Informationstechnik und Technische Informatik

it+ti

5/96

Schwerpunktthema: Informatik in den Biowissenschaften

Organ der Fachbereiche 3 „Technische Informatik und Architektur von Rechensystemen“ und 4 „Informationstechnik und Technische Nutzung der Informatik“ der GI e.V. und des Fachbereichs 4 „Technische Informatik“ der ITG

ITG

Oldenbourg

it + ti - Informationstechnik und Technische Informatik

it + ti veröffentlicht Themenhefte aus den Bereichen Informationstechnik und technische Informatik. Jedes Themenheft wird von einem kompetenten Gastherausgeber zusammengestellt. Mit Autoren aus dem jeweiligen Fachgebiet stellt er sein Thema als Sammlung der ausgewählten wesentlichen Aspekte dieses Bereiches dar.
it + ti informiert den Leser über den "state of the art" in der Forschung und stellt Verfahren und Anwendungen in verschiedenen Branchen aus technischer und wirtschaftlicher Sicht vor.

Das Themenheft **Informatik in den Biowissenschaften** (it+ti 5/96) haben PD Dr. Ralf Hofestädt, Universität Leipzig, Prof. Dr. Thomas Lengauer, GMD-Forschungszentrum Informationstechnik GmbH, und Prof. Dr. med. Markus Löffler, Universität Leipzig, zusammengestellt. In diesem Heft präsentieren sie als Gastherausgeber einige der zentralen Fragestellungen und Lösungswege sowie neuere Forschungsergebnisse aus dem Bereich der Molekularen Bioinformatik.

Bestellcoupon Fax 089 - 45051 - 204

R. Oldenbourg Verlag, Postfach 80 13 60, 81613 München, Tel. 089/45051 - 292, Fax - 204

☐ Schicken Sie mir bitte **Heft 5/96** Informatik in den Biowissenschaften der **it + ti** *Informationstechnik und Technische Informatik* zum Einzelpreis von DM 52,-- plus Versandspesen

☐ Ich bestelle zunächst ein **Probeabonnement** (2 Hefte) zum Vorzugspreis von DM 58,-/ÖS 423,-/SFR 58,- plus Versandspesen. Wenn ich nicht innerhalb von 3 Wochen nach Erhalt des 2. Heftes abbestelle, wünsche ich **it+ti** weiterhin im normalen Abonnement zu beziehen (Preisangaben gültig für 1997).

Name/Vorname	Firma (falls Lieferanschrift)
Straße/Nr./Postfach	PLZ/Ort
Datum	Unterschrift

Mir ist bekannt,daß ich diese Bestellung innerhalb von 10 Tagen (rechtzeitige Absendung genügt) schriftlich widerrufen kann bei R. Oldenbourg Verlag, Rosenheimer Str. 145, 81671 München.

Datum: Unterschrift:

it+ti – Informationstechnik und Technische Informatik

38. Jahrgang 1996 · Heft 5

Schwerpunktthema:
Informatik in den Biowissenschaften

Gastherausgeber:
R. Hofestädt
M. Löffler
T. Lengauer

EDITORIAL

BEITRÄGE · PAPERS

NACHRICHTEN · NEWS

übrigens...

it + ti Informationstechnik und Technische Informatik
(vormals: it – Informationstechnik)

Organ der Fachbereiche 3 „Technische Informatik und Architektur von Rechensystemen" und 4 „Informationstechnik und Technische Nutzung der Informatik" der Gesellschaft für Informatik e.V. Unter Mitwirkung des Fachbereichs 4 „Technische Informatik" der Informationstechnischen Gesellschaft im VDE (ITG).

Herausgeber:
Prof. Dr. rer. nat. O. Abeln, Karlsruhe,
Prof. Dr.-Ing. G. Färber, München,
Dr. rer. nat. Rainer Janßen, Heidelberg,
Dipl.-Ing. P. G. Jilek, München,
Prof. Dr.-Ing. H. M. Lipp, Karlsruhe,
Prof. Dr. rer. nat. Fritz Lehmann, München,
Prof. Dr.-Ing. D. Monjau, Chemnitz,
Dipl.-Ing. H.-R. Schuchmann, München,
Prof. Dr. techn. H. Zemanek, Wien.

Redaktion:
PD Dr.-Ing. Günter Söder (verantwortlich),
Technische Universität München, Lehrstuhl für Nachrichtentechnik, Arcisstr. 21
80290 München, Tel: (089) 2 89-2 34 86, Fax: -2 34 90
Email: guenter@LNT.e-technik.tu-muenchen.de

PD Dr. Ralf Hofestädt (Nachrichten GI/FB 4),
Universität Leipzig, Inst. f. Med. Inform., Stat. u. Epidem. Liebigstr. 27, 04103 Leipzig, Tel: (03 41) 97-1 61 20
Email: ralf@imise.uni-leipzig.de.

Prof. Dr.-Ing. D. Monjau (Nachrichten GI/FB 3, ITG/FB 4),
TU Chemnitz-Zwickau, Fakultät für Informatik, Straße der Nationen 62, 09111 Chemnitz, Tel: (03 71) 5 31-14 67, -14 69
Email: monjau@informatik.tu-chemnitz.de

Redaktionsbüro im Verlag:
Ursula Killguß
Telefon (0 89) 4 50 51-2 92
Telefax (0 89) 4 50 51-2 04
Email: it-ti@verlag.oldenbourg.de

Verlag:
R. Oldenbourg Verlag GmbH, Rosenheimer Straße 145,
D-81671 München, Telefon (0 89) 4 50 51-0
Oldenbourg im Internet: http://www.oldenbourg.de

Anzeigenverwaltung:
R. Oldenbourg Verlag GmbH.
Verantwortlich für den Anzeigenteil: Konrad Haslbeck,
Telefon (0 89) 4 50 51-2 06, -2 23, Telefax (0 89) 4 50 51-2 07,
Anschrift siehe Verlag. Zur Zeit gilt Anzeigenpreisliste Nr. 35.

Bezugsbedingungen:
„it + ti – Informationstechnik und Technische Informatik" erscheint jeden 2. Monat. *Jahresabonnementpreis* inc. gesetzlicher Mehrwertsteuer: Inland DM 302,20 (DM 292,– + DM 10,20 Versandspesen)
Österreich öS 2263,– (öS 2161,– + öS 102,– Versandspesen)
Schweiz und übriges Ausland sFr/DM 305,80 (sFr/DM 292,– + sFr/DM 13,80)
Einzelpreis: DM/sFr 52,–, öS 385,– + Versandspesen,
Studentenpreis: 20% Ermäßigung gegen Nachweis.
Persönliche Mitglieder der Gesellschaft für Informatik oder der Informationstechnischen Gesellschaft können im Rahmen ihrer Mitgliedschaft **it + ti** zu besonderen Bedingungen über die GI bzw. die ITG beziehen. Bestellungen über jede Buchhandlung oder direkt an den Verlag. Abonnements-Kündigung 8 Wochen zum Ende des Kalenderjahres möglich.

ISSN 0944-2774

Druck: R. Oldenbourg, Graphische Betriebe GmbH,
Rosenheimer Straße 145, D-81671 München.

Gedruckt auf chlor- und säurefreiem Papier.

Informatik in den Biowissenschaften

PD Dr. Ralf Hofestädt studierte Informatik mit Nebenfach Biologie an der Universität Bonn. Von 1985 bis 1990 war er wissenschaftlicher Mitarbeiter am Institut für Theoretische Informatik der Universität Bonn und von 1990 bis 1994 Hochschulassistent an der Universität Koblenz-Landau. Seit 1995 ist er wissenschaftlicher Mitarbeiter am Institut für Medizinische Informatik, Statistik und Epidemiologie mit den Forschungsschwerpunkten: Medizinische Wissensverarbeitung, Formale Systeme, Metabolic Engineering, Modellierung und Simulation sowie Parallele Datenverarbeitung.

Prof. Dr. Thomas Lengauer studierte Mathematik an der FU Berlin und Informatik an der Stanford-University. Von 1984 bis 1992 war er Professor für Informatik an der Universität Paderborn; seit 1992 ist er Professor für Informatik an der Universität Bonn und Leiter des Institutes für Algorithmen und wissenschaftliches Rechnen bei der GMD. Seine Forschungsinteressen sind algorithmische Anwendungen in Naturwissenschaft und Technik; Probleme der molekularen Analyse und Modellierung in Chemie und Molekularbiologie sowie diskrete Optimierungsprobleme in der Fertigungstechnik.

Prof. Dr. med. Markus Löffler habilitierte sich 1990 in Köln und ist seit 1994 Direktor des Institutes für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig. Zudem ist er Leiter der Arbeitsgrupe „Mathematische Modelle in Medizin und Biologie" der GMDS und der Internationalen Biometrischen Gesellschaft.

Die Bioinformatik verfolgt heute im wesentlichen zwei Ziele. Auf der Anwendungsseite sind Werkzeuge in den unterschiedlichsten Bereichen der Biowissenschaften zu entwickeln und zu implementieren, die dazu dienen, biologische Fragestellungen zu klären. Andererseits sind die biologischen Paradigmen (Lernen von der Natur) für die Informatik von Bedeutung. Dieses Arbeitsgebiet führt zu Innovationen, wie die Methoden der Genetischen Algorithmen, der Neuronalen Netze sowie des DNA-Computing verdeutlichen. Aber auch die Entstehungsgeschichte der endlichen Automaten, der zellularen Automaten und der Lindenmayer-Systeme veranschaulicht das Spektrum der Möglichkeiten.

Im Mittelpunkt des vorliegenden Sonderheftes steht ein Teilaspekt der Bioinformatik, der in Deutschland mit der Bezeichnung Molekulare Bioinformatik (Computational Molecular Biology) in Beziehung gesetzt wird. Dazu werden Überblicke und neuere Forschungsergebnisse aus dem Bereich der Molekularen Bioinformatik präsentiert. Dieses interdisziplinäre Forschungsgebiet mit Anteilen aus den Bereichen der Informatik, Statistik, Genetik, Molekularbiologie und Biochemie existiert, seit dem man vor etwa 30 Jahren begonnen hat, biomolekulare Polymersequenzen im Rechner zu analysieren.

Zwei Jahrzehnte lang führte dieses Gebiet ein peripheres Dasein, zum einen, weil molekularbiologische Daten nur in relativ kleinem Umfang existierten, zum anderen, weil die Leistungsfähigkeit der Rechner den anstehenden Aufgaben bei weitem nicht gewachsen war.

Seit Mitte der achtziger Jahre ist eine rasante Veränderung dieser Situation eingetreten. Auf der einen Seite hat die in den siebziger Jahren erfundene Technik der Genrekombination zu der Möglichkeit geführt, Proteine in großem Reinheitsgrad und in vertretbaren Mengen wirtschaftlich herzustellen. Auf der anderen Seite ist mit der Mitte der achtziger Jahre erfundenen PCR (Polymerase Chain Reaction)-Technik die Entschlüsselung der Polymerketten von Desoxyribonukleinsäure (DNA)-Sequenzen möglich geworden. Schon bald danach wurde die Entschlüsselung ganzer Genome von einfachen Organismen (Bakterien, Hefe), aber auch von komplexen Lebewesen (Maus, Mensch) zu einem zentralen wissenschaftlichen Ziel erklärt. Die Vision, die sich als Fokus bildete, war die Entschlüsselung des menschlichen Genoms mit $34 \cdot 10^9$ Basenpaaren. Es wird erwartet, daß diese Aufgabe bis zum Jahre 2003 (oder 2005) abgeschlossen sein wird. Neben dem menschlichen Genom werden auch Genome anderer Organismen vollständig sequenziert. Schon heute liegen vollständige Genome für Bakterien wie Haemophilus influenzae (1,8 Millionen Basenpaare) und Synechocystis (3,5 Millionen Basenpaare) sowie für einen Eukaryonten (Hefe, 12,6 Millionen Basenpaare) vor.

Als Folge dieser Entwicklungen gab es eine Explosion molekularbiologischer Datenbestände (DNA-Sequenzen, Proteinsequenzen, Proteinstrukturen etc.), deren Speicherung allein schon Computer zwingend erforderte. Glücklicherweise wuchs der Speicherumfang der Rechner in einem Maße, das den Aufbau von Computerdatenbanken für diese molekularbiologischen Daten ermöglichte. Die Schnelligkeit der Rechner wuchs jetzt auch in Bereiche, die komplexe Analysen dieser Daten zulassen. Schließlich ermöglichten es Fortschritte in der Computergraphik, die molekularen Strukturen im Detail und in vielen Darstellungsweisen zu visualisieren. Alle diese Elemente führten zu einem rasanten Wachstum der Molekularen Bioinformatik.

Ziel der Molekularen Bioinformatik ist es, mit Methoden der Informatik die Aufklärung der atomaren Zusammensetzung von Biomolekülen, ihrer dreidimensionalen Strukturen sowie der Wechselwirkungen zwischen diesen Molekülen zu unterstützen. Diese Wechselwirkungen und ihr Zusammenspiel bilden die Grundlage für die

Zelldifferenzierung und damit für alle Prozesse des Lebens.

Die Molekulare Bioinformatik kann grob in folgende Bereiche unterteilt werden:

1. Genomsequenzierung: Unter diesem Begriff versteht man das Lesen genomischer Information im molekularbiologischen Labor. Die Sequenzierungsraten von Genomen wachsen auch heute noch stark an. Man muß auf einen Zuwachs von Millionen von Basenpaaren pro Tag kommen, um das gesteckte Ziel der Entschlüsselung des menschlichen Genoms innerhalb der nächsten zehn Jahre zu erreichen. Die in den Experimenten anfallenden Daten sind fragmentiert, interpretationsbedürftig und fehlerbehaftet. Rechnergestützte Methoden helfen sowohl bei der Übersetzung experimenteller Daten (etwa von Hybridisierungskarten) in Sequenzdaten als auch bei der Assemblierung der Sequenzdaten und der Kartographierung der Genome.

2. Sequenzanalyse: Nach erfolgter Genomsequenzierung werden die Wissenschaftler mit einem Umfang von DNA-Text konfrontiert, der beim menschlichen Genom etwa einem dreißigbändigen Lexikon entspricht. Diese Daten sind zunächst uninterpretiert. In den Genomsequenzen gilt es, die Gene (Proteine kodierende Regionen) zu finden, die zu allem Übel meist noch zerstückelt sind. Auch nichtkodierende Teile der DNA wollen in ihrer Funktion analysiert werden. Dazu sind rechnerbasierte umfangreiche Sequenzanalysen durchzuführen.

3. Molekulare Strukturvorhersage: Hat man die Gene in den Genomsequenzen identifiziert, so kann man aus ihnen die entsprechenden Proteinsequenzen ablesen. Hierzu bedient man sich des seit etwa vierzig Jahren bekannten genetischen Codes. Die nächste Herausforderung ist, die dreidimensionale Struktur des Proteins zu entschlüsseln. Man weiß, daß diese Struktur für die von der Natur verwendeten Proteine eindeutig ist. Sie ist ferner die Grundlage für die Funktion des Proteins.

Auch Strukturen anderer Moleküle als Proteine sind wesentlich. RNA ist eine Nukleinsäure, die eine wesentlich größere Strukturvielfalt aufweist als die praktisch immer als Doppelhelix gewundene DNA. Da RNA eine Zwitterrolle als Speichermolekül für genetische Information aber auch als Stoffwechselagent ausübt, ist auch die Kenntnis der 3D-Struktur von RNA- Molekülen wesentlich.

Selbst die 3D-Struktur von DNA ist wichtig, spielen doch die kleinen strukturellen Unterschiede in der Doppelhelix, die sich durch unterschiedliche Abfolgen von Nukleotidsequenzen ergeben, die entscheidende Rolle bei der Übersetzung der genetischen Information in molekulare Lebensbausteine.

4. Molekulare Wechselwirkungen: Prozesse des Lebens bestehen aus Abfolgen molekularer Wechselwirkungen. Zwei Moleküle binden sich aneinander, werden modifiziert und lösen sich wieder. Abermillionen solcher Reaktionen finden in unserem Körper in jeder Sekunde statt. Sie zu verstehen bedeutet, die Bindungsvorgänge zwischen Molekülen aufzuklären. Die Bereitstellung geeigneter Moleküle zur Modifikation solcher Prozesse ist die Grundlage für einen Zugang zum Wirkstoffentwurf.

5. Metabolische Netzwerke: Schon eine einzige Wechselwirkung zwischen zwei Biomolekülen zu verstehen ist eine große Herausforderung. Das Verständnis metabolischer Vorgänge bedingt jedoch häufig die Kenntnis ganzer Pfade oder sogar komplexer regulatorischer Netzwerke. Der Zelldifferenzierung und biologischen Strukturbildung liegt eine Koordination der metabolischen Netzwerke zugrunde.

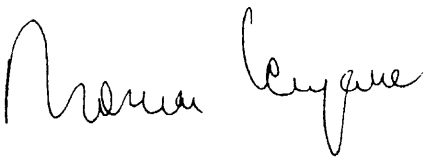
Jede der oben beschriebenen Aufgaben ist eine große Herausforderung für die Wissenschaft. Vielfältige experimentelle Methoden der Molekularbiologie bilden das Rückgrat für diesen Prozeß. So werden etwa Proteinstrukturen mit Mitteln der Röntgenkristallographie oder der Kernresonanzspektroskopie (NMR) entschlüsselt. Aber oft sind die Experimente schwierig und liefern noch bruchstückhafte und mit Ungenauigkeiten behaftete Informationen. Deshalb ist eine Unterstützung der Experimente durch rechnerbasierte Methoden dringend erforderlich. An solchen Verfahren wird fieberhaft gearbeitet, und sie machen den wesentlichen Teil der Molekularen Bioinformatik aus.

Dies beinhaltet die Bereitstellung von molekularbiologischen Datenbanken, geeigneten Browser durch solche Datenbanken, die Fehlersuche in den Datenbeständen sowie die Verbindung der Datenbanken untereinander. Der Artikel von A. Kaps et al. beschreibt, wie man Datenbanken und effiziente Zugriffsmethoden als Basis für Genomanalysen verwenden kann und wie das WWW als effektives Medium zur Datenvisualisierung eingesetzt werden kann. Der Artikel von O. Ritter und S. Suhai konzentriert sich auf die Probleme der Verbindung heterogener und räumlich getrennter Datenbanken im Bereich der Molekularbiologie. P. Bork beschreibt in seinem Artikel, wie man codierende, also den Aufbau von Proteinen beschreibende Sequenzen in Genomen findet, sie in Proteinsequenzen übersetzt und diese dann auf die Funktion der entsprechenden Proteine hin untersucht. Die Aussagen über die Funktion von Proteinen ist leichter, wenn man etwas über deren dreidimensionale Struktur weiß. Die Proteinstruktur-Vorhersage ist ein wichtiges Problem der Molekularen Bioinformatik, das jedoch in dem vorliegenden Band aus Platzgründen nicht besprochen wird. In einer der nächsten Ausgaben wird zu diesem Thema ein Artikel erscheinen. Kenntnisse über die molekulare Struktur erleichtern Analysen über Wechselwirkungen zwischen Molekülen wie etwa Proteinen (molekulares Docking) erheblich. Biologische Prozesse entstehen aus einer Synthese von metabolischen Aktionen von Proteinen (und von Ribonukleinsäure (RNA)) untereinander sowie mit DNA. Dabei werden auch nichtcodierende Regionen der DNA einbezogen. Sie üben zum Teil regulatorische Funktionen aus. Methoden für die Auffindung solcher genregulatorischer DNA-Sequenzen werden in

dem Artikel von R. Knüppel und E. Wingender beschrieben. Die Zelldifferenzierung basiert schließlich auf einem komplexen Netzwerk von kontrollierten metabolischen und regulatorischen Wechselwirkungen und führt letztlich über geeignete biochemische Mechanismen der Zell-Zell-Kommunikation von der befruchteten Eizelle zum komplexen Organismus. Der Artikel von H. Meinhardt beschreibt eine Methode zur Modellfindung für kommunikative Prozesse zwischen Zellen, die der Suche nach der molekularbiologischen Grundlage für biologische Strukturbildungsprozesse – wie solche der Zelldifferenzierung – eine Richtung weist. Schließlich berichtet der Artikel von C. Büschking und R. Giegerich über einen WWW-basierten Kursus zur biologischen Sequenzanalyse, der den Einsatz dieses neuen Mediums in der Lehre diskutiert.

Die Molekularbiologie hat heute einen Zustand erreicht, der des massiven Einsatzes der Methoden und Konzepte der Informatik bedarf. Die Bundesregierung hat in ihrem Programm „Biotechnologie 2000" betont, daß für Deutschland, insbesondere auf dem Gebiet der Bioinformatik, Nachholbedarf besteht. Die Gesellschaft für Informatik e.V. hat den interdisziplinären Bestrebungen Rechnung getragen, indem neben der bereits existierenden Arbeitsgruppe „Simulation in Biologie und Medizin" 1992 die Fachgruppe 4.0.2 „Informatik in den Biowissenschaften" gebildet wurde. Die Aufgaben der Fachgruppe liegen u.a. in der Verflechtung moderner biotechnologischer Forschung mit anwendungsorientierter Entwicklung von Methoden und rechnergestützten Verfahren der Informatik. Seit dem Gründungsworkshop in Bonn im Jahre 1992 hat die FG 4.0.2 eine Vielzahl von nationalen und internationalen Workshops unter dieser Zielsetzung veranstaltet. Eine Übersicht über die Aktivitäten der Fachgruppe findet man unter (http://www.imise.uni-leipzig.de/org/gi/fachgruppe4_0_2.html). In dem vorliegenden Sonderheft präsentieren wir nur einige der zentralen Fragestellungen und Lösungswege der Molekularen Bioinformatik in der Hoffnung, Informatiker für dieses interessante Anwendungsgebiet gewinnen zu können. An herausfordernden Aufgaben mangelt es nicht.

Dr. habil. Ralf Hofestädt



Prof. Dr. Thomas Lengauer

Prof. Dr. med. Markus Löffler

Dr.-habil. Ralf Hofestädt
Universität Leipzig, Inst. f. Med. Inform., Stat. u. Epidem.
Liebigstr. 27, 04103 Leipzig
Email: ralf@imise.uni-leipzig.de

Professor Dr. Thomas Lengauer
GMD-Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
SCAI, 53754 St. Augustin
Email: Thomas.Lengauer@gmd.de

Prof. Dr. med. Markus Löffler
Universität Leipzig, Universitätsklinikum
Liebigstr. 27, 04103 Leipzig
Email: loeffler@imise.uni-leipzig.de

Genomanalyse und WWW: Vom Klon zum Klick

Andreas Kaps, Klaus Heumann, Andreas Maierl, Hans-Werner Mewes,
Martinsrieder Institut für Proteinsequenzen (MIPS),
Max-Planck-Institut für Biochemie, Martinsried

Von links nach rechts: *H.-W. Mewes, A. Kaps, K. Heumann* und *A. Maierl.*
Dr. Hans-Werner Mewes, Arbeitsgruppenleiter von MIPS, studierte Chemie an der Philipps-Universität Marburg und promovierte über die Identifizierung von Proteinen durch computergestützte Aminosäurenanalyse. Seit Beginn der Achtziger Jahre beschäftigt er sich mit Konzepten der Informatik und ihrer Anwendung auf biologische Problemstellungen. **Dipl.-Inform. Andreas Kaps, Dipl.-Inform. Klaus Heumann** und **Dipl.-Inform. Andreas Maierl** studierten Informatik an der Technischen Universität München. Sie sind als wissenschaftliche Mitarbeiter bei MIPS beschäftigt. *K. Heumann* beschäftigt sich mit Algorithmen und Datenstrukturen zur biologischen Sequenzdatenanalyse, *A. Kaps* mit Geschäftsprozeßmodellierung und Workflow-Management im Bereich biologischer Sequenzdatenbanken sowie verteilten Anwendungen, und *A. Maierl* mit objektorientierter Analyse und Design sowie objektorientierten Datenbanken.

Die systematische Analyse komplexer Genome, die mehrere Millionen Basenpaare enthalten, und die Organisation dieser Informationen in objektorientierten Datenbanken benötigt die profunde Unterstützung einer leistungsfähigen Informatik. Die von der Arbeitsgruppe MIPS am Max-Planck-Institut für Biochemie entwickelten Konzepte werden vorgestellt und die eingesetzten Technologien beschrieben. Ausgehend von der Darstellung der Proteinsequenzdatenbank (PIR-International) werden die Ziele der systematischen Genomanalyse aufgezeigt sowie die dazu notwendigen Algorithmen und Datenstrukturen dargestellt. Dabei wird ein Schwerpunkt auf eine Variante des Positionsbaums, den Hashed Position Tree, *gelegt. Die Visualisierung biologischer Information innerhalb des Genoms eines Organismus wird durch den* Genombrowser *ermöglicht. Im Rahmen von Funktionsanalyseprojekten finden Konzepte aus dem Bereich CSCW Anwendung. Die Integration dieser vielschichtigen Datensammlungen und Dienste unter einer einheitlichen intuitiven Bedienoberfläche basiert auf dem World Wide Web. Die Integrität vollständig replizierter Datensätze wird durch ein Synchronisationsprotokoll sichergestellt.*

Genome Analysis and WWW: From Clone to Click

The systematic analysis of complete genomes containing millions of base pairs and the organization of these information in object–oriented databases requires profound support from informatics. The concepts developed by MIPS at the Max-Planck-Institute for Biochemistry are introduced and the used technologies are described. Starting with the description of the protein sequence database (PIR-International) the aims of systematic genome analysis and the necessary algorithms and data structures are described. We will focus on a variant of the position tree, the Hashed Position Tree. *The* Genomebrowser *facilitates the visualization of biological information within a genome. Within the bounds of functional analysis projects concepts of CSCW are used. These heterogeneous data collections and services are integrated under a uniform and intuitive user interface that is based on the World Wide Web. A synchronizaton protocol ensures the integrity of fully replicated data sets.*

1 Einleitung

Die moderne Biologie ist in der Entwicklung den klassischen Naturwissenschaften Physik und Chemie gefolgt und hat atomare Auflösung erreicht. Es ist möglich geworden, biologische Zusammenhänge auf molekularer Ebene zu untersuchen. Molekularbiologische Experimente generieren heute Daten, deren Auswertung und Organisation nur durch moderne Informationstechnologie möglich ist.

Das vorhandene Wissen in den Biowissenschaften ermöglicht die industrielle Nutzung. Besonders der Bereich der Biotechnologie hat sich zu einem schnell wachsenden und wichtigen Wirtschaftssektor entwickelt. Fortschritte in der biomedizinischen Forschung, speziell in der Diagnose und Therapie genetischer Defekte und Karzinome, wecken Hoffnungen, eine unmittelbare Behandlung auf molekularer Ebene durchführen zu können. Grundvoraussetzung dazu ist das Wissen um die Funktionalität biologischer Makromoleküle. Nukleinsäuren und Proteine sind die

wichtigsten Bestandteile der Zelle, der elementaren biologischen Organisationsform lebender Organismen. Jedes Protein wird durch genomische Nukleinsäuren (DNA) kodiert, d.h. der dreistellige 4-Zeichen Code der DNA[1] wird abgebildet auf den einstelligen 20-Zeichen Code der Proteine[2].

Die Aminosäuresequenz eines Proteins bestimmt dessen dreidimensionale Struktur und damit die Funktionsweise des Proteins im Organismus. Solche Sequenzen werden seit den frühen siebziger Jahren gesammelt und seit Beginn der achtziger Jahre in Sequenzdatenbanken organisiert. Genomanalyse, medizinische Diagnostik, Molekularbiologie, Gentherapie und Biotechnologie nutzen diese biologischen Rohdaten in vielfältiger Weise. Voraussetzung für eine sinnvolle und effiziente Exploration der Daten ist ihre Strukturierung und Interpretation in Gestalt vollständiger, konsistenter und redundanzfreier Sequenzdatenbanken. Diese Anforderungen konnten bisher von existierenden Datenbanken nicht erfüllt werden.

Heute führen weltweit Tausende von Laboratorien DNA-Sequenzierung durch, so daß genomische Datensammlungen exponentielle Zuwachsraten aufweisen. Derzeit haben die molekularen Sequenzdatenbanken eine Verdoppelungsrate von 2,5 Jahren. Anfang 1996 waren etwa 120 000 Proteine mit insgesamt 40 Millionen Aminosäuren und etwa 270 000 Nukleinsäuren mit insgesamt 350 Millionen Basen bekannt.

Die Komplexität dieser Datenstrukturen und der mit ihrer Auswertung verbundenen Anfragen erfordern den Einsatz von theoretisch evaluierten Methoden der Informatik. Dazu gehören der Aufbau von objektorientierten Datenbanken sowie der Einsatz effizienter Algorithmen. Biologische Aussagen werden aus automatisierten experimentellen Abläufen durch elektronische Prozessierung des Datenmaterials abgeleitet. Die rechnergestützte Genomanalyse gemeinsam mit anderen Problemstellungen der Biologie (z.B. Strukturaufklärung und -vorhersage, Berechnung molekularer Dynamik) haben ein neues Arbeitsgebiet der Informatik, die Bioinformatik, geschaffen.

Unabhängige Institutionen entwickelten Datenbankschemata und Software-Pakete zur Sequenzdatenanalyse. Diese Entwicklungen geschahen isoliert. Es wurde nicht beachtet, Beziehungen zwischen unterschiedlichen Datenbanken und Diensten herstellen zu können. Ansätze, eine standardisierte Sprache zur Beschreibung biologischer Datenbanken einzuführen, schlugen fehl [6]. Diese Vielzahl verschiedener Informationsrepräsentationen und Bedienoberflächen erfordert zur effizienten Nutzung einen zu großen Einarbeitungsaufwand.

Die Integration dieser heterogenen Ressourcen unter einer einheitlichen intuitiven Benutzungsoberfläche wird durch Verwendung des World Wide Web ermöglicht. Wir entwickelten eine hierarchische Software-Architektur, die eine einheitliche Sicht auf unterschiedliche Datenbanken und Programme ermöglicht. Die Kommunikation mit diesem System erfolgt mit Hilfe des gewohnten WWW-Browsers.

Diese Dienstintegration kann jedoch nur eine Zwischenlösung sein auf dem Weg, existierende Datenbanken in systematischer Weise zu einem vollständigen, konsistenten und redundanzfreien Datensatz zusammenzufassen.

2 Die Proteinsequenzdatenbank als Beispiel biologischer Sequenzdatenbanken

Die ersten biologischen Sequenzen wurden 1953 bestimmt. Da die chemische Abbaureaktion nicht mit absoluter Zuverlässigkeit durchgeführt werden konnte, überstieg auch nach technischer Verfeinerung der Methode die Länge der bestimmten Sequenz nur unter idealen Bedingungen 40-50 Aminosäuren[3]. Die prinzipiellen Nachteile der chemischen Proteinsequenzierung konnten erst durch DNA-Sequenzierungsmethoden kompensiert werden.

Die erste systematische Analyse der bekannten Proteinsequenzen findet sich in den Arbeiten von M. Dayhoff *et al.* an der National Biomedical Research Foundation (NBRF). Die ersten Ausgaben der Datenbank erschienen in gedruckter Form als „ATLAS of Protein Sequences" [3]. Die elektronische Sequenzdatenanalyse hat sich aus sehr einfachen Datenverarbeitungsmethoden entwickelt. Am Anfang stand die Prozessierung unstrukturierter Textdateien mit Hilfe von Editoren. Seit den ersten Versuchen in den späten siebziger Jahren, evolutionäre Beziehungen mit Hilfe von Rechnern zu ermitteln, hat sich die Informatik zu einem unverzichtbaren Werkzeug der Molekularbiologie weiterentwickelt. In den frühen Achtzigern wurden die ersten Programme für den Zugriff auf Sequenzdaten veröffentlicht [1]. Der Sequenzteil der Information wurde durch eine Drei-Zeichen-Indextabelle indiziert, so daß die Anfrage: *„Ist eine gegebene Sequenz bereits in der Datenbank vorhanden?"* bereits damals sehr zeiteffizient beantwortet werden konnte. Es wurden Softwarepakete zur Sequenzdatenanalyse entwickelt und Datenbankzentren für Nukleinsäure- und Proteinsequenzen gegründet.

Die Proteinsequenzdatenbank *PIR-International* [7] ist eine Faktendatenbank. Die gespeicherten Proteinsequenzen sind jedoch mit experimentellen Fehlern und Fehlern der Interpretation der experimentellen Daten behaftet. Sequenzdaten unterliegen der ständigen Veränderung und müssen daher entsprechend dem sich verändernden biologischen Wissen angepaßt werden. Eine Proteinsequenz wird mit interpretativer Information versehen, die mit Hilfe eines breiten Spektrums von experimentellen Techniken und Schlußfolgerungen aus Sequenzvergle-

[1] bestehend aus den vier Basen Adenin, Thymin, Guanin und Cytosin.

[2] bestehend aus den 20 in der Natur auftretenden Aminosäuren.

[3] Die durchschnittliche Länge eines Proteins in der Sequenzdatenbank beträgt ca. 400 Aminosäuren.

chen erzeugt wird. Diese Annotation beschreibt die biologischen Eigenschaften der Sequenz. Annotation erfordert die kritische Beurteilung durch entsprechend biologisch geschultes wissenschaftliches Personal, und ist daher zeitintensiv, teuer und schwierig zu automatisieren.

Die Entwicklung der Sequenzdatenbanken ist nicht abgeschlossen. Neben der Zunahme der Sequenzinformation unterliegt die biologische Interpretation der Daten einem ständigen Wandel. Vor ca. fünfzehn Jahren hat man begonnen zu verstehen, daß Bereiche der DNA, die für ein Protein kodieren (*exons*), häufig unterbrochen sind von nicht übersetzten Bereichen (*introns*). Da ein erheblicher Teil der Proteinsequenzen aus DNA-Sequenzen abgeleitet ist, hat dieses geänderte Verständnis den Faktenbestand der Proteinsequenzdatenbank fundamental verändert. Jedes Modell zur Repräsentation biologischer Daten muß daher in hohem Maße erweiterbar und anpassungsfähig sein, um neuen Erkenntnissen folgen zu können. Das gegenwärtige biologische Wissen und Verständnis ist unvollständig, viele wichtige Eigenschaften der Moleküle sind nur in groben Zügen bekannt. Andere Eigenschaften beruhen auf Vorhersagen, empirischen Beobachtungen oder vergleichenden Methoden. Die Daten sind den Einschränkungen dieser Methoden unterworfen und ausgesetzt. Daher müssen alte Daten fortlaufend überprüft und reprozessiert werden, um den Datensatz hinsichtlich des aktuellen Kenntnisstandes konsistent zu halten. Jedes Schema zur Prozessierung muß daher dynamisch sein.

