

Bioinformatik & Bioinformation

Braunschweig spielt eine führende Rolle

Dietmar Schomburg, Edgar Wingender

Abteilung Bioinformatik & Biochemie
IB³, Institut für Biochemie, Biotechnologie & Bioinformatik
Langer Kamp 19B
38106 Braunschweig
D.Schomburg@tu-bs.de

Abstract: Die Bioinformatik ist eine vor ca. 20 Jahren entstandene wissenschaftlicher Disziplin, die Informatik-Werkzeug für die Bearbeitung biologischer Fragestellungen einsetzt. Insbesondere die Entwicklung neuer Algorithmen und Datenbanken und die Integration außergewöhnlich umfangreicher Datenmengen spielen hier eine entscheidende Rolle. Die schnelle Entwicklung der Bioinformatik hat die heutige biowissenschaftlicher Forschung erst ermöglicht. Bei der Etablierung des Gebietes in Deutschland und im Hinblick auf die internationale Sichtbarkeit hat Braunschweig eine führende Rolle gespielt.

1 Die ersten Jahre bis 1990 – Etablierung einer neuen Disziplin

Die Entwicklung und Anwendung von Algorithmen und Datenbanken im Bereich der molekularen Biowissenschaften begann im Bereich der Strukturbiologie in den frühen 1960er Jahren und im Bereich der Biochemie/Molekularbiologie mit der Bestimmung der ersten Proteinsequenzen im Jahrzehnt zwischen 1970 und 1980. Sowohl die Verarbeitung der Röntgenbeugungsdaten als auch eine Auswertung und ein Vergleich der Protein- und Nucleotidsequenzen war ohne den Einsatz von Computern nicht möglich. Hier setzte sich im englischen Sprachgebrauch der Begriff “Biocomputing” durch, in Deutschland sprach man anfangs von “Computereinsatz in den Biowissenschaften”.

1.1 Datenbanken

Schnell wurde deutlich, dass nicht nur die rechnergestützte Auswertung der experimentellen Daten ein Gebot der absoluten Notwendigkeit war, sondern dass auch die weltweite Zugänglichkeit der biomolekularen Daten gesichert werden musste. So entstanden weltweit die Datenbanken für Nucleotidsequenzen, Proteinsequenzen,

Protein-3D-Strukturen und Enzymdaten in der zweiten Hälfte der 1980er Jahre, letztere im Jahr 1987 im Arbeitskreis Schomburg der damaligen Gesellschaft für biotechnologische Forschung, dem heutigen HZI. Die ersten Medien der Verbreitung dieser Daten vor Existenz des Internets, in Form von Büchern, später über Magnetbänder, "Floppy Discs", CDs etc. erscheint heute schon vorsintflutlich.

1.2 Algorithmen

Neben dem Start der Enzymdatenbank konzentrierte man sich in der von Schomburg geleiteten GBF-Abteilung "Molekulare Strukturforshung" auf Algorithmen- und Software-Entwicklung im Bereich Protein-3D Strukturvorhersage, molekulare Graphik und Computer-unterstütztes Protein-Design. Jahrelang war das Protein-Modelling Programm BRAGI eines der am meisten verwendeten Programme auf diesem Gebiet. Die entwickelten Methoden wurden in einem gemischt theoretisch/experimentellen Ansatz in der gleichen Gruppe erfolgreich z.B. In der Entwicklung von neuen Biokatalysatoren und potentiellen Medikamenten z.B. Zur Behandlung von septischem Schock eingesetzt.

1.4 Start der jährlichen Bioinformatik-Tagungsreihe in Braunschweig

Schon frühzeitig wurde erkannt, dass bei den wenigen zur Verfügung stehenden Wissenschaftlern auf diesem Gebiet, das in den nächsten Jahren die weitere Entwicklung der modernen Biowissenschaften dominieren und es erst ermöglichen sollte, eine enge nationale und internationale Vernetzung essentiell ist. So wurden im Jahr 1985 eine Tagungsreihe ins Leben gerufen, die 2011 zum 26. mal stattfand, anfangs in deutscher Sprache mit 1985 vierzig Teilnehmern, ab 1993 dann als "German Conference on Bioinformatics" (GCB) international aufgestellt und normalerweise mit ca. 250-300 Teilnehmern. Die ersten vier Jahrestagungen fanden in Braunschweig statt, seitdem wechseln die Tagungsorte, insgesamt gelang es 8mal, die Jahrestagungen nach Braunschweig zu holen.

