organelles are present and even more proteomic fractions can be generated. After fractionation gel-based separation is applied routinely. This includes one-dimensional and two-dimensional gel-electrophoresis.

Although being able to separate many proteins by their isoelectric points and their sizes, 2D-gel electrophoretic separation has limitations with respect to proteins having an isoelectric point above 7.0 and (hydrophobic) membrane proteins. Specifically for membrane proteins, special techniques for solubilisation are required and only 1D-gels are suitable for membrane protein separation.

After staining, proteome or sub-proteome gels can be compared and stained spots representing single protein species can be identified on the basis of a tryptic fingerprint and subsequent MALDI-TOF mass spectrometry (Bruker ultrafleXtreme). By this method, the masses of trypsin-generated peptides obtained from proteins that have been eluted from gel slices are determined exactly, and comparison of these masses to peptide mass databases generated from genome sequences allow the easy and sure identification of the respective protein.

Alternatively, proteins can be identified also by liquid-chromatography (LC) separations. In this method, proteins are trypsin-digested first, then the peptide mixture is separated by HPLC and later injected into a mass spectrometer. In this mass spectrometer, the exact masses of the peptide itself (mother ion) and those of specific fragment ions were recorded and used for identification of the peptide and the respective protein. Mass spectrometers applicable for such analyses at the TPG comprise an ESI ion trap mass spectrometer (Thermo LCQ Deca) and a triple quadrupol mass spectrometer (Bruker qTOF).

Metabolomics

Metabolomics at the TPG comprise metabolic profiling and flux analysis by gas-chromatography (GC) or liquid-chromatography (LC) coupled to mass spectrometry (MS). Metabolites are chemically even more diverse than proteins. Therefore, no single method is available to separate and analyse all metabolites from a cell or a tissue. In addition, all methods have their specific advantages and drawbacks. At the TPG, routinely GC-MS analysis is applied to samples obtained from bacteria, plants and even from humans. GC-MS is very sensitive and able to separate hydrophobic molecules. Hydrophilic molecules can be made more hydrophobic by chemical derivatization and then also be separated by GC-MS. The coupling to mass spectrometry again allows the determination of masses of specific ions and comparison with mass spectral databases helps in identification of a specific substance. The TPG is running two GC ion trap mass spectrometers (Thermo ITQ) and a special two-dimensional GC-TOF MS machine (LECO Pegasus IV).

However, a number of very hydrophilic molecules can not be separated by GC-MS. In addition, derivatization quickly increases the size of a molecule to a level that can no longer be analysed by mass spectrometry. Therefore, LC-MS is an ideal complement to GC-MS, since it enables the separation of underivatized and water-soluble molecules.

Figure 3: A scientist inspecting a two-dimensiol protein gel in front of an opened MALDI mass spectrometer.

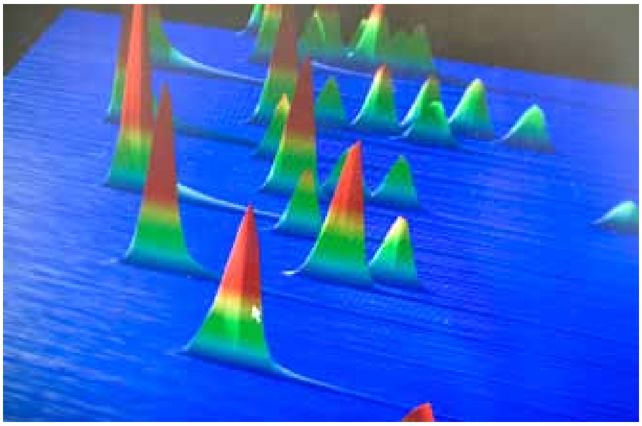


Figure 4: A two-dimensional GC-MS run showing the abundant metabolites as peaks.