3 Genomanalyse: Algorithmen, Datenstrukturen und Visualisierung

3.1 Was ist Genomanalyse?

Die systematische Erforschung genomischer Information erreichte die öffentliche Diskussion, als das Human-Genom-Projekt zur Analyse der 3,3 Milliarden menschlichen Basenpaare in Angriff genommen wurde. Erste Ansätze, das komplette Genom eines Modell-Organismus systematisch zu erforschen, stammen von A. Goffeau, der 1988 in einer Studie die Sequenzierung des Hefe-Genoms der Europäischen Kommission vorgeschlagen hat [4]. Am 24. April 1996 wurden die 16 Chromosomen[4] der Bäckerhefe (*S. cerevisiae*) als Ergebnis eines gemeinsamen Projektes von Europa, Kanada, USA und Japan der Öffentlichkeit vorgestellt, in dem MIPS als *informatics coordinator* fungierte.

Ziel der Genomanalyse ist es, zunächst eine vollständige Karte genetischer Elemente eines Organismus auf DNA-Sequenz-Ebene zu erstellen. Liegen diese Daten als Rohdaten vor, wird unter Anwendung biologischen Wissens die Information strukturiert. Ein wesentlicher Schritt in diesem Prozeß ist die Identifizierung von DNA-Bereichen, die für ein Protein kodieren (offene Leserahmen[5]). Dazwischen können Bereiche lokalisiert werden, die an der Steuerung und Regulation zellulärer Abläufe in der Zelle beteiligt sind.

Von kodierenden Bereichen, die durch Start- und Stoppsignale gekennzeichnet sind, werden hypothetische Proteine abgeleitet. Zu einem neuen hypothetischen Protein werden ähnliche Proteine gesucht, von denen die Funktionsweise bekannt ist. Handelt es sich dabei um ein Protein aus einem nahe verwandten Organismus, werden aufgrund evolutionärer Abstammung die Eigenschaften dieses Homologs[6] auch für das unbekannte Protein angenommen. Etwas mehr als die Hälfte der im Rahmen systematischer Sequenzierung identifizierten offenen Leserahmen sind eindeutig ähnlich zu Sequenzen mit bekannter Funktion. Dennoch ist lediglich in 30% der Fällen eine befriedigende funktionelle Zuordnung der gefundenen Proteine möglich.

3.2 Alignment von Sequenzen

Im Rahmen der Sequenzdatenanalyse ist man mit dem Problem konfrontiert, Teilsequenzen in einer Sequenzdatenbank zu finden. Diese Teilsequenzen erlauben eine gewisse Variabilität der Sequenz, da Homologie bereits bei einer signifikanten Sequenzähnlichkeit angenommen wird. Es werden Sequenz-Alignments durchgeführt, bei denen Ähnlichkeiten in der Abfolge der Aminosäuren gesucht werden.

Es existieren Bewertungsmatrizen[7], um in der Natur auftretende physikalisch-chemische Ähnlichkeiten von Aminosäuren modellieren und gewichten zu können. Es wird eine Alignmentmatrix aufgebaut, in der die Reihen und Spalten durch die beiden zu vergleichenden Sequenzen gegeben sind. Diese Matrix wird anhand der gegebenen Austauschmatrix gefüllt. Der Pfad durch die Alignmentmatrix, der den höchsten Wert liefert, repräsentiert das optimale Alignment. Dabei müssen folgende Bedingungen erfüllt sein:

- jede Aminosäure muß Bestandteil des Alignments sein,
- jede Aminosäure der beiden Sequenzen darf nur genau einmal im Alignment vorkommen, und
- die Reihenfolge der Aminosäuren muß erhalten bleiben.

Needleman und Wunsch lösten 1970 mit Hilfe eines *dynamic programming*-Algorithmus das Problem, den besten Pfad durch die Alignmentmatrix zu finden [14].

Die Ähnlichkeit von Sequenzen wird zusätzlich zur Austauschmatrix durch Einfügen von Lücken (*gaps*) in das Alignment modelliert. Im Extremfall kann ein Alignment einer kurzen Sequenz gegen eine lange sehr leicht eine 100%-ige Sequenzidentität ergeben, wenn nicht entsprechende Parameter gesetzt sind. Das Einfügen solcher *gaps* wird daher negativ bewertet.

[4] Ein Chromosom ist eine genetische Einheit in der Zelle, die Erbinformation enthält. Die Anzahl der Chromosomen ist organismusspezifisch (z.B. 16 bei Hefe, 46 beim Menschen).

[5] *open reading frame (ORF)*, kodierende Bereiche, die kein Stoppsignal enthalten.

[6] Homologie beschreibt das Vorhandensein einer evolutionären Beziehung.

[7] Eine 20×20-Matrix im Fall der Proteine.

Wird eine Sequenzähnlichkeit angezeigt, muß anhand des aktuellen Wissensstands der Biologie entschieden werden, ob tatsächlich eine funktionelle Verwandtschaft vorliegt. Aufgrund der unscharf trennenden Algorithmen kann dieser Schritt in kritischen Fällen nicht automatisiert werden.

3.3 Musterorientierte Sequenzdatenanalyse

Wenn die zu durchsuchende Datenmenge eine kritische Größe[8] übersteigt, kann der beschriebene Ansatz zur Homologiesuche interaktiv nicht mehr angewandt werden.

Aussagen über Sequenzobjekte beziehen sich auf Sequenzeigenschaften, die durch charakteristische Teilsequenzen repräsentiert werden. Dadurch werden Mustern Eigenschaften zugeordnet, die dann in einem Text identifiziert werden können. Ein anderer Ansatz zum Sequenzvergleich liegt damit im Suchen aller Vorkommen eines signifikanten Musters der Länge m in einem Text der Länge n, der in einer Datenbank abgelegt ist. Als geeignete Datenstruktur für solche Problemstellungen werden Positionsbaumvarianten verwendet. Eine besondere Variante, der von uns entwickelte *Hashed Position Tree (HPT)* [13], wird im folgenden kurz vorgestellt.

Die Effizienz von Positionsbäumen ist begrenzt durch die Größe des vorhandenen Hauptspeichers. Der *HPT* erlaubt im Mittel *Finde Teilwort* Operationen angewandt auf große Datensätze in einem Diskzugriff durchzuführen. Entgegen natürlichen Sprachen, in denen die Menge der vorkommenden Wörter über einem Alphabet nur eine kleine Teilmenge der kombinatorisch möglichen Varianten ausmacht, sind im Fall der Proteinsequenzen alle möglichen Kombinationen bis zur Länge fünf in den aktuellen Datensammlungen enthalten. Die evolutionär notwendige Konservierung biologischer Funktionalität drückt sich im Positionsbaum durch tiefe und weitverzweigte Teilbäume aus, die einzelne Formen konservierter Teilsequenzen repräsentieren.

Der *HPT* ist eine hybride Datenstruktur, die ein Hash-Directory, eine Menge von Kollisionsklassen-Positionsbäumen und eine Menge von Datenseiten kombiniert. Er ist eine Verallgemeinerung des Standard-Positionsbaumes, in dem nicht Positionsidentifikatoren, sondern partielle Positionsidentifikatoren[9] kodiert sind. Die Kanten sind mit Elementen aus dem Alphabet markiert. Positionsidentifikatoren, die einen gemeinsamen partiellen Positionsidentifikator haben, werden im *HPT* zusammengefaßt. Die einzelnen Komponenten des *HPT* haben folgende Bedeutung:

- Das *Hash-Directory* repräsentiert den Baum von der Wurzel bis zu der Tiefe, in der er voll besetzt ist (fünf für Proteine). Die Einträge zeigen entweder auf ein Segment, das auf einer Datenseite lokalisiert ist, oder sie referenzieren auf einen Kollisionsklassen-Positionsbaum.
- Die *Kollisionsklassen-Positionsbäume* separieren grob die Daten in Klassen. Die Blätter verweisen auf Datenseiten, die Segmente enthalten.
- Die *Datenseiten* stellen eine betriebssystemabhängige Transporteinheit dar. Sie sind in Segmente unterteilt, die Positionsidentifikatoren des Textes enthalten. Per Definition darf ein Segment nicht größer sein als eine Datenseite.

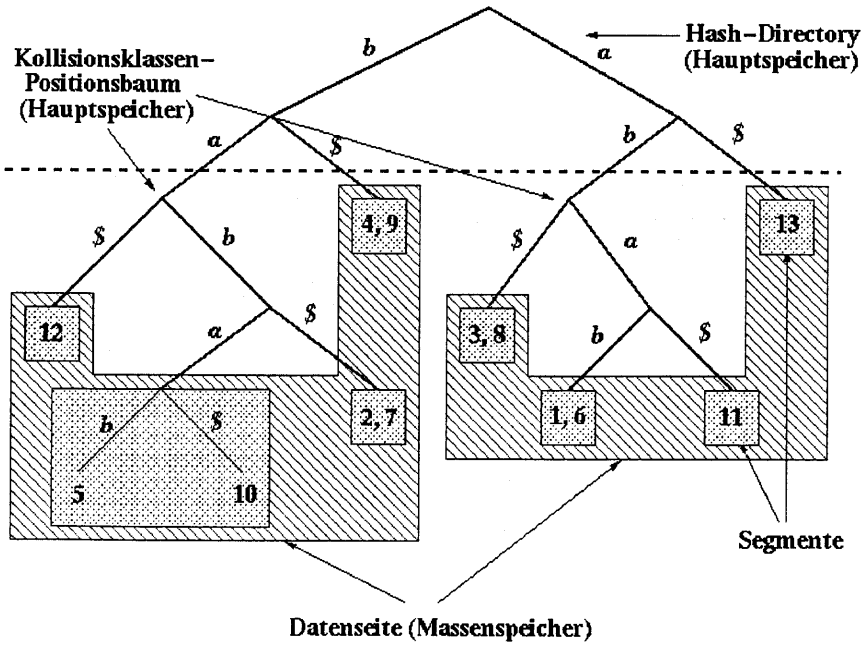
Es kann gezeigt werden, daß jeder *HPT* in einen Standard-Positionsbaum konvertiert werden kann [9]. Bild 1 zeigt ein Beispiel für einen *HPT*. Will man die Aktualität eines Datensatzes gewährleisten, ist ein inkrementelles Aktualisieren des großen Datensatzes erforderlich. Das Löschen eines Eintrages im Standard-Positionsbaum kann nur aufwendig gelöst werden, da das Entfernen eines Zeichens aus dem Positionsbaum die Modifikation einer unter Umständen sehr großen Zahl von Positionsidentifikatoren erfordert.

Dagegen erlaubt der *HPT* ein effizientes Aktualisieren ganzer Sequenzen. Die vorgestellte Indexdatenstruktur hat die Aufgabe, Beziehungen zwischen semantischen Einheiten herzustellen. Bei den Basisoperationen auf dem *HPT*, bei denen immer genau eine Einheit verändert wird, ergeben sich

[8] Die derzeitige Grenze liegt bei ca. 10 000 Sequenzen auf einer üblichen Workstation.

[9] Jedes nicht-leere Präfix eines Positionsidentifikators ist ein partieller Positionsidentifikator.

Bild 1: HPT für den Eingabetext: *abab$babab$baba$.*



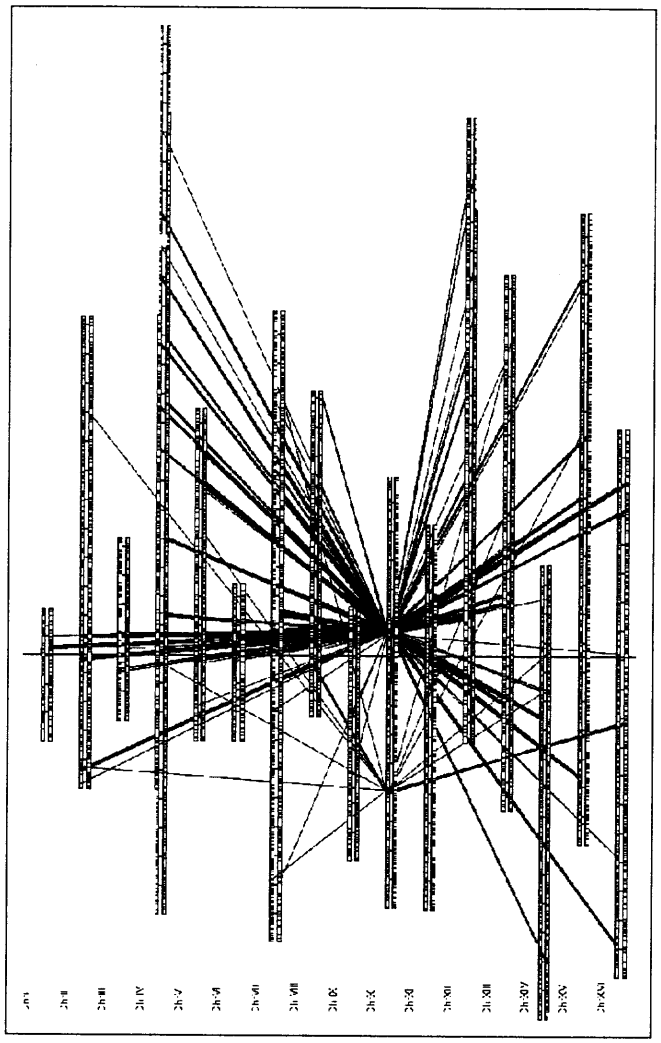


Bild 2: Beispiel für einen *GSG* (Hefegenom).

Komplexitäten von $O\ (m \cdot log(n))$ für das Einfügen und Löschen einer Sequenz[10]. Während sich eine Insertion in den *HPT* hinsichtlich der Zeitkomplexität nicht vom Standard-Positionsbaum unterscheidet, wird beim Entfernen aus dem *HPT* ausgenutzt, daß nicht einzelne Positionen, sondern semantische Einheiten[11] gelöscht werden. Positionsidentifikatoren können sich nicht über semantische Einheiten hinweg erstrecken, d.h. sie sind voneinander unabhängig.

3.4 Visualisierung intragenomischer Homologien

Der *Genombrowser* ist ein graphisches Werkzeug, das den vollständigen Vergleich jedes Abschnitts eines Genoms, wie z.B. der Hefe, mit jedem anderen des gleichen Genoms erlaubt. Dies kann dazu dienen, evolutionäre Strukturen und Duplikationen innerhalb eines Genoms aufzuklären.

Die Menge aller Ähnlichkeitsbeziehungen zwischen Blöcken auf dem Genom wird als Graph repräsentiert. Relationen zwischen Blöcken auf unterschiedlichen Chromosomen werden durch Kanten angezeigt. Es entsteht der *Genome Similarity Graph (GSG)* (Bild 2).

Die Gestalt des Graphen ist von unterschiedlichen Parametern abhängig, die interaktiv verändert werden können. So eignet sich z.B. eine Blockgröße von 500, um Ähnlichkeiten zwischen Genen zu identifizieren[12]. Für die Suche nach intergenen Bereichen empfiehlt sich eine Blockgröße von 50. Weitere Filterparameter sind der Molekültyp (DNA oder Protein), die Sensitivität der Ähnlichkeit, die Kategorisierung nach bekannten genetischen Elementen, usw. Es können dadurch bestimmte Bereiche, die nicht von Interesse sind, ausgeblendet werden.

Der Genombrowser kann darüberhinaus eingesetzt werden, die Vollständigkeit und Korrektheit der Annotation zu überprüfen. Damit wird die Qualität der gesamten Datenmenge erhöht. Das Finden von Information, die ansonsten in der Datenflut verlorengehen würde, wird durch den Genombrowser maßgeblich unterstützt.

4 CSCW in der Biologie: Das Funktionsanalyseprojekt

Umfangreiche Projekte wie z.B. die systematische Sequenzierung ganzer Genome lassen sich nur durch Zusammenarbeit mehrerer Laboratorien bewältigen. Da sich die Projektteilnehmer an verschiedenen Orten befinden und zu verschiedenen Zeitpunkten miteinander kommunizieren, eignet sich das *bulletin board*-Konzept [5] zur Koordination des Gruppenprozesses. Das World Wide Web ermöglicht allen Projektbeteiligten den Zugriff auf die gemeinsame Datensammlung (*repository*). Auf diese Weise findet eine Kooperation unmittelbar bei der Planung, Durchführung und Auswertung der Experimente statt.

Bei der Funktionsanalyse versucht man, experimentell die Funktion von Proteinen zu bestimmen. Dazu züchtet man neue Hefestämme aus einem Wildtypstamm, in dessen DNA jeweils ein offener Leserahmen gezielt zerstört wurde (*Disruption*). Anschließend vergleicht man das Wachstumsverhalten der neuen Stämme mit dem des Wildtypstamms.

Im deutschen Hefefunktionsanalyseprojekt übernimmt jedes Labor zwei Aufgaben: es führt einen Teil der erforderlichen Disruptionen durch und testet das Verhalten der

[10] m ist Länge des zu löschenden Sequenzeintrages, n die Länge des Textes.

[11] Komplette Einträge der Sequenzdatenbank.

[12] Die durchschnittliche Länge eines Gens in Hefe beträgt 1200 Nukleotide.

neuen Stämme. Jedes Labor kann zusätzlich spezielle Experimente an ausgewählten Stämmen durchführen. Um ein effizientes Arbeiten innerhalb des Projektes zu ermöglichen, müssen die Disruptionsergebnisse einer Gruppe möglichst schnell allen Beteiligten bekannt gemacht werden.

In Deutschland arbeiten elf Gruppen an 160 Disruptionen. Die Erfahrungen aus diesem Projekt fließen ein in das europäische Funktionsanalyseprojekt, an dem ca. 130 Laboratorien beteiligt sind. In beiden Projekten erstellt unsere Gruppe die gemeinsame Arbeitsumgebung, deren Hauptbestandteil die objektorientierte Datenbank der Disruptionsergebnisse darstellt. Darüberhinaus haben alle Gruppen Zugriff auf Daten und Dienste zur Visualisierung der ausgewählten Leserahmen der Hefesequenzen. Die Ergebnisse der einzelnen Projektabschnitte werden unmittelbar in das World Wide Web übernommen. Die Erfahrungen aus dem Funktionsanalyseprojekt haben gezeigt, daß nach anfänglichen Vorbehalten das World Wide Web von allen Projektmitgliedern als Arbeitsmedium akzepiert wurde.

5 Dienstintegration mit Hilfe des WWW

Die Nutzung biologischer Datenbanken ist gekennzeichnet durch eine stetig wachsende Menge von Rohdaten, unterschiedlichen Datenbanken und Datenbankanbietern sowie uneinheitlichen Zugängen zu diesen Ressourcen. Die Notwendigkeit von Datenbankinteroperabilität und effizienter Methoden zur Interdatenbankkommunikation gewinnt zunehmend an Bedeutung [8]. Solange diese Forderungen nicht erfüllt werden können, wird eine Integration dieses heterogenen Systems unter einer einheitlichen und intuitiven Benutzungsoberfläche durch eine hierarchische Software-Architektur erreicht [11]. Existierende Dienste müssen dabei nicht verändert oder angepaßt werden. Allerdings müssen dienstspezifische Charakteristika, wie z.B. unterschiedliches zeitliches Verhalten oder zustandslos versus sitzungsorientiert, durch eine übergeordnete Schicht einheitlich in das globale System integriert werden. Die Kommunikation mit diesem System erfolgt durch gewohnte WWW-Browser, wodurch die Akzeptanz des Systems erhöht werden kann. Die aktuelle Statistik unseres WWW-Servers[13] mit 100 000 Anfragen von 2 000 unterschiedlichen Institutionen pro Monat bestätigt dies.

Das Schichtenmodell erlaubt Verbindungen zwischen Diensten herzustellen, die sich im Modell auf einer Ebene befinden. Limitierend bei der Modellierung von Beziehungen zwischen Diensten ist das Fehlen formaler Definitionen von Datenbankressourcen. Die Konvertierung von Information in konsistenter und eindeutiger Weise in verschiedene physikalische Repräsentationen wird dadurch erschwert. Das im folgenden dargestellte System löst nicht das Problem semantischer Inkonsistenzen, erlaubt aber einen einheitlichen Zugriff auf benötigte Information.

Das hierarchische Modell beinhaltet folgende Schichten mit folgenden Aufgaben:

- *Interface Layer:* Es wird eine homogene intuitive graphische Benutzungsoberfläche angeboten, die einen plattform- und lokalitätsunabhängigen einheitlichen Zugang zu den angebotenen Diensten des Systems ermöglicht. Aufgrund der hohen Verbreitung und Akzeptanz entschieden wir uns für das World Wide Web.
- *Link Layer:* Diese Schicht dient als Zwischenschicht, um Verbindungen zwischen unabhängigen Diensten herstellen zu können. Alle zur Verfügung gestellten Dienste sind in dieser Schicht bekannt. Die Ausgabe von Informationen bzw. Ergebnissen von Anfragen muß abhängig von der eingesetzten Benutzungsoberfläche im entsprechenden Format erfolgen, wie z.B. in *HTML*.
- *Gateway Layer:* Diese Schicht muß die Reglementierungen kompensieren, die aufgrund der Dienstintegration auftreten. Existierende Programme, die Bestandteil des Systems sind, sollen nicht verändert werden. Sind die Dienste über unterschiedliche Plattformen verteilt, wird diese Heterogenität im *Gateway Layer* behandelt und nach oben hin verdeckt.
- *Service Layer:* Hier sind alle Dienste lokalisiert, die nach außen hin angeboten werden. Insbesondere auch die unterschiedlichen Sequenzdatenbanken mit dazugehörgien Anfragemaschinerien.
- *Synchronization Layer:* In dieser Schicht wird die Integrität von Daten überwacht. Sie ist orthogonal zu den übergeordneten Schichten.

6 Synchronisation biologischer Datenbanken

Für eine effiziente Sequenzdatenanalyse wird der komplette konsistente Datensatz lokal benötigt. Die Replikation einer Datenbank an mehrere Orte über ein Netzwerk erfordert ein Synchronisationskonzept, das globale Konsistenz gewährleistet.

Hier wird als zugrundeliegendes Netzwerk das öffentliche Internet angenommen. Dabei handelt es sich um ein unzuverlässiges Weitverkehrsnetz, in dem Verbindungen nicht zustande kommen können bzw. nicht die benötigte Zeit verfügbar sind. Es ist nicht realistisch anzunehmen, daß alle Replikate über eine solche Kommunikationsinfrastruktur nach einer Zustandsänderung an einer Kopie sofort synchronisiert werden können. Die Unerreichbarkeit eines Knotens in diesem verteilten Datenbanksystem soll aber nicht zu einem Stop der gesamten Synchronisation führen. Das von uns entwickelte Konzept garantiert, daß alle Datenbankzustände zu einem gemeinsamen Zustand konvergieren [12].

Konzeptionell wird die Menge aller Knoten des verteilten Sy-

[13] http://www.mips.biochem.mpg.de/

stems unterteilt in privilegierte, die den Zustand der Datenbank ändern dürfen, und nicht-privilegierte. Jeder privilegierte Knoten ist Ausgangspunkt für ein Datenverteilsystem [10]. Transaktionen, die potentiell den Datenbankzustand verändern, werden an eine virtuelle Primärdatenbank geschickt, prozessiert und im Erfolgsfall mit einer globalen Transaktionsnummer versehen. Diese Transaktion wird anschließend zusammen mit der globalen Transaktionsnummer an alle privilegierte Knoten weitergesandt. Nach dortiger lokaler Prozessierung werden sie dem jeweiligen Datenverteilsystem übergeben, das das Versenden an alle Knoten des Systems garantiert. Fehlende Transaktionen aufgrund von Netzwerk- oder Systemfehlern können anhand der Transaktionsnummer lokal erkannt und vom übergeordneten Nachbarn angefordert werden. Als Kommunikationsmechanismus werden *Remote Procedure Calls (RPCs)* [2] verwendet.

Um dieses Konzept realisieren zu können, wird innerhalb des *Synchronization Layers* ebenfalls eine hierarchische Software-Architektur verwendet. Server, die Datenbankzugang besitzen, werden in einer *Service-Schicht* zusammengefaßt. Verwaltet werden diese Dienste von einem Server in der benachbarten *Management-Schicht*, dem *Multiserver*. Dieser spezielle Server, von dem in jedem Knoten genau eine Instantiierung existiert, erfüllt folgende Aufgaben:

- Der *Multiserver* akzeptiert Anfragen von Clients und überprüft sie auf Gültigkeit. Dazu sind ihm alle angebotenen Dienste der *Service-Schicht* bekannt. Gültige Anfragen werden an den entsprechenden Server in der *Service-Schicht* weitergereicht. Ist dieser temporär nicht verfügbar, wird eine Warteschlange verwaltet.
- Die Menge der Server in der *Service-Schicht* kann vom *Multiserver* dynamisch verändert werden. Dadurch können Engpässe bei der Bearbeitung von Anfragen gleichen Typs vermieden werden, indem die Menge der dienstanbietenden Server vergrößert wird. Globale Datenintegrität wird beachtet, indem eine Parallelisierung nur bei idempotenten Operationen zugelassen wird.
- Der *Multiserver* kann eigenständig Anfragen an Server initiieren. Dies wird z.B. im Rahmen der Synchronisation eingesetzt.

7 Ausblick

Das schnell wachsende Interesse an biologischer Information, vor allem im Bereich der biomedizinischen und biotechnologischen Forschung, erfordert eine Dateninfrastruktur, die die gestellten Anforderungen nur durch den Einsatz einer leistungsfähigen Informatik bewältigen kann. Effiziente Algorithmen in der Sequenz- und Strukturanalyse komplexer biologischer Moleküle, Datenbanken und verteilte Anwendungen sind Einsatzgebiete von Konzepten der Informatik in der modernen molekularen Biologie.

Die international verteilten Ressourcen konnten durch das World Wide Web mit den dazugehörigen Browsern als graphische Bedienoberflächen allgemein zugänglich gemacht werden. Die Umstellung von hypertext- zu funktionsorientierten Applikationen, basierend auf der WWW-Programmiersprache Java[14], ist die logische Weiterentwicklung. Diese Technologie erhöht die plattformunabhängige Funktionalität. Die Entwicklung und Wartung bestehender Programme und intuitiver Benutzungsschnittstellen wird dadurch schneller und robuster. Betriebssystemspezifische Portierungen von Software können reduziert werden und damit die Fehleranfälligkeit des gesamten Systems verringern.

Die Bereitstellung biologischer Datenbanken mit den dazugehörenden Zugriffsmechanismen auf CD-ROM erlaubt eine effiziente Installation auf lokalen Netzwerken. In Kürze wird das komplette Hefegenom mit Java-gesteuerter Software auf CD-ROM verfügbar sein. Die Produktion und der Vertrieb von Datensätzen und Software auf optischen Medien ist dann unzulänglich, wenn unter Wettbewerbsgesichtspunkten biologische Datenbanken ständig in ihrer aktuellen Form vorliegen müssen. Daher ist die Synchronisation von kompletten Datenbankkopien über ein gemeinsames Netzwerk ein wichtiger Akzeptanzfaktor.

Die Modellierung biologischer Daten in relationalen Schemata ist nicht flexibel und leistungsfähig genug. Die moderne Genomforschung beschäftigt sich mit komplexen Objekten des Lebens bzw. des Organismus und ihren Beziehungen. Eine objektorientierte Datenmodellierung und die Verwendung objektorientierter Datenbanken erlaubt eine direktere Abbildung der Wirklichkeit. Zusammen mit industriellen und akademischen Partnern realisieren wir in einem europäischen Verbundprojekt die Migration der Proteinsequenzdatenbank *PIR-International* zu einer objektorientierten Datenrepräsentation. Als Datenbankmanagementsystem verwenden wir dabei *ObjectStore* von Object Design [15].

In Ergänzung untersuchen wir, ob der Industriestandard CORBA[15] zur Integration existierender Datenbanken über ein Weitverkehrsnetz eingesetzt werden kann[16].

Die Informatik hat in den Biowissenschaften ein neues faszinierendes Anwendungsgebiet gefunden. Es gilt, eine Vielzahl von Herausforderungen zu bestehen. Noch können wir nur hoffen, daß diese Entwicklung von den deutschen Hochschulen aufgegriffen und in ihr bestehendes Lehrgebäude integriert wird, doch sind schon vorbildliche Ansätze zu erkennen (z.B. der Studiengang naturwissen-

[14] Java wurde von Sun Microsystems entwickelt (http://java.sun.com/).

[15] Common Object Request Broker Architecture, entwickelt von der Object Management Group (OMG), siehe z.B. http://www.acl.lanl.gov/CORBA/

[16] Projekt gemeinsam mit dem European Institute of Bioinformatics, Cambridge.

schaftliche Informatik, FB Informatik, Univ. Bielefeld, Prof. Giegerich).

Literatur

[1] *Barker, W. C., George, D. G., Hunt, L. T.*: Protein Sequence Database, volume 183, pp. 31–49. Academic Press, New York, 1990.

[2] *Birrell, A., Nelson, B.*: Implementing remote procedure calls. ACM Transactions on Computer Systems, 2(1), February 1984, pp. 39–59.

[3] *Dayhoff, M.*: Atlas of protein sequences and structure. National Biomedical Research Foundation, Silver Spring, Maryland, 1978.

[4] *Goffeau, A. (ed.)*: Sequencing the yeast genome. A detailed assessment. Technical report, Commision of the European Communities, 1988.

[5] *Ellis, C. A., Gibbs, S. J., Rein, G. L.*: Groupware: Some issues and experiences. Communications of the ACM, 34(1), January 1991, pp. 38–58.

[6] *George, D. G., Orcutt, B. C., Mewes, H. W., Tsugita, A.*: An object-oriented sequence database definition language (SDDL). Prot. Seq. Data Anal., 5, 1993, pp. 357–399.

[7] *George, D. G., Barker, W. C., Mewes, H. W., Pfeiffer, F., Tsugita, A.*: The pir-international protein sequence database, Nucleic Acids Research, 24(1), 1996, pp. 17–20.

[8] *George, D. G., Maierl, A., Heumann, K., Mewes, H. W.*: The quest of a common data model. In: Second Meeting on the Interconnection of Molecular Biology Databases. Cambridge, United Kingdom, July 1995.

[9] *Heumann, K.*: Biologische Sequenzdatenanalyse großer Datensätze basierend auf Positionsbaumvarianten. Dissertation, Technische Universität München, 1996.

[10] *Heumann, K., George, D. G., Mewes, H. W.*: A new concept of sequence data distribution on wide area networks. Comput. Appl. Biosci., 10, 1994, pp. 519–526.

[11] *Heumann, K., Harris, C., Kaps, A., Liebl, S., Maierl, A., Pfeiffer, F., Mewes, H. W.*: An integrated services approach to biological sequence databases. In: Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism. GBF Braunschweig, 1996.

[12] *Kaps, A.*: Konsistenzsicherung in einem verteilten objektorientierten Datenbanksystem. Diplomarbeit, Technische Universität München, 1995.

[13] *Mewes, H. W., Heumann, K.*: Genome analysis: – pattern search in biological macromolecules. Technical Report 7, Max-Planck-Inst. f. Biochemie, 82152 Martinsried, Germany, May 1995.

[14] *Needleman, S. B., Wunsch, C. D.*: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48, 1970, pp. 443–453.

[15] Object Design, Inc., Twenty Five Mall Road, Burlington, MA 01803-4194, USA. ObjectStore Reference Manual, release 4.0 beta edition, March 1995.

Anschrift der Autoren:
Max-Planck-Institut für Biochemie,
Abteilung MIPS, Am Klopferspitz,
D-82152 Martinsried,
Email: {kaps,heumann,maierl,mewes}
@mips.embnet.org

Gigabytes über Gigabasen – Informationsintegration in der Genomforschung

Otto Ritter, Sándor Suhai, Deutsches Krebsforschungszentrum, Heidelberg

Dr. habil. Sándor Suhai studierte Physik, Mathematik und Chemie in Budapest, Göttingen und Erlangen und wurde in Theoretischer Physik an der Universität Budapest und in Theoretischer Chemie an der Friedrich-Alexander-Universität in Erlangen promoviert. 1984 habilitierte er sich in Erlangen. Dr. Suhai ist Leiter der Abteilung Molekulare Biophysik des Deutschen Krebsforschungszentrums in Heidelberg. Seine Forschungsschwerpunkte liegen hauptsächlich auf dem Gebiet der computergestützten Modellierung von biomolekularen Phänomenen und in der Entwicklung mathematischer und informatorischer Methoden für die Analyse der Genome höherer Organismen.

Dr. Otto Ritter studierte Mathematik und wurde 1985 an der Charles-Universität in Prag in Mathematik (Theoretische Kybernetik, mathematische Informatik und Systemtheorie) promoviert. Er ist wissenschaftlicherphysik des Deutschen Krebsforschungszentrums (DKFZ). Sein Forschungsschwerpunkt beinhaltet mathematisches Modellieren in der Genetik sowie Darstellungen und Entwicklung von Softwaresystemen für biologische Anwendungen.

Die Genomforschung ist in hohem Maße eine Informationswissenschaft. Der Zugang zu den relevanten Informationen sowie zu Softwarewerkzeugen für die Verarbeitung, die Kombination, den Vergleich, die Analyse und die Darstellung der Daten und des Wissens spielen eine immer wichtigere Rolle in der gesamten Molekularbiologie – in der Genomforschung sogar eine absolut zentrale Rolle. Dieser Artikel berichtet über offene Fragen im Zusammenhang mit dem Aufbau von übergreifenden Informations-Managementsystemen, die auf bestehenden heterogenen und autonomen Datenbanken und analytischen Softwarewerkzeugen aufbauen.

Merging Gigabytes about Gigabases – On Information Integration in Genome Research

Genome research is to a large extent information science. Access to relevant information in context, and to software tools for processing, combining, comparing, analyzing and rendering the data and knowledge, plays an increasingly important role in the whole molecular biology, and an absolutely critical one in genome research. The paper reports on open issues in building comprehensive information management systems on top of preexisting heterogeneous and autonomous databases and analytical software tools.

1 Einleitung

Die molekularbiologische und die genetische Forschung werden immer datenintensiver und informationsabhängiger. Der schnelle Zugriff auf die relevanten Informationen in ihrem Gesamtkontext sowie auf Softwarewerkzeuge für die Verarbeitung, die Kombination, den Vergleich, die Analyse und die Darstellung der Informationen und des Wissens spielen eine absolut zentrale Rolle für das Humangenomprojekt.