2 1990-1996 – Rasante Entwicklungen

Nachdem Ende der 1980er Jahre die Entscheidung gefallen war, das menschliche Genom vollständig zu sequenzieren, wurde die entscheidende Bedeutung der Bioinformatik für die weitere Entwicklung der Lebenswissenschaften einer Vielzahl von Wissenschaftlern und auch den Forschungs-fördernden Institutionen deutlich.

Initiiert durch Prof. Schomburg und von diesem die ersten 13 Jahre lang geleitet, entstand in Deutschland ein zentraler Bioinformatik Koordinations-Kreis, der durch delegierte Bioinformatik Experten aus den wissenschaftlichen Organisationen GBM, GDCh, Dechema, GI und GMDS besetzt war und bis heute existiert und bei der Dechema angesiedelt ist. Der Arbeitskreis spielte in den folgenden Jahren eine entscheidende Rolle bei der Initiierung neuer Förderkonzepte des BMBF und der DFG.

In Zusammenarbeit mit dem Bundesministerium für Forschung & Technologie (BMFT, heute BMBF) entwickelte der Koordinationskreis ein zukunftsweisendes Förderkonzept "Molekulare Bioinformatik", in dem Kooperationsprojekte gefördert wurden, die von je einem Arbeitskreis aus den Lebenswissenschaften und einem Arbeitskreis aus Informatik/Mathematik bearbeitet wurden. Dieses Förderkonzept wurde in seiner Durchführung von Prof. Schomburg GBF Braunschweig und Prof. Lengauer, Universität Bonn, koordiniert. Hier spielte Deutschland eine absolute Vorreiterrolle, und das Konzept wurde später von einer Reihe von Staaten, unter anderem Großbritannien, kopiert.

Katalysiert durch die Möglichkeit der gemeinsamen Projektfinanzierung entstand dann die seitdem in Deutschland sehr starke Bioinformatik-Forschungsszene. Es entstanden enge Kooperationen der von Prof. Schomburg geleiteten GBF-Abteilung mit den Informatik-Arbeitskreisen Ehrich an der TU Braunschweig, Kriegel an der Universität München und Sagerer in Bielefeld. Gemeinsames Forschungsthema mit den letzten beiden war die Vorhersage von funktionellen Protein/Protein Interaktionen unter Entwicklung neuer Algorithmen. Mit Prof. Ehrich wurde im Rahmen der Doktorarbeit von Karl Teille eine Pilotstudie für die Anwendung von objektorientierten Datenbankansätzen auf die Enzymdatenbank BRENDA erarbeitet.

Die methodischen Entwicklungen und erfolgreichen Projekte auf dem Gebiet des Computer-unterstützten Protein-Design wurden von der Industrie aufmerksam registriert und es kam zur Bildung eines von GBF und Industrie kofinanzierten Konsortiums, die das CAPE (Centre for Applied Protein Engineering), lokalisiert an der GBF als Zentrum für die gemeinsame Methoden- und Software-Entwicklung mit Prof. Schomburg als Leiter gründet. Hier arbeiteten zwischen 1989 und 1995 acht Firmen aus dem Pharma- und Biotechnologie-Bereich mit zwei Computer-Hardware Firmen und der GBF an innovativer Methodenentwicklung im Bereich der Struktur-Bioinformatik.