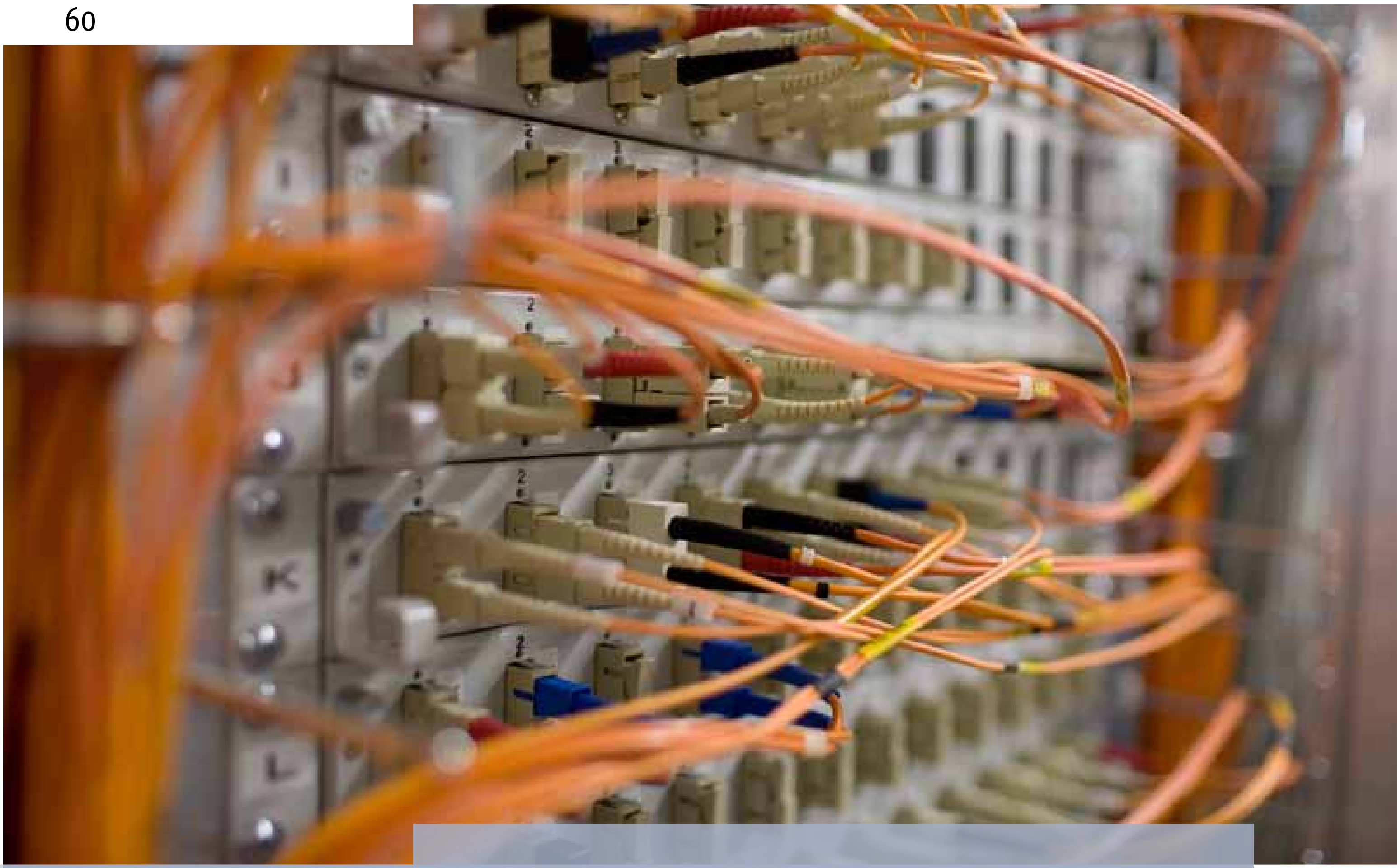
Currently, the main obstacle in metabolomics is the lack of pure reference substances that can be used for calibration and to identify metabolites in a complex sample obtained from cells or tissue. Today, only around 150 molecules from cellular metabolism can be identified unequivocally. The rest of peaks in the metabolome samples can only be approximately determined by comparison to MS databases. There is an urgent need to improve this situation by enabling *de novo* identification of metabolites. A second problem consists in the preparation of the metabolome samples. Since key metabolites have half-lives in the sub-second range, cellular metabolism has to be stopped (quenched) immediately. This issue is not been solved completely at the moment, and quenching methods need to be adapted to the individual biological system.