Die genomischen Informationen sind auf hunderte heterogener und autonomer Informationssysteme verteilt [12]. Diese Systeme besitzen ihre eigenen (oftmals idiosynkratischen) Interfaces und Kontrollsprachen und stellen die Informationen mit inkompatiblen Datenmodellen und Formaten dar. Diese Heterogenität und Autonomie bestehen aufgrund von soziologischen, wissenschaftlichen und technologischen Unterschieden, und sie sind in vielen Fällen angemessen und sinnvoll. Gleichzeitig stellen sie jedoch ein großes Problem für die Endbenutzer dar und beschränken den potentiellen synergetischen Gebrauch der verwalteten Informationen. Die Benutzer einschließlich der Anwendungsprogramme bevorzugen einen logisch integrierten und konsistenten Zugang zu allen Informationen in ihrer Domäne und zu allen assoziierten Operationen.

2 Das Interconnection-Problem

Das Problem der Informations-Interconnection läßt sich erklären als ein Problem des effizienten und konsistenten Zugangs zu molekularbiologischen Datenbanken (MBDs) und ihren Operationen sowie zu den domänenspezifischen externen Anwendungen [1].

In technischer Hinsicht lassen sich die Probleme bei der Entwicklung und Systematisierung vergleichen mit denen bei der Geschäfts- oder Büroautomatisierung. Es bestehen jedoch einige spezifische Unterschiede:

- große Datenmengen (Terabytes, exponentieller Anstieg),
- sehr große und sehr komplexe Metadaten,
- große Anzahl von komplexen Operatoren,
- große Anzahl von heterogenen und autonomen Komponenten (Legacy-Datenquellen und analytische Werkzeuge),
- hohe Dynamik (Metadaten, Daten und Methoden ändern sich häufig und neue werden hinzugefügt),

- inherente Unvollständigkeit und Verschwommenheit der Informationen und das Fehlen von Erklärungsmodellen hinter den Daten.

3 Primärsysteme

Viele Primär-MBD-Systeme wurden anfänglich als strukturierte Textdatei-Bibliotheken (z.B. Tagged-ASCII-Dateien) verwaltet und verteilt. Beispiele sind PDB [16], SWISS-PROT [19] und OMIM [14]. Einige dieser Systeme sind mittlerweile zu einem weiterentwickelten Managementmechanismus übergegangen (PDB zu OPM/Sybase [15; 20], OMIM zu einem auf SGML basierenden Dokument-Management-System), aber sie alle verteilen die Daten immer noch hauptsächlich als ASCII-Bibliotheken. Es gibt viele Legacysysteme, bei denen die Daten in Dateien aufbewahrt werden, und in einigen Fällen fehlen die Metadaten, sind überaltet oder unvollständig. Die dort aufbewahrten Informationen sind jedoch trotz alledem wertvoll. Auch heute noch sind viele kleinere MDBs oder zu einem früheren Zeitpunkt generierte Datensätze (z.B. die physikalische CEPH-Genethon [3] Humangenomkarte von 1993) ASCII-Dateien. Bei diesen Systemen und Datensammlungen ist es generell sehr schwierig, sie miteinander zu verbinden.

Größere und in jüngerer Zeit entstandene MDBs werden in vielen Fällen als relationale Datenbanken verwaltet (GDB, FlyBase, EMBL) [18]. Sie bieten Zugang zu den Daten durch mehrere Kanäle, so z.B. direktes SQL, indirektes SQL durch eine Bibliothek von gespeicherten Prozeduren oder 4GL-Benutzeroberflächen oder durch Hypertext-Oberflächen wie z.B. WWW-Clients. Darüber hinaus exportieren diese Systeme gewöhnlich Daten in genau definierte Dateien im ASCII-Format und machen sie so verfügbar. Einige von diesen auf RDBMS basierenden MDBs – vor allem solche mit Objektschichten wie z.B. OPM – wollen in der Zukunft den Datenzugang durch Objektbroker anbieten, die kompatibel oder ähnlich zu formalen Modellen wie z.B. CORBA sind.

Andere Beispiele für relationale MBDs sind Datenbanken, die die Genomkartierung und die Sequenzierung auf längeren Abschnitten oder für größere Kooperationen (z.B. LLNL, TIGR, CEPH) [2; 11; 21] unterstützen. Der Zugang zu diesem Datentyp ist beschränkt, die öffentlichen Daten werden gewöhnlich als ASCII-Dateien verteilt.

Viele der Primär-Genom-MBDs sind Datenbanken, die auf ACEDB[5] basieren, z.B. Gemeinschaftsdatenbanken für den Fadenwurm, Hefe sowie verschiedene Pflanzen und Tiere. Mehrere Labors und Einzel-Chromosom-Konsortien, die das menschliche Genom erforschen, verwenden ACEDB zur Verwaltung ihrer experimentellen und gemeinsam genutzten Daten – in den meisten Fällen sind diese Systeme Ableitungen oder Anwendungen von IGD/ACEDB[10; 17].

Verhältnismäßig wenige MBDs sind deduktive Datenbanken, Experten- oder wissensbasierte Systeme. Sie liefern die Daten entweder in traditionellen Formaten ihrer relationalen oder OO-Komponenten oder als strukturierte ASCII-Transkription ihrer Fakten und Regeln (z.B. Knowledge Interchange Format (KIF)).

Viele bibliographische und elektronische Publikationssysteme liefern die Daten in einem Markup-Sprachenformat (gewöhnlich SGML oder HTML) oder als Compound-Dokumentformat (gewöhnlich PDF). Verschiedene Primär- und Sekundär-MDBs liefern die Daten im ASN.1-Format, auch wenn sie kein auf ASN.1 basierendes Informations-Management verwenden. Erwähnenswerte Beispiele sind die NCBI-Datenbanken [13] wie z.B. GenInfo, ENTREZ, GenBank. Außer in einfachen, strukturierten und Hypertext-Formaten sind wichtige Daten immer häufiger in diversen graphischen, Still-Image-, Audio/Video- und Multimedia-Formaten verfügbar.

In technischer Hinsicht sind die Daten und Metadaten der MBDs entweder über das Internet oder als Datenpaket auf magnetooptischen Medien (Bänder, CDs, Disketten) verfügbar. Ob es leicht oder schwer ist, eine Verbindung zwischen den Inhalten einer MBD und anderen MBDs herzustellen, hängt hauptsächlich von mindestens drei Faktoren ab:

- Art der Anfrage/Update Operationen, die die MBD unterstützt.
- Qualität und Analysierbarkeit der exportierten Metadaten und Daten.
- Standards und Konventionen bei der Benennung der Objekte und Elemente der (Meta)daten. Eindeutige und dokumentierte Semantik aller Datentypen, Operationen und Beschränkungen.

Update-Operationen sind nur bei operationellen Interconnections wichtig. Anfrageoperationen sind wichtig, wenn Exporte der Datenpakete nicht verfügbar oder nicht auf dem neuesten Stand sind. Qualität und Analysierbarkeit der Metadaten sind von höchster Wichtigkeit. Konventionen bei der Benennung und die Dokumentation der Semantik spielen eine wichtige Rolle bei der Bestimmung der Beziehungen zwischen den MBDs sowohl auf der Daten- als auch der Metadatenebene.

4 Externe Anwendungen

Externe Anwendungen sind gemäß unserer Definition domänenspezifische Programme – aber keine MBDs – zusammen mit nichtspezifischen Werkzeugen und Serviceeinrichtungen.

Domänenspezifische Anwendungen ermöglichen die Analyse, den Vergleich, die Visualisierung, die Durchsicht und/ oder Umwandlung von molekularbiologischen Daten. Kennzeichnend ist, daß sie Funktionen beinhalten, die sich mit internen (vom Programmieraspekt her unvollständigen) Operationen der MBDs nicht durchführen lassen, oder sie tun es auf effizientere Weise. In den meisten Fällen sind

diese Anwendungen autonome Programme oder Programmkollektionen, unabhängig von einer bestimmten MBD. Daten-I/O-Formate und Kontrolloberflächen variieren beträchtlich. Als Beispiele genannt werden können Editoren und Browser von molekularen Strukturen und Sequenzen, genetische und physikalische Karten, Stammbäume und metabolische Pfade. Andere Anwendungen zielen ab auf die Vorhersage von bestimmten Eigenschaften von biologischen Datenstrukturen, z.B. Exons und Introns auf genomischer DNA, funktionelle Domänen oder Sekundärstrukturen von Proteinen, etc. sowie auf die Berechnung von abgeleiteten Strukturen, z.B. bearbeitete Sequenzen oder Sequenzprofile.

Zu den externen Anwendungen, die nicht spezifisch für die MBD-Domäne sind, gehören verschiedene Arten von Editoren, Werkzeuge für die persönliche Effizienz, Formatkonvertern, Display- und Kommunikationswerkzeuge, Werkzeuge zur Komprimierung und Verschlüsselung, etc. Diese nicht-spezifischen Anwendungen haben gewöhnlich bekannte (manchmal urheberrechtlich geschützte) I/O-Formate und Kontrolloberflächen.

Um MBDs mit externen Anwendungen zu verbinden, benötigt man eine formalisierte Repräsentation der Anwendungsfunktionen, der Kontroll- und I/O-Formate, der operationellen Beschränkungen und der Zugangsmethoden. Da sie alle Daten oder Metadaten produzieren bzw. konsumieren, können sie als eine spezielle Art von MBDs angesehen werden – allerdings mit idiosynkratischen und manchmal recht konfusen Informationsmodellen. Einige dieser Anwendungen wurden bereits in Sekundärsystemen eng verknüpft mit richtigen MBDs, z.B. die GCG- oder HUSAR-Pakete [8; 9] für molekulare Sequenzdaten und -analyse.

5 Derzeitige Lösungen

In den letzten Jahren wurden verschiedene Systeme entwickelt, um die Heterogenität und Inkompatibilität zu überwinden. Zentren für Biocomputing wie z.B das EDR-Konsortium (DKFZ, HGMP und INSERM), das EBI und die EMBnet-Knoten in Europa oder das NCBI [13] in den USA, arbeiten mit mehr oder weniger einheitlichen Zugangsoberflächen zu einigen replizierten oder Ferndatenbanken, analytischen Werkzeugen oder kombinierten Paketen. Index- und Abrufsysteme wie z.B. IRx, WAIS oder SRS [6] bieten einheitliche Oberflächen für Anfragen und Browsing auf unterschiedlichen Datenformaten sowie locker verbundene Netzwerke von Datenbanken mit Hypertext (WWW). Cross-Referenzen zwischen Datenbanken ermöglichen eine gewisse Navigations-Verbindung ohne globales Schema, Anfragenoptimierung, Entdeckung und Behebung von Konflikten auf Daten- oder Metadaten-Ebene etc. Die beiden besten und vollständigsten Systeme sind zur Zeit das europäische IGD-Projekt [10] und der US-amerikanische GenomeTopographer [4].

6 Die integrierte genomische Datenbank (IGD)

IGD ist ein internationales Kollaborationsprojekt mit der Zielsetzung, ein offenes Informations-Management-System für (hauptsächlich menschliche) genombezogene Daten und analytische Werkzeuge zu entwickeln. IGD integriert Informationen aus öffentlichen Datensammlungen in eine einzige logische Datenbank mit Zugang über das Internet und bietet ein graphisches Front-End zur Verwaltung und Analyse von aus öffentlichen Daten und/oder lokalen experimentellen Datensätzen gewonnenen Datenmengen.

Als Datenbank integriert IGD und verweist auf genombezogene Daten aus öffentlichen Quellen. Dies sind die sogenannten ‚Resource End Databases' (IGD-REDs). Zu ihnen gehören GDB, OMIM, EMBL, SWISS-PROT, RLDB, DNA Probe Bank und viele andere Datenbanken, auch über Mäuse und landwirtschaftlich genutzte Tiere. RED-Datenbanken sind sehr heterogen hinsichtlich der ihnen zugrundeliegenden Datenmodelle und Datenbank-Management-Systeme, und sie sind völlig autonom – es ist keinerlei Zusammenarbeit von seiten der potentiellen RED-Stelle notwendig, um in IGD integriert zu werden.

Daten aus den IGD-REDs werden in regelmäßigen Abständen gesammelt, reformatiert und auf verschiedene IGD-Server, die sogenannten ‚Target End Databases' (IGD-TEDs) exportiert. Alle diese IGD-TEDs haben das gleiche konzeptuelle Schema, aber sie unterscheiden sich zum Teil hinsichtlich der physikalischen Implementierung. Zur Zeit werden zwei Implementierungen unterstützt: die rationale (Sybase) und die nicht-standardisierte (ACEDB).

Als Analysewerkzeug bietet IGD eine einheitliche Oberfläche zu existierenden Programmen und Programmpaketen für Struktur- und Sequenzanalyse, genetische und physikalische Kartierung und Analyse, etc.

Die Benutzer interagieren mit dem IGD-System mittels verschiedener lokal installierter ‚Front End'-Werkzeuge (IGD-FRED). Die wichtigsten Teile des IGD-FRED sind der lokale Datenbankmanager und Oberflächen für Kommunikation und Analyse. Die Benutzer können an den IGD-TED Anfragen richten und die Daten, die sie daraufhin erhalten, in ihre lokale Datenbank laden. Sie speichern auch private Daten und Analyseergebnisse in der lokalen Datenbank ab. Das IGD Front-End kann beliebige Untermengen dieser Daten bearbeiten, und es kann auch lokale Daten verwalten. Das Schema unterstützt die Speicherung von detaillierten experimentellen Daten bei Sequenz- und Kartierungsprojekten und ist lokal erweiterbar. Datenobjekte haben Verbindungen zu externen Methoden. Hierzu gehören URLs zu ihren Herkunftsquellen (so können z.B. Objekte aus GDB [7] in ihren Originalversionen auf den Bildschirm geholt werden, indem man den WWW-

Server bei GDB anwählt) und Methoden, um lokale oder entfernte analytische Programme aufzurufen, so z.B. das HUSAR/GCG-Paket [9] bestehend aus 140+ Sequenz- und Strukturanalyse-Werkzeugen, einem Stammbaum-Displaytool, einer Oberfläche zur genetischen Analyse (Cri-Map) sowie in Kürze auch eine Oberfläche zur Assemblierung/Analyse von physikalischen Karten.

Alle IGD-TEDs pflegen mehr oder weniger die gleichen Daten (unter schwachen Konsistenz-Beschränkungen), die Daten sind „Read-Only" für alle FRED-Clients. Neue, bei den FRED-Stellen gewonnene Daten, werden (im geeigneten Format) an die relevanten REDs weitergeleitet mittels Übertragungskanälen, mit dem die jeweilige RED ausgestattet wurde.

7 Schlußbemerkungen

Genomforscher haben große Schwierigkeiten mit dem immer weiter wachsenden Umfang und der Komplexität von relevanten Daten, die über das Internet verfügbar sind. Eine effiziente Infrastruktur für den Austausch und den Zugriff auf dieses Wissen wird gebraucht, um einen konsistenten deklarativen Zugang zu den Datenquellen und Funktionen zu gewährleisten. Trotz ersten Annährungen, wie mit IGD, sind derzeit keine befriedigenden Lösungen verfügbar oder in Sicht.

Danksagung

Die Arbeit an IGD wird von der Europäischen Kommission mit Fördermitteln unterstützt (GENE-CT93-0003 / DG 12 SSMA). Die Autoren danken Frau Anke Retzmann für die Übersetzung dieses Artikels aus dem Englischen.

Literatur

[1] Abstracts of the Second Meeting on Interconnection of Molecular Biology Databases; abrufbar unter http://www-genome.wi.mit.edu/informatics/abstracts.html.

[2] CEPH – Fondation Jean Dausset; http://www.cephb.fr/

[3] CEPH-Genethon integrated map; http://www.cephb.fr/ceph-genethon-map.html

[4] *Cozza, S., Reed, E. C., Salit, J., Chang, W., Marr, T.*: Genome Topographer: A Next Generation Genome Database System (Abstract), vorgelegt bei dem „Meeting on Genome Mapping and Sequencing", Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 1994.

[5] *Durbin, R., and Thierry-Mieg, J.*: Syntactic Definitions for the ACeDB Data Base Manager, 1992; abrufbar unter http://probe.nalusda.gov:8000/acedocs/.

[6] *Etzold, T., Argos, P.*: SRS, An Indexing and Retrieval Tools for Flat File Data Libraries. Computer Applications of Biosciences, 9, 1993, pp. 49–57.

[7] *Fasman, K.*: Restructuring the Genome Data Base: A Model for a Federation of Biological Databases. Journal of Computational Biology, 1(2), 1994, pp. 165–171.

[8] Genetics Computer Group; http://www.gcg.com/

[9] HUSAR – Heidelberg Unix Sequence Analysis Resources; http://genome.dkfz-heidelberg.de/biounit/

[10] The Integrated Genomic Database project; http://genome.dkfz-heidelberg.de/igd/

[11] Lawrence Livermore National Laboratory; http://www.llnl.gov/llnl/04bio.html

[12] *Markowitz, V. M., Ritter, O.*: There's Never Time to Do It Right, But Always Time to Do It Over, Characterizing Heterogeneous Molecular Biology Database Systems, J Comp Biol Vol 2, Nr 4, 1995.

[13] National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/

[14] Online Mendelian Inheritance in Man (OMIM); http://www.ncbi.nlm.nih.gov/omim/

[15] The Object Protocol Model; http://gizmo.lbl.gov/opm.html

[16] Protein Data Bank; http://www.pdb.bnl.gov/

[17] *Ritter, O.*: The Integrated Genomic Database. In Suhai, S. (ed.): Computational Methods in Genome Research. Plenum, New York, 1994, pp. 57–73.

[18] Special GENOME Issue of Nucleic Acids Research, September 1994; http://www-lmmb.ncifcrf.gov/NAR-sept94.html

[19] The SWISS-PROT protein sequence data bank; http://expasy.hcuge.ch/sprot/sprot-top.html

[20] Sybase Inc.; http://www.sybase.com/

[21] The Institute for Genomic Research; http://www.tigr.org/

Dr. habil. Sándor Suhai, Dr. Otto Ritter
Abt. Molekulare Biophysik,
Deutsches Krebsforschungszentrum,
69120 Heidelberg,
Email: {O.Ritter,S.Suhai}@DKFZ-Heidelberg.de

Von Genomsequenzen zu Proteinfunktionen

Peer Bork, Europäisches Laboratorium für Molekularbiologie, Heidelberg

Dr.-habil. Peer Bork studierte von 1983 bis 1988 Biochemie und wurde 1990 über die Mustererkennung in Sequenzen promoviert. 1995 habilitierte er sich in theoretischer Biophysik. Seit 1990 ist er Wissenschaftler am MDC für Molekulare Medizin in Berlin und seit 1991 Gastwissenschaftler am Europäischen Laboratorium für Molekularbiologie (EMLB) in Heidelberg. Dr. Bork beschäftigt sich hauptsächlich mit der vergleichenden Sequenz- und Genomanalyse, molekularer Evolution sowie mit Funktions- und Stoffwechselwegvorhersagen.

Mit dem Vorliegen kompletter Genome in Textform bricht für die Molekularbiologie und Molekularmedizin ein neues Zeitalter an, da das Erbmaterial im Prinzip alle Informationen für Lebenszyklus und Vermehrung von Zellen und Organismen enthält. Auf dem Weg zur Entschlüsselung dieser Information müssen zuerst die Informationsträger, die Gene, erkannt und die entsprechenden Funktionsträger, die Genprodukte (Proteine) charakterisiert werden. In diesem Beitrag wird in vergleichende sequenzanalytische Methoden im Prozeß der Funktionsvorhersage von Proteinen eingeführt und ihre Stärken und Schwächen beleuchtet.

From Genom Sequences to Protein Functions

The successful sequencing of complete genomes is the beginning of a new age in Molecular Biology and Molecular Medicine. Encoding the information of a complete genome should enable us to understand fundamental processes of life. A first step in the understanding of these molecular processes is the identification of the genes and their gene products, the proteins, as well as their functional characterization. This introduction into the prediction of protein function using comparative sequence analysis will also deal with the powers and pitfalls of the currently used methodology.

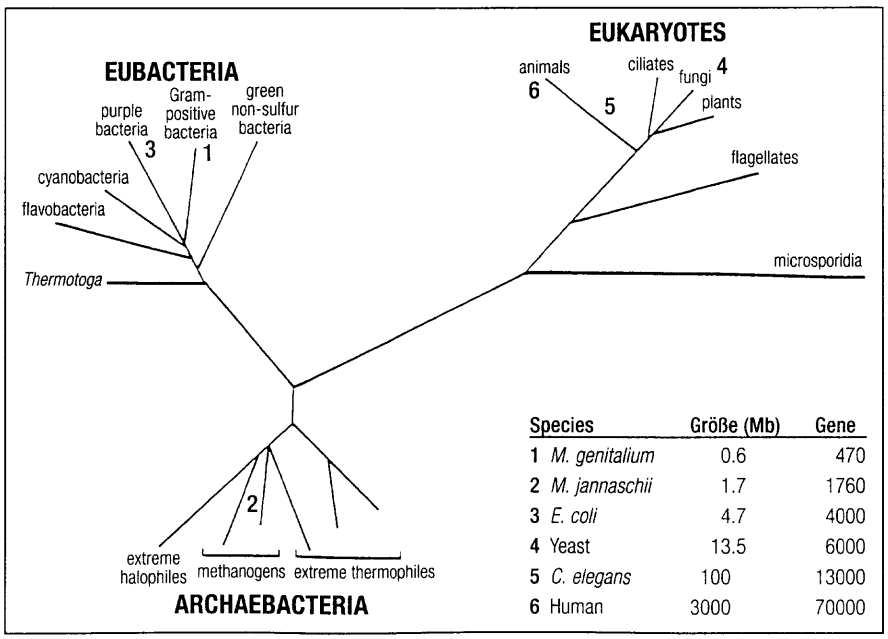
1 Einleitung

1.1 Die „Postgenom"-Ära in der Biologie

In der Molekularbiologie ist eine neue Ära angebrochen – die der komplett sequenzierten Genome zellulärer Organismen. Am 28. Juli 1995 wurde das Erbmaterial des parasitisch lebenden Bakteriums *Haemophilus influenza* veröffentlicht [1]. Neben den Genomen einiger Eubakterien liegt nun seit 1996 auch von Archaebakterien und mit der Bäckerhefe auch von Eukaryoten das komplette Erbmaterial in Form von 4-Buchstabentexten vor (für die 4 Basen **A**denine, **G**uanin, **C**ytidin und **T**hymidin). Schon bald werden komplexere Genome wie die des Nematoden *C.elegans* (im Jahre 1998) und auch des Menschen (derzeitige Schätzungen gehen von 2003 aus) folgen (Bild 1). Damit eröffnet sich ein völlig neuer Zugang für das Verständnis der Grundstrukturen des Lebens, den Zellen [2].

Das Erbgut (die komplette DNA eines Organismus) sollte alle Informationen enthalten, die eine Zelle entstehen, sich entwickeln, mit ihrer Umwelt interagieren, und sie als ein offenes Fließgleichgewicht bis zu ihrem vorprogrammierten Tod erhalten lassen. Die Entschlüsselung dieser Information ist eine gewaltige Aufgabe, bei der wir erst am Anfang stehen (Bild 2). Wie kann man aus der Erbin-

Bild 1: Speciesevolution und Genomdaten. Ein Stammbaum basierend auf RNA-Daten (aus [19]) ist mit Proteinanzahl und Genomgrößen von Modellorganismen verglichen, die bereits oder in naher Zukunft komplett vorliegen werden.



Species	Größe (Mb)	Gene
1 *M. genitalium*	0.6	470
2 *M. jannaschii*	1.7	1760
3 *E. coli*	4.7	4000
4 Yeast	13.5	6000
5 *C. elegans*	100	13000
6 Human	3000	70000

formation auf den Lebenszyklus einer Zelle oder gar eines multizellulären Organismus mit seinen speziellen Umweltbedingungen schließen? Wie erkennt man Dysfunktionen, d.h. krankhaft veränderte Gene? Wie kann man eine Zelle simulieren und bestimmte Eigenschaften vorhersagen?

Dies sind Fragen, die man in naher Zukunft sicher nicht so einfach beantworten kann – trotzdem ist es an der Zeit, sich solchen Aufgaben zu stellen. Der **Bioinformatik** wird hierbei eine Schlüsselrolle zugeordnet. Dies drückt sich nicht nur in sprunghaft gestiegenen Forschungsaktivitäten aus, sondern auch in neuen Konzepten und Investitionen in der pharmazeutischen Industrie und in einer neuen Generation von Biotechnologiefirmen.

1.2 Bioinformatik von Genomanalyse bis Gewebemodellierung

Das Erbgut besteht aus DNA (Desoxyribonucleinsäure), dem „Informationsträger" der Zelle. Zelluläre Prozesse werden aber hauptsächlich von Proteinen, den „Funktionsträgern" gesteuert. Diese Steuerung realisiert sich erst im gefalteten Zustand über die dreidimensionale Struktur und Dynamik von Proteinen und ihren Reaktionspartnern. Die Zelle besteht auch aus anderen Molekülen, z.B. Fetten, Zuckern, Wasser, Metallionen etc., die durch ihre Wechselwirkungen die Zelle am Leben erhalten. Neben den im Ergbut kodierten Informationen hängt der Zustand einer Zelle aber auch von äußeren Faktoren, wie dem Nahrungs- und Energieangebot ab, und ist je nach Zell-Zyklus-Stadium verschieden. In komplexeren Organismen sind Zellen spezialisiert und bilden gemeinsam mit ähnlich spezialisierten Zellen übergeordnete Struktur- und Funktionseinheiten, z.B. die Gewebe. Die **Bioinformatik** ist in allen diesen verschiedenen Ebenen (von der DNA bis zur Dynamik eines Organismus) von enormer Bedeutung.

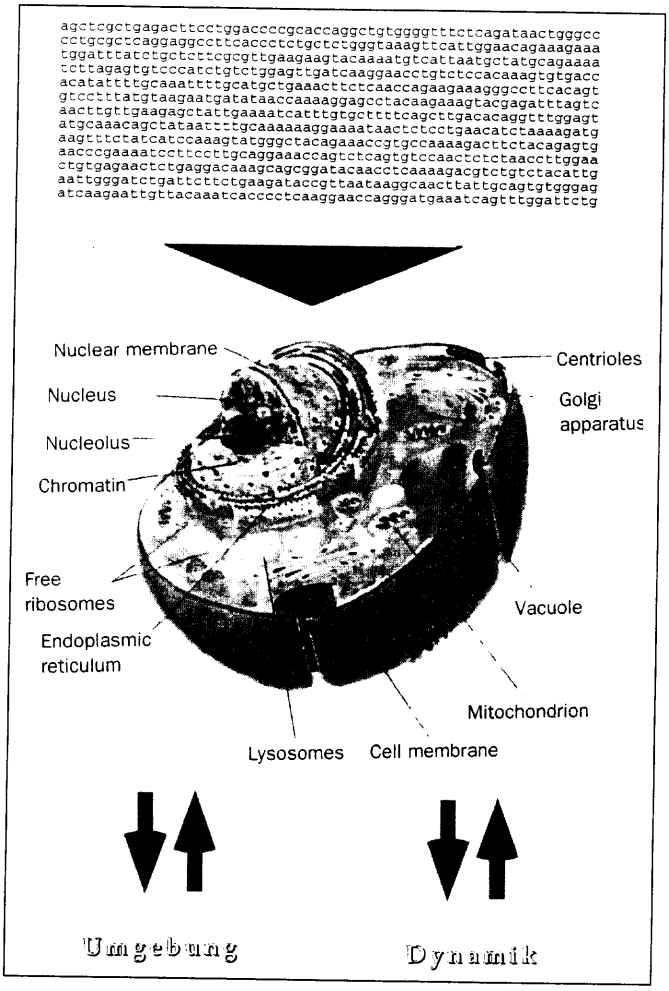


Bild 2: Ein Bruchteil der Nukleotidsequenz des Brustkrebsgenes BRCA1 ist dargestellt (einige Hundert Basen; das menschliche Genom enthält 3000 Megabasen dieser Buchstaben), um die Aufgabe zu verdeutlichen, Information aus der DNA zwecks Vorhersage zellulärer Funktionen zu extrahieren. Die Zellen (Abbildung einer Zelle mit Erlaubnis aus [20]) sind allerdings keine statischen Gebilde – sie wechselwirken mit der Umgebung (z.B. Nahrungsaufnahme) und durchlaufen verschiedene Entwicklungsstadien.

2 Proteinfunktion: Was wollen wir woraus vorhersagen?

2.1 Was ist Proteinfunktion?

Hier soll in den Teil der Bioinformatik eingeführt werden, der sich mit der Genomanalyse und Proteinfunktionsvorhersage beschäftigt, dem gegenwärtigen Hauptziel bei der Aufarbeitung der genomischen Daten [2]. Computergestützte Funktionsvorhersagen erfolgen gewöhnlich mit Hilfe von Proteinsequenzvergleichen. Proteinsequenzen, die durch den genetischen Code bestimmte lineare Anordnung von 20 verschiedenen Aminosäuren, sind durch ihr größeres Alphabet (20 Buchstaben) für die Erkennung entfernter Ähnlichkeiten besser geeignet als DNA (4-Buchstaben-Alphabet). Normalerweise wird eine Zielsequenz mit Tausenden anderen, in Datenbanken gespeicherten Sequenzen verglichen. Ein einfacher Fall ist eine hohe Ähnlichkeit mit einem funktionell beschriebenen Protein, dessen Funktion dann auf das Zielprotein übertragen werden kann.

Leider ist das Wort „Proteinfunktion" nicht sehr gut definiert, da die Funktion im Zusammenspiel mit anderen Molekülen ausgeübt wird, sich je nach Status der Zelle verändern kann und man funktionelle Details noch von keinem Protein genau kennt. Proteinfunktionsvorhersagen können sich aber nur auf Information stützen, die zusammen mit den Sequenzen in Datenbanken abgespeichert ist. Selbst für die am besten charakterisierten Proteine, die metabolischen Enzyme, deren Stoffwechselwege bekannt und von denen Raumstruktur und Substrat-Bindungsverhalten aufgeklärt sind, bleiben noch viele Fragen offen: Wie finden auf der molekularen Ebene Regulationen statt? Bilden die Enzyme eines Stoffwechselweges einen Komplex oder werden die Substrate in Lösung weitergereicht? Bei den mei-

sten Proteinen ist die sogenannte molekulare Funktion aber viel schlechter beschrieben, z.B. „Dieses Protein bindet ATP". Wenn man nun über Ähnlichkeitssuchen in Datenbanken versucht, eine molekulare Funktion vorherzusagen, kann man für das vorherzusagende Protein im besten Fall diese sehr vage Aussage übernehmen. Leider sind in den Sequenzdatenbanken aber molekulare Funktionen mit Funktionsaussagen auf anderen Ebenen vermischt, z.B. mit phänotypischen Aussagen wie „Dieses Protein spielt eine Rolle in der Entwicklung von Flügeln" oder zellulären Erkenntnissen wie z.B. „involviert in Proteinfaltung". Viele Aspekte der Funktion, wie z.B. die zeitliche gesteuerte Expression des entsprechenden Proteins, sind überhaupt noch nicht berücksichtigt. Hier muß also noch erhebliche Arbeit in die Annotation der Datenbanken investiert werden, um die durch komplizierte Methoden gewonnenen Ähnlichkeitsaussagen zur Funktionsvorhersage nutzen zu können.

2.2 Wie gut sind die Sequenzdaten und wie verläßlich die Datenbanken?

Ein weiteres Problem, das ständig berücksichtigt werden muß, ist die Qualität der Daten und ihrer Annotation [3]. Obwohl die 4-Buchstabentexte (DNA) oder das 20-Buchstabenalphabet (Proteine) eindeutig interpretierbar sind, ist mit einer hohen Fehlerrate zu rechnen. Bedingt durch den genetischen Code kann das Fehlen einer einzigen Base in der DNA zu Leserahmenverschiebungen bei der Übersetzung in Proteinsequenzen führen und die korrespondierende Proteinsequenz wird nach der fehlerhaften Stelle total verschieden von der wirklichen Sequenz sein. Bei Stichprobenuntersuchungen in dem Bakterium *E.coli* fanden wir z.B., daß fast 10% der Gene (für Proteine codierende DNA Bereiche) zu Leserahmenverschiebung führende Fehler enthielten. Bei höheren Organismen, in denen Gene durch nicht-codierende Einschübe (Introns) unterbrochen sind, ist die Übersetzung in Proteinsequenzen komplizierter. In derzeit fast 30% aller Fälle werden die codierenden Abschnitte des Genes (Exons) fehlerhaft zusammengefügt [4], d.h. trotz korrekter DNA-Sequenz ist die Proteinsequenz mit groben Fehlern behaftet. Zu den durch technische Schwierigkeiten bei der Sequenzierung bedingten Datenfehlern kommen noch Fälle, in denen durch experimentelle Probleme DNA von einem falschen Organismus sequenziert wird (ein bekanntes Beispiel ist eine kontaminierte menschliche DNA-Bibliothek der Firma Genethon, in der viele Hefesequenzen enthalten sind, die auch heute noch in Datenbanken als „vom Menschen stammend" geführt werden). Eine weitere häufig vorkommende fehlerhafte Annotation ist eine funktionelle Beschreibung, die für die gegebene Sequenz nicht zutrifft – Ursachen können im Experiment, aber auch in der halbautomatischen Annotation selber liegen. Durch die tägliche Datenflut ist manuelle Annotation nicht mehr in vollem Umfange möglich. Da Fehler in den Daten, der Interpretation der Daten und ihrer Annotation akkumulieren, ist bei Genomanalysen Vorsicht in der Interpretation der Resultate geboten, und Rückschlüsse auf einzelne Gene oder Proteine bedürfen deshalb derzeit noch der manuellen Kontrolle. Das Ausmaß zweifelhafter molekularbiologischer Daten wurde kürzlich anhand der weitverbreiteten Proteinstrukturdatenbank (PDB) ermittelt: über eine Million wahrscheinlicher Fehler unterschiedlichster Ursache wurden Anfang 1996 in den nur knapp 3500 verschiedenen Datenbankeinträgen ermittelt [5].