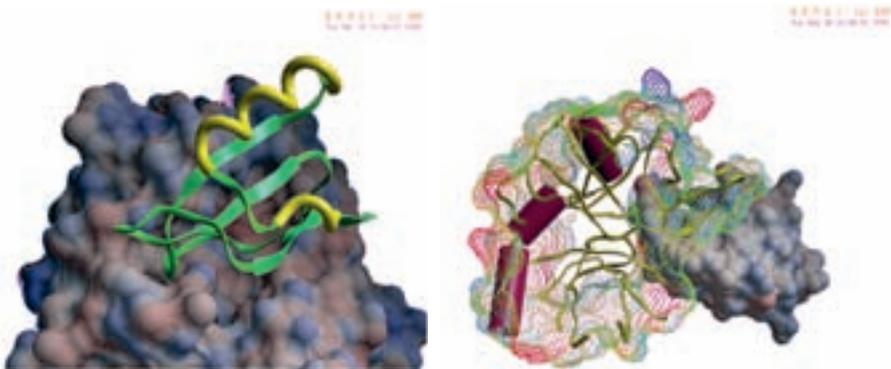


Abb1: Vorhersage und Optimierung von Protein/Protein Wechselwirkungen, Modellierung und Darstellung mit BRAGI

Eine bereits 1988 von E. Wingender (damalige GBF, jetziges HZI) publizierte und kontinuierlich erweiterte tabellarische Übersicht über eukaryote Transkriptionsfaktoren und deren genomische Bindungsstellen wurde Anfang der 90er Jahre als Datenbank implementiert und unter dem Namen TRANSFAC verbreitet [Wi88; WHL91]. Im Rahmen eines der ersten BMBF-geförderten Bioinformatik-Projekte (GENUS, 1993-1997; koordiniert von E. Wingender) wurde diese Datenbank quantitativ wie konzeptionell deutlich erweitert und entwickelte sich zu einem „Gold Standard“ auf ihrem Gebiet [Wi08]. Sie ist bis heute die wichtigste Quelle zur Identifizierung bekannter und potentieller Transkriptionsfaktor-Bindungsstellen in eukaryoten Genomen.

3 1997 – 2007 Braunschweig wird Standort eines der Bioinformatik-Zentren

Zum Ende des Jahres 1996 wechselte Prof. Schomburg an die Universität zu Köln. Seine Nachfolger an der GBF, später HZI, konzentrierten sich auf die experimentelle Strukturforschung, so dass dieser Forschungsbereich in Braunschweig ein abruptes Ende fand. Auch die Entwicklung der Enzymdatenbank BRENDA wurde nach Köln verlagert.

Wieder initiiert durch Prof. Schomburg und Prof. Lengauer wurde ein DFG Schwerpunktprogramm „Informatikmethoden zur Analyse und Interpretation großer genomischer Datenmengen“ 1996 beantragt und 1997 bis 2005 durchgeführt. Schließlich, wieder durch den Dechema-Koordinierungsausschuss dem BMBF vorgeschlagen, gab es eine Ausschreibung des Bundesministerium für Bildung und Forschung (BMBF) zur Einreichung von Vorschlägen zur Einrichtung von Forschungszentren auf dem Gebiet der Bioinformatik. Nach hochkarätiger internationaler Begutachtung wurden sechs Anträge ausgewählt und es wurden Zentren u.a. an der Universität Köln (CUBIC, Cologne University Bioinformatik Centre, Direktor Prof. Schomburg) und als Kooperationsprojekt zwischen GBF/HZI in Braunschweig „Intergenomics“. Die Zentren wurden 2002-2006 vom BMBF großzügig finanziert. Des Weiteren gab es 5 DFG-geförderte Bioinformatik Zentren, die sich auf den Aufbau neuer Studiengänge konzentrierten.