Data Integration and Systems Biology

Data Integration and Systems Biology represent the major goals to be achieved by using the high-throughput data generated by genomics, transcriptomics, proteomics and metabolomics. An important prerequisite is a maximally controlled cultivation. For microorganisms and cell cultures, fermentation under conditions of optimal temperature, aeration, pH and nutrient supply assures a very reproducible sampling prior to -omics analyses. With the data generated, Bioinformatics is the key to analyse and integrate the data from genome, gene expression and metabolism into models of cellular systems and subsystems. With these integrated models, simulations can be performed and predictions allow to extend our knowledge on the interactions and dynamics of living systems (Systems Biology). ■

cDNA	(copy DNA); a single-stranded, complementary copy of an original RNA molecule
GC	(gas chromatography); separation of hydrophobic molecules; can be coupled to mass spectrometry
ESI	(electrospray ionization); method for generating ions of the analytes prior to mass spectrometry
HPLC	(high-pressure liquid chromatography); method for separation of molecules; can be coupled to mass spectrometry
MALDI	(matrix-assisted laser desorption ionization); method for generating ions of the analytes prior to mass spectrometry
MS	(mass spectrometry); method for identifying molecules on the basis of their molecular mass
MS/MS	(tandem mass spectrometry); consecutive rounds of analyses of fragment and sub-fragment ions
LC	(liquid chromatography); separation of hydrophilic molecules; can be coupled to mass spectrometry
PCR	(polymerase chain reaction); method for exponential amplification of DNA and cDNA
RT-PCR	(real-time PCR); method for quantification of DNA or cDNA through kinetic analysis of the PCR process

Glossary: The text contains a number of abbreviations that are explained above.



Central Hardware and Software Infrastructure of the Bioinformatics Resource Facility

Central Hardware Resources and Data Management

The Bioinformatics Resource Facility (BRF) at the CeBiTec provides general hardware and software support for research groups within genome and post genome projects. Therefore, the BRF has continuously extended the required data storage and compute capacities as illustrated in Figure 1. Today, the current BRF storing capacities add up to 165 Terabytes disk storage and 280 Terabytes tape storage. A high-performance compute cluster with more than 700 CPUs (more than 1470 CPU cores) and an overall capacity of 9.5 Teraflops is available for large-scale computations such as whole genome annotations.

All generated raw experimental data sets e.g. from ultrafast sequencing or high-throughput metabolomics and results derived from their bioinformatics analysis are stored redundantly on the central bioinformatics platform in a well-structured and systematic way. Regular backups of all data and long-term archival using a Hierarchical Storage Management (HSM) system guarantee the availability and persistence of all valuable raw data sets and results derived from their analyses. An experienced team of six expert IT system administrators maintains and optimizes all hardware components. They also integrate new hardware components into the existing platform and adapt the configuration in order to optimize the overall performance.

In addition to the above mentioned general data management tasks, the BRF provides local access to more than 300 bioinformatics tools, 30 sequence databases, and other biological data repositories that are updated on a regular basis. Access to the Bielefeld infrastructure is granted and controlled via the General Project Management System (GPMS) employing a role based access mechanism. Access to data is only possible with individual user name and password authentication and all data transfer is SSL-encrypted (https).

The Bielefeld software suite for genome research

During the GenoMik funding period of the competence network 'Genome Research of Bacteria for the Environment, Agriculture, and Biotechnology', Bielefeld University's competence center has played a crucial role when processing the network's genome and post-genome projects. Bielefeld has become a national resource for all GenoMik-Plus projects during the period of time. The BMBF has provided the required funding for both staff and equipment. Furthermore, the group participated in many other national and international research projects funded by the DFG, BMBF, and EU.

Genome Annotation, Sequence Analysis and Metabolic Reconstruction

Services of the BRF platform concerning genomics comprise the maintenance and annotation of complete genomes as well as the evaluation of other sequence data employing the *GenDB* and *SAMS* software. While the GenDB system was successfully used for the automatic and manual annotation of more than 50 microbial genomes, SAMS is routinely applied for detailed analyses of large sequence sets (ESTs or metagenome reads). Based on the genome annotation data provided by GenDB, the *CARMEN* application can perform a metabolic reconstruction. CARMEN provides two main applications: the generation of *de novo* models based on KEGG database information and the creation of template based SBML models for comparative genomics. An organism-specific network model of individual or large-scale metabolic pathways can be reconstructed based on genome annotation data that can be obtained from GenDB or alternatively loaded from NCBI GenBank files. The typical workflow for analyzing a novel genome is illustrated in Figure 3.