3 Computermethoden in der Funktionsvorhersage

3.1 Erstanalyse genomischer Sequenzen

Der erste Schritt bei der Funktionsvorhersage ist die Analyse genomischer Sequenzen, hauptsächlich zur Identifizierung der Gene. Da in der Praxis ein Genom in der Größenordnung von 3000 Megabasen (z.B. Mensch) nicht in einem Stück sequenziert wird, sondern über Jahre hinweg kleine Bruchstücke von 300–600 Basen mit (je nach Methodik) einer 3–10-fachen Redundanz (zur Reduzierung der Fehler) erzeugt und gesammelt werden, benötigt man Computermethoden schon für das Zusammensetzen („assembly") und Korrigieren der Sequenzen. Da im Erbgut oftmals identische oder fast identische Wiederholungen vorkommen, ist selbst diese mathematisch triviale Aufgabe nicht einfach zu bewältigen. Die Methoden sind recht komplex, da nicht nur durch Überlappungen Sequenzregionen aneinandergefügt werden, sondern auch externe Information herangezogen wird, z.B. Marker (kurze unterscheidbare Sequenzstücke, deren Reihenfolge bekannt ist, und mit denen längere Genomstücken vor der Sequenzierung kartiert werden). Oftmals werden bestimmte Regionen von verschiedenen Arbeitsgruppen unabhängig sequenziert und es gilt, diese Fremdinformation mitzubenutzen, um das Zusammensetzen zu beschleunigen.

Ein nächster Schritt ist die Identifizierung der Gene, die die Information für die Proteinsequenz enthalten. Dies ist speziell bei höheren Eukaryoten ein schwieriges Problem, da der Gehalt an Genen teilweise weniger als 3% des gesamten Erbgutes ausmachen kann und die Protein-kodierenden Regionen der Gene, die Exons, durch die nicht-kodierenden Introns unterbrochen sind. Die Zellen produzieren ihre Proteine, indem sie eine Kopie der gesamten Genregion erstellen und in mehreren späteren Schritten die Exons zusammenfügen („splicing"). Die Stellen, an denen das „splicing" erfolgt, enthalten schwache Signale; viele Computerprogramme versuchen, diese bei der Genvorhersage mitzubenutzen. Erschwerend kommt allerdings hinzu, daß viele Proteine in Varianten existieren, je nach Zellstatus also aus demselben Pool von Exons verschiedene mRNAs (die von Introns befreiten Kopien der Genabschnitte der DNA) geschaffen werden („alternative splicing"), die dann die eigentliche In-

formation zur Proteinproduktion enthalten. Das Hauptsignal in der Genidentifizierung ist aber der Gebrauch der Codons (DNA Triplett, das für eine bestimmte Aminosäure kodiert). Häufig benutzte Aminosäuren werden durch mehrere Codons kodiert, jedoch nutzen die Organismen nicht alle Varianten gleichmäßig und jeder Organismus hat ein bestimmtes Erwartungsmuster für die Benutzung dieser Codons. Dieses Signal ist aber nicht stark genug, was sich darin zeigt, daß derzeit neuronale Netze bessere Resultate liefern als regelbasierte Methoden (für einen Überblick gängiger Methoden und ihrer Stärken und Schwächen siehe [4]). Die Vorhersagegüte derzeitiger Methoden ist aber trotz besser klingender Angaben der jeweiligen Autoren immer noch im 70%-Bereich [4], was nicht ausreichend ist in der automatischen Analyse großer Genomdatenmengen.

In zunehmendem Maße werden schon in die Genvorhersage Sequenzdatenbanksuchen integriert: Ähnlichkeiten zu charakterisierten Proteinen sind ein starkes Indiz für das Vorhandensein von Exons. Leider bietet die Biologie wie immer viele Ausnahmen, die bedacht sein müssen, wie z.B. Pseudogene (duplizierte Gene, die nicht exprimiert werden und oftmals durch Leserahmenverschiebungen und eine hohe Mutationsrate charakterisiert sind). Neuere Überlegungen gehen davon aus, Genvorhersagen mit der Erkennung anderer Elemente in der DNA zu koppeln, z.B. mit RNA-Vorhersage und der Identifizierung von nichtkodierenden Wiederholungen (im menschlichen Erbmaterial befindet sich z.B. eine Vielzahl von sogenannten „Alu repeats"). Die Erkennung solcher Elemente schließt das Vorhandensein kodierender Bereiche aus. Gene sind umgeben von regulatorischen nichtkodierenden Elementen, z.B. Ribosomenbindungsstellen und Promotoren. An letztere binden regulatorische Proteine, um die Genregion zum Kopieren freizugeben. Auch diese Elemente tragen zur Signalverschärfung bei. Deren Erkennung erfordert aber andere Methoden – auch hier liegen neuronale Netze derzeit vorn. Die zur Zeit am häufigsten benutzten Genvorhersageprogramme sind GRAIL-Varianten [6] (ursprünglich ein reines neuronales Netz, später aber mit zahlreichen Expertenregeln ergänzt), GENEFINDER (P. Green, unveröffentlicht), das in verschiedenen Sequenzierungszentren eingesetzt wird, und GENEMARK [7], ein auf Markov-Modellen basierendes System, das sich bei Prokaryoten bewährt hat. Trotz des Nachholbedarfes in der Analyse nichtkodierender DNA sind wissensbasierte Systeme im Vormarsch, die Einzelmethoden für die Erkennung verschiedener DNA Elemente kombinieren.

3.2 Analyse intrinsischer Eigenschaften von Proteinen

Mit der Identifizierung eines Genes und der Vorhersage der Exons ist die Übersetzung der DNA in eine Proteinsequenz mittels des genetischen Codes möglich. Obwohl die erfolgversprechendste Methode zur Funktionsvorhersage eine Ähnlichkeitssuche in Datenbanken ist, sollten möglichst unabhängige Informationsquellen erschlossen werden, die auf verschiedenste funktionelle und strukturelle Eigenschaften des Zielproteins hinweisen können. Einige dieser Eigenschaften beeinflussen entscheidend die Parameter von Datenbankähnlichkeitssuchen und müssen deshalb zuerst ausgewertet werden [2; 8]. Ein wichtiger Test ist die Analyse der Aminosäurezusammensetzung von Proteinen. Nicht alle der 20 Aminosäuren kommen gleichhäufig vor – bestimmte Abweichungen von der Normalverteilung geben schon erste funktionelle Hinweise. Es gibt viele biologisch relevante Regionen, die mit einer Reduktion des 20-Aminosäuren-Alphabets auskommen [8; 9]. Ein typisches Beispiel sind membrandurchdringende Abschnitte, die wegen der Lipidumgebung meist hydrophob sein müssen und somit über mehr als 20 Positionen nur hydrophobe Aminosäuren enthalten – ein Charakteristikum, das bei normaler löslicher Umgebung unmöglich ist. Zur Ermittlung solcher Transmembranregionen gibt es mittlerweile Dutzende Programme, basierend auf neuronalen Netzen, sowie regelbasierten, statistischen und informationstheoretischen Ansätzen. Obwohl sich fast genausoviele Aminosäuren in Regionen mit einer anderen ungewöhnlichen Zusammensetzung befinden, den „Tripelhelices" („coiled coil"), dominiert hier derzeit nur ein Programm für die Erkennung solcher Regionen [10], die außer einer Reduzierung des Alphabetes zudem noch eine Positionsabhängigkeit bestimmter Aminosäuren aufweisen. Obwohl von vielen Extremen der Aminosäurekomposition die Funktion und Struktur noch weitgehend unbekannt ist, kann die einfache Ermittlung solcher Abweichungen zumindest bei Datenbanksuchen behilflich sein, da diese Regionen die Bewertungschemata von Ähnlichkeitssuchen unterlaufen und, wenn nicht vorher „maskiert", zu falschen Ergebnissen führen. Inzwischen gibt es mehrere Algorithmen, die versuchen, solche Regionen zu erkennen. Am häufigsten wird zur Zeit das Programm „SEG" eingesetzt, das verschiedene Komplexitätsmessungen beinhaltet [9].

Eine weitere Eigenschaft von vielen Proteinen sind interne Duplikationen, d.h. nicht-identische, oftmals nur sehr entfernt ähnliche, sich wiederholende Abschnitte. Obwohl hier Algorithmen von Laplacetransformationen bis zu Markovmodellen reichen, sind Programme zur Identifizierung solcher „repeats" nicht weitverbreitet, da ihre Identifizierungsrate noch relativ gering ist.

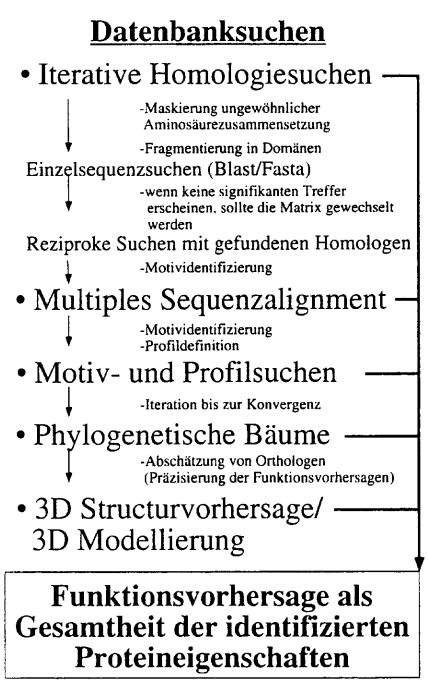
Eine weitere Kategorie von Signalen in Proteinsequenzen, die nicht durch Ähnlichkeitssuchen erfaßt werden können, sind posttranslationale Modifikationen, d.h. kurze Abschnitte, die auf eine Modifizierung einer Aminosäure hinweisen z.B. die kovalente Bindung von Zuckerresten. Diese Modifizierungen sind im 20-Buchstabenalphabet nicht direkt sichtbar, sind oftmals aber entscheidend für die Funktion des Proteins. Inzwischen gibt es Datenbanken solcher Modifikationen und Sammlungen kurzer Aminosäurekonsensusmuster, die auf bestimmte Modifikationen hin-

weisen. Leider sind diese in den meisten Fällen nur hinreichend, aber nicht notwendig – im Kontext anderer Eigenschaften eines Proteines verschärft sich aber das Signal. So kommen bestimmte Glykosylierungen nur in extrazellulären Proteinbereichen vor und spielen erst eine Rolle, wenn für das Zielprotein die zelluläre Lokalisierung bekannt ist. Für die Bestimmung der Lokalisierung eines Proteines (s. Bild 2) existieren verschiedene Verfahren. Allerdings ist bei allen Verfahren die Vorhersagegüte unter 80%, d.h. für das einzelne Protein können keine sicheren Angaben gemacht werden. Auch diese Eigenschaft muß in einem automatischen System zur Funktionvorhersage mit Wahrscheinlichkeiten behandelt werden. Der Fall vom Brustkrebsgen BRCA1 macht die Notwendigkeit guter Vorhersagen deutlich: Keine klaren Funktionsaussagen sind möglich; nicht einmal die Lokalisierung ist bekannt, um weitere Experimente planen zu können. Zwischen Ende 1995 und Mitte 1996 haben mehrere Gruppen in führenden Journalen Experimente publizert, die gegensätzliche Aussagen liefern und BRCA1 sowohl im Zellkern als auch im Cytoplasma oder im extrazellulären Raum ansiedeln. Die vergleichende Sequenzanalyse konnte auch hier einen entscheidenen Beitrag zur Klärung dieser Frage liefern und darüberhinaus funktionelle Eigenschaften von BRCA1 vorhersagen: Es ist mit hoher Wahrscheinlichkeit ein nukleares Protein, das einen Zellzyklus-Kontrollpunkt darstellt (s. [11] und Zitate in [11]).

3.3 Funktionsvorhersage durch vergleichende Sequenzanalysen

Wie kommt man vom Sequenzvergleich zu diesen funktionellen Aussagen? In den meisten Fällen wird eine Datenbanksuche durchgeführt (Bild 3) und die Funktion vom ähnlichsten Protein übernommen. Dies ist aber ein gefährliches Unterfangen, da es in einem Organismus viele Proteine gibt, die ähnlich zueinander sind, aber unterschiedliche Funktionen haben. Es ist noch nicht klar, welche Proteinfamilie im Menschen am größten ist, aber allein zwei aussichtsreiche Kandidatenfamilien (Proteinkinasen und an G-Proteine gekoppelte Rezeptoren) schätzt man auf ungefähr je 1000 Mitglieder. Obwohl diese bestimmte Teilfunktionen gemeinsam haben, hat jedes Mitglied einer Proteinfamilie eine spezielle Aufgabe. Somit ist die Vorhersage der Funktion durch Übernahme der Beschreibung eines ähnlichen Proteins, so wie es derzeit noch üblich ist, in den meisten Fällen nicht korrekt. Eine weitere Komplikation ist der modulare Aufbau vieler Proteine. Es kann eine kurze Region (Baustein, Domäne) einer Datenbanksequenz durchaus ähnlich zum Zielprotein sein, aber der überwiegende Teil der beiden Proteine ist völlig verschieden und dementsprechend auch ihre Funktion (obwohl die ähnliche Domäne eine gemeinsame Teilfunktion ausüben kann, aber nicht muß). Solche Überlegungen muß man sich bei der Analyse ganzer Chromosomen oder Genome immer vor Augen halten, genauso wie die Stärken und Schwächen der Datenbanksuchmethoden. Auch hier kann es leicht zu Überinterpretationen kommen, d.h. aus zufälligen Ähnlichkeiten werden evolutionäre Verwandtschaften, Homologien, gedeutet. Die Erkennung von Homologien zwischen zwei Sequenzen, d.h. der Annahme einer gemeinsamen Vorläufersequenz, bildet die Grundlage der Datenbanksuchen. Im Normalfall ist die gemeinsame Wurzel zweier Sequenzen viele Millionen Jahre alt und Mutationen haben die beiden zu untersuchenden Sequenzen stark verändert – ihre Buchstabenabfolge läßt also kaum noch Ähnlichkeiten erkennen. Trotzdem ist die Information über ähnliche Raumstruktur und Funktion in den Buchstabenketten verschlüsselt und kann über Mustererkennungsverfahren oftmals noch identifiziert werden. Eine Hilfe dabei sind Ähnlichkeitsmaße zwischen den einzelnen Buchstaben, den Aminosäuren, die teilweise ähnliche sterische und physikochemische Eigenschaften besitzen und somit nicht unabhängig voneinander sind. Allerdings ist es nach wie vor ein noch nicht vollständig geklärtes Problem, Sequenzen optimal gegeneinander auszurichten, so daß die Ähnlichkeit maximiert wird, da die Komplexität durch Einschübe (eine Sequenz ist länger als die andere) noch erhöht wird. Die Behandlung von Einschüben erfordert sogenannte dynamische Programmierung, kostet Rechenzeit und wird angesichts der riesigen Datenmengen derzeit in der Genomanalyse nur in Einzelfällen angewandt.

Bild 3: Iterative Datenbanksuchen und daran ansetzende Methoden zur Funktionsvorhersage.



International werden schnelle Programme der Blast-Serie [8] am häufigsten benutzt, die im ersten Schritt nur Teilstringe nach Identitäten durchsuchen, danach mit Hilfe von Ähnlichkeitsmaßen und Statistiken die Länge der Treffer optimieren und sie nach bestimmten Kriterien sortieren. Gerade in großen Sequenzierungsprojekten wird Blast oft als einziges Datenbanksuchverfahren eingesetzt, teilweise noch durch den Vergleich mit in speziellen Datenbanken abgespeicherten Sequenzmustern für charakteristische funktionelle Regionen erweitert. Diese Muster sind meist konservierte Regionen (d.h. bestimmte Buchstabenfolgen oder unterliegende Aminosäureei-

genschaften sind in sonst stark unterschiedlichen Sequenzen invariant), die durch Studium einzelner Proteinfamilien charakterisiert werden konnten.

Der Erwartungswert für einen Treffer bei einer Blast-Datenbanksuche mit einem signifikanten Schwellenwert hängt von der Species ab: bei Bakterien ist er im Bereich von 80%, bei Hefe um 70% und beim Menschen bei 50%. Hierbei muß aber berücksichtigt werden, daß viele der Datenbanksequenzen nicht oder nur ungenügend funktionell charakterisiert sind, d.h. der Prozentsatz für Funktionsvorhersagen ist gerade bei menschlichen Sequenzen noch ungenügend. Die Trefferausbeute kann aber durch Anwendung von komplexeren Suchverfahren und sensitiveren Methoden um 10-20% gesteigert werden [12, 13]. Weiterhin kann auch die Anzahl der funktionellen Vorhersagen prozentual gesteigert und präzisiert werden, wenn eine bessere Analyse der gefundenen Ähnlichkeiten erfolgt [13]. Die Ausbeuterate verbessert sich auch mit der Zeit, da der „Sequenzraum" immer dichter und die funktionelle Charakterisierung immer besser wird.

Ein typischer Ablauf einer iterativen Suche (Bild 3) ist der Start mit einer Standarddatenbanksuche (z.B. Blast), und bei fehlenden Treffern die Variation der unterliegenden Ähnlichkeitsmatrizen. Ein nächster Schritt ist die Identifizierung von Mustern im Grauzonenbereich der Standardsuchprogramme, d.h. der ausführlichen Analyse von Treffern, die nicht als signifikant eingestuft werden können, aber trotzdem Kandidaten für entfernte Homologien darstellen. Die erkannten Muster (gleiche Abfolgen von Buchstaben sind identisch in mehreren, nicht klar verwandten Proteinen) können dann für Motivsuchverfahren eingesetzt werden, in denen man sich nur auf diese lokalen Muster konzentriert [12]. Oftmals sind solche Muster der einzige sichtbare Hinweise auf eine funktionelle Region, die einer Proteinfamilie gemeinsam ist und auf die man aus der (derzeit noch manuellen) Analyse der biochemischen Charakterisierung einzelner Mitglieder schließen kann. Um die Proteinfamilie möglichst komplett zu erkennen, sind Iterationen nötig: neu gefundene Mitglieder müssen in die Berechnung von Matrizen/Mustern für eine erneute Datenbanksuche einfließen. Reziproke Suchen mit gefundenen Kandidaten sind essentiell. Oftmals müssen auch noch Profilsuchen mit globaleren Ausgangsregionen durchgeführt werden [12].

Das ganze Arsenal von Methoden, Parametern und die Reihenfolge ihrer Anwendung reicht für wochenlange Beschäftigung mit einer einzigen Sequenz aus, die sich allerdings in besonderen Fällen (z.B. Krankheitsgenen) auszahlen kann. Im Zeitalter der Genomsequenzierung sollten aber Tausende Sequenzen in wenigen Tagen analysiert werden. Hier müssen Kompromisse in der Senistivität gemacht werden und es Bedarf einer genauen Kenntnis der Prozedur, um trotzdem eine hohe und korrekte Funktionsvorhersagerate zu erreichen. Wir haben z.B. 1992 [14; 15] mit der Analyse des ersten komplett sequenzierten eukaryotischen Chromosomes (Hefechromosom III [16]) begonnen, ein automatisiertes Verfahren zur Funktionsvorhersage zu entwickeln. Trotz vieler Mannjahre Arbeit, die in ein erfolgversprechendes System, GENEQUIZ, mündeten [13], sind noch nicht alle der genannten Punkte integriert und besonders die letzten Schritte der präzise Vorhersage von Funktionen aus ermittelten homologen Proteinen in der Datenbank bereiten noch Schwierigkeiten – Sequenzdatenbanken sind noch zu inhomogen und funktionelle Informationen zu ungenau gespeichert.

3.4 Funktionsvorhersagen durch Analyse kompletter Genome

Mit dem Vorhandensein kompletter sequenzierter Genome eröffnen sich aber weitere indirekte Möglichkeiten zur Funktionsvorhersage: die Ausnutzung von Wissen über Stoffwechselwege. Wenn z.B. bestimmte metabolische Reaktionsketten hinreichend bekannt sind und durch Homologiesuchen einige Mitglieder der Kette zweifelsfrei identifiziert worden sind, müssen entweder die weiteren Mitglieder der Kette auch vorhanden sein, oder es können in limitiertem Umfange Abweichungen des Stoffwechselweges auftreten, die mit Zusatzinformation über benachbarte Wege vorhersagbar sind [17]. Auch hier gibt es natürlich verschiedene Schwierigkeiten zu überwinden. Durch die Analyse kompletter Genome konnte kürzlich gezeigt werden, daß in verschiedenen Organismen unterschiedliche, nicht-ähnliche Gene für die gleiche Funktion kodieren und daß dieses Phänomen relativ häufig vorzukommen scheint [18]. Das Nichterkennen eines Proteins einer Kette bedeutet somit noch lange nicht, daß die Kette nicht geschlossen und funktionstüchtig ist.

Weitere Aussagen können aus der Anordnung von Genen im Genom getroffen werden. In Bakterien sind z.B. häufig Stoffwechelwege in einer Einheit, dem Operon, kodiert, d.h. sie werden als ein gemeinsames Stück von der DNA abgelesen und dann erst später in die eigentlichen Gene zerlegt. In Eukaryoten sind solche Abhängigkeiten kaum noch zu finden, doch sind z.B. Untereinheiten eines Proteines (unabhängige Proteinketten, die aber nur im Komplex ihre Funktion ausüben können) oft noch in benachbarter Position im Genom.

3.5 Methodenentwicklung und Datenverwaltung im Internet

Es ist in diesem Artikel bewußt wenig auf spezielle Methoden eingegangen worden, da sich das gesamte Gebiet rasant verändert. Durch das Medium Internet werden fast täglich neue Programme für Teillösungen angeboten, die wieder ein paar der Probleme reduzieren. Neue, besser annotierte Spezialdatenbanken entstehen, die je nach Aufgabenstellung mitgenutzt werden können. Neben der Datenexplosion gibt es also auch eine Methodenexplosion, und durch die freie Verteilung im Internet ist es oftmals schwer, die Güte der Methoden und ihre Kompatibilität genau zu bestimmen. Da zur

Zeit gerade die ersten kompletten Genome fertig sequenziert werden, ist ein gewisser Zeitdruck für deren Auswertung gegeben. Oftmals leidet darunter die Qualität der Analysen und fehlerhaft interpretierte Information wird in die Datenbanken eingespeist.

Gegenwärtige Methoden richten sich auf die Verbindung zwischen Datenbanken, man kann sich manuell seine Informationen in einem Netz von mehreren hundert Spezialdatenbanken zusammensuchen. Trotz „links", d.h. einer Verbindung von Datenbanken, ist derzeit jedoch kaum eine gezielte Extraktion unterschiedlicher Daten möglich. Viele der sogenanntemn Datenbanken sind eigentlich nur Datensammlungen – selbst die häufig benutzten Sequenzdatenbanken stellen erst jetzt auf Datenbanksysteme um. Besonders schwierig erweist sich die erwähnte schlecht strukturierte funktionelle Information, „data mining"-Technologien werden aber hoffentlich schon bald auch diese Lücke schließen

Ein weiteres Problem sind die oftmals unzureichenden Netzwerkverbindungen, bei denen hinsichtlich detaillierter Analyse oft Kompromisse gemacht werden müssen. Da es sich aber in der Genom und Sequenzanalyse vorwiegend um lineare Information handelt, machen diese Daten nur einen Bruchteil des internationalen Datentranfers aus. Rechnerleistungen und Speicherkapazitäten sind aus ähnlichen Gründen ebenfalls keine limitierenden Faktoren.

4 Wo sind die Engpässe?

Viele der Algorithmen zur Lösung von Teilproblemen scheinen an ihre Grenzen gekommen zu sein. So sind existierende Ähnlichkeitssuchen schon sehr ergiebig und die vielen neuen Daten machen die Suchen auch einfacher. Wenige Lösungen sind für die Automatisierung verschiedener Schritte in der Funktionsvorhersage vorhanden. Die Datenerzeuger haben meist ihre eigene Software und entwickeln sie weiter, und das Problem ist zu komplex, als daß es von kleineren Arbeitsgruppen allein bewältigt werden kann. Ein Hauptproblem besteht in der Extraktion von Daten aus existierenden, nicht immer gut strukturierten Datensammlungen und Datenbanken. Hier sind eindeutig neue Technologien gefordert. Integrierte Systeme mit anspruchsvollen und dennoch nutzerfreundlichen grafischen Oberflächen, in denen nicht nur alle Daten verfügbar gemacht, sondern auch aufbereitet werden, sind für die nahe Zukunft zu erwarten. Hier macht besonders die Aufarbeitung experimenteller Arbeiten Schwierigkeiten, da fehlerhafte Information bisher kaum abgeschätzt werden kann und strukturierte Informationsaufarbeitung von funktionellen Daten bislang kaum möglich ist. Selbst bei Details wie Gen- oder Proteinnamen gibt es nur wenige Konventionen und keine Systematik.

Eine gute Funktionsvorhersage ist natürlich Voraussetzung für die Lösung von globaleren Aufgaben wie der Identifizierung von Stoffwechselwegen und der Modellierung von Regulationsmechanismen. Hier wird sich ein weiterer Schwerpunkt in der Bioinformatik herausbilden.

Literatur

[1] *Fleischmann et al.*: In: Science 269 (1995), S. 496–512.

[2] *Bork, P., Ouzounis, C. Sander, C.*: Curr. Opinion. Struct.Biol. 4. (1994) S. 393–403.

[3] *Bork, P., Bairoch, A.*: Trends Genet. (1996), Oktober-Ausgabe.

[4] Burseé, M., *Guigo, R.*: Genomics 34 (1996) S. 353–367.

[5] *Hooft, R. W. W., Vriend, G., Sander, C., Abola, E. E.*: Nature 381 (1996) S. 272.

[6] *Xu, Y., Einstein, J. R., Mural, R. J., Shah, M., Uberbacher E. C.*: ISMB 2 (1994), S. 376–384.

[7] *Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Medigue, C., Danchin, A.*: Nucl.Acid. Res. 23 (1995), S. 3554–3562.

[8] *Altschul, S. F., Boguski, M. S., Gish, W., Wootton, J. C.*: Nature Genet. 6 (1994), S. 119–129.

[9] *Wootton. J. C.*: Curr. Opin. Struct. Biol. 4 (1994), 413–421.

[10] *Lupas, A., Van Dyke, M., Stock, J.*: Science 252 (1991), S. 1162–1164.

[11] *Koonin, E. V., Altschul, S. F., Bork, P.*: Nature Genet. 13, (1996), S. 266–268.

[12] *Bork, P., Gibson, T.*: Meth. Enzymol. 266, (1996), S. 162–184.

[13] *Casari, G., Andrade, M., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, A., Valencia, A., Sander, C.*: Nature 376 (1995), S. 647–648.

[14] *Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E.*: Protein Science 1 (1992), S. 1677–1690.

[15] *Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E.*: Nature 358 (1992), S. 358.

[16] *Oliver, S. G. et al.*: Nature 357 (1992), S. 38–46.

[17] *Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E., Koonin, E. V.*: Curr. Biol. 6 (1996), S. 279–291.

[18] *Koonin, E. V., Mushegian, A. R., Bork, P.*: Trends Genet. (1996), September-Ausgabe.

[19] *Woese, C. R.*: Microbiol. Rev. (1987), S 221–271.

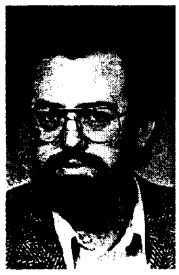
[20] *Voet, D., Voet, J.*: Biochemie; Weinheim, New York, Basel, VCH 1992.

Dr. habil. Peer Bork
Max-Delbrück-Centrum für Molekulare Medizin, 13122 Berlin-Buch und EMBL, Meyerhofstr. 1, 69012 Heidelberg,
Email: bork@embl-heidelberg.de

Modellierung der Genregulation in Eukaryonten[1]

Rainer Knüppel, Gesellschaft für Biotechnologische Forschung mbH, Braunschweig,
Edgar Wingender, Sys Design, Kressborn

Dipl.-Biol. Dipl.-Inform. Rainer Knüppel hat an der FU Berlin Biologie und an der TU Berlin Informatik studiert. Von 1993–1996 arbeitete er am GSF-Zentrum für Umwelt und Gesundheit (Oberschleißheim) und an der GBF im BMBF-finanzierten Verbundprojekt GENUS (Genregulatorische Nukleinsäuresequenzen) mit.



Dr. rer. nat. Dipl.-Chem. Edgar Wingender hat an der TU Braunschweig Chemie studiert. Nach Mitarbeit in einer Reihe experimenteller Forschungsprojekte an der Universität Marburg und der GBF leitet er seit 1994 an der GBF das Projekt „Bioinformatik der Genregulation" und koordiniert die Verbundprojekte GENUS (BMBF) und TRADAT (EU).

Die Aufklärung der Mechanismen der Genregulation[2] stellt eines der fundamentalen Probleme der modernen Molekularbiologie dar. Insbesondere die Identifizierung regulatorischer Signale in kompletten Genomen[3] und die Erforschung ihrer Semantik und Grammatik sind von hoher medizinischer und biotechnologischer Relevanz. Angesichts der Komplexität des bereits verfügbaren und des zu erwartenden Datenmaterials sind rechnergestützte Methoden erforderlich, um ein Gesamtbild von der Realisierung genetischer Information zu gewinnen.

Modelling of Gene Regulation in Eukaryotes

Elucidation of the mechanisms of gene regulation is one of the most fundamental problems of modern molecular biology. In particular the identification of regulatory signals in complete genomes and the investigation of their semantics and grammar is of high medical and biotechnological relevance. Considering the complexity of available and expected data, computer-based tools are needed to gain an integrated view of the realization of genetic information.

1 Einführung

Ein fundamentales Problem in der Biologie, das erst mit den Methoden der modernen Molekularbiologie erklärend angegangen werden konnte, ist die Frage nach der Entwicklung und Differenzierung von Zellen in einem hochkomplexen Verband, wie ihn ein höherer Organismus darstellt. Wie schafft es ein bestimmter Zelltyp wie etwa eine menschliche Leberzelle, gezielt den kleinen Teil aller ca. 100000 Gene zu aktivieren, der für die spezifischen Funktionen dieses Organs erforderlich ist? Und wie erreicht sie es, einen weiteren Teil in eine Art „Wartestellung" zu bringen, so daß diese Gene nur in der Leber, aber auch hier nur unter einem gewissen Stimulus (etwa einer Hormonwirkung) aktiv werden? Die Modellierung der Mechanismen, die dieser Genregulation zugrunde liegen, ist von hoher medizinischer und biotechnologischer Relevanz.

Alle höheren multizellulären Organismen gehören ebenso wie auch schon die Hefe zu den Eukaryonten. Sie haben die Tatsache gemeinsam, daß ihre Erbsubstanz in einem besonderen Kompartiment, dem Zellkern, enthalten ist. Das Genom des Organismus, also die Gesamtheit aller Gene, enthält dessen vollständige Bau- und Betriebsanleitung in einer speziell kodierten Form. Der materielle Träger dieser Information ist die DNA (deoxyribonucleic acid). Die in ihr niedergelegte Information wird realisiert (exprimiert), indem die DNA zunächst in RNA umgeschrieben (transkribiert) und dann in ein Protein übersetzt (translatiert) wird. Nach unserem heutigen Wissensstand findet die Regulation dieser Genexpression[4] hauptsächlich auf der Ebene der Transkription[5] statt. Zu diesem Zweck hat die Natur molekulare Schalter entwickelt, die aus relativ kurzen Sequenzelementen von etwa 5–25 Basenpaaren und Proteinen bestehen, die an diese Elemente binden

[1] Eukaryo(n)ten: Höhere Organismen mit Zellkern (Nucleus), im Gegensatz zu den Prokaryonten (z. B. Bakterien), die keinen Zellkern besitzen.

[2] Genregulation: Kontrolle der Genexpression; sie kann an jedem beteiligten Informationsträger (DNA, RNA, Protein) sowie an den Übertragungsmechanismen (Transkription, Translation) ansetzen; nach gegenwärtigem Wissenstand ist die Transkriptionskontrolle der wichtigste Mechanismus.

[3] Genom: Die Gesamtheit aller Erbanlagen eines Organismus bzw. die Gesamtheit seiner DNA.

[4] Genexpression: Realisierung genomischer Information über Transkription und Translation.

[5] Transkription: Umschreiben von DNA- in RNA-Sequenzen durch eine bestimmte Klasse von Enzymen (RNA-Polymerasen). Der Anheftungsort der RNA-Polymerase an die DNA wird durch einen kompliziert aufgebauten „basalen Transkriptionskomplex" festgelegt, der sich an der DNA im Bereich 30 Nukleotide vor bis zum Transkriptionsstartpunkt eines Gens bildet. Der Aufbau und die Aktivität dieses Komplexes wird durch weiter vor dem Transkriptionsstart oder weit entfernt bindende Transkriptionsfaktoren reguliert. – Die in der Transkription synthetisierte RNA kann bereits das Endprodukt darstellen, wird aber in den meisten Fällen translatiert (s. u.).

[9; 10]. Diese Proteine werden Transkriptionsfaktoren[6] genannt. Die Eigenschaften dieser Wechselwirkungen und ihrer Komponenten determinieren letztlich die Expression eines Gens in einem raumzeitlichen und konditionalen Koordinatensystem des Organismus. Ihre Identifikation ist daher von entscheidender Bedeutung für das Verständnis genomischer DNA-Sequenzen, wie sie zur Zeit und in der nahen Zukunft drastisch zunehmend im Rahmen von Genomforschungsprojekten ermittelt werden.

Unser eigener Ansatz besteht nun darin, zunächst ein „Wörterbuch" dieser regulatorischen Elemente in Form einer Datenbank (TRANSFAC) zu erstellen, um auf dieser Grundlage und im Rahmen eines nationalen und eines europäischen Verbundes computergestützte Werkzeuge zu entwickeln, die eine Beschreibung dieser Elemente und ihre Erkennung im Genom ermöglichen.

2 Die Datenbank TRANSFAC

Die Speicherung der reinen DNA- und RNA-Sequenzinformation wird im internationalen Rahmen im wesentlichen von drei eng kooperierenden Datenbanken (EMBL[7], GenBank, DDBJ) durchgeführt. Hier werden auch ungeprüft von den Autoren der Sequenzen vorgeschlagene Annotationen aufgenommen, die dementsprechend sehr unterschiedliche Qualität haben können. Wir haben daher vor einigen Jahren begonnen, eine Datenbank über regulatorische Genomelemente und Transkriptionsfaktoren (TRANSFAC) aufzubauen, die die Qualität der experimentellen Evidenz berücksichtigt [8; 11]. Der Natur der Sache gemäß ist der zentrale Inhalt dieser Datenbank als eine n:n-Relation zwischen den Hauptkomponenten modelliert, den Sequenzelementen und den Faktoren [4].