2001 wurde in Braunschweig das vom BMBF geförderte Bioinformatik-Kompetenzzentrum „Intergenomics“ gegründet. Im Zentrum des wissenschaftlichen Interesses stand die Bioinformatik der Interaktionen genomkodierter Prozesse von miteinander wechselwirkenden Organismen. Ziel des Zentrums war die in der Region bioinformatisch tätigen Organisationen zusammenzuführen und die Bioinformatik auch als Lehrfach an der TU Braunschweig und der FH Braunschweig-Wolfenbüttel zu etablieren. Die Partnerschaft des Zentrums umfasste neben TU (drei Institute) und FH noch die GBF (fünf Arbeitsgruppen), bei der auch die Koordination lag, sowie die BIOBASE GmbH (s.u.). Der Koordinator E. Wingender wechselte während der Förderperiode an die Universität Göttingen. Als Ergebnis des Intergenomics-Zentrums wurden nicht nur bestehende bioinformatische Aktivitäten insbesondere an der GBF und der TU deutlich verstärkt und eine wissenschaftliche Infrastruktur für den Bereich Bioinformatik aufgebaut, sondern auch ein neuer Lehrstuhl für Bioinformatik an der TU geschaffen, auf den 2007 D. Schomburg berufen wurde (s.u.).

4 2007-2012 Etablierung der Bioinformatik an der TU

Im Rahmen der BMBF-Förderung des Braunschweiger Biozentrums hatte die TU die Einrichtung von zwei Bioinformatik-Lehrstühlen, je einem in der Gauß Fakultät und der Fakultät für Lebenswissenschaften, zugesagt. Als erster wurde an der Fakultät für Lebenswissenschaften Prof. Schomburg für den Bereich Bioinformatik und Biochemie unter gleichzeitiger Einrichtung einer namensgleichen Abteilung am Institut für Biochemie, Biotechnologie und Bioinformatik berufen.

Die Forschungsaktivitäten des Schomburg Arbeitskreises sind im weitesten Sinne auf dem Gebiet des Stoffwechsels angesiedelt. Dies beinhaltet sowohl die Weiterentwicklung des Enzym-Informationssystems BRENDA sowie die Entwicklung und Anwendung von Methoden zur Modellierung von metabolischen Netzen. Unterstützt werden die Arbeiten, wie in der Schomburg Gruppe üblich, durch experimentelle Arbeiten.

4.1 Das Enzym-Informationssystem BRENDA

Die molekularen Biowissenschaften stellen heute den Bereich der Forschung dar, in dem mit Abstand die größten Datenmengen entstehen und verarbeitet und zugänglich gemacht werden müssen. Nachdem im Jahr 2001 eine Rohversion des entschlüsselten menschlichen Genoms veröffentlicht wurde, ist auf diesem Gebiet eine Explosion der Daten eingetreten. Die Kosten dieses ersten menschlichen Genoms mit ca. 3 Milliarden Basen als Informationseinheiten beliefen sich auf mindestens 3 Milliarden US Dollar. Während es unter Beteiligung einer Vielzahl von Laboratorien aus 40 Ländern in ca. 11 Jahren entschlüsselt wurde, ist für Ende des Jahres 2012 eine Technik angekündigt, die die Sequenzierung eines menschlichen Genoms für unter 1000 Dollar erlaubt.

Parallel zur schnellen Entwicklung der experimentellen Möglichkeiten entstanden eine Vielzahl von über das Internet zugänglichen Datenbanken, die diese der wissenschaftlichen Community zugänglich machten. Nur sehr wenige davon konnten aber über längere Zeit überleben und haben internationale Bedeutung erlangt. Hierzu gehört ohne Zweifel BRENDA. In einer vom „European Bioinformatics Institute“ in Auftrag gegebenen Umfrage unter Lebenswissenschaftlern bezeichneten ca. 45% die Datenbank als essentiell für ihre Forschung. Wer „BRENDA AND enzym“ googelt, findet ca. 2.5 – 3 Millionen Verweise im Internet auf die Datenbank. Pro Monat werden bis zu 200 000 Nutzer registriert.

Die Datenbank wird heute in einer Kombination von manueller Literatursuche und Text-Mining Verfahren laufend aktualisiert. Weiter wurde eine komplexe Software für die Query-Engine entwickelt, die z.B. strukturbasierte Anfragen erlaubt. Die integrierten Daten werden über Data Mining Verfahren für den Benutzer aufgearbeitet und mit Vorhersage Tools ergänzt. Die Gruppe ist über die Entwicklung der BTO (BRENDA Tissue Ontology) eng in die internationale Entwicklung biologischer Ontologien eingebunden.