Software support for transcriptomics, proteomics, and metabolomics

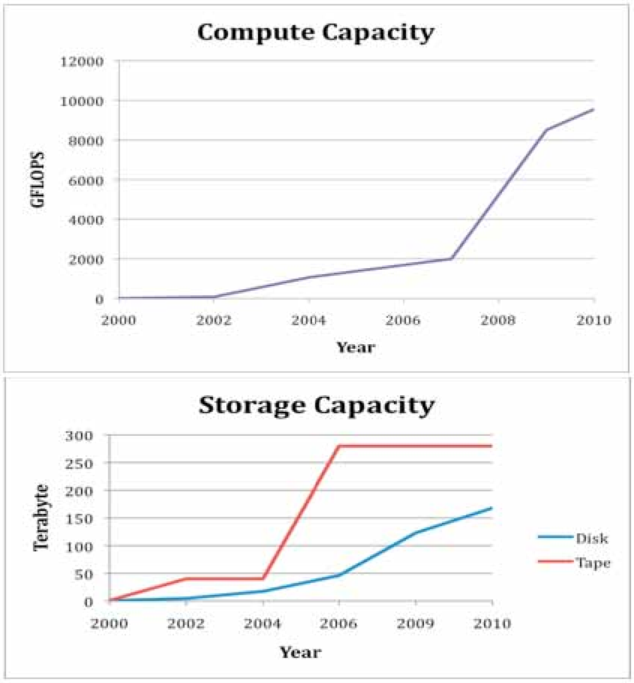
In the field of transcriptomics, the *EMMA* software is provided as a MAGE-compliant software platform for the evaluation of data resulting from genome-wide transcriptomics studies. So far it has been used for analyzing microarray data of more than 20 organisms. Detailed experimental setups and protocols as well as all raw data sets are stored in a separate LIMS component (*ArrayLIMS*). EMMA 2 allows mapping of gene expression data onto proteome data or pathways and vice versa. It provides extensible analysis and visualization plug-ins via the R-language. Normalization of single and multiple microarrays and statistical tests for inferring differentially expressed genes are provided. Co-regulated genes can be detected easily through the integrated cluster analysis methods.

Dr. Alexander Goesmann

Alexander Goesmann studied informatics in the natural sciences at Bielefeld University, focusing early on bioinformatics and genome research. After completing his PhD at the Technical Faculty of Bielefeld University in 2004, he continued to work within the BMBF funded competence centre of the national network 'Functional genome research on bacteria relevant for agriculture, environment and biotechnology. He is now head of the junior research group for Computational Genomics at the Bielefeld Center for Biotechnology (CeBiTec) and executive director of the Bioinformatics Resource Facility (BRF).

In proteomics, tandem mass spectrometry (LC-MS/MS) in combination with isotopic labeling techniques provides a common way to obtain a direct insight into regulation at the protein level. The Internet application *QuPE* provides comprehensive data management and analysis functions for these experiments. Starting with the import of mass spectra data the system guides the experimenter through the process of protein identification by database searches, the calculation of protein abundance ratios, and, in particular, the statistical evaluation of the quantification results including multivariate analysis methods such as analysis of variance or hierarchical cluster analysis. Furthermore, a well-defined programming interface facilitates the integration of novel approaches.

To facilitate the systematic storage, analysis and integration of metabolomics experiments, we have implemented *MeltDB*, a web-based software platform for the analysis and annotation of datasets from metabolomics experiments. MeltDB supports open



F igure 1: Since its start in the year 2000, the BRF has continuously expanded the central data storage and compute capacities. As illustrated above, the online disk storage capacities roughly doubled every 2 years, while the compute capacities even quadrupled over the last 2 years.