Die Sequenzelemente selbst, oft nur wenige Basenpaare lang, sind Teil sehr großer, bis zu einigen 100 Millionen Basenpaaren langer DNA-Sequenzen, deren Basenabfolgen heute bis auf wenige Ausnahmen (z.B. die Chromosomen der Hefe) nur abschnittsweise bekannt sind. So sind die Sequenzelemente neben der Sequenz selbst charakterisiert durch deskriptive Merkmale ihrer Herkunft (Spezies), des Gens, dem sie zugehören, ihrer Position innerhalb des Gens (z. B. relativ zum Transkriptionsstart) und der Position innerhalb bekannter Sequenzen, die in den Sequenzdatenbanken – z. B. der EMBL-Sequenzdatenbank – verwaltet werden, sowie funktionalen Merkmalen wie ihrer Einstufung als Teile eines Promotors / Enhancers und des Namens des Elements, der oft einen Hinweis auf einen bindenden Faktor oder die vermittelte Reaktivität enthält.

Diese Faktoren sind Proteine oder Proteinkomplexe, die in ihren physiko-chemischen Eigenschaften, ihrer Struktur und ihrem modularen Aufbau in Domänen detaillierter beschrieben werden können. Da den Domänen oft eine funktionelle Eigenschaft, z.B. die Art der DNA-Bindung oder Faktordimerisierung[8], zugeordnet werden können, werden Transkriptionsfaktoren gemäß ihrem modularen Aufbau klassifiziert.

Für das Verständnis der Genregulation ist die Regulation der Faktoren selbst, aus phänomenologischer Sicht ihr Auftreten im räumlich-zeitlichen Koordinatensystem des Organismus und die Vermittlung externer und interner Signale (Signaltransduktion[9]) von besonderer Bedeutung. Informationen dieser Art sind bisher nur in geringem Maße vorhanden, gewinnen jedoch zunehmend an Bedeutung und werden durch methodische Weiterentwicklung experimentell zugänglich [6; 7].

Transkriptionsfaktoren sind in der Lage, an verschiedene Sequenzelemente mit voneinander abweichenden Basenfolgen zu binden. Das Charakteristikum „Sequenzspezifität" eines Faktors kann neben der Aufzählung aller bekannten gebundenen Sequenzen in davon abgeleiteter Weise, z.B. IUPAC-Consensi (eine spezielle Notation regulärer Ausdrücke für Nukleinsäuren) oder Nukleotidverteilungsmatrizen[10], beschrieben werden. Diese Beschreibungen sind ebenfalls in der Datenbank enthalten und können zum Auffinden potentieller Faktorbindungsstellen in noch nicht annotierten Sequenzen benutzt werden.

Die Daten in TRANSFAC sind weitgehend experimentell gewonnene, publizierte Daten, so daß auch Informationen über experimentelle Systeme und Methoden sowie Verweise auf die Originalliteratur in die Datenbank aufgenommen werden.

Für die Verwaltung und den internen Gebrauch der Daten verwenden wir ein relationales Datenbankmanagementsystem. Über ein WWW-Interface und als strukturierte Textfiles sind die Daten öffentlich zugänglich (Bild 1).

3 Verknüpfung mit anderen Datenbanken

Wie bereits weiter oben angedeutet, sind verwandte Inhalte in zahlreichen anderen Datenbanken gespeichert. Da dem Nutzer – zumal dem nur gelegentlich mit diesen Datenquellen umgehenden Experimentalwissenschaftler - nicht zuzumuten ist, sich mit der Struktur einer Vielzahl von relevanten Datenbanken vertraut zu machen, um dort die für ihn wichtigen In-

[6] Transkriptionsfaktor: Protein, das direkt oder indirekt an DNA bindet und die Transkription bestimmter Genabschnitte reguliert.

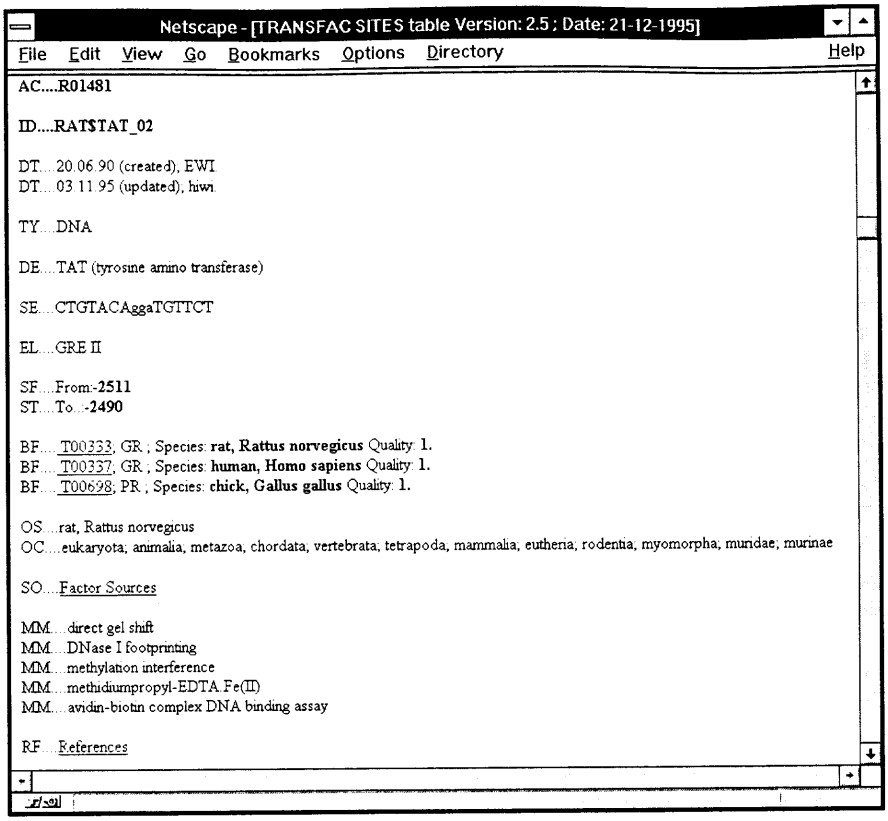
[7] EMBL: European Molecular Biology Laboratory, Sitz in Heidelberg, mit dem EBI (European Bioinformatics Institute) als Außenstation in Hinxton, England; pflegt die europäische Datenbank für DNA- und RNA-Sequenzen (EMBL Data Library).

[8] Dimerisierung: Wechselwirkung von zwei Polypeptid-Molekülen zur Bildung eines mit neuen Eigenschaften ausgestatteten Proteinkomplexes.

[9] Signaltransduktion: Übertragung von außen kommender Signale innerhalb einer lebenden Zelle.

[10] Nukleotid: Baustein der DNA und der RNA, bestehend aus einer Base, dem Zucker Desoxyribose (DNA) bzw. Ribose (RNA).

A



B

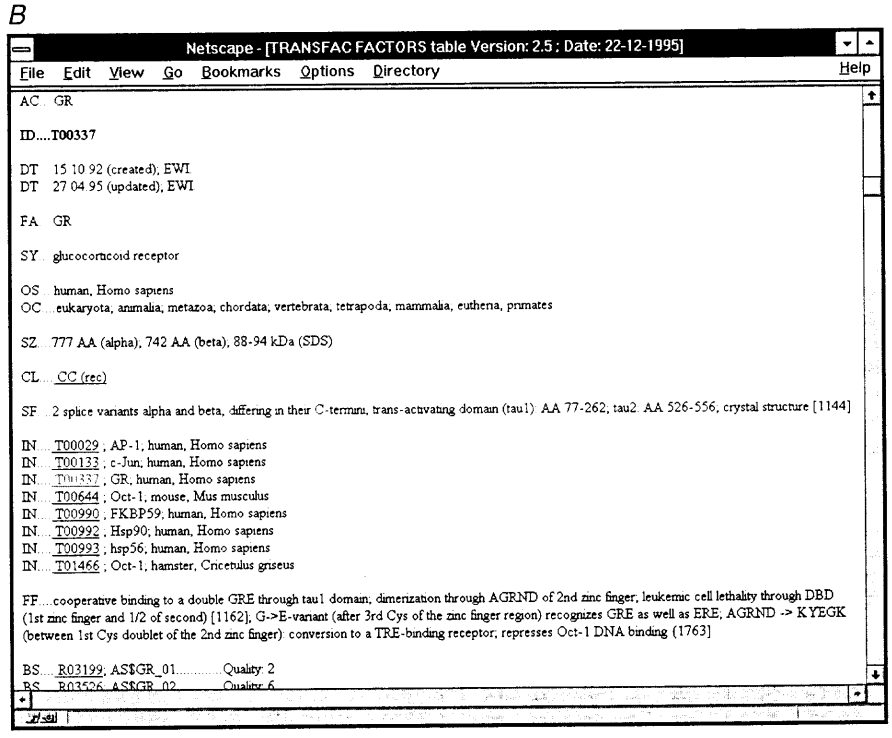


Bild 1: Nutzeroberfläche der TRANSFAC-Datenbank im WWW: (A) Ein Eintrag aus der SITE-Tabelle für das Tyrosinaminotransferase-Gen der Ratte. (B) Ein Eintrag aus der FACTOR-Tabelle, der den Glucocorticoid-Rezeptor beschreibt.

formationen abzurufen, müssen ihm übergreifende Zugriffsmöglichkeiten an die Hand gegeben werden. Ein solches Datenbankübergreifendes Zugriffssystem ist SRS (Sequence Retrieval System) von T. Etzold [1; 2]. Die Voraussetzung dafür, daß solche Systeme verschiedene Datenbanken erfolgreich verknüpfen können, ist allerdings, daß die einzelnen Datenbankentwickler die erforderlichen Querverweise zur Verfügung stellen.

Dies ist für TRANSFAC in Hinblick auf eine Reihe wichtiger Datenbanken geschehen. So enthält TRANSFAC Referenzen auf Einträge in TRRD (Transcription Regulatory Region Database), EMBL-Sequenzdatenbank, SwissProt und PIR (Proteindatenbanken), Prosite (Proteinmotivdatenbank) und Flybase (Drosophila-Gendatenbank). Diese Verweise können weitere Daten enthalten, wie die oben erwähnten Positionsangaben der Sequenzelemente in EMBL-Sequenzeinträgen. Diese Informationen sind somit weitere Annotationen der Sequenzeinträge und haben auch für die Sequenzanalyse eine besondere Bedeutung, da bei dieser die Umgebung der Sequenzelemente oft miteinbezogen wird.

Da diese Referenzen nur zeitlich begrenzte Gültigkeit besitzen, bedürfen sie bisher einer regelmäßigen Überprüfung seitens der Einzeldatenbanken. Andere Mechanismen, wie Eintragsversionen, sind eine Möglichkeit, Referenzen langfristige Gültigkeit zu verleihen.

Solchen Datenbanken, die viele derartige Referenzen enthalten, kommt in diesem Zusammenhang eine besondere Bedeutung zu, da diese benutzt werden können, um weitere Beziehungen zwischen Einträgen in unterschiedlichen Datenbanken zu generieren (transitive Hülle der vorhanden Referenzen). Dies wird bereits in SRS getan, wobei jedoch weitere Kontrollmechanismen, z. B. die Symmetrie der Referenzen, einbezogen werden.

Diese etablierten Referenzen erzeugen ein Netz von inhaltlichen Bezügen zwischen in ihrer Bedeutung und Repräsentation unterschiedlichen Daten, welches über WWW-Server in einer bisher nur rudimentären Form nutzbar ist. Für eine effektivere Nutzung dieses Datennetzes sind jedoch andere Mechanismen, die komplexere Anfragen erlauben, notwendig. Bisherige Bemühungen, Daten in einem einzigen integrierten Datenbanksystem zusammenzuführen, waren

bisher wenig erfolgreich, so daß zunehmend föderierte Datenbanksysteme, z.B. unter Verwendung von CORBA (Common Object Request Broker Architecture), favorisiert werden.

4 Verknüpfung mit Sequenzanalyse-Routinen

Eine Aufgabe der TRANSFAC-Datenbank ist somit, die verfügbaren experimentellen Daten über regulatorische Regionen (Promotoren, Enhancer) zu strukturieren und zu systematisieren. Über diesen deskriptiven Ansatz hinaus können diese Daten aber auch dazu genutzt werden, neue Genomsequenzen hinsichtlich ihres regulatorischen Potentials zu analysieren.

Im einfachsten Fall werden dazu alle in TRANSFAC gespeicherten Sequenzelemente mit einer neuen Sequenz, die einige Megabasen (= Millionen von Nukleotiden) umfassen kann, abgeglichen. Wir machen dabei Gebrauch von der speziellen Notation, in der TRANSFAC viele der bekannten Elemente wiedergibt. Von den jeweiligen Autoren für besonders wichtig erachtete Nukleotide sind in Groß-, die flankierenden in Kleinbuchstaben angegeben, so daß der Nutzer die Gewichtung zwischen beiden wählen kann.

Aufwendigere Verfahren benutzen als Suchpattern sogenannte Consensus-Sequenzen oder Nukleotidverteilungsmatrizen (Bild 2), die aus selektierten TRANSFAC-Datensätzen generiert werden können oder die bereits in TRANSFAC enthalten sind.

Die Verwendung experimentell verifizierter Faktorbindungsstellen als Suchmuster ist zunächst naheliegend. Es werden jedoch keine Muster gefunden, die nicht bereits in der Ausgangsmenge enthalten sind, es sei denn, bei der Suche sind „mismatches" erlaubt. Diese können jedoch in gleicher Weise ‚wichtige' Positionen im Suchmuster betreffen und verstärkt zu falsch-positiv-Fundstellen führen. Consensus-Sequenzen sind hier selektiver, da „mismatches" positionsabhängig bewertet werden können. Allerdings sind die Regeln zur Aufstellung solcher Consensus-Sequenzen teilweise willkürlich gewählt, und jeder Consensus vernachlässigt ausdrücklich Teile des experimentellen Datenmaterials (Bild 2). Auch Nukleotidverteilungsmatrizen bewerten positionsabhängig, allerdings ohne eine derartige Vernachlässigung. Sie sind der Suche mit Consensus-Sequenzen eindeutig überlegen, können jedoch wie diese keine Nachbarschaftsbeziehungen darstellen. Nukleotidverteilungsmatrizen müßten daher zu Oligonukleotidverteilungsmatrizen erweitert oder Positionskorrelationen in die Gewichtung von „matches" einbezogen werden, entsprechende Programme sind allerdings noch nicht verfügbar. Dagegen wurde die Matrixsuchroutine MatInspector an die TRANSFAC-Datenbank angebunden [5], ein PatternSearch-Modul wird demnächst zur Verfügung stehen.

In vielen Fällen sind diese Verfahren, die ihre Beschreibung nur aus der ‚Basenbuchstabenabfolge' ableiten, nicht geeignet, um zuverlässige Vorhersagen von Bindungsstellen zu machen. In diesen Fällen können Strukturparameter der DNA, wie Tiefe und Weite der Furchen der DNA-Doppelhelix oder Ladungsverteilungen auf der Oberfläche, als Suchmuster dienen. Hierfür mußten effiziente Verfahren entwickelt werden, um für große Sequenzen diese Muster berechnen zu können. Ein Verfahren, das hierbei zur Anwendung kommt, beruht auf einer vollständigen Bibliothek von Mustern für relativ kleine Sequenzen (Hexanukleotiden), die für große Sequenzen in geeigneter Weise aneinandergereiht werden [3].

Ein weiterer Aspekt neben der Rolle als Datenquelle für die Charakterisierung der Faktorbindungsspezifität ist der Rückverweis auf TRANSFAC-Einträge in den Resultaten der Sequenzanalyse. Diese erlauben den Zugriff auf TRANSFAC und mittelbar auf alle verknüpften Einträge in weiteren Datenbanken.

Bild 2: A) 8 experimentell gefundene, alinierte Bindungsstellen für den Transkriptionsfaktor E2F; B) die aus diesen 8 Sequenzen abgleitete Consensus-Sequenz, wobei folgende (nicht allgemeingültige) Regeln in der angegebenen Reihenfolge zur Ableitung des Consensus in den einzelnen Positionen verwendet wurden: (1) ist die Häufigkeit eines Nukleotids >50% und ist es mindestens doppelt so oft vertreten wie das Zweithäufigste, so wird dieses gewählt; (2) ist die Häufigkeit der beiden häufigsten Nukleotide >75%, so werden diese beiden gewählt (hier: Y=C/T, K=G/T); (3) alle Nukleotide (N=A/C/G/T); C) Nukleotidverteilungsmatrix der 8 angegebenen Sequenzen.

A

C	C	G	C	G	C	G	A	A	A	A	T	T
T	C	C	C	G	C	G	A	A	A	A	T	G
A	A	G	C	G	C	G	A	A	A	A	C	T
G	G	G	C	G	C	G	A	A	A	C	T	A
T	G	G	C	G	C	G	A	A	A	T	T	G
T	G	G	C	G	C	G	A	A	A	T	T	G
T	G	G	C	G	G	G	A	A	A	A	A	G
C	A	G	C	G	C	G	A	T	C	C	C	T

B

N	N	G	C	G	C	G	A	A	A	N	T	K

C

A:	1	2	0	0	0	0	0	8	7	7	4	1	1
C:	2	2	1	8	0	7	0	0	0	1	2	2	0
G:	1	4	7	0	8	1	8	0	0	0	0	0	4
T:	4	0	0	0	0	0	0	0	1	0	2	5	3

5 Ausblick

Über die Beschreibung regulatorischer Signale und die Nutzung dieser Information für ihre Erkennung in genomischen DNA-Se-

quenzen hinaus wird die TRANSFAC-Datenbank in Zukunft zu einem Werkzeug zu entwickeln sein, das vollständige Regulationsmechanismen zu modellieren vermag. Hier sind insbesondere kompliziert verzweigte Signaltransduktionskaskaden zu nennen, über die von außerhalb der lebenden Zelle kommende Signale, z. B. hormonelle Impulse, komplette genetische Programme einzuschalten vermögen. Oder die komplexe Quervernetzung der Transkriptionsfaktoren und ihrer Gene, die ja selbst unter der Kontrolle derartiger Faktoren stehen. Diesen Fragen nachzugehen wird uns schließlich zu einem umfassenden Schema derjenigen regulatorischen Prozesse führen, die die Basis für die Morphogenese[11] eines Organismus darstellen.

[11] Morphogenese: Gestaltbildung, Entwicklung der Form eines Organismus.

Literatur

[1] *Etzold, T., Argos, P.*: SRS an indexing and retrieval tool for flat file data libraries. Comput. Appl. Biosci. 9 (1993), S. 49–57.

[2] *Etzold, T., Argos, P.*: Transforming a set of biological flat file libraries to a fast access network. Comput. Appl. Biosci. 9 (1993), S. 59–64.

[3] *Karas, H., et al.*: Manuskript eingereicht.

[4] *Knüppel, R., Dietze, P., Lehnberg, W., Frech, K., Wingender, E.*: TRANSFAC Retrieval Program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol 1 (1994), S. 191–198.

[5] *Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T.*: MatInd and MatInspector – New fast and sensitive tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23 (1995), S. 4878–4884.

[6] *Schena, M., Schalon, D., Davies, R. W., Brown, P. O.*: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. In: Science 270 (1995), S. 467–470.

[7] *Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K.W.*: Serial analysis of gene expression. In: Science 270 (1995), S. 484–487.

[8] *Wingender, E.*: Compilation of transcription regulating proteins. Nucleic Acids Res. 16 (1988), S. 1879–1902.

[9] *Wingender, E.*: Transcription regulating proteins and their recognition sequences. CRC Crit. Rev. in Eukaryotic Gene Expression 1 (1990), S. 11–48.

[10] *Wingender, E.*: Gene Regulation in Eukaryotes. Weinheim: VCH, 1993.

[11] *Wingender, E., Dietze, P., Karas, H., Knüppel, R.*: TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24 (1996), S. 238–241.

Dr. Edgar Wingender
GBF, Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, 38124 Braunschweig,
Email: EWI@GBF-Braunschweig.de

Dipl.-Biol. Dipl.-Inform. Rainer Knüppel
Sys Design,
Technische EDV-Systeme,
Hauptstr. 48, 88079 Kressbronn,
Email: knueppel@ets.sel.alcakl.de

Wie Zellen miteinander reden, um die Strukturbildung eines sich entwickelnden Organismus zu organisieren

Hans Meinhardt, Max-Planck-Institut für Entwicklungsbiologie, Tübingen

Prof. Dr. Hans Meinhardt studierte Physik in Köln und Heidelberg und wurde 1966 über ein Thema zur Schwachen Wechselwirkung promoviert. Von 1966 bis 1968 Fellowship am Hochenergie-Forschungszentrum CERN in Genf. Seit 1969 ist er am Max-Planck-Institut, zunächst für Virusforschung, später für Entwicklungsbiologie. Dort befaßt er sich mit theoretischen Arbeiten zur Steuerung der Strukturbildung. 1982 habilitierte er sich im Fach Entwicklungsbiologie an der Universität Tübingen. Sein Forschungsschwerpunkt liegt im Bereich der theroretischen Aspekte der Musterbildung in höheren Organismen.

Höhere Organismen entwickeln sich aus einer einzigen Zelle, der befruchteten Eizelle. Wie kann sich der ganze Reichtum an Strukturen aus dieser einen Zelle entwickeln? Es werden Modelle diskutiert, die wichtige Schritte darin beschreiben. Primäre Musterbildung erfordert lokale Selbst-Verstärkung und langreichweite Inhibition. Stabile differenzierte Zustände werden erreicht durch eine Rückkopplung von Genen auf ihre eigene Aktivation zusammen mit einer Kompetition zwischen alternativen Genen. Das hat zur Folge, daß in einer bestimmten Zelle nur eines der alternativen Gene aktiviert werden kann. Unter dem Einfluß gradierter Konzentrationsverteilungen können ortsabhängig bestimmte Gene aktiviert werden. Es entstehen scharf begrenzte Regionen, in denen jeweils nur eines der alternativen Gene aktiviert ist. Durch Kooperation zwischen zwei solchen Bereichen können an der gemeinsamen Grenze neue Moleküle synthetisiert werden, die dann wieder als Positionsinformation für einen kleineren Bereich gebraucht werden können. Die korrekte Anlage von Armen, Beinen und Flügel an definierter Stelle, mit richtiger Orientierung und Händigkeit, wird durch das Modell verständlich. Eine solche iterative Musterbildung ermöglicht eine immer feinere Unterteilung in reproduzierbarer Weise. Die Modelle haben inzwischen direkte Unterstützung durch molekular-genetische Untersuchungen erfahren. Es werden auch Modelle diskutiert, bei denen mehr zufällige Ereignisse eine Rolle spielen. Die Bildung netzartiger Strukturen (z.B. Blattadern) und die Musterbildung auf Schneckenschalen sind Beispiele für Muster, die eine große Variabilität zeigen.

How the Cells Communicate, to Organize Pattern Formation

The generation of the complex structure of an organism in each life cycle is one of the most fascinating aspects of biological systems. Models of biological pattern formation are discussed and compared with experimental observations. Local autocatalysis and long ranging inhibition play a decisive role for primary pattern formation. Gradients appropriate to supply positional information as well as periodic distributions can be generated in this way. Cells obtain a stable state of differentiation by direct or indirect autoregulation of genes accompanied by a mutual competition among alternative genes. In this way, only one of several alternative genes can remain active within a particular cell. Boundaries between regions in which different genes are active obtain organizing properties for the formation of substructures such as legs or wings. Also discussed are processes in which random fluctuations play a crucial role. Models for leaf venation and pigmentation pattern on shells of tropical mollusks are examples. The simulations shown are numerical solutions of partial differential equations.

1 Einleitung

Ein faszinierender Aspekt biologischer Systeme ist die Fähigkeit, die komplexe Struktur eines Organismus in jedem Generationszyklus neu aufzubauen. Die Struktur des fertigen Organismus ist sicher nicht in latenter Form bereits im Ei vorhanden. Sie muß während der Entwicklung gebildet werden. Es hat den Anschein, als ob die belebte Natur da etwas Unmögliches möglich macht. Naiv würde man nach den Gesetzen der Physik statt der Bildung von geordneten Strukturen eine stetige Zunahme an Unordnung erwarten.

Die Entwicklung muß in den Genen kodiert sein. Die Ähnlichkeit eineiiger Zwillinge zeigt, bis in welche Details die genetische Festlegung geht. Wir wissen, daß in der Regel bei jeder Teilung beide Tochterzellen die gleiche Information erhalten. Es stellt sich daher die Frage, wie die Information auf der DNA in räumlich-zeitliche Muster übersetzt werden kann. Grundlegende Einsichten wurden durch Experimente gewonnen, bei denen die normale Entwicklung

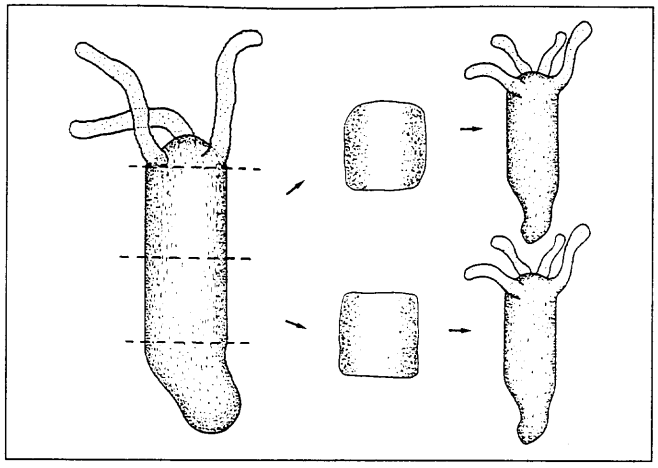


Bild 1: Ein Beispiel für die Stabilität eines entwicklungsbiologischen Systems: kleine Stücke einer Hydra regenerieren den vollständigen Süßwasserpolypen. Dabei bleibt die ursprüngliche Kopf-Fuß-Polarität erhalten.

gestört wurde. Entfernt man z.B. bei einem Süßwasser-Polypen Hydra den Kopf, so regeneriert ein neuer (Bild 1). Das zeigt, daß im Organismus ein Kommunikationssystem vorhanden sein muß, das in der Lage ist, das Fehlen eines Teiles festzustellen und vorhandene Zellen so umzuprogrammieren, daß ein normaler und voll funktionsfähiger Organismus wiederhergestellt werden kann. Erst in neuerer Zeit ist ein direkter Zugang zu den die Entwicklung steuernden Molekülen möglich geworden.

Wir haben Theorien für diese Kommunikations-Systeme entwikkelt [5]. Durch Computer-Simulationen haben wir gezeigt, daß die Theorien die beobachteten Phänomene sehr genau beschreiben. Die Modelle haben inzwischen mehrfach eine direkte Bestätigung auf molekularem Niveau erhalten. Hier soll eine kurze Übersicht gegeben werden. (Für eine ausführliche Darstellung und Referenzen siehe [10], für einen Vergleich mit neueren Experimenten siehe [13]).

2 Primäre Musterbildung durch lokale Selbstverstärkung und langreichweitige Inhibition

Die Bildung von Strukturen ist kein Privileg der Biologie. Auch in der unbelebten Natur entstehen strukturierte Gebilde aus homogenen Anfangsverteilungen, natürlich ohne daß die Gesetze der Physik verletzt werden. So kann aus einer diffusen Wolke ein fein strukturierter Blitz schlagen, oder gleichmäßig über das Land verteilter Regen kann im Laufe der Zeit durch Erosion zu scharf begrenzten Flüssen führen. Diesen strukturerzeugenden Prozessen ist gemeinsam, daß kleine Störungen eine so starke Rückwirkung auf sich selbst haben, daß die Störungen weiter anwachsen. Betrachten wir z.B. die Geschichte einer Sanddüne. Vielleicht war ein Stein der Ausgangspunkt. Der Stein erzeugte einen Windschatten, in dem sich Sand ablagern konnte. Der abgelagerte Sand vergrößerte den Windschatten, noch mehr Sand wird abgelagert. Ein sich selbst verstärkender Prozeß.

Neben der Selbstverstärkung muß aber noch eine andere Bedingung erfüllt sein, wenn Strukturbildung stattfinden soll. Selbstverstärkung allein würde nur zu einer immer weiteren Ausbreitung dieser Reaktion führen, so wie sich etwa ein Waldbrand ausweitet. Dadurch würde nur ein Zustand in einen anderen, wieder strukturlosen Zustand übergehen. Strukturbildung impliziert aber, daß an einem Ort etwas geschieht, was in einer weiteren Umgebung nicht geschieht. Eine Inhibition (Hemmung) muß dieser Selbstverstärkung entgegenwirken, und diese Inhibition muß sich schneller ausbreiten als die autokatalytische Reaktion selbst. Eine lokale Begrenzung der Selbstverstärkung wird demnach dadurch erreicht, daß diese nur auf Kosten einer größeren Umgebung möglich ist.

Dieses Grundprinzip, lokale Selbstverstärkung und langreichweitige Inhibition, ist leicht auf molekulare Wechselwirkungen übertragbar [5; 11; 14]. Denkbar ist ein „Aktivator", der seine eigene Synthese direkt oder indirekt verstärkt. Für die erforderliche inhibitorische Reaktion soll der Aktivator die Synthese eines Inhibitors steuern, der seinerseits die Selbstverstärkung hemmt. Da sich die Inhibition schneller ausbreiten muß, wird angenommen, daß der Inhibitor schneller als der Aktivator diffundiert. In einem kleinen Areal von Zellen, in dem sich eine lokale Schwankung der Aktivator-Konzentration schnell ausgleichen kann, sollen sich beide Stoffe in einem stabilen Gleichgewicht befinden. Zum Beispiel soll eine Erniedrigung der Aktivator-Konzentration zu einer solchen Erniedrigung der Inhibitor-Produktion führen, daß das System durch die autokatalytische Aktivator-Produktion wieder in das Gleichgewicht zurückkehrt.

Wenn aber ein solches „Feld" wächst, werden die homogenen Aktivator- und Inhibitor-Verteilungen instabil. Eine kleine, z. B. durch eine statistische Schwankung bedingte lokale Erhöhung des Aktivators erzeugt zwar nach wie vor eine erhöhte Inhibitor-Produktion. Jetzt ist aber genug Raum vorhanden, in den der überschüssige Inhibitor diffundieren kann. Die erhöhte Aktivator-Konzentration wird durch den Inhibitor nicht vollständig zurückgeregelt, sondern wird durch die Autokatalyse weiter anwachsen. Der zusätzlich produzierte Inhibitor unterdrückt aber in der weiteren Umgebung dieses sich herausbildenden Maximums mehr und mehr die Aktivator-Produktion. Es entsteht eine zeitlich stabile, inhomogene Verteilung von beiden Substanzen (Bild 2), die als Signalsystem für die beteiligten Zellen verwendet werden kann.

Musterbildungen, die auf dem gleichen Prinzip basieren, wurden auch in Gasentladungen [1] und in entsprechenden elektronischen Schaltungen beobachtet [2; 3].

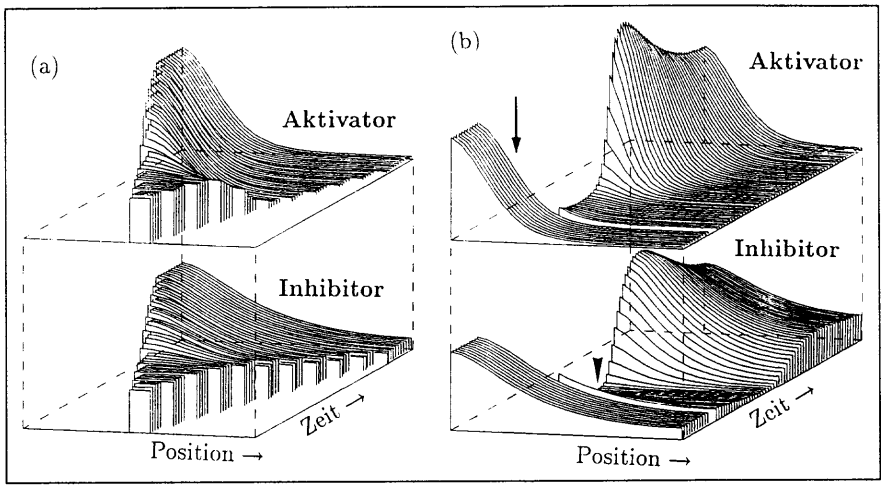


Bild 2: Bildung und Regeneration einer Aktivator- und Inhibitor-Verteilung. Die Konzentrationen sind als Funktion der Zeit dargestellt. Der Aktivator hat eine positive Rückwirkung auf seine eigene Produktion und auf die des sich schnell ausbreitenden Inhibitors [5; 10]. (a) Angenommen ist eine wachsende Kette von Zellen. Wenn das Feld eine bestimmte Größe überschritten hat, setzt die Musterbildung ein: es bildet sich eine hohe Konzentration auf der einen und eine niedrige Konzentration auf der anderen Seite. (b) Regeneration: Nach Entfernen der aktivierten, d.h. der Inhibitor-produzierenden Zellen (Pfeil) zerfällt der verbliebene Inhibitor (Pfeilspitze) bis die Selbstverstärkung der Aktivator-Produktion wieder einsetzt. Die ursprüngliche Verteilung wird wiederhergestellt. Die Abbildung zeigt numerische Lösungen von zwei gekoppelten, nichtlinearen Differentialgleichungen.

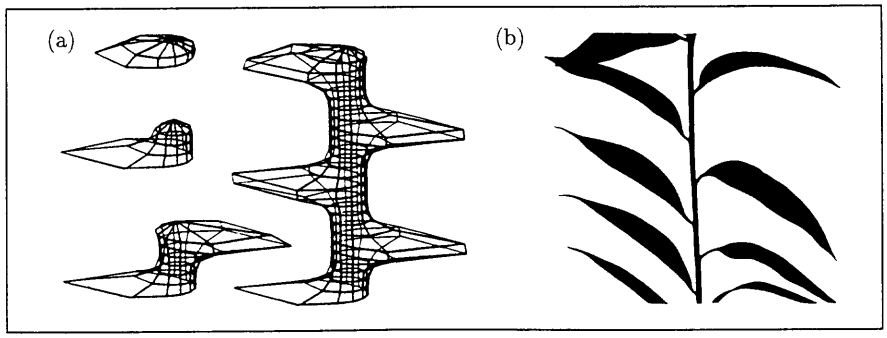
3 Morphogenetische Gradienten

Eine klassische Vorstellung in der Entwicklungsbiologie ist die der Steuerung der Entwicklung durch gradierte Konzentrations-Verteilungen. Wolpert [19] hat dafür den Begriff der „*Positional Information*" geprägt. Abhängig von der Konzentration werden positionsabhängig verschiedene Gene aktiviert. Das beschriebene Modell zeigt, wie solche Gradienten gebildet werden können. Für biologische Anwendungen des obigen Modells ist wichtig, daß in einem kleinen Feld eine Schwankung am Rand schon zur Ausbildung eines Maximums führen kann, während eine Schwankung im Zentrum noch zurückgeregelt wird (Bild 2). Es entstehen also polare Strukturen: durch die hohe Signal-Konzentration kann auf der einen Seite eine andere Struktur angelegt werden als auf der anderen. Das ist für die Ausbildung embryonaler Achsen (z. B. der Kopf-Schwanz- oder der Rücken-Bauch-Achse) eines sich entwickelnden Embryos von zentraler Bedeutung. Das bisher eindeutigste Beispiel für einen solchen „Morphogen-Gradienten" ist die Verteilung des sogenannten „*bicoid*"-Proteins im Ei der Fruchtfliege *Drosophila* [4] (die Bildung dieses Gradienten geschieht aber auf eine viel komplexere Weise). Wenn das Feld von Zellen dagegen sehr viel größer ist als die Reichweite des Inhibitors, so entstehen viele Maxima, die voneinander einen bestimmten maximalen und minimalen Abstand haben. Solche Signale sind für periodische Strukturen wie z.B. Haare, Borsten, etc. notwendig. Die Simulation in Bild 3 zeigt als Beispiel die Bildung von Blattanlagen auf einem wachsenden Sproß.