Die Datenbank wurde in enger Kooperation mit dem European Bioinformatics Institute in Hinxton, England von der EU finanziert. Die weitere Existenz ist durch BMBF und niedersächsische Förderung für die nächsten Jahre gesichert. Eine kommerzielle IN-House Version für die Industrie wird von der Firma Enzymeta in Erfstadt/Rheinland entwickelt und durch Biobase, Wolfenbüttel vertrieben.

4.2 Systembiologie und die Modellierung zellulärer Netzwerke, Algorithmenentwicklung und Anwendung

Beginnend im Jahr 2003 wurde in der Schomburg Gruppe erst in Köln, dann in Braunschweig Software für die Systembiologie entwickelt und in Projekten angewandt. Das wissenschaftliche Ziel besteht in der Anwendung von metabolischen Netzen auf die Entwicklung von Medikamenten gegen bakterielle Infektionen, Optimierung von biotechnologischen Produktionsprozessen und dem Verständnis von stabilen Zuständen der Zellen z.B. von thermophilen und Meeresbakterien.

Hierzu werden Methoden zur Konstruktion von genom-weiten metabolischen Netzen aus der Kenntnis des Genoms sowie Methoden zur eigentlichen Modellierung entwickelt und angewandt. Die verwendete Methode der „Flux Balance Analyse“ erlaubt z.B. die Vorhersage der Reaktion der Zelle auf ihre Umgebung, z.B. Sauerstoffmangel, Änderung der Nahrungsquelle, Antibiotika Gabe und Analyse von genetischen Störungen und Veränderungen.

Die Modelle werden durch das Zusammenspiel von experimentellen Daten (Genomsequenzen, Transkriptom, Proteomdaten sowie Metabolomdaten aus der eigenen Gruppe) entwickelt, zum Design von Experimenten verwendet und zyklisch optimiert.

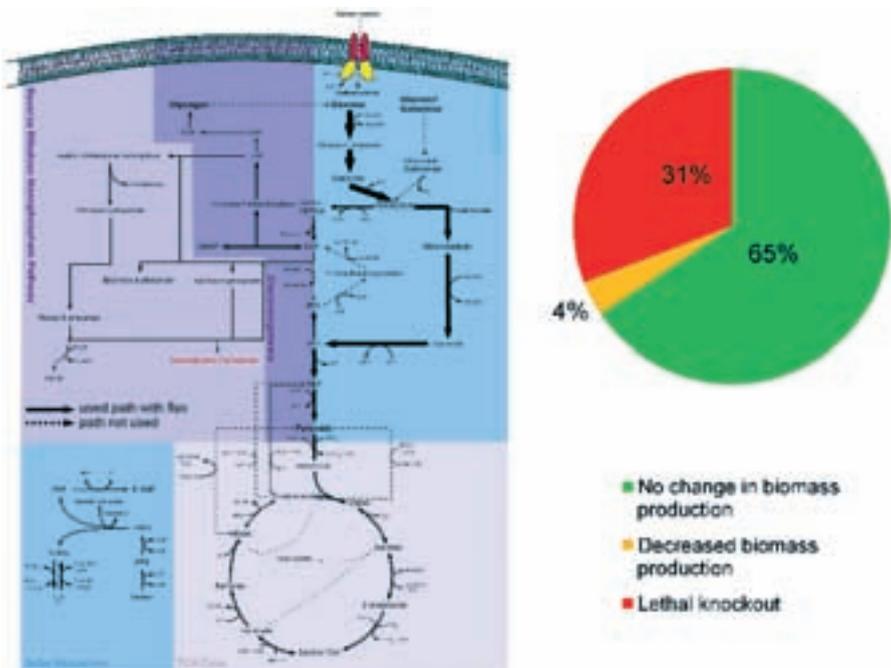


Abbildung 2: Vorhersage von Stoffwechselsflüssen und Vorhersage des Einflusses von genetischen Veränderungen in einem hyperthermophilen Organismus bei 80° C.