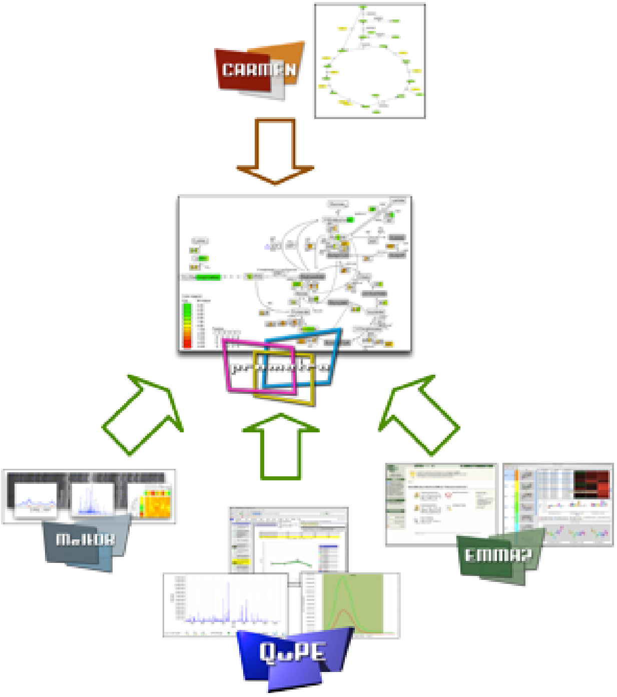


Figure 2: The integrated functional genomics software platform established at the BRF allows to combine and visualize quantitative data from transcriptomics, proteomics, and metabolomics experiments on user defined or reconstructed metabolic pathway maps. The CARMEN applicaton delivers the reconstructed metabolic pathways that are visualized and enriched via the ProMeTra application. ProMeTra directly accesses the quantitative data from metabolomics experiments stored in MeltDB and can combine the results with data from quantitative proteomics experiments organized in QuPE. The third functional genomics platform Emma2 for transcriptomics experiments is integrated as well and shares the experimental data via web services technology.

file formats (netCDF, mzXML, mzDATA) and facilitates the integration and evaluation of existing preprocessing methods. The system provides researchers with means to consistently describe and store their experimental datasets. Comprehensive analysis and visualization features of metabolomics datasets are offered to the community through a web-based user interface. The system covers the process from raw data management to the visualization of results in a knowledge-based background and is integrated into the context of existing software platforms of the Bioinformatics Resource Facility at Bielefeld University.

All components are linked via the *BRIDGE* integration layer providing an interface for further data mining, modeling, simulation, and visualization approaches. The *ProMeTra* application is one example for the ongoing development of novel software tools to support researchers in the field of systems biology. ProMeTra uses web services technology to enrich metabolic pathway maps with quantitative results stored in the omics platforms MeltDB, QuPE, and Emma2 as presented in Figure 2.

High-throughput analysis and data integration

The strength of Bielefeld's Bioinformatics lies in various areas of genome and post-genome research of microorganisms. Strict and comprehensive quality assurance mechanisms that are applied throughout the analysis process are the prerequisite for the creation of high-quality genome sequences. In particular, the efforts of optimizing and automating high-throughput analysis have to be stressed in this context. Using the current version 2.4 of the GenDB software, the automatic annotation of an average-sized bacterial genome can be completed within a few hours on the current compute cluster. As illustrated in Figure 3 the automatic annotation rapidly delivers the core information about a genome and serves the required enzymatic data for the CARMEN software to reconstruct major metabolic pathways automatically. As a result, new knowledge about the key features of a newly sequenced organism can be directly inferred by further inspecting the reconstructed pathway charts.

As more and more experimental results are generated for each organism under investigation, the integration of the individual, specialized software components is getting increasingly important. In addition to the software packages described above, the enhancement of existing tools as well as new software development are essential for focussing research and project work towards systems biology. We are therefore adapting the GenDB