4 Regeneration

Eine solche Wechselwirkung zwischen zwei Stoffen hat nicht nur die Eigenschaft, Muster zu bilden, sie kann auch Regelprozesse wie die oben erwähnte Regeneration einer Hydra (Bild 1) erklären. Nehmen wir an, eine hohe Aktivator-Konzentration ist das Signal zur Kopfbildung. Entfernt man den Kopf, so entfernt man damit auch die den Inhibitor produzierenden Zellen. Der verbleibende Inhibitor zerfällt, bis eine erneute Autokatalyse des Aktivators möglich wird. Ein neues Maximum und damit ein neues Signal für eine Kopfbildung bildet sich. Durch neu synthetisierten Inhibitor wird das System wieder in ein Gleichgewicht gebracht (Bild 2b). Die tatsächliche Musterbildung in der Hydra ist aber noch komplizierter, da nicht nur ein Kopf-Signal, sondern auch Signale für Fuß-, Tentakel- und Knospen-Bildung erzeugt werden müssen. Ein entsprechendes Modell [12] hat in der Zwischenzeit direkte experimentelle Unterstützung erfahren [17].

Bild 3: Bildung einer periodischer Struktur während des Wachstums. (a) Die Simulation zeigt die Bildung von Signalen auf einem wachsenden Sproß, die eine Blattbildung auslösen. Angenommen ist ein Zylinder, der durch Proliferation der obersten Zellreihe wächst. Ein existierendes Maximum erzeugt eine inhibitorische Zone um sich (nur die Aktivator-Verteilung ist gezeigt). Eine neue Aktivation wird ausgelöst, wenn durch Wachstum eine genügende Distanz zu einem existierenden Maximum erreicht worden ist. Es entsteht eine regelmäßige Struktur, obwohl nur kleine statistische Schwankungen angenommen wurden. (b) Eine natürliche Struktur (nach [10]).



5 Gen-Aktivierung: molekulargenetische Analog-Digital-Konvertierung

Die Bildung eines Kopfes, Rumpfes oder Schwanzes erfordert die stabile Aktivierung jeweils verschiedener Gene. Ein auf Diffusion basierendes Kommunikationssystem kann nur über kurze Distanzen aufrechterhalten werden. Für den (ungerichteten) Transport von Molekülen über größere Strecken wäre die Zeit, die für die Kommunikation benötigt würde, viel zu lang.

Gen-Aktivierung und räumliche Musterbildung haben formale Aspekte gemeinsam. In der räumlichen Musterbildung sollen z.B. nur die Zellen auf einer Seite aktiviert werden. Bei der Gen-Aktivierung soll z.B. das Gen *3* in einer Zelle aktiviert werden, die alternativ in dieser Entwicklungssituation möglichen Gene *1, 2* und *4* aber nicht. Gen-Aktivierung erfordert also eine Musterbildung zwischen alternativen Genen. In einer entsprechenden Theorie habe ich postuliert, daß eine stabile Gen-Aktivierung durch eine direkte oder indirekte Rückwirkung eines Gen-Produkts auf die Aktivierung des eigenen Genes erreicht wird [9]. Heute sind viele Beispiele für eine solche positive Auto-Regulation bekannt, wie etwa das Gen *Deformed* in *Drosophila* [6].

Durch die nichtlineare Rückwirkung eines Gen-Produktes auf seine eigene Aktivierung entsteht eine scharfe Schwelle. Wenn diese überschritten ist, entsteht eine Gen-Aktivierung, die von dem auslösenden Signal unabhängig ist. Wenn das Signal eine gradierte Verteilung war, so entstehen scharf abgegrenzte Regionen, in denen bestimmte Gene aktiv sind.

Für die positionsabhängige Aktivation mehrerer Gene unter dem Einfluß einer gradierten Konzentrationsverteilung konnte ich zeigen, daß eine Reihe von experimentellen Ergebnissen nur mit der Annahme erklärbar ist, daß die Determination der Zellen schrittweise geändert wird, bis die erreichte Determination der lokalen Morphogen-Konzentration entspricht. Die Zellen werden sozusagen „befördert"; wie weit, bestimmt die lokale Konzentration des Morphogens. Jeder dieser Schritte ist irreversibel. Durch die Selbst-Aktivierung der Gene bleibt das System auch dann stabil, wenn das Signal nicht mehr vorhanden ist.

Eine Analogie soll den Vorgang anschaulicher machen: Nehmen wir ein Holzstück an, das am Fuße einer Kellertreppe liegt. Durch eine Überschwemmung kann es auf eine höhere Stufe gehoben werden, auf der es dann liegenbleibt, auch wenn die Flut wieder abläuft. Diese Stufe ist ein Maß für den höchsten Wasserstand. Eine spätere noch höhere Flut kann das Stück noch um weitere Stufen hochsetzen, eine spätere Flut mit niedrigerem Maximalwert ist aber ohne Einfluß auf ein schon höher liegendes Holzstück.

Der Grund für diese Art der Regelung ist leicht zu verstehen. Wenn eine gradierte Verteilung durch eine lokale Quelle und Diffusion aufgebaut wird, dann führt jedes Wachstum zu einer Vergrößerung des Abstandes einer Zelle von der Quelle und damit zu einer Erniedrigung der lokalen Konzentration. Wenn nun eine einmal erreichte Gen-Aktivierung stabil gegenüber einer Erniedrigung der Signal-Konzentration ist, dann bleibt eine einmal erreichte Aufteilung erhalten.

6 Bildung neuer Strukturen an oder um die Grenzen verschiedener Gen-Aktivität

Die Komplexität eines höheren Organismus ist zu groß, als daß sie durch zwei orthogonale Gradienten erreicht werden könnte. Experimentell hat sich gezeigt, daß Unterstrukturen wie Arme, Beine und Flügel ihr eigenes Koordinaten-System haben. Im Axolotl, einem mexikanischen Höhlenmolch, kann man zum Beispiel von einem bestimmten Stadium an das Gewebe, das den Arm machen wird (von dem aber zu diesem Zeitpunkt noch nichts sichtbar ist), auf die Kopf-Kapsel transplantieren, und der vollständige Arm entwickelt sich dort [16]. Ich habe vorgeschlagen, daß sekundäre Strukturen um existierende Grenzen angelegt werden. Nehmen wir an, daß durch die primäre Unterteilung eine Reihe von scharf voneinander abgegrenzten Regionen gebildet wurden, darunter auch die benachbarten Regionen *P* (posterior) und *A* (anterior). Wenn ein Molekül *m* nur durch eine Kooperation der *A*- und der *P*-Zellen produziert werden kann, so ist die Produktion nur an der *P/A*-Grenze möglich. Diffundiert dieses Molekül in die Umgebung, so ist die sich einstellende lokale *m*-Konzentration ein Maß für die Entfernung von der *P/A*-Grenze (Bild 4a). Obwohl das Signal symmetrisch ist, kann die entstehende Struktur asymmetrisch sein, denn die *A*- und die *P*-Zellen können sehr verschieden auf das Signal reagieren. Besonders ausgeprägt ist dieses in der Arm-Bildung von Vertebraten, wo nur die *A*-Zellen auf das Signal reagieren. Die Abfolge unserer Finger vom Daumen bis zum kleinen Finger entsteht durch die abnehmende Distanz zu einer solchen primären Grenze und damit zu ansteigender Morphogen-Konzentration. Bild 4 zeigt, daß das Modell die Bildung eines weiteren Beines mit umgekehrter Polarität nach einer bestimmten Operation zwanglos erklären kann. Dieses Ergebnis war lange Zeit sehr rätselhaft, denn bei der Transplantation wurde nur Gewebe aus einer mehr kopfnahen in eine mehr schwanznahe Position gebracht. Dabei wurde das Gewebe nicht gedreht, aber das entstehende zusätzliche Bein hat trotzdem die umgekehrte Ausrichtung.

Eine solche *A/P*-Grenze umspannt einen (zylinderförmigen) Embryo wie ein Gürtel. Um z. B. die Lage eines Bein-Paares entlang einer solchen Grenze zu bestimmen, ist ein Schnittpunkt mit einer zweiten Grenze notwendig, die eine Rücken- gegen eine Bauchregion abgrenzt. Es ist also eine dorsoventrale Grenze *(D/V)* erforderlich. Solche *A/P*- und *D/V*-Schnittpunk-

te treten notwendigerweise immer als Paare auf, einer auf der linken und einer auf der rechten Seite des Organismus (Bild 4b). Beide haben entgegengesetzte Händigkeit. Ein Arm vom linken Typus entsteht auf der linken Körperseite nicht, weil dort z.B. ein Gen „links" aktiviert worden ist, sondern weil die vier Regionen, die zur Beinbildung notwendig sind, eine Anordnung mit Drehsinn haben. Beim linken Bein ist die Reihenfolge gegen den Uhrzeigersinn gerichtet (*AD, AV, PV* und *PD*). Damit erklärt das Modell die paarige Anlage der Extremitäten und ihre Symmetrie. Im Gegensatz zum klassischen Modell ist bei dem Anlegen einer neuen Struktur um eine Grenze nie ein strukturloser Zustand vorhanden, der dann organisiert werden muß. Das Modell hat in der Zwischenzeit viel Unterstützung von molekulargenetischer Seite erfahren [7; 13; 18].

7 Die Bildung netzartiger Strukturen

Aus dem bisher Gesagten könnte man den Eindruck bekommen, daß die Entwicklung eines Organismus bis in alle Details durch die Gene festgelegt ist. Das ist aber sicher nicht der Fall. Die Adern eines jeden Blattes einer Pflanze sind etwas unterschiedlich angeordnet (Bild 5), obwohl sich natürlich alle Blätter einer Pflanze unter dem Einfluß der gleichen genetischen Information entwickelt haben.

Solche filamentartigen verzweigten Strukturen sind Bestandteil wichtiger Organe in allen höheren Organismen, wie Blutgefäße, Lymphgefäße, Tracheen oder Nierentubuli. Sie dienen zur Versorgung oder Drainage des Gewebes. Der für die Netzbildung vorgeschlagene Mechanismus beruht auf folgender Idee [8; 10]: Ein lokales Signal, z.B. durch ein Aktivator-Inhibitor-System gebildet, bewirkt eine lokale Differenzierung in einen zum Netzwerk gehörenden Zelltyp. Hat sich aber eine Zelle unter dem Einfluß dieses Signals differenziert, so wird das Signal unterdrückt. Eine differenzierte Zelle stößt daher das Signal ab, es wird in eine Nachbarzelle verschoben, die sich nun ebenfalls differenziert. Das Verschieben geschieht in eine Richtung, die noch nicht ausreichend durch Adern versorgt worden ist. Differenzierte Zellen entstehen quasi als eine Spur hinter einem wandernden Signal. Lange Filamente von differenzierten Zellen entstehen gleichsam als Spur hinter wandernden Aktivator-Maxima. Sind die Wachstumspunkte – in ihnen wird Inhibitor ausgeschüttet – weit genug voneinander entfernt, können neue Maxima entlang existierender Filamente entstehen, die wiederum von den differenzierten Zellen wegstreben. Das führt zu Verzweigungen (Bild 5).

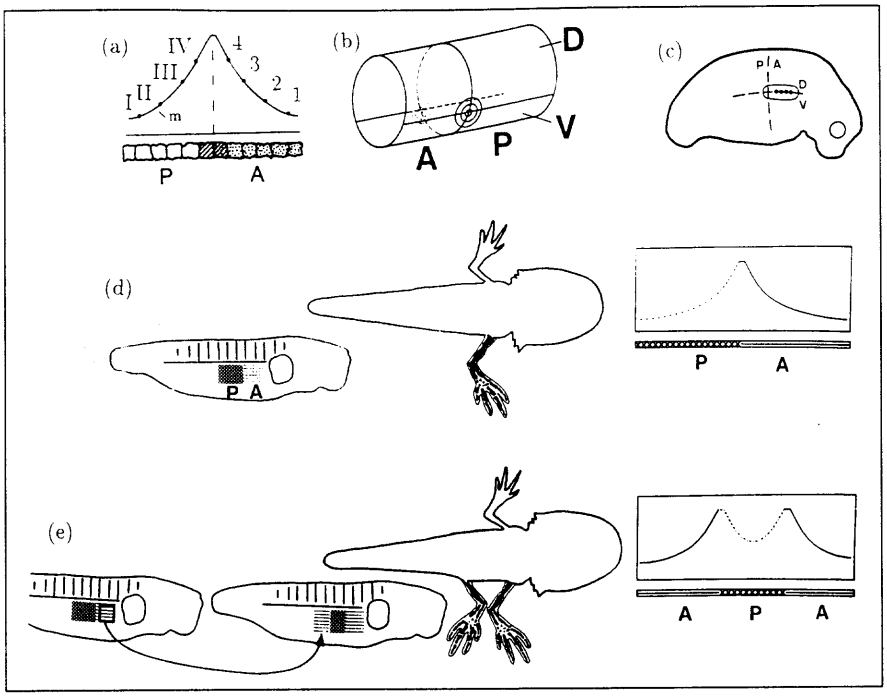


Bild 4: Modell zur Bildung einer Bein-Anlage. (a) Wenn zwei verschieden determinierte Regionen (*A* und *P*) zusammenarbeiten müssen, um eine neue Substanz *m* zu produzieren, so kann deren Produktion nur an der gemeinsamen Grenze stattfinden. Die lokale Konzentration ist ein Maß für die Entfernung von der Grenze. (b) Wenn die Kooperation von zwei Paaren von differenzierten Zellen (*A/P* und *D/V*) erforderlich ist, so entstehen die Organisator-Regionen an den Schnittpunkten der beiden Grenzen (konzentrische Kreise). In einem zylindrischen Embryo entstehen diese immer in Paaren, eines auf der linken, das andere auf der rechten Seite. (c) Lage des Bein-Feldes im Embryo. Die Finger entstehen entlang der *D/V*-Grenze, der Fingertyp ist von der Entfernung zur *A/P*-Grenze abhängig. (d, e) Bildung eines überzähligen (rechten) Beines nach Transplantation von (linkem) Gewebe an eine mehr posteriore Stelle. (d) normale Situation; nur der Gradient in der A-Region wird verwendet. (e) Nach der Transplantation einer *A*-Region hinter eine *P*-Region entsteht ein zusätzliches Bein. Nach dem Modell entsteht ein zweiter *A/P*-Schnittpunkt und damit ein zweiter Gradient in einer *A-Region*. Er hat eine umgekehrte *A/P*-, aber gleiche *D/V*-Polarität. Bei einer Operation auf der linken Seite hat das zusätzliche Bein also die Struktur eines rechten Beines (Experiment: [16], Modell: [10; 11]).

8 Die Bildung von Pigmentmustern auf Schnecken- und Muschelschalen

Die Adern waren ein Beispiel für ein nicht vollständig determiniertes Muster. Noch offensichtlicher ist das bei den Farbmustern auf Schalen tropischer Meeresschnecken. Auch innerhalb der gleichen Specie gleicht kein Muster genau dem eines anderen Tieres. Diese Muster können aus Linien bestehen, die parallel, senkrecht oder geneigt zur Richtung des Schalenwachstums angeordnet sind. Sie können sich verzweigen, kreuzen oder zu Punktreihen aufgelöst sein. Es war für uns eine Überraschung, daß diese Muster auch durch Wechselwirkungen des

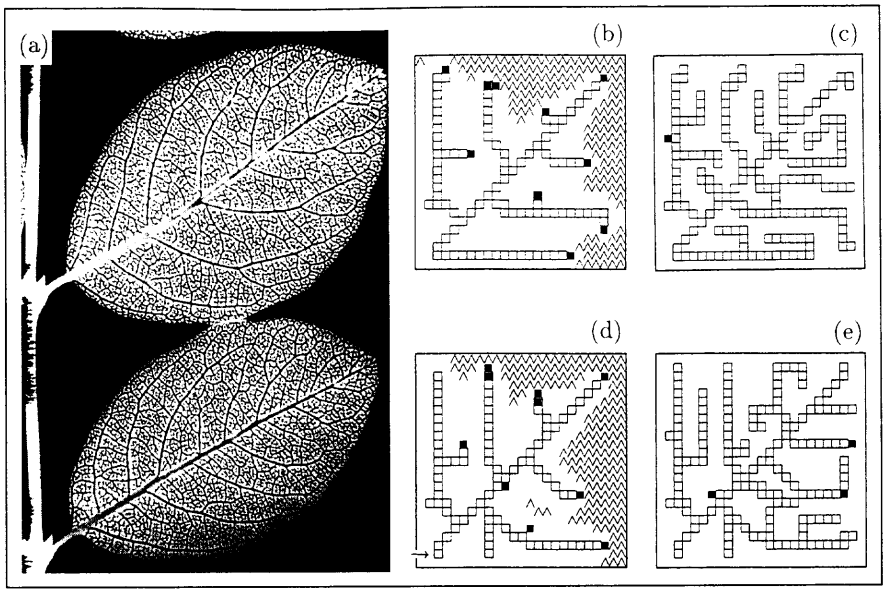


Bild 5: (a) Zwei Blätter eines Baumes sind nicht identisch, obwohl sie unter Kontrolle der gleichen genetischen Information entstanden sind. Diese Verschiedenheit zeigt sich auch in der Simulation [8; 10]. Kleine zufällige Schwankungen beeinflussen die Entscheidung, welche weitere Zelle aktiviert werden. Das hat jeweils starke Auswirkungen auf weitere Entscheidungen. Die globale Struktur ist sehr ähnlich, die Details sind aber verschieden. (b) und (c) sowie (d) und (e) zeigen zwei Simulationen, (b) und (d) sind Zwischenzustände, (c) und (e) die stabilen Endzustände. Die Adern-Bildung wurde jeweils durch die gleiche Zelle (Pfeil) initiiert; schwarz: hohe Aktivierung, umrandete Quadrate: differenzierte Zellen.

oben genannten Typs – Selbstverstärkung und antagonistische Reaktion – simuliert werden könnten [14; 15]. Wie können Muster, die so verschieden aussehen, auf dem gleichen Prinzip beruhen?

Um die Bildung dieser Muster auf den Gehäusen von Schnecken oder Muscheln zu verstehen, muß man sich zunächst vor Augen halten, daß sie in einer besonderen Weise entstehen. Die Schale kann naturgemäß nur durch Anlage von Material an der äußersten Kante vergrößert werden; das Gehäuse selbst ist starr. In der Regel findet nur in dieser Wachstumsregion Pigmenteinbau statt. Die Muster sind also eine zeitliche „Aufzeichnung" von Prozessen, die an der wachsenden Kante stattgefunden haben. Denken wir uns eine Schale flach ausgebreitet, so bedeutet eine Achse die Position entlang der Kante, die andere die Zeit-Koordinate.

Es kann hier nur beispielhaft eine Erklärung für die Entstehung von zwei bestimmten Schalenmustern gegeben werden. Bild 6 zeigt die Schale der Schnecke *Oliva porphyria*. Man sieht Linien ungefähr diagonal zur Wachstumsrichtung, die sich häufig verzweigen. An einer Verzweigung bildet sich eine neue, ebenfalls diagonal verlaufenden Linie, die aber die entgegengesetzte Orientierung hat.

Wie entstehen diagonale Linien? Offenbar lagern kleine Gruppen von Zellen jeweils nur für eine kurze Zeitspanne Pigment ein. Etwas verzögert beginnt eine benachbarte Zelle, ebenfalls nur für kurze Zeit, mit Pigmenteinlagerung usw. Es entsteht eine Wanderwelle von Pigmentproduktion. Unmittelbar nach einer Pigmenteinlagerung sind die Zellen jedoch für eine bestimmte Zeit „immun" gegen eine weitere Infektion. Nach dem oben beschriebenen Aktivator-Inhibitor Modell entstehen solche Wanderwellen, wenn der Aktivator schneller diffundiert und kürzer lebt als der Inhibitor, also die Situation im Vergleich zur oben beschriebenen Bildung stabiler Muster (Bild 2) gerade umgekehrt ist. Durch den diffundierenden autokatalytischen Aktivator entsteht die Infektiosität. Durch die auf die Aktivierung folgende Produktion des Inhibitors und seines langsameren Abbaus wird die Aktivierung nach kurzer Zeit wieder unterdrückt, und es entsteht die immune Phase. Nach dem Abbau des Inhibitors ist eine erneute Autokatalyse möglich.

Das System verhält sich so, als ob die Pigmentproduktion ein ansteckender Prozeß sei, wie etwa bei einer Grippewelle. Diese entsteht auch durch die Kopplung eines sich selbst-verstärkenden und eines antagonistischen Prozesses. Die Viren sind selbstvermehrend. Deshalb genügen wenige Viren, um uns krank zu machen. Aber die Viren lösen auch eine antagonistische Reaktion aus, die Immun-Antwort, die das Abtöten der Viren bewirkt.

Die kurzen, in Nachbarzellen aufeinanderfolgenden Phasen von Pigmentproduktion führen in der zeitlichen Aufzeichnung zu den schrägen Linien. Die Neigung der Linien ergibt sich aus dem Verhältnis von Wachstumsgeschwindigkeit der Schale und der Geschwindigkeit der Infektions-Welle. Wenn eine Zelle spontan aktiviert wird, können beide Nachbarzellen angesteckt werden, denn keine Nachbarzelle ist im Zustand der Immunität. Eine solche Zelle wird zum Ausgangspunkt von zwei Linien mit entgegengesetzter Neigung. Wenn zwei Wellen aufeinandertreffen, sind alle in Frage kommenden Zellen immun, und die Wellen löschen sich gegenseitig aus (V-artiges Musterelement).

Wodurch entstehen aber Verzweigungen? Da sich zwei ineinander laufende Wellen gegenseitig auslöschen, wird die Zahl der Wellen, die entlang der Wachstumskante laufen, immer kleiner; es sei denn, ein separater Mechanismus sorgt für die Bildung neuer Wellen. Ein solcher Mechanismus ist die Bildung von Verzweigungen, wie in *Oliva porphyria* (Bild 6) realisiert ist. Eine Verzweigung bedeutet, daß sich eine Welle aufgespalten hat in eine normal weiterlaufende Welle und eine rückwärts laufende Welle. Das geschieht, wenn eine Zelle so lange aktiviert bleibt, bis die immune Phase ihrer Nachbarzelle vorüber ist und diese also wieder angesteckt werden kann. In der Simulation in Bild 6

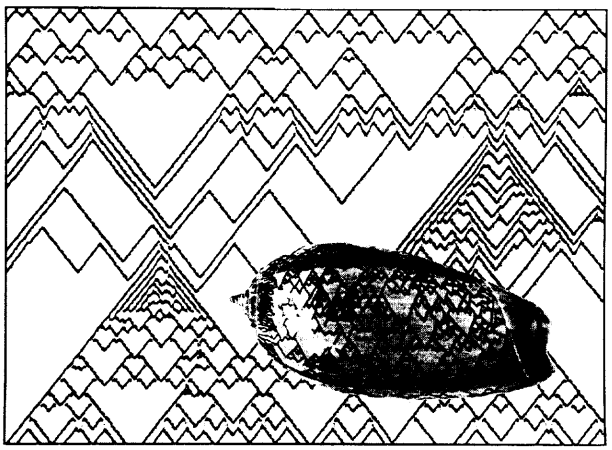


Bild 6: Ein Gehäuse der Schnecke *Oliva porphyria* und, im Hintergrund, eine Computer-Simulation des Pigment-Musters. Die schrägen Linien werden durch pigmentauslösende Wanderwellen erzeugt. Abzweigungen sind möglich, wenn die Zahl der Wellen (d.h. die Zahl der Linien zu einem bestimmten Zeitpunkt) zu gering geworden ist. Regionen mit dichter Linienbildung können entstehen, wenn durch Zufall große nicht-pigmentierte Regionen entstanden sind (aus [14]).

wurde angenommen, daß jede Wanderwelle zur Produktion einer hormonartigen Substanz beiträgt, die den Abbau des Inhibitors blokkiert. Mit abnehmender Anzahl der Wanderwellen wird also die Hormonkonzentration immer kleiner und die Lebensdauer des Inhibitors immer kürzer. Wenn die Zahl der Wanderwellen und damit die Hormonkonzentration einen bestimmten Wert unterschreitet, so stellt sich die Inhibitorkonzentration so schnell auf eine veränderte Aktivatorkonzentration ein, daß die pulsartige Aktivierung zugunsten einer permanenten Aktivierung aufgegeben wird. Die neu aktivierten Zellen bleiben so lange aktiv, bis Rückwärtswellen initiiert worden sind. Damit steigt die Hormonkonzentration wieder an, es tritt wieder eine pulsartige Aktivierung auf und normale Wanderwellen werden wieder gebildet. In der zeitlichen Aufzeichnung auf der Schale führt dieser Prozeß zu einer schrägen Linie mit Verzweigung.

Andere Muster deuten auf zwei miteinander wechselwirkende Systeme hin. Dabei ist in der Regel nur ein System sichtbar, das zweite System modifiziert die Parameter des ersten. Eine Vielzahl solcher Schalenmuster läßt sich mit kleinen Veränderungen dieses Prinzips deuten und durch Computersimulationen reproduzieren. Unser Modell macht damit den Reichtum an Mustern auf Schnecken- und Muschelschalen verständlich. Weil diese Muster offenbar keinem starken selektiven Druck ausgesetzt waren, konnte diese Vielzahl von Mustern gebildet werden. Da die räumliche-zeitliche Entwicklung im Muster konserviert ist, sind sie ein wunderbares Bilderbuch zum Studium dynamischer Systeme der Natur. Ein Buch mit vielen solchen Simulationen ist kürzlich erschienen [14]. Es enthält auch eine Programm-Diskette (mit Quellcode), mit der die Simulationen auf einem PC wiederholt und neue Modelle integriert werden können.

9 Schlußbetrachtung

Sich selbst verstärkende Reaktionen, die mit antagonistischen Reaktionen gekoppelt sind, können eine Vielzahl von Mustern bilden. Breitet sich die antagonistische Reaktion schnell aus und hat eine kurze Zeitkonstante, so bilden sich stabile Muster im Raum. Selbst nach einer Störung kann das normale Muster wieder gebildet werden – eine Stabilität, die für sich entwickelnde Organismen sehr wichtig ist. Hat dagegen die antagonistische Reaktion eine lange Zeitkonstante, so treten Oszillationen auf. Wenn, wie bei der Pigmentbildung auf den Schalen tropischer Meeres-Schnecken, eine Vielzahl solcher Oszillatoren gekoppelt sind, so kann chaotisches Verhalten stattfinden mit der Folge, daß alle Muster voneinander verschieden sind. Ob also robuste Muster gebildet werden, oder solche, die auf kleine Variationen äußerst empfindlich reagieren, ist nicht eine Frage der Wechselwirkung, sondern eine Frage der Lebensdauern und der Ausbreitung der beteiligten Stoffe. Die Natur macht bei passender Gelegenheit von allen diesen Möglichkeiten Gebrauch.

Danksagung

Viele dieser Modelle sind in einer für mich sehr schönen Zusammenarbeit mit Alfred Gierer entstanden, für die ich mich hier herzlich bedanken möchte.

Literatur

[1] *Astrov, Y.u., Ammelt, S., Teperick, S., Purvin, H.G.*: Phys. Let. A 211 (1995), S. 184–190.

[2] *Berkemeier, J., Dirksmeyer, T., Klempt, G., Purwins, H.G.*: Z. Phys.B – Condensed Matter 65 (1986), S. 255–258.

[3] *Chua, O.L., Hasler, M., Moschytz, G.S., Neirynck, J.*: IEEE Tansactions on Circuits and Systems 42, (1995), S. 559–577.

[4] *Driever, W., Nüsslein-Volhard, C.*: Cell, 54, (1988), S. 83–93.

[5] *Gierer, A., Meinhardt, H.*: 12 (1972), S. 30–39.

[6] *Kuziora, M.A., McGinnis, W.*: Mechanisms Development 33, (1990), 83–94.

[7] *Martin, G.R.*: Nature 374 (1995), S. 410–411.

[8] *Meinhardt, H.*: Differentiation 6 (1976), S. 117–123.

[9] *Meinhardt, H.*: J. theor. Biol. 74 (1978), S. 307–321.

[10] *Meinhardt, H.*: Models of biological pattern formation. Academic Press, London, (1982).

[11] *Meinhardt, H.*: J. Embryol exp. Morphol 76 (1983), S. 115–137.

[12] *Meinhardt, H.*: Dev. Biol. 157 (1993), S. 321–333.

[13] *Meinhardt, H.*: Bioessays 16 (1994), S. 627–632.

[14] *Meinhardt, H.*: The Algorithmic Beauty of Sea Shells. Springer, Heidelberg, New York 1995.

[15] *Meinhardt, H., Klingler, M.*: J. theor. Biol 126 (1987), S. 63–69.

[16] *Slack, J. M. W.*: Nature 261 (1976), S. 44–46.

[17] *Technau, U., Holstein, T. W.*: Development 121 (1995), S. 1273–1282.

[18] *Vincent, J. P., Lawrence, P. A.*: Nature 372 (1994), S. 132–133.

[19] *Wolpert, L.*: J. theor. Biol. 25 (1969), S. 1–47.

Prof. Dr. Hans Meinhardt
Max-Planck-Institut für Entwicklungsbiologie,
Spemannstraße 35, 72076 Tübingen,
E-mail meinh@mpib-tuebingen.mpg.de

Sequenzanalyse mit verteilten Ressourcen: Ein WWW-basierter Kurs[1]

Christian Büschking, Robert Giegerich, Technische Fakultät, Universität Bielefeld

Dipl.-Inform. Christian Büschking studierte an der Technischen Fakultät in Bielefeld Naturwissenschaftliche Informatik mit Schwerpunkt Molekularbiologie und ist heute wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Praktische Informatik an der Technischen Fakultät der Universität Bielefeld. Die Forschungsschwerpunkte liegen in der Visualisierung von Sequenzanalysedaten, sowie in der Vorhersage von RNA-Sekundärstrukturen.

Prof. Dr. Robert Giegerich studierte Informatik an der TU München und in Stanford, USA. Heute ist er Professor für Praktische Informatik an der Technischen Fakultät der Universität Bielefeld. Zu seinen wichtigsten Forschungsgebieten gehören Programmiersprachen und Compiler sowie Algorithmen und Werkzeuge für die Analyse von Biosequenzdaten. In der Lehre entwickelt er neue Lehrformen für die Bioinformatik im Rahmen des Bielefelder Studiengangs Naturwissenschaftliche Informatik. Aus dieser Tätigkeit ist der vorliegende Beitrag entstanden.

Die Sequenzanalyse beginnt gleich, nachdem die ersten 300 bis 400 Basen sequenziert sind. Antworten zu Fragen, wie: „Gibt es irgendwelche homologen Sequenzen in den Datenbanken?", „Kann man aus der Ähnlichkeit zu anderen Sequenzen auf die Funktion schließen?", „Gibt es Verweise zu homologen Sequenzen?" können Werkzeuge auf dem World Wide Web (WWW) geben. Beispielsweise kann man aus einem Labor in Europa einen Homologievergleich in Japan durchführen und ein multiple sequence alignment *in den USA berechnen lassen. Molekularbiologen arbeiten mit diesen Werkzeugen, die auf dem WWW frei verfügbar sind, fast täglich. Wenn Sequenzanalyse-Werkzeuge auf dem WWW benutzt werden, dann ist natürlich auch für den Kurs* ***Sequenzanalyse mit verteilten Ressourcen*** *das WWW das Mittel der Wahl. Um am Kurs teilzunehmen, werden weder WWW- noch spezielle Bioinformatik-Kenntnisse vorausgesetzt. Es wird der Umgang mit Werkzeugen, die für die Sequenzanalyse benutzt werden, an praktischen Beispielen geübt. Das Hauptaugenmerk wird dabei auf das Benutzen der Werkzeuge gelenkt, nicht auf ihre Algorithmen. In den letzten beiden der insgesamt neun Kapitel zeigen wir an wenig untersuchten Sequenzen, wie man bei der Sequenzanalyse vorgeht. Dabei wenden wir alle bis dahin gelernten Sequenzanalyse-Werkzeuge an.*

Sequence Analysis with Distributed Resources: A WWW-Based Course

The process of sequence analysis starts after a few hundred bases are sequenced. Questions arise like: Are there any homologous sequences? Can functions be derived from homologies? Are there any references available on homologous sequences? These questions can be answered using the tools available on the Internet via the World Wide Web (WWW). For example, while sitting in their lab in Europe, people can search for homologies in Japan and do a multiple sequence alignment in the US. For molecular biologists, these tools are indispensable for daily sequence analysis work. For teaching sequence analysis with distributed resources, the WWW itself is the natural medium. Our WWW-based course assumes a minimal background in bioinformatics. It does not explain the algorithms behind the tools. Instead, it emphasizes practical tool usage in connection with the biological considerations that guide the process of sequence analysis. 7 of the 9 course chapters introduce various software tools. In the final two chapters all knowledge about sequence analysis acquired in the previous chapters can be applied in a small but realistic sequence analysis project.