5 1998 – 2012 Bioinformatik-Unternehmen in der Braunschweiger Region

Nach insgesamt zwei BMBF- und einer EU-Förderung waren die TRANSFAC-Datenbank und die assoziierten Programme 1998 ein akzeptierter Bestandteil der internationalen Bioinformatik-Infrastruktur geworden. Um ein langfristig tragfähiges Konzept für die Pflege und Weiterentwicklung dieser Ressource zu entwickeln, wurde im Oktober 1997 von vier Mitgliedern der Arbeitsgruppe Bioinformatik der GBF die Firma BIOBASE GmbH gegründet. Über ein Technologietransfer-Abkommen mit der damaligen GBF wurden die Datenbank TRANSFAC und TRANSPATH zur Verwertung an BIOBASE übertragen. Das Unternehmen war zunächst auf dem GBF-Campus ansässig, bis es 2001 nach Wolfenbüttel umzog.

Von der Gründung an war BIOBASE international aktiv, so dass es nach der Einwerbung von Risikokapital (2000) noch zu einem japanischen Investment (2002) kam. Im Jahr 2005 erfolgte die Akquisition der Datenbankbranche von Incyte (Beverly/MA, USA) und die Gründung der US-amerikanischen Tochter BIOBASE Corp. Eine deutliche Erweiterung der Annotationskapazitäten wurde durch die Gründung eines Tochterunternehmens in Bangalore, Indien, ermöglicht (2006). Der besonderen Bedeutung des japanischen Marktes wurde durch die Gründung von BIOBASE Japan, KK (2008) Rechnung getragen.

Während BIOBASE sein Kerngeschäft im Bereich der Aufbereitung molekularbiologischer und –genetischer Informationen in entsprechenden Datenbanken hat, widmet sich die geneXplain GmbH in erster Linie bioinformatischer, systembiologischer und cheminformatischer Software. Dieses Unternehmen wurde 2010 vom BIOBASE-Gründer E. Wingender und dem früheren Forschungsleiter von BIOBASE, A. Kel, ins Leben gerufen. Es ist derzeit Partner in vier großen Forschungsverbänden.

Literaturverzeichnis

- [Ba10] Bannert C, Welfle A, Aus dem Spring C, Schomburg D: BrEPS: a flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation. *BMC Bioinformatics*. 2010 Dec 1;11:589.
- [Gr10] Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C & Schomburg D: The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res*. 39, D507-13, 2010.
- [Le95]. Lessel, U.; Schomburg, D.: Comparison, Assessment and Classification of Protein 3D Structures. *Theochem* 336, 261-267, 1995.
- [Ma08] Martin O, Schomburg D.: Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins*. 2008 Mar;70(4):1367-78.
- [Pa07] Parthiban V, Gromiha MM, Hoppe C & Schomburg D: Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins* 66, 41-52, 2007.
- [Qu11] Quester, S.; Schomburg, D.: EnzymeDetector: an integrated enzyme function prediction tool and database, *BMC Bioinformatics* 12, 376, 2011.
- [Ra06] Rahman SA & Schomburg D: Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics* 22, 1767-1774, 2006.
- [Sch10] Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J & Schomburg D: BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 39, D670-, 2010..
- [Sch88] Schomburg, D; Reichelt J.: BRAGI: a comprehensive protein modeling program system. *J. Mol. Graphics* 6, 161-5, 1988
- [WHL91]Wingender, E.; Heinemeyer, T.; Lincoln, D. Regulatory DNA sequences: predictability of their function. *Genome Analysis - From Sequence to Function; BioTechForum-Advances in Molecular Genetics* (J. Collins, A.J. Driesel, eds.) Vol. 4, S. 95-108, 1991
- [Wi08] Wingender, E.: The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* Vol. 9, S. 326-332, 2008.
- [Wi88] Wingender, E.: Compilation of transcription regulating proteins. *Nucleic Acids Res.* Vol. 16, S. 1879-1902, 1988.