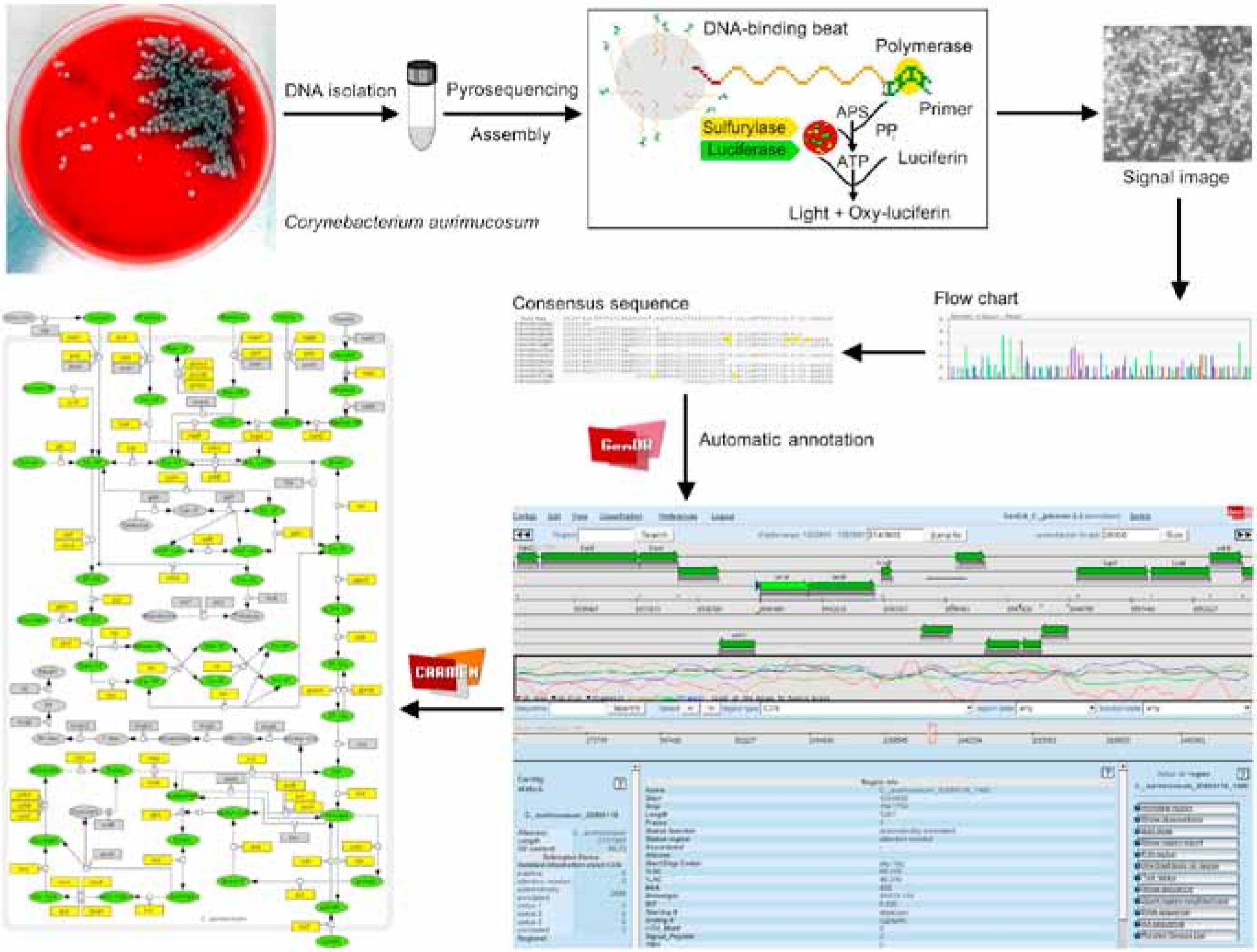


Figure 3: Schematic representation of an ultrafast genome analysis workflow. Subsequent to DNA isolation and purification, the pyrosequencing technology by Roche/454 is used to generate a large number of short sequence reads. After assembling the genome sequence an automated genome annotation is performed using the GenDB genome annotation system. Finally, the obtained information can be used to reconstruct metabolic pathways with CARMEN.

genome annotation platform for the analysis and annotation of eukaryotic genomes and we will continue our efforts to develop novel tools for the detailed analysis of metagenome data sets.

Furthermore, we have recently started to utilize modern graphic adapters that provide a compute power that exceeds the performance of modern CPUs by orders of magnitude. The special architecture of graphic processing units (GPUs) allows calculating hundreds of identical operations in parallel. Technologies like CUDA (Compute Unified Device Architecture) or OpenCL (Open Computing Language) provide an interface for the implementation of scientific algorithms on graphics hardware. As an example, this approach is ideally suited to process millions of small-scale alignments, as needed for the mapping of short reads to a reference sequence. To solve this task we have implemented the algorithm SARUMAN using GPU programming. As SARUMAN is not using a heuristic approach, it finds all possible alignment positions and always returns the perfect alignment.

In total, more than 2,000 registered internal and exernal users have currently access to the hardware and software resources of the BRF. Additionally, tools of the Bielefeld bioinformatics software are deployed at over 30 sites, both national and international.

A list of publications that document the successful cooperations with international partners can be found on our homepage at *http://www.cebitec.uni-bielefeld.de/brf*. ■

Location

→ By train:

Bielefeld is easy to reach both by car and by train: every hour an intercity train on the route from Cologne/Bonn to Berlin stops at Bielefeld Hbf. Then you take *Stadtbahn Linie 4* towards *Universität* (7 minutes).