1 Bioinformatik-Studium an der Universität Bielefeld

1.1 Bioinformatik im Studiengang Naturwissenschaftliche Informatik

Seit dem Studienjahr 1989/90 gibt es an der Technischen Fakultät der Universtät Bielefeld den Diplomstudiengang *Naturwissenschaftliche Informatik* (NWI) [2]. Die Mehrzahl der Studierenden entscheidet sich darin für die Fachkombination Informatik/Biologie, kurz Bioinformatik genannt. Anders als in einem Informatik-Studium mit Nebenfach stehen hier Informatik und Biologie als zwei gleichberechtigte Hauptfächer nebeneinander. Anders als bei einem Aufbau- oder Doppelstudium sind diese beiden Fächer von vornherein aufeinander bezogen. Nur so ist es möglich, unter guter Annäherung an die Regelstudienzeit von 9 Semestern eine wissenschaftlich aktuelle und ausreichende Qualifikation zu vermitteln. Es erfolgt recht früh eine Orientierung auf ein bestimmtes Teilgebiet der Biologie – am häufigsten werden Molekularbiologie, Neurobiologie und Ökologie gewählt – und die Informatik-Ausbildung im Hauptstudi-

[1] http://www.techfak.uni-bielefeld.de/techfak/persons/chrisb/biocourse/welcome.html

um konzentriert sich auf diejenigen Techniken und Gebiete der Informatik, die für die gewählte Studienrichtung relevant sind. Die Diplomarbeit liegt in der Regel im Überschneidungsbereich der beiden Fächer. Eine kleine Auswahl von Diplomarbeits-Themen von 1995 und 1996 mag eine Vorstellung davon geben, was man als Absolvent dieses Studiengangs gelernt hat:

- „Ein adaptives neuronales Modell zum Bewegungssehen bei der Erdkröte,"
- „Informationsspeicherung durch synaptische Modulation",
- „Interaktive Vermessung von Fluß- und Teichgebieten aus Grundbuchkarten zur ökologischen Auswertung",
- „Automatische Detektion von Wurzelsystemen in Endoskopie-Bildern",
- „Korrespondenz zwischen Aminosäuresequenz und geometrischen Segmenten biomolekularer Oberflächen",
- „GeneFisher – ein Programm zum Design degenerierter PCR-Primer",
- „Aufbau einer Datenbank mit Restriktionsmustern [...] und Entwicklung eines Algorithmus zur Identifizierung von Bakterienstämmen".

Insgesamt gibt es heute (Juni 1996) etwa 70 Absolventen aus diesem Studiengang, davon etwas über die Hälfte in der Fachkombination Informatik/Biologie. Durch das Aufleben der molekularen Bioinformatik weltweit durch die großen Genom-Projekte und die auch in Deutschland einsetzende Förderung hat dieser Studiengang besondere Aktualität erhalten.

1.2 Entwicklung der Lehre in der Bioinformatik

Einer der Grundsätze im Selbstverständnis des Studiengangs NWI ist, daß die Lehrveranstaltungen der beiden Fächer – hier Informatik und Biologie – nicht beziehungslos nebeneinander stehen dürfen. Dies beginnt bei den Anwendungsbeispielen in den klassischen Lehrgebieten der Informatik. Bei uns beschreiben kontextfreie Grammatiken nicht nur Programmiersprachen, sondern auch Sekundärstrukturen von RNA-Sequenzen, endliche Automaten erkennen (auch) Promotorabschnitte in der DNA, und semantische Netze aus der Bild- und Sprachverarbeitung dienen (auch) zur Strukturerkennung auf molekularen Oberflächen. Weiterhin werden Vorlesungen entwickelt, die von vornherein interdisziplinär definiert sind – „Algorithmen zur Sequenzanalyse" oder „Regelungstechnik biotechnischer Prozesse". Solche Durchdringung in der Lehre ist nur möglich auf der Basis ausgeprägter Kooperation in der Forschung. Es gibt oft gemeinsame Seminare zwischen Informatik, Biologie und (Bio-)Mathematik, und schließlich gibt es Praktika, die von Studierenden der Diplom-Biologie und der Naturwissenschaftlichen Informatik gemeinsam besucht und die von Biologen und Informatikern gemeinsam betreut werden. Zu den letzteren gehört ein Sequenzierpraktikum und der Kurs über **Sequenzanalyse mit verteilten Ressourcen**, der in diesem Beitrag beschrieben wird (Bild 1).

Die Molekularbiologie, und insbesondere die Sequenzanalyse, arbeitet heute netzbasiert. Immer seltener werden Softwarewerkzeuge lokal installiert. Ihre Nutzung über Internet und WWW ist die Regel nicht nur für Anfragen an die großen DNA-, Protein- und Strukturdatenbanken. Für Genetiker, Biomediziner, Molekularbiologen und natürlich auch Bioinformatiker ist die Nutzung verteilter Ressourcen und das Verständnis der ihnen zugrundeliegenden Konzepte ein wesentlicher Bestandteil des täglichen Geschäfts.

Nun ist es fast zwingend, diese Übereinstimmung von Form und Inhalt zu erkennen und die Lehrveranstaltungen zum Erlernen dieser Techniken selbst netzbasiert durchzuführen. Unser erster Schritt war ein „Virtual Course on Biocomputing"[2], der mit einem internationalen Teilnehmerkreis auf dem Internet durchgeführt wurde [5]. Darin ging es um die mathematischen und algorithmischen Grundlagen der Werkzeuge zur Sequenzanalyse. Das Pendant dazu, ein *hands-on*-Praktikum zur Sequenzanalyse mit verteilten Ressourcen, wurde im Anschluß an und basierend auf Materialien und Erfahrungen aus dem virtuellen Kurs von Biologen und Informatikern gemeinsam entwickelt.

[2] http://www.techfak.uni-bielefeld.de/bcd/welcome.html

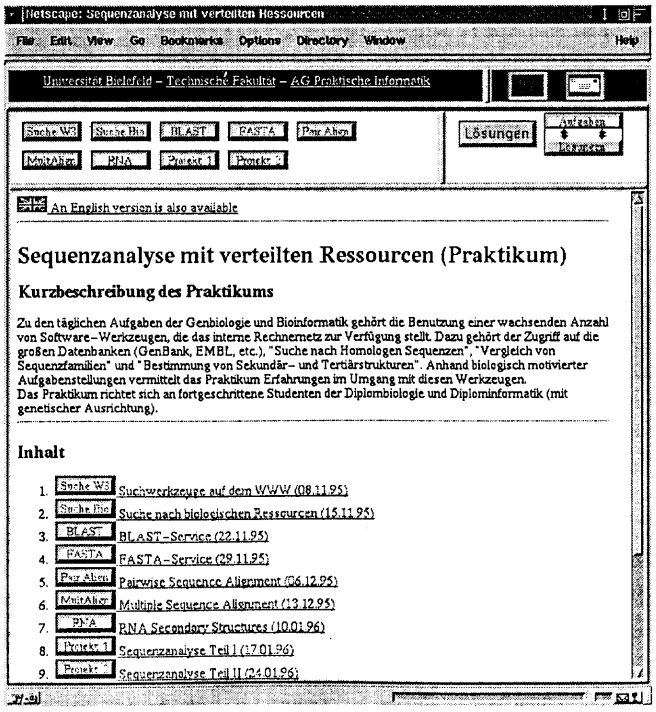


Bild 1: Die Homepage unseres Kurses *Sequenzanalyse mit verteilten Ressourcen;* er ist in 9 Abschnitte unterteilt. Zur besseren Übersicht wird das Netscape-Feature *frames* verwendet (siehe Text).

2 Wann beginnt die Sequenzanalyse?

Die Sequenzanalyse beginnt direkt nach dem Sequenzieren der ersten 300 bis 400 Basen. Es wird ein Vergleich dieser kurzen Abschnitte mit allen bisher in eine Datenbank (GenBank, EMBL) eingetragenen Sequenzen über das WWW durchgeführt. So möchte man feststellen, ob der Sequenzieransatz die gewünschte DNA enthält. Falls bei der Datenbankanfrage unerwartete Ergebnisse zurückgeliefert werden, kann die Fortsetzung der Sequenzierung gestoppt werden. Somit kann sich ein einfacher Homologievergleich auf dem WWW für eine kostenintensive Sequenzierung kostensenkend bemerkbar machen. Nachdem der richtige Sequenzieranszatz gefunden ist, fährt man mit der Sequenzierung fort, bis die Sequenz vollständig ermittelt ist, wobei auch hier immer wieder zwischendurch Homologievergleiche durchgeführt werden. Durch diese Vergleiche werden sehr oft sogenannte unsichere Stellen in einer Sequenz identifiziert, die dann überprüft und ggf. korrigiert werden können. Für diesen wesentlichen Bestandteil der Sequenzierung gibt es einige hervorragende Dienste auf dem World Wide Web; einen davon werden wir in Abschnitt 4 ausführlich erläutern. Nach Fertigstellung der Sequenz und deren Vergleich mit den Einträgen einer Nucleinsäurendatenbank ist es erwünscht, die dazugehörige Proteinsequenz an eine Proteindatenbank zu schicken, wobei das nur funktioniert, wenn die DNA-Sequenz vorher in eine Proteinsequenz übersetzt wurde. Auch dafür stehen Programme auf dem WWW zur Verfügung. Vielleicht möchte man danach ein Alignment zwischen zwei zurückgelieferten ähnlichen DNA- oder Protein-Sequenzen durchführen? Solche Werkzeuge – *pairwise sequence alignment tools* genannt – sind ebenso übers Web erreichbar, wie *multiple sequence alignment tools*, mit denen mehr als zwei Sequenzen auf Ähnlichkeiten bzw. Unterschiede untersucht werden können.

Bei einzelsträngiger RNA möchte man oft die Sekundärstruktur vorhergesagt bekommen. Man denke dabei an die tRNAs, die längeren rRNAs oder die snRNAs, die wahrscheinlich alle nur ihrer Struktur ihre Aufgaben zu verdanken haben. Wir werden ein auf dem WWW verfügbares Programm hierzu in Abschnitt 4.2 vorstellen.

Man könnte noch andere Werkzeuge auf dem Web aufführen, die zur Sequenzanalyse benutzt werden. Da gibt es Programme zur Vorhersage von Proteinfaltungen, Programme zur Erstellung von phylogenetischen Bäumen, Programme zum Auffinden von geeigneten Primern für die Polymerasekettenreaktion [3], etc. Eine Auflistung der meisten Werkzeuge findet man beispielsweise auf der WWW-Seite: *Pedro's BioMolecular Research Tools*[3].

Anhand der genannten Beispiele wird die Relevanz von verteilten Ressourcen auf dem WWW für Molekularbiologen deutlich. Dementsprechend hat die Bioinformatik durch den Einsatz von WWW-angepaßten Werkzeugen in den letzten 10 Jahren an Bedeutung gewonnen. Für die Sequenzanalyse ergeben sich dadurch genauso große Vorteile wie für andere Bereiche der Biologie, z.B. für die Phylogenetik[4], die Botanik[5] oder die Ökologie[6]. Man frage sich nur, wie man heute ohne Internetanbindung oder die auf dem Web verfügbaren Werkzeuge auf die riesigen Datenmengen der verschiedenen Bereiche zugreifen bzw. durchsuchen sollte?

Wenn schon Molekularbiologen und Bioinfomatiker einen großen Teil ihrer Arbeit an oder mit Werkzeugen im WWW verbringen, so liegt es nahe, auch Lehrveranstaltungen netzbasiert durchzuführen, insbesondere dann, wenn es um ein Praktikum zum Erlernen dieser Arbeitstechniken geht. Aus diesem Grund haben wir, gemeinsam mit Kollegen aus der Bielefelder AG Genetik, den hier beschriebenen Kurs entwickelt.

3 Aufbau des Kurses

Der Kurs **Sequenzanalyse mit verteilten Ressourcen**[7] richtet sich an Molekularbiologen, die die oben genannten Werkzeuge noch nicht oder nicht gut genug kennen, und an Informatiker, die sich für die Molekularbiologie interessieren. Grundkenntnisse der jeweils anderen Disziplin sollten von beiden Seiten mitgebracht werden, wobei genetische Grundkenntnisse wichtiger sind als informatische.

Es hat sich bei der Durchführung des ersten Kurses im letzten Wintersemester bewährt, daß die Studierenden Zweiergruppen mit je einem Biologen und einem Informatiker bilden. Unbedingt sollte jemand den Kurs betreuen, der schon reichlich Erfahrung mit dem Auswerten von Datenbankanfragen hat, wie im allgemeinen ein erfahrener Molekularbiologe. Die Zusammenarbeit zwischen Informatikern und Biologen ist also fast obligatorisch.

Die Gliederung des Kurses beruht auf dem in Abschnitt 2 genannten Ablauf der Sequenzanalyse. Bevor wir allerdings im Kurs auf die eigentliche Sequenzanalyse eingehen, beginnen wir mit einer allgemeinen Einführung in das WWW und in Suchwerkzeuge, mit denen man sowohl etwas über Autohersteller, Computerfirmen, Verlage, Städte, usw., als auch speziell über biologische Ressourcen (Membranen, Mutation, Centriolen, Baumarten, phylogenetische Bäume, etc.) finden kann. Danach gehen wir dazu über, uns mit einer der häufigsten Arbeiten eines Molekularbiologen am Computer zu beschäftigen: die Suche nach Homologien. Für diese schon in Abschnitt 2 erwähnte Tätigkeit gibt es zwei Methoden, die auf den BLAST- [1] bzw. FASTA-Algorithmus [4] zurückgehen. Dann

[3] http://www.biophys.uni-duesseldorf.de/bionet/research_tools.html

[4] http://www.ucmp.berkeley.edu/subway/phylogen.html

[5] http://herb.biol.uregina.ca/liu/bio/idb.html

[6] http://www.econet.apc.org/econet/

[7] http://www.techfak.uni-bielefeld.de/techfak/persons/chrisb/biocourse/welcome.html

folgen die angesprochenen *pairwise sequence alignment tools* und *multiple sequence alignment tools*. Um Strukturvorhersagen zu treffen, haben wir uns entschieden, *mfold* [6] zu studieren. Schließlich wenden wir all das Gelernte auf Sequenzen an, die nur wenig untersucht worden sind. Dabei handelt es sich um Sequenzabschnitte, die Mitarbeiter der Bielefelder AG Genetik sequenziert und uns freundlicherweise zur Verfügung gestellt haben.

An der Technischen Fakultät findet der Kurs einmal jährlich statt und ist ungefähr auf 11 Wochen begrenzt. Dieses läßt einen Spielraum sowohl im Winter- als auch im Sommersemester zu, um ggf. zusätzliches Kursmaterial einzubinden oder Abschnitte über zwei Wochen zu behandeln. Als Zeitaufwand für den Betreuer sind 3 Semesterwochenstunden anzusetzen, die je nach Studentenanzahl und Semesterlänge variieren können.

Bild 1 zeigt die Startseite[7] unseres Kurses. Wir haben uns entschieden, den Kurs mit dem Netscape-Feature *frame*[8] zu gestalten. Das hat den großen Vorteil, daß man auf bestimmte Bereiche im gewünschten Kontext immer zugreifen kann, entsprechend einer Toolbar, wie man sie aus vielen Programmen – auch Netscape hat eine Toolbar – kennt. Leider hat das auch einen Nachteil: der Kurs ist somit nur mit einer Netscape-Version ab 2.0[9] durchführbar. Wir hoffen allerdings, daß zukünftig auch andere Browser dieses Feature unterstützen, um eine vollständige Kompatibilität des Kursmaterials zu gewährleisten.

Das linke obere Fenster der Kurs-Homepage enthält drei Hyperlinks, welche zu den Startseiten der Universität Bielefeld, der Technischen Fakultät und zu unserer Arbeitsgruppe Praktische Informatik führen. Durch die Icons im Fenster rechts oben gelangt man entweder zum Hilfebildschirm des Kurses oder man kann uns eine Mail schicken. Diese beiden Fenster sind während des gesamten Kurses anwählbar.

Direkt darunter befinden sich zwei weitere Fenster. Das linke davon (Aufgaben-Fenster) beinhaltet alle *Icons*, über die man die einzelnen Übungsabschnitte erreichen kann, die dazugehörigen Aufgaben werden dann in dem größten Fenster (Hauptfenster) gezeigt. Das Fenster rechts neben dem Aufgaben-Fenster enthält zwei *Icons*, über die man die **Lösungen**-Seite oder **Aufgaben und Lösungen**-Seite erreichen kann. Wenn das **Lösungen**-*Icon* auswählt wird, werden die *Icons* der Aufgaben durch die der Lösungen ersetzt und im Hauptfenster erscheint die Einleitungsseite der Lösungen. Möchte man Aufgaben und Lösungen parallel betrachten, sollte man das *Icon* **Aufgaben und Lösungen** anklicken (siehe Bild 2); das Hauptfenster wird dann geteilt, die Aufgaben im oberen und die Lösungen im unteren Bereich des Hauptfensters dargestellt. Der Inhalt des Aufgaben-Fensters wird ebenfalls ersetzt. Klickt man darin auf ein *Icon*, werden die Inhalte der beiden Fenster im Hauptfenster durch die jeweiligen Aufgaben bzw. Lösungen ersetzt. Nur so ist eine parallele Betrachtung von Aufgaben und Lösungen möglich.

4 Technische Umsetzung

Die Werkzeuge, die wir im Kurs besprechen, sind meistens intuitiv zu bedienen, aber nicht ganz so einfach zu verstehen. Daher üben wir, mit den Werkzeugen umzugehen, und versuchen ebenso verständlich zu machen, wie die gelieferten Ergebnisse zu interpretieren sind. Um einen kleinen Einblick zu geben, beschreiben wir im folgenden zwei der im Kurs behandelten Kapitel. Zum einen ist das der Abschnitt über den BLAST-Service, der in Japan zur Verfügung gestellt wird, um z.B. selbstsequenzierte Sequenzen mit allen bekannten Sequenzen auf Homologien hin zu

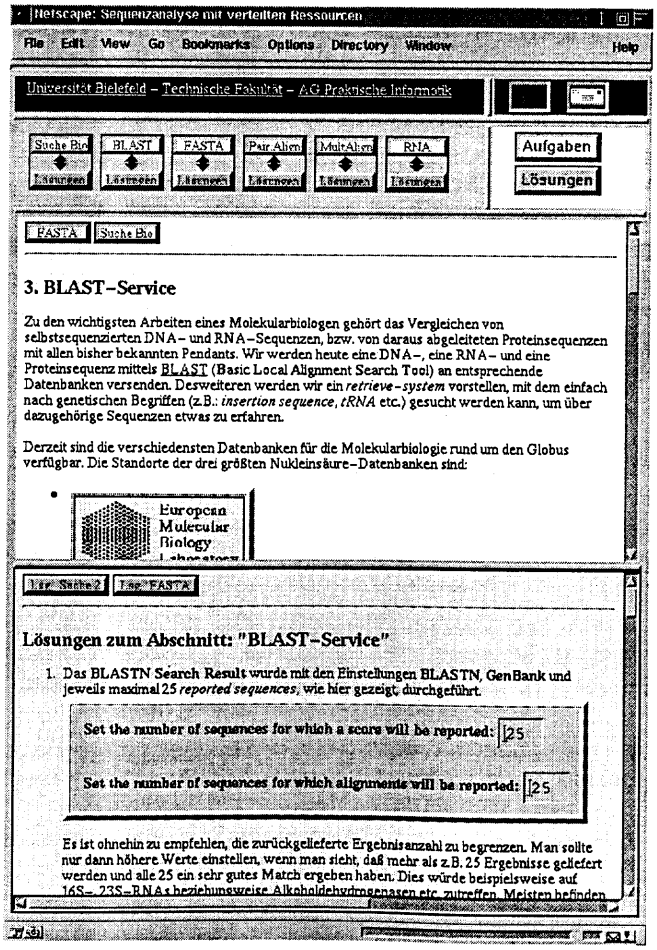


Bild 2: Die *Aufgaben und Lösungen*-Seite des dritten Kurstages. Wenn man einen der sechs Icons in dem Aufgaben-Fenster anklickt, werden die dazugehörigen Aufgaben ins obere und die entsprechenden Lösungen ins untere Fenster geladen.

[8] http://www.netscape.com/assist/net_sites/frames.html

[9] http://home.netscape.com/cgi-bin/123.cgi

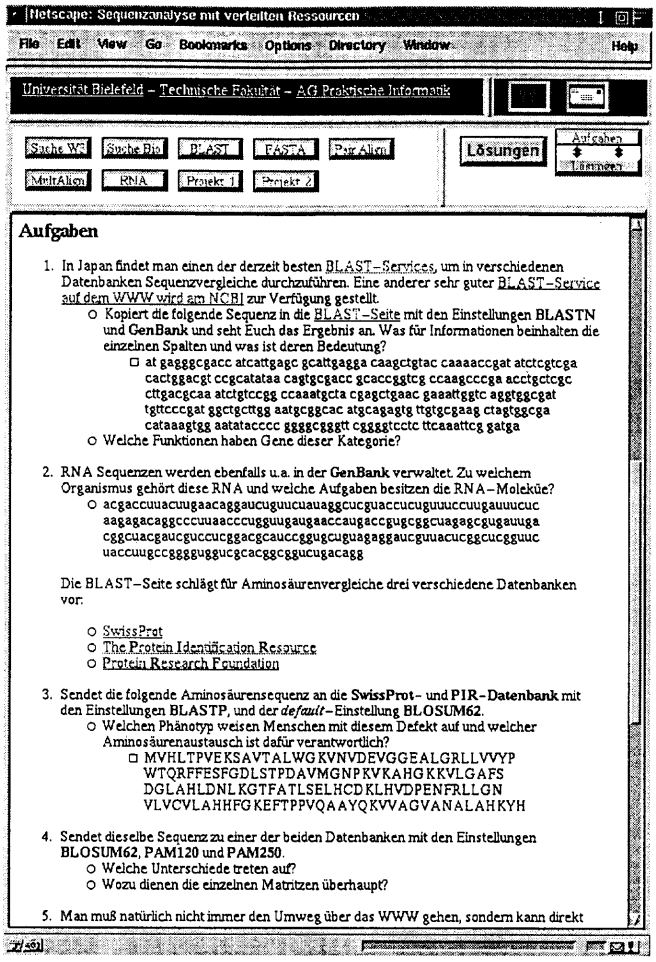


Bild 3: Ein Ausschnitt der Aufgaben des dritten Kurstages. Es wird der BLAST-Service behandelt, wobei verschiedene Sequenzarten auf Homologien hin überprüft werden.

überprüfen. Zum anderen werden wir uns den Abschnitt zur Sekundärstrukturvorhersage von RNA-Sequenzen genauer ansehen.

Da wir hier nicht davon ausgehen wollen, daß unsere Leser selbst Molekularbiologen sind, werden wir die Funktion der behandelten Werkzeuge in aller Kürze miterklären.

4.1 Übungen mit einem BLAST-Service

In dem dritten Abschnitt unseres Kurses (siehe Bild 3) behandeln wir primär den eben erwähnten BLAST-Service[10], denn zu den wichtigsten Arbeiten eines Molekularbiologen gehört das Vergleichen von DNA- und RNA-Sequenzen bzw. die daraus resultierenden Proteinsequenzen mit allen bisher bekannten Pendants. Für diese fast tägliche Aufgabe bietet der BLAST-Service eine komfortable Möglichkeit, wobei verschiedene Einstellungen, welche einfach per Mausklick auszuwählen sind, vom Benutzer vorgenommen werden müssen, bevor ein Homologievergleich durchgeführt werden kann. Auf der BLAST-Seite muß zunächst bestimmt werden, welches Programm auf die Sequenz angewendet werden soll; es gibt vier Möglichkeiten: **BLASTP**, **BLASTN**, **TBLASTN** oder **BLASTX**. Beispielsweise kann man mit **BLASTP** Proteinsequenzen mit den Einträgen verschiedener Proteindatenbanken vergleichen oder mit **TBLASTX** kann eine DNA-Sequenz mit den Sequenzen einer Proteindatenbank verglichen werden. Hierbei wird die DNA-Sequenz zunächst in ihre sechs möglichen Leseraster übersetzt, dann werden alle sechs daraus resultierenden Proteinsequenzen mit der ausgewählten Datenbank verglichen. Weiter muß eine dem Programm entsprechende Protein- oder Nukleinsäuren-Datenbank gewählt werden. Darüberhinaus kann bei Proteinsequenzen zwischen verschiedenen *scoring matrices* gewählt werden. Eine *scoring matrix* gibt ein Maß für den Abstand zweier Aminosäuresequenzen an. Zusätzlich kann ein Filter bei Proteinsequenzen angegeben werden, der es ermöglicht, eine durch eine bestimmte Maske veränderte Anfrage-Sequenz zu untersuchen und gekennzeichnete Bereiche entsprechend zu berücksichtigen. Häufig verwendet man einen Filter, um Proteinsequenzen, die lange saure, basische oder prolinreiche Regionen besitzen, zu kennzeichnen, denn gerade diese Sequenzabschnitte liefern häufig BLAST-Ergebnisse zurück, die nicht von Interesse sind. Nachdem die Haupteinstellungen vorgenommen worden sind, empfiehlt es sich, die Ergebnis-Anzahl der zurückgelieferten Sequenzen von 500 auf 50 zu reduzieren, was meistens einen Geschwindigkeitsgewinn bedeutet. In mindestens 80 Prozent aller Homologievergleiche findet man nämlich unter den ersten 50 Einträgen die für die Auswertung Gewünschten. Schließlich muß noch die Sequenz in das dafür vorgesehene Kästchen kopiert werden und durch einen Mausklick auf **Exec** die Anfrage gestartet werden. Optional kann man einen Kommentar zur Sequenz in eine andere Box eintragen.

Das Benutzen dieser WWW-Seite ist sehr intuitiv, doch benutzt man sehr oft die vorgegebenen Einstellung und macht sich keine Gedanken darüber, was verschiedene Parametereinstellungen bewirken können. Selten wird eine andere als die *default scoring matrix* verwendet und der Filter sogar nie.

Wie können wir nun dazu beitragen, daß die BLAST-Seite nicht nur oberflächlich benutzt wird, sondern ihre zahlreichen Möglichkeiten auch ausgenutzt werden?

Im Kurs wird erklärt, wo die Unterschiede bei der Benutzung von den verschiedenen *scoring matrices* liegen, wie man den Filter einsetzen kann usw. Ebenso wird eine Auswertung der Ergebnisse vorgenommen, die von den ver-

[10] http://www.genome.ad.jp/SIT/BLAST.html

schiedenen Programmen geliefert werden. Kann man beispielsweise ein Ergebnis von dem eben beschriebenen BLAST-Service sofort verstehen (vergl. Bild 4)? Wir machen deutlich, wie die verschiedenen Werte zu interpretieren sind. Wie berechnet man den **High Score**? Was ist **P(N)**? Was **N**? Was sind **Positives**? ...

Zusätzlich zum BLAST-Service wird in der dritten Lehreinheit ein *retrival-service*[11] vorgestellt, mit dem man in Datenbanken nach *Accession-numbers*, wie **M38616** und **K01557** oder bestimmten Begriffen, wie *insertion sequences oder human tRNAs* suchen kann.

4.2 Übungen zur Sekundärstrukturbestimmung mit *mfold*

Es soll noch ein weiterer Abschnitt (siehe Bild 5) aus unserem Kurs vorgestellt werden. Es handelt sich dabei um das Programm *mfold* von Michael Zuker[12] [6]. Wie schon oben angesprochen, sind Sekundärstrukturen von RNA-Strängen deshalb von Interesse, weil ihre Funktion auf ihrer Struktur beruht. Zumindest alle strukturellen RNAs – vielleicht auch einige kodierende RNAs (mRNAs) – gehen aufgrund ihrer Struktur Interaktionen mit anderen Molekülen, wie Proteinen, ein.

Das bedeutet, daß bestimmte Sequenzabschnitte zu Stukturelementen geformt werden, die eine große Bedeutung bei einer Interaktion spielen. Es gibt verschiedene Stukturelemente, die durch *mfold* vorhergesagt und dargestellt werden können: *stacking regions, bulges, hairpin loops, interior loops, multi loops* und *dangling ends* (siehe Bild 6).

Wenn eine RNA-Sequenz mit *mfold* gefalten werden soll, ruft man zunächst die RNA-Faltungsseite[13] auf, und kopiert die Sequenz in das entsprechende Eingabefeld. Darüberhinaus muß ein Name für die Sequenz eingegeben werden, sonst wird man kurze Zeit später dazu aufgefordert. Danach sollte noch eine der Umgebungstemperatur der RNA-Sequenz ungefähr entsprechende Temperatur gewählt werden. Die Temperatur spielt eine Rolle, denn je höher sie gewählt wird, desto geringer ist die Anzahl möglicher stabiler Strukturen. Dies hängt mit der Destabili-

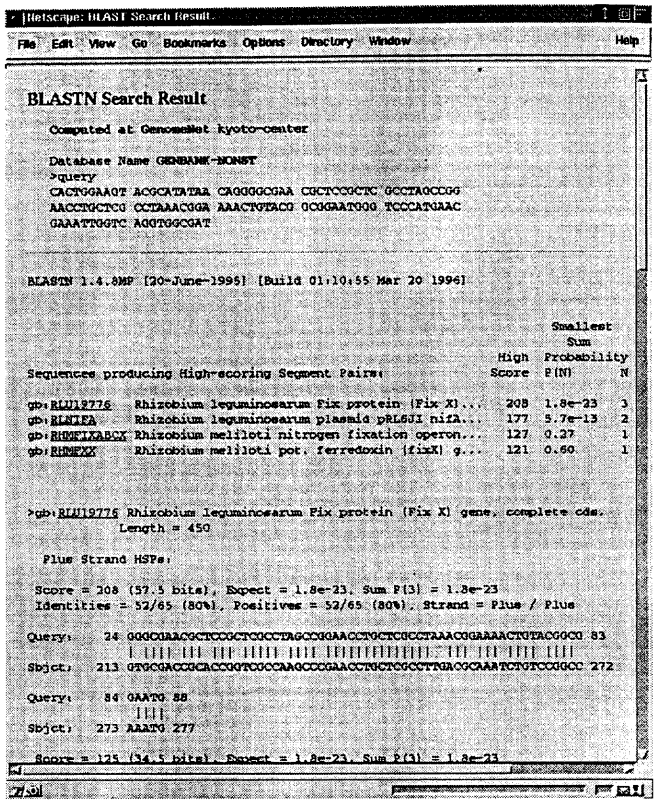


Bild 4: Hier wird ein Auszug einer zurückgelieferten Datenbankanfrage gezeigt. Wie berechnet man den *High Score*, was ist P(N), was N, was sind Positives ... Die Bedeutung dieser und anderer Kürzel des BLAST-Ergebnisses werden im Kurs erläutert.

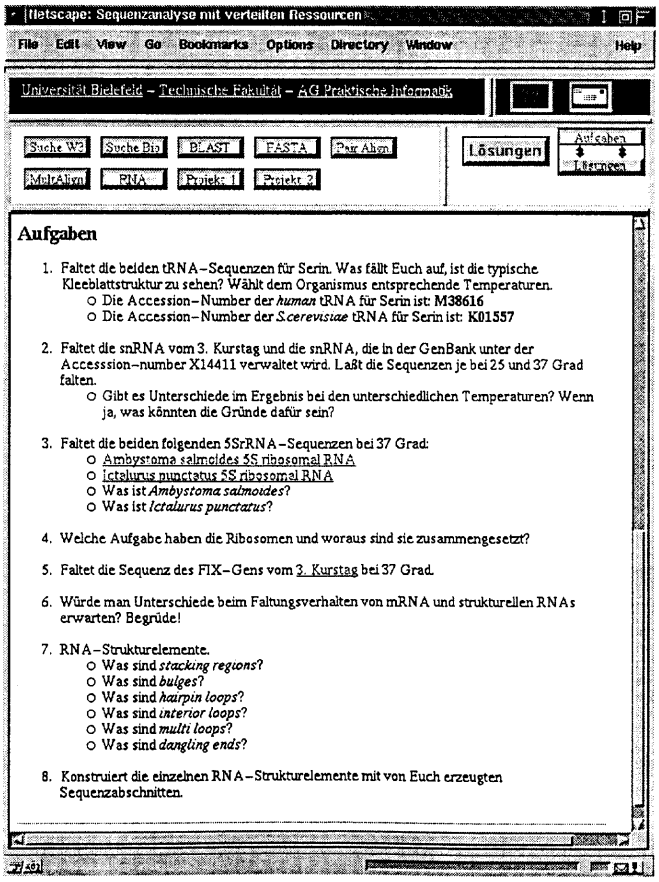


Bild 5: Ein Ausschnitt der Aufgaben des siebten Kurstages. Wir berechnen mit *mfold* Sekundärstrukturen und diskutieren die Ergebnisse.

[11] http://www.ebi.ac.uk/queries/queriesex.html

[12] http://ibc.wustl.edu/~zuker/.cgi

[13] http://ibc.wustl.edu/~zuker/rna/form1.cgi

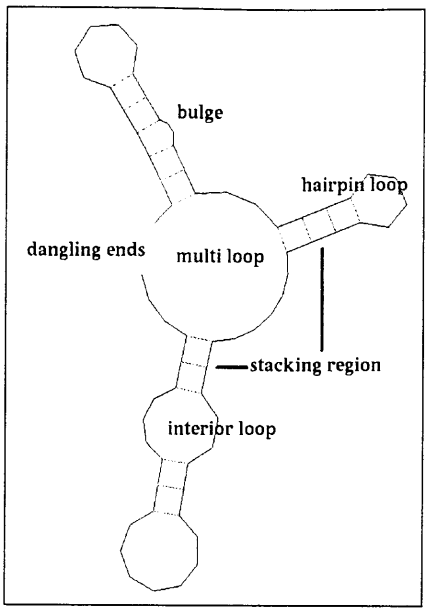


Bild 6: Eine mit *mfold* berechnete Sekundärstruktur kann verschiedene Strukturformen annehmen. Die verschiedenen Formen sind hier abgebildet.

sierungsenergie einiger Strukturelemente zusammen (*bulges, hairpin loops, interior loops, multi loops*, siehe Bild 6). Man kann sich gut vorstellen, daß Energie aufgewendet werden muß, um z.B. einen RNA-Strang zu einem *hairpin loop* zu formen, und je höher die Umgebungstemperatur ist, desto höher ist die destabilisierende Energie solcher Strukturelemente, bzw. desto geringer wirken die stabilisierenden Energien der *stacking regions*. Zusätzlich kann noch eingestellt werden, inwieweit eine suboptimale Struktur von der optimalen prozentual abweichen darf, damit sie in der Ergebnismenge vorkommt. Dazu sollte man wissen, daß bei diesem Programm davon ausgegangen wird, daß die Struktur einer Sequenz als optimal bezeichnet wird, bei der die meiste freie Energie entsteht. Energie wird frei, wenn stabile Strukturen, wie *stacking regions*, gebildet werden oder allgemein, wenn Watson-Crick Basenpaare (A-U, G-C), sowie bei RNA G-U-Paare sich zu einer stabilen Struktur formen. In unserem Kurs lernt man aber auch, daß diese Annahme nur mangels besserer Kenntnisse gemacht wird und häufig nicht ganz zutrifft.