→ By plane:

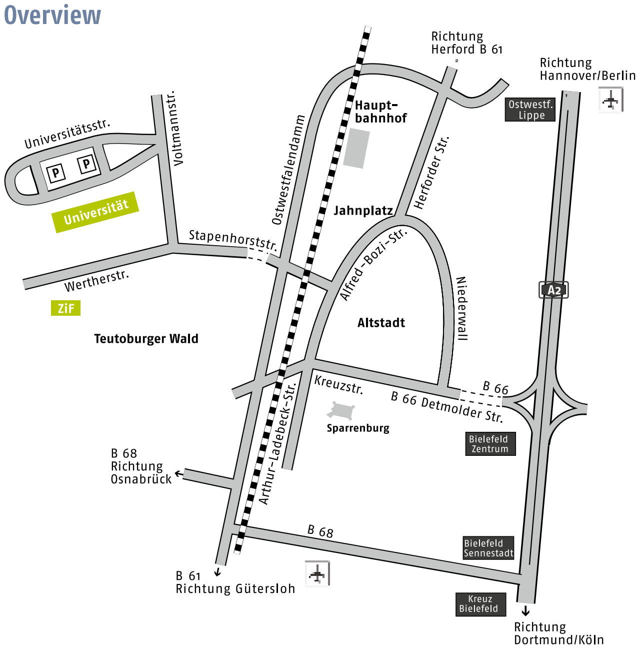
The closest airports to Bielefeld are Paderborn/Lippstadt and Hannover.

→ By car:

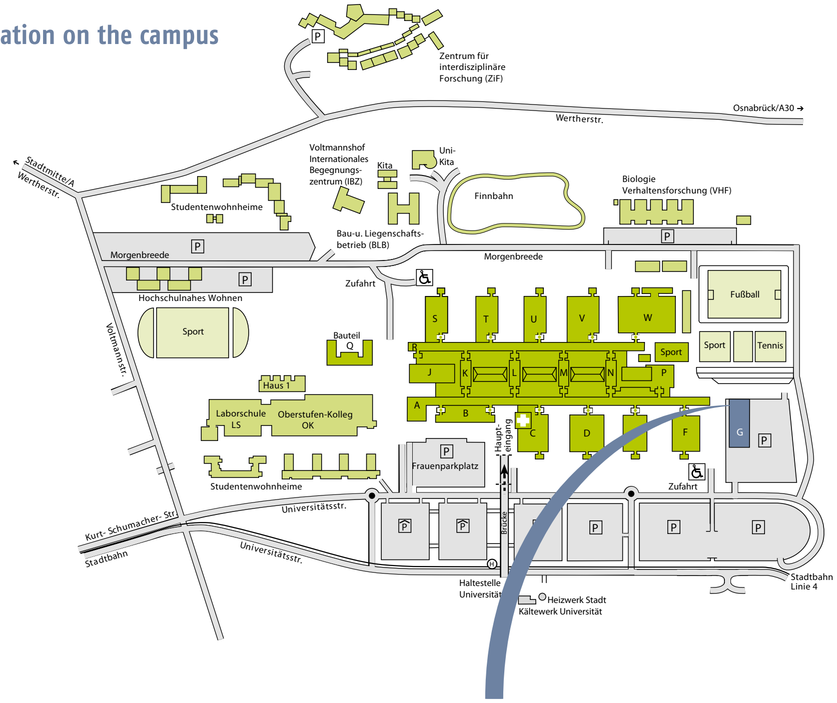
You can take the A2 from Dortmund to Hannover, exit at *Bielefeld-Zentrum*, follow the street signs towards the centre (*Zentrum*), and from there the University (*Universität*) is signposted.

The CeBiTec laboratory building is labeled as part G on the plans of the University.

Overview



Location on the campus



Entrance of the CeBiTec building

Editor
Bielefeld University
Center for Biotechnology
Universitätsstraße 27
D-33615 Bielefeld

Conceptual Design
Prof. Dr. Alfred Pühler
Dr. Stefan Weidner

Layout/Design
Susanne Konermann
Kim-Christian Meyer

Photos
Martin Brockhoff
Peter Husemann
Kim-Christian Meyer
Olena Perlova (Genomxpress)
Stefan Pelzer (Genomxpress)
Bielefeld University

Date of publication
May 2010

Universität Bielefeld
Centrum für Biotechnologie
Universitätsstraße 27
D-33615 Bielefeld
fon   +49 521.106-87 52
fax   +49 521.106-89 041

→ **www.cebitec.uni-bielefeld.de**