Zuletzt kann noch die Anzahl der zurückgelieferten Ergebnisse verändert werden, wobei der Default-Wert 50 gut gewählt ist. Durch Klicken auf **Send data for processing** wird der Job gestartet.

Sequenzen der Länge >500 können nicht mehr *online* gefaltet werden. Dazu muß **A batch** ausgewählt und danach noch die eigene Email-Adresse eingegeben werden. Kurz darauf wird eine Mail zurückgeschickt, die die Adresse der WWW-Seite mit den Lösungen enthält.

Was haben wir uns zur Aufgabe in dieser Lehreinheit gesetzt? Ein Hauptaugenmerk muß zunächst auf die einzelnen oben genannten Strukturelemente gelegt werden. Wovon hängt es ab, ob Strukturen stabil oder instabil sind? Womit sollte man rechnen, wenn man eine falsche Temperatur wählt und warum kommt beispielsweise nicht die berühmte Kleeblattstruktur heraus, wenn man eine tRNA faltet? Antworten auf solche Fragen sollen dem Benutzer zum besseren Verständnis über die Arbeitsweise des Werkzeuges verhelfen, was dann auch das Interpretieren der Ergebnisse wesentlich erleichtert, womit wir schon beim nächsten Punkt angelangt wären. In Bild 7 ist die Ergebnis-Seite zu sehen, die geliefert wird, wenn man die Sequenz: „AA CCCA..“ zum Falten geschickt hat. Die Schwierigkeit hierbei ist, daß man wissen muß, was *different structure file formats* sind und wie man sie lesen muß, was ein *energy dot plot* ist, und so weiter. Eine Vielzahl von weiteren Fragen treten bei der Benutzung des Werkzeuges und bei der Ergebnisinterpretation auf, diese werden im Kurs besprochen und diskutiert.

5 Perspektiven

Der Web-basierte Kurs **Sequenzanalyse mit verteilten Ressourcen** wird regelmäßig in Bielefeld an der Technischen Fakultät für die Studiengänge Naturwissenschaftliche Informatik und Diplom-Biologie angeboten, wobei der nächste Kurs im Wintersemester 1996 stattfindet. Der Kurs wird noch in einigen Punkten verändert und ergänzt werden. Für die Lehre in Deutschland ist es allerdings

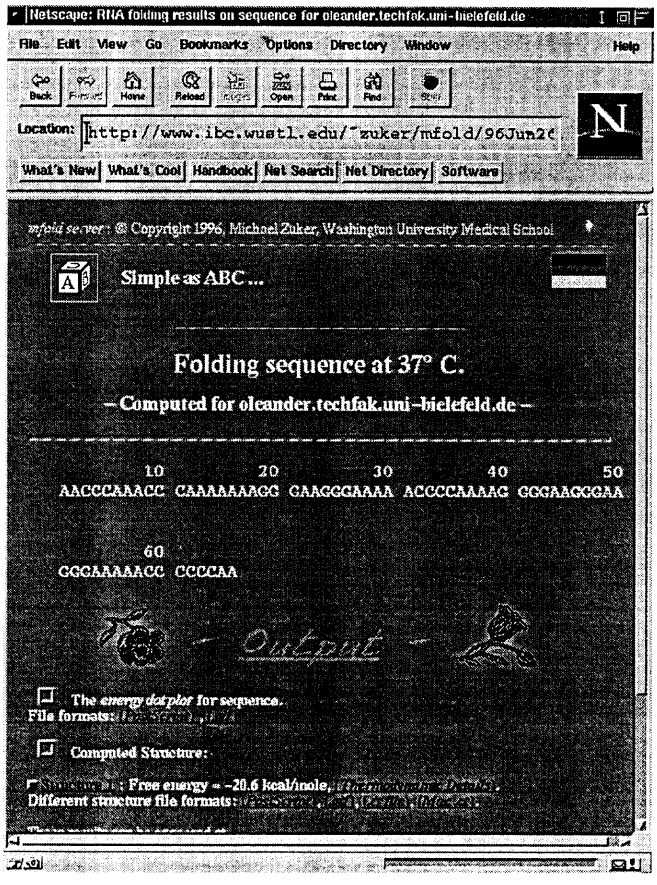


Bild 7: Das gelieferte Ergebnis von *mfold* mit der Sequenz „AACCCAAACC CAAAAAAAGG GAAGGGAAAA ACCCCAAAAG GGGAAGGGAA GGGAAAAACC CCCCA“. Was sind *different structure file formats*, wie muß man sie lesen, was ist ein *energy dot plot*, und so fort. Solche Fragen werden im Kurs erläutert.

wichtig, daß der Kurs aufgrund der Semesterlänge 11 Wochen nicht überschreitet. Selbstverständlich stellen wir das Kursmaterial im deutschsprachigen Raum zur Verfügung. Wir würden uns freuen, wenn der Kurs mit dem gleichen oder leicht verändertem Inhalt an anderen Orten Deutschlands ebenfalls abgehalten wird. Lehrmatrial, das so untrennbar mit Verweisen auf verteilte Ressourcen verknüpft ist, bedarf ständiger Pflege, um konsistent und aktuell zu bleiben.

Ferner haben wir eine Version des Kurses in englischer Sprache erstellt und suchen zur Zeit nach Partnern, die diese gemeinsam mit uns weiterentwickeln. Auch eine Version in spanischer Sprache ist in Aussicht.

Literatur

[1] *Altschul, S. F., Gish, W., Miller, W., Lipman, D. J.*: Basic Local alignment Search Tool. In: J. Mol. Biol., 215 (1990), pp. 403–410.

[2] *Crus, H., Giegerich, R., Hinze, J., Schepper, W.*: Der Studiengang Naturwissenschaftliche Informatik an der Universität Bielefeld. Technische Fakultät, Universität Bielefeld, 1992.

[3] *Giegerich, R., Meyer, F., Schleiermacher, C.*: GeneFisher – Software Support for the Detection of Postulated Genes. In: Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISBM), 1966.

[4] *Pearson, W. R., Lipman, D. J.*: Improved Tool for Biological Sequence Comparison. In: Biochemistry, 85 (1988), pp. 2444–2448.

[5] *De La Vega, F. M., Giegerich, R., Fuellen, G.*: Distance Education Through the Internet: The GNA-VSNS Biocomputing Course. In: Proceedings of Pacific Symposium on Biocomputing, Hawaii, 1966.

[6] *Zuker, M.*: Prediction of RNA Secondary Structure by Energy Minimization. In: Humana Press Inc., (1994), pp. 267–294.

Dipl.-Inform. Christian Büschking, Prof. Dr. Robert Giegerich
Technische Fakultät, Universität Bielefeld,
Postfach 100131, 33501 Bielefeld,
Email: {giegerich;chrisb}@techfak.uni-bielefeld.de

Nachrichten aus der Informationstechnischen Gesellschaft und der Gesellschaft für Informatik

Das Thema: Mobile Softwareagenten

Das Konzept des Softwareagenten, das bereits schon in den 80er Jahren eingeführt wurde, hat in den vergangenen Jahren in verschiedenen Bereichen der Informatik steigendes Interesse gefunden. Abhängig vom jeweiligen Gebiet werden mit dem Begriff „Agent" ganz unterschiedliche Eigenschaften und Funktionalitäten verbunden: Das Spektrum reicht hier von adaptiven Benutzerschnittstellen, über mobile Objekte bis hin zu „intelligenten Prozessen", die zur Erfüllung ihrer Aufgaben mit anderen Prozessen kooperieren. Angesichts dieses Spektrums läßt sich eine allgemeine und zugleich aussagekräftige Definition für den Begriff „Softwareagent" nicht angeben, weshalb wir uns auf die Beschreibung von Symptomen agentenbasierter Systeme beschränken.

Betrachtet man heutige Agentenansätze, so lassen sich die folgenden Symptome identifizieren:

Intelligenz: Dieses Merkmal bezieht sich primär auf die Methoden, die zur Entwicklung eines Agenten herangezogen werden. Als „Intelligente Agenten" werden üblicherweise die Agenten bezeichnet, die mit Techniken der Künstlichen Intelligenz realisiert werden.

Kooperationsfähigkeit: Zur Erfüllung ihrer Aufgabe sind Agenten in der Lage, mit anderen Agenten zu kooperieren. Hier reicht das Spektrum von Client/Server-artigen Interaktionen bis hin zu komplexen Verhandlungsprotokollen, wie sie im Bereich der Verteilten Künstlichen Intelligenz erforscht werden.

Autonomie: Ein Agent kann seine Aufgabe völlig entkoppelt von seinem Nutzer und anderen Agenten ausführen. Er agiert dabei völlig autonom und kontaktiert seinen Auftraggeber erst dann wieder, wenn die Aufgabe erfüllt ist.

Mobilität: Zwei Arten von Mobilität können unterschieden werden, nämlich Fernausführung und Migration. Im Falle der Fernausführung kann ein Agent (Programm) auf einen entfernten Rechner transferiert und dort zur Ausführung gebracht werden. Auf diesem Rechner bleibt er bis zu seiner Terminierung. Im Falle der Migration kann ein Agent während seiner Ausführung mehrfach von System zu System wandern. Bei jedem Migrationsschritt muß der Zustand des Agenten auf das Zielsystem transferiert werden.

Agentenbasierte Systeme lassen sich grob in drei Klassen unterteilen: Multiagentensysteme, persönliche Agenten und mobile Agenten. Multiagentensysteme (MAS) haben ihren Ursprung in der Verteilten Künstlichen Intelligenz. In MAS steht die verteilte Koordination des Verhaltens einer Menge autonomer intelligenter Agenten im Vordergrund: Agenten verhandeln miteinander mittels höherer Protokolle, wie etwa KIF und KQML, und stimmen dadurch ihr Wissen, ihre Ziele, Fähigkeiten und Pläne ab. MAS können beispielsweise Planungs- und Steuerungsaufgaben übernehmen, etwa im Bereich der Logistik oder Robotik. Während Intelligenz, Kooperationsfähigkeit und Autonomie in Multiagentensystemen sehr ausgeprägte Eigenschaften darstellen, spielt die Mobilität in MAS bisher keine Rolle.

Persönliche Agenten (PA) sind Agenten, die Nutzer in ihrer täglichen Arbeit unterstützen. PAs realisieren in der Regel adaptive Schnittstellen, die in der Lage sind, sich im Laufe der Zeit auf die Gewohnheiten und Präferenzen ihrer Nutzer einzustellen. Obwohl PAs auch komplexe Aufträge ausführen können, steht die Interaktion zwischen Nutzer und Agent im Vordergrund der Forschungsbemühungen. Die Anwendungsmöglichkeiten von PAs sind vielfältig: Informationsbeschaffung und -filterung, Mail-Management, Terminplanung etc. Bei PAs spielen Mobilität sowie Kooperationsfähigkeit (zwischen Agenten) eine eher untergeordnete Rolle, während die Eigenschaften Intelligenz und Autonomie in der Regel sehr ausgeprägt sind.

Bei mobilen Agenten spielen neben der Mobilität die Eigenschaften Autonomie und Kooperationsfähigkeit eine wichtige Rolle. Zur Realisierung mobiler Agenten stehen heute objektorientierte Programmiersprachen, wie etwa Telescript und Java, oder Skriptsprachen, wie etwa Tcl und Perl, zur Verfügung. Zur Unterstützung der Kooperation zwischen Agenten werden Kommunikationsmechanismen, wie etwa RPC, Tuple Spaces oder Message Queues angeboten; höhere Verhandlungsprotokolle werden bisher systemseitig nicht unterstützt. Das Hauptanwendungsfeld, das im Zusammenhang mit mobilen Agenten genannt wird, ist „Electronic Commerce". Die hinter diesem Begriff stehende Vision ist die eines elektronischen Marktes, in dem Agenten die Käufer und Verkäufer von Waren und Dienstleistungen sind. Mobile Agentensysteme bieten eine Um-

gebung, in der ein Händler das Äquivalent eines Einkaufladens zur Verfügung gestellt bekommt, in dem die Agenten von Benutzern Geschäfte tätigen können.

Technologie

Im Bereich der mobilen Agentensystem gibt es mehrere kommerziell vertriebene Systeme (z.B. CyberAgents, Telescript u.a.). Wir werden Telescript als den erfolgreichsten Vertreter näher betrachten. Außerdem gibt es verschiedene Programmiersprachen, die durch verfügbare Erweiterungen als Agententechnologie verwendbar erscheinen. Alle werden entweder interpretiert (Skriptsprachen), oder in eine Zwischendarstellung überführt, die dann von einer virtuellen Maschine ausgeführt wird. Beide Ansätze bieten im Gegensatz zu in Maschinencode übersetzten Programmen die Möglichkeit, jeden Befehl vor der Ausführung zu überprüfen. Im Bereich der prozeduralen Sprachen sind dies Perl und Tcl/Tk mit den Varianten SafeTcl (das inzwischen in Tcl integriert worden ist) und AgentTcl. Im Bereich der objektorientierten Sprachen dominiert Java durch seine extrem hohe Akzeptanz im Bereich des Internet und des WWW. Auf Tcl und Java werden wir näher eingehen.

Telescript wurde von General Magic entwickelt, finanziell unterstützt durch verschiedene große Firmen im Bereich von Telekommunikation und Informationsverarbeitung (u.a. Apple, AT&T, France Telecom, Fujitsu, Matsushita, Motorola, Philips, Sony und Toshiba), die die Hoffnung haben, daß die von General Magic zur Verfügung gestellte Technologie eine Hinwendung zur netzwerkorientierten Verarbeitung nach sich ziehen würde.

Die Technologie gründet sich auf der Idee der Mobilität der Agenten, von General Magic das „Remote Programming Paradigm" genannt. Die damit verbundenen Konzepte von Plätzen, Verbindungen und Meetings folgen daraus. Plätze können verschachtelt sein, das heißt, innerhalb eines Platzes, der eine Bank repräsentiert, können durchaus verschiedene Plätze für Kredite, Immobilien und Kontoführung existieren. Agenten können miteinander auf zwei Arten kommunizieren, entweder durch Verbindungen (entfernt) oder mit Hilfe von Meetings (lokal).

General Magics neuester Ansatz stellt das Produkt „Tabriz Agentware and Agenttools" dar. Es wird momentan kostenfrei zur Verfügung gestellt, und umfaßt die Telescripttechnologie, WWW-Server, Schnittstellen zum World Wide Web und eine Entwicklungsumgebung. Tcl („Tool Command Language") ist eine einfache, leicht erweiterbare Skriptsprache zur Einbettung in Applikationen, die, zusammen mit ihrer ersten Erweiterung, dem Tk-Toolkit, ursprünglich ein System für die Entwicklung graphischer Benutzerschnittstellen darstellte. Nachdem Tcl/Tk 1990 der Öffentlichkeit vorgestellt wurde, avancierte es schnell zu einem der meistbenutzten Werkzeuge zur Entwicklung von Benutzerschnittstellen. Aufgrund der Tatsache, daß Tcl/Tk inzwischen für eine Vielzahl von Plattformen (u.a. Windows) verfügbar ist, stellt es außerdem eine Plattform für die Entwicklung von Anwendungen dar, die in einem heterogenen Umfeld entwickelt werden.

Aufgrund der Einfachheit bietet Tcl nur einen einzigen Datentyp, die Zeichenkette, an. Der Vorteil ist eine einfache Integrierbarkeit in Applikationen. Dieser Aspekt ist aber inzwischen weitgehend uninteressant, da die Überwindung der Heterogenität als der Hauptvorteil gesehen wird. Und hier stellt das Nichtvorhandensein weiterer Datentypen einen schwerwiegenden Nachteil für Programmierer dar, die größere Applikationen entwickeln wollen. In diesem Zusammenhang ist auch die sehr gewöhnungsbedürftige Syntax von Tcl hinderlich. Ein weiterer Nachteil ist die Tatsache, daß Tcl eine prozedurale Sprache ist. Objektorientierte Sprachen scheinen im Gegensatz zu prozeduralen Sprachen die Problemdomäne, die mit Agenten bearbeitet wird, besser abbilden zu können.

SafeTcl, eine Erweiterung von Tcl, wurde ursprünglich von Nathaniel Borenstein zur Unterstützung von „Active Mail" entwickelt. „Active Mail" sind Nachrichten, die außer normalem Text auch Programmtext enthalten, der beim Empfänger ausgeführt werden kann. Diese Idee wurde im Prinzip bei der WWW-Integration von Java durch Sun wieder aufgegriffen.

Inzwischen wird auch ein Plugin (ein Modul) für den Netscape-Browser, den meistverbreiteten WWW-Browser, angeboten, das es ermöglichen soll, in direkter Konkurrenz zu Java Tcl-Programme in WWW-Dokumente einzubetten.

Java wurde im März 1995 als eine Sprache für aktive Webseiten vorgestellt. Diese aktiven Webseiten können Javaprogramme, sogenannte Applets, enthalten, die dann auf der Maschine des Betrachters ablaufen. Hiermit verwandelte sich das bis dahin statische WWW in eine hochdynamische Anwendung, die einen ersten Schritt in Richtung netzwerkorientierte Anwendungen und Agentensysteme darstellte. Die meisten großen Software- und Hardwarehersteller lizenzierten bereits 1995 die Technologie und portierten Java auf ihre Plattformen. So ist Java inzwischen auf fast allen Unix-Varianten (inklusive Linux), Windows 95 und Windows NT, OS/2 und anderen verfügbar. Jeder Hersteller von Browsern, Microsoft eingeschlossen, baut die Möglichkeit, Java interpretieren zu können, in seine Browser ein, und trägt damit zum Erfolg von Java bei. Java soll unter anderem auch eine Standardsprache für „Stored Procedures" bei Datenbanken werden, eine Anwendung, an die sicher noch vor einem Jahr niemand gedacht hätte. Man kann Java schon heute als den kommenden Standard für verteilte Anwendungen in verschiedensten Bereichen sehen. Weitere Schritte in diese Richtung, neben verschiedenen auf Java basierenden Agentensystemen, stellt zum Beispiel die Entwicklung des RMI- Paketes (Remote Method Invocation) dar. Es bietet die Möglichkeit des entfernten Aufrufes von Objektmethoden. Ist das Ob-

jekt auf der Zielseite des Aufrufes nicht vorhanden, wird es dort erzeugt. Dies stellt nichts anderes dar als die entfernte Ausführung von Programmen.

RMI soll in den Standardumfang von Java aufgenommen werden und bietet damit eine überall verfügbare Möglichkeit, über Heterogenitätsgrenzen hinweg Kommunikation und Fernausführung von Javaprogrammen transparent zu ermöglichen.

Telescript als Technologie ist zwar schon einige Zeit verfügbar, der kommerzielle Durchbruch ist aber bisher nicht gelungen. Vielleicht wird sich das neue Produkt Tabriz, das als direkter Konkurrent zu Java entworfen scheint, durchsetzen können. Während Tcl/Tk im Bereich der Benutzerschnittstellen unbestritten seinen wohlverdienten Platz hat, wird es sich im Bereich der Agentensysteme aufgrund der oben skizzierten Nachteile wohl nicht durchsetzen.

Die hohe Qualität von Java und der Erfolg der Sprache im Bereich des WWW unterstützen Entwicklungen im Gebiet der mobilen Agenten hingegen sehr stark. Die mit RMI zur Verfügung stehende Möglichkeit zur Remote Execution ist frei verfügbar, und es gibt bereits verschiedene auf Java aufbauende Agentensysteme. Diese implementieren über die von RMI gebotene Fähigkeit zur Fernausführung von Programmen die Migration von Agenten, eine der Grundvoraussetzungen für den angestrebten Einsatz mobiler Agenten in netzwerkorientierten Anwendungen. Exemplarisch erwähnt seien hier „Java-to-Go" der University of Berkeley und das Agentensystem „Mole" der Universität Stuttgart, die bereits heute einen mit dem von General Magic angebotenen System vergleichbaren Stand erreichen beziehungsweise in einigen Bereichen darüber hinausgehen.

Chancen und Risiken

Nachfolgend werden wir uns auf Chancen und Risiken beschränken, die direkt aus der Mobilität von Agenten erwachsen. Die Vorteile der Agententechnologie kann man insbesondere im Anwendungsbereich „Electronic Commerce" deutlich machen. Voraussetzung für Vitalität und damit Attraktivität eines elektronischen Marktes ist, daß prinzipiell jedermann die angebotenen elektronischen Dienste nutzen sowie Dienste in diesem Markt anbieten kann. Mobile Agenten können hierfür die technischen Grundlagen schaffen.

Sofortige Nutzung von Diensten und aktives Trading: Um in einem elektronischen Markt auf einen Dienst zugreifen zu können, ist auf der Klientenseite ein Client-Programm erforderlich, das den entfernten Dienst beispielsweise einem Endanwender über eine adäquate Benutzerschnittstelle zugänglich macht. Heute werden diese Programme oftmals per Post auf Disketten zugestellt und müssen vom Empfänger vor der Nutzung des Dienstes installiert werden. Selbst wenn das Client-Programm (in Maschinencode) über das Netz heruntergeladen werden kann, bleiben immer noch mehr oder weniger aufwendige Installationsschritte durchzuführen. Außerdem ist das Herunterladen von Maschinencode aus Sicherheitsgründen außerordentlich bedenklich.

Werden Client-Programme als Agenten implementiert und beispielsweise von einem Verzeichnisdienst verwaltet, so hat jeder Nutzer die Möglichkeit, sämtliche im System verfügbaren Dienste sofort und ohne zusätzlichen Managementaufwand zu nutzen. Wählt ein Nutzer mit Hilfe des Verzeichnisdienstes einen Dienst aus, so wird automatisch das zugehörige, als Agent realisierte Client-Programm auf das System des Nutzers transferiert. Der Agent ist ohne weiteres Zutun des Nutzers sofort ablauffähig, und die Ausführung des Client-Programms geschieht in der Agentenumgebung und ist daher sicher. Ein Vorteil aus Sicht des Dienstanbieters ist, daß das Client-Programm plattformunabhängig ist und daher nur einmal entwickelt werden muß.

Ein wichtiges Instrument eines realen Marktes ist die Werbung. Im elektronischen Markt kann mit Hilfe von Agenten ein sogenanntes aktives „Trading" durchgeführt werden, bei dem von den Dienstanbietern „Werbeagenten" zu potentiellen Kunden transferiert werden. Ein solcher Agent kann für den angebotenen Dienst werben, indem er beispielsweise dem potentiellen Nutzer anbietet, das Angebot sofort probeweise zu nutzen. Es ist vorstellbar, daß es für einen Dienst mehrere Typen von Werbeagenten gibt, die jeweils auf ein spezielles Nutzerprofil zugeschnitten sind.

Sofortige Installation von Diensten: Mit der Möglichkeit, Agenten auf Server-Maschinen zu transferieren, kann prinzipiell jedermann auf einfache Weise elektronische Dienste sofort anbieten. Geht man davon aus, daß auf einem Server Basisdienste, wie etwa Datenbank-, Kommunikations- oder OLTP-Dienst verfügbar sind, so können auf dieser Grundlage neue Dienste mit Agententechnologie realisiert werden. Ein neuer, als Agent realisierter Dienst kann sofort installiert werden, indem man den Agenten auf eine Server-Maschine transferiert. Beispielsweise kann auf einem Server, der einen allgemeinen Informationsdienst anbietet, durch Plazierung entsprechender Agenten ein Wetterdienst oder ein Börseninformationsdienst sofort realisiert werden. Unter Ausnutzung eines Datenbank-, Informations- und OLTP-Dienstes könnte ein Agent einen Broker-Service realisieren. Hierbei kann man zwei Arten von Dienstanbietern unterscheiden, nämlich solche, die Server-Infrastruktur und die Basisdienste zur Verfügung stellen, und solche, die auf dieser Grundlage Dienste für die Endnutzer anbieten. Durch die Realisierung in Agententechnologie lassen sich Endnutzerdienste prinzipiell von jedermann sofort und problemlos installieren.

Für beide oben beschriebenen Anwendungen der Agententechnologie ist die Möglichkeit der Fernausführung von Agenten ausreichend. Kommt die Fähigkeit der Migration hinzu, so ergibt sich ein außerordentlich flexibles Verarbeitungsmodell. Agenten können zur

Erledigung ihrer Aufgaben von Knoten zu Knoten migrieren und neue Agenten erzeugen, die völlig autonom die ihnen übertragenen Teilaufgaben ausführen. Agenten können sich bei Bedarf treffen, Informationen austauschen und miteinander kooperieren. Welche Vorteile dieses flexible Verarbeitungsmodell und insbesondere die Möglichkeit der Migration in der Praxis bringen wird, ist heute mangels Erfahrung mit dieser Technologie schwer einzuschätzen. Die Akzeptanz dieser Technologie hängt sicher auch davon ab, was für Sicherheits- und Kontrollmechanismen zukünftige Agentensysteme bereitstellen werden (s. u.).

Die Mobilität von Agenten eröffnet nicht nur interessante Möglichkeiten, sondern birgt auch Risiken in sich. Bedingt durch die Mobilität spielen Fragen der Sicherheit eine zentrale Rolle. Es ergeben sich hierbei zwei Fragestellungen:

1. Wie kann ein System vor böswilligen Agenten geschützt werden? Hier sind Mechanismen gefragt, die sicherstellen, daß ein Agent nur auf die Ressourcen des Systems zugreifen kann, für die er eine Berechtigung hat. Die interpretierte Ausführung von Agenten in einer sicheren Agentenumgebung ist ein erster wichtiger Schutzmechanismus. Zusätzlich sind leistungsfähige Authentifikations- und Autorisierungsmechanismen erforderlich. Es ist zu erwarten, daß man hier weitgehend auf bekannten Verfahren aufbauen kann.

2. Wie kann ein Agent vor böswilligen Agenten und Systemen geschützt werden? Es muß sichergestellt werden, daß das Programm eines Agenten nicht manipuliert werden kann. Außerdem muß verhindert werden, daß sein Zustand nicht böswillig verändert und/oder ausspioniert werden kann. Führt beispielsweise ein Agent elektronisches Geld mit sich und wird er von dem System, das er gerade besucht, ausgeraubt, so entspricht dies einer böswilligen Änderung seines Zustands. Die Sicherheit von Agenten wirft eine Reihe interessanter Probleme auf, zu deren Lösung Forschungsbedarf besteht.

Ein weiterer Problembereich ist das Fehlen von adäquaten Kontrollmechanismen in heutigen Agentensystemen. Für praktische Anwendungen sind Mechanismen zur Terminierung von Agentengruppen sowie zur automatischen Waisenerkennung und -beseitigung erforderlich. Betrachtet man die typischen Anwendungen eines elektronischen Marktes, so wird deutlich, daß eine Integration von transaktionalen Mechanismen geboten ist. Insbesondere muß es möglich sein, den Zustand eines Agenten als Teil einer Transaktion zu verändern.

Zusammenfassend kann man sagen, daß mobile Agenten eine Reihe interessanter Eigenschaften aufweisen, die insbesondere für den Anwendungsbereich „Electronic Commerce" sehr attraktiv sind. Für einen breiten Einsatz der Agententechnologie sind jedoch weitere Forschungs- und Entwicklungsaktivitäten insbesondere in den Bereichen Sicherheit und Kontrollstrukturen erforderlich.

Prof. Rothermel

GI/ITG-Fachgruppe FG 3.3.1: Kommunikation und Verteilte Systeme

Der Fachausschuß 3.3.1 „Kommunikation und Verteilte Systeme" ist ein gemeinsamer Fachausschuß der GI und ITG und versteht sich als anwendungs- und grundlagenorientierte Plattform von Forschern, Herstellern und Anwendern im Bereich der Kommunikationssysteme und -anwendungen. Dabei stehen folgende Themenbereiche im Vordergrund:

- „Architektur Verteilter Systeme" mit den Arbeitsbereichen Kommunikationsmodelle, Entwurf und Spezifikation, Analyse und Bewertung.
- „Netze" mit den Arbeitsbereichen private und öffentliche Netze, Hochgeschwindigkeitsnetze und Breitbandkommunikation.
- „Kommunikationsdienste und -anwendungen" mit den Schwerpunkten Telematikdienste, Multimediakommunikation, Mehrwertdienste sowie Netzwerkmanagement und -sicherheit.

Um den Informationsaustausch zwischen den Fachgruppenmitgliedern anzuregen und zu pflegen, werden von der Fachgruppe folgende Aktivitäten durchgeführt:

- Mitherausgabe der Fachzeitschrift PIK – Praxis der Informationsverarbeitung und Kommunikation (Saur Verlag, München) als Organ der Fachgruppe. Jedes Fachgruppenmitglied erhält die viermal jährlich erscheinende Fachzeitschrift.
- Konferenz KIVS'xy – Kommunikation in Verteilten Systemen. Diese von der Fachgruppe organisierte größte deutsche Kommunikationskonferenz findet alle zwei Jahre an wechselnden Orten statt. Die nächste KIVS'97 findet vom 17. – 22. Februar 1997 in Braunschweig statt. Infos über: http://ibr.cs.tu-bs.de/ KiVS/
- häufige Workshops und Arbeitsgespräche zu den oben genannten generelleren Themenbereichen und zu ausgewählten Spezialthemen und Technologietrends.

Die Fachgruppe 3.3.1 wird von einem 10köpfigen Leitungsgremium geführt. Es setzt sich aus jeweils drei benannten Mitgliedern der GI und der ITG und aus vier von der Mitgliederversammlung gewählten Mitgliedern zusammen. Derzeit gehören dem Leitungsgremium an:

B. Butscher, GMD-FOKUS Berlin, Prof. Effelsberg, Uni Mannheim, Fr. Gerner, Siemens München, Dr. Holler, KfK Karlsruhe, Prof. Kühn, Universität Stuttgart, Prof. Raubold, Deutsche Telekom TZ Darmstadt, Prof. Spaniol, RWTH Aachen, Dr. Stöttgen, IBM-ENC Heidelberg (kommissarisch), Prof. Swoboda, TU München, Fr. Prof. Zitterbart, TU Braunschweig.

Zu den 10 Leitungsmitgliedern sind noch für jeweils 4 Jahre weitere 10 Fachleute berufen, die jeweils ein aktuelles Arbeitsgebiet vertreten. Zusammen bilden sie das erweiterte Leitungsgremium der Fachgruppe. Sprecher ist seit 1991 B. Butscher aus Berlin, stellvertretender Sprecher W. Effelsberg.

Die Fachgruppe 3.3.1 umfaßt derzeit ca. 1100 Mitglieder, der Jahresbeitrag (incl. Fachzeitschrift) beträgt 55,- DM.

Mitteilungen aus den Gesellschaften

Gründung des GI-Fachausschusses 4.3/1.4 „Robotersysteme"

Am 20. Juni 1996 haben die beiden GI-Fachgruppen 1.0.3 und 4.0.1 einen Fachausschuß gegründet. Dieser Fachausschuß besteht gemeinsam im FB1 (1.4) und FB4 (4.3) unter dem Namen „Robotersysteme"; die Federführung liegt im FB4. Mit der kommissarischen Leitung des neuen Fachausschusses wurden beauftragt: Prof. Dr. R. Dillmann (Sprecher), Dr. T. Lüth (stellv. Sprecher) und Dr. E. Prassler (Schatzmeister). Die früheren Fachgruppen 1.0.3 und 4.0.1 werden sich Ende des Jahres neu gruppieren. Der Vollzug der Umstrukturierung in Form von Neuwahlen ist im Rahmen der AMS '96 in München geplant.

Weitere Informationen: Dr. E. Prassler, FAW Ulm, (0731) 501-621, Email: prassler@faw.uni-ulm.de

Studien- und Forschungsführer Robotik

Der Aufruf zur Mitwirkung an dem Studien- und Forschungsführer „Robotik", der über die GEMROB Mailing List und die GEMROB Homepage verbreitet wurde, ist auf sehr große Resonanz gestoßen. Bisher liegen 60 Interessensbekundungen vor. Die erste Redaktionssitzung fand am 15. Juli 1996 in Karlsruhe statt. Bei diesem Treffen wurden erste Vorschläge für die Struktur des Führers und für die Form der Einzelbeiträge diskutiert. Als Termin für die Erscheinung des Führers wird das Frühjahr 1997 angestrebt.

Weitere Informationen: Dr. E. Prassler, FAW Ulm, (0731) 501-621, Email: prassler@faw.uni-ulm.de

Das CAD-Referenzmodell auf dem Wege zum Anwender

Das vom FB4 der GI initiierte und forcierte Verbundprojekt CAD-Referenzmodell zwischen acht Forschungsinstituten aus den Bereichen Informatik, Maschinenbau und Arbeitswissenschaft sowie fünf Firmen aus unterschiedlichen Branchen hat die Halbzeit der Entwicklung überschritten. Die Ergebnisse gehen jetzt zügig in die Anwendungsumgebung der beteiligten Industrien.

Neben verbesserten Ablauforganisationen sind es in erster Linie neue Werkzeuge und Methoden für die Konstruktion wie Aufbau und Einführung von Produktmodellen, das Arbeiten mit Lösungskatalogen und die Einbindung ins weltweite Internet als Engineering-Netzwerk. Weitere Aktionen sind eine intensive Nutzung und Einbindung von Werkzeugen der Wissensverarbeitung, die Prozeßintegration von Elektronik und Mechanik, ein PPS-getriebener Konstruktionsablauf und simultanes Arbeiten von unterschiedlichen Firmenbereichen entlang einer Auftragsabwicklung.

Die Forschungsinstitute liefern dazu eine Reihe von innovativen Basisdiensten und Vorschläge für eine neue CAD-Architektur. Die beteiligten CAD-Anbieterfirmen werden die Ergebnisse in ihr Produktangebot einbinden und damit die Marktchancen ihrer CAD-Programme erhöhen. Das Projekt wird bis Ende 1997 laufen und einen gesamten Entwicklungsaufwand von 25 Mio DM erreichen.

Prof. Dr. Olaf Abeln, Projektleiter des CAD-Referenzmodells

VHDL-A – Kurze Informationen zum aktuellen Stand der Standardisierungsbemühungen

VHDL, der IEEE-Standard 1076, wurde als Beschreibungssprache für digitale Hardware entworfen und hat sich durchgesetzt. Fast alle Anbieter von Logiksimulatoren lassen in VHDL geschriebene Modelle zu. Seit langem besteht die Notwendigkeit, analoge und digitale elektrische Modelle in einem Modell zu simulieren. Um dies zu ermöglichen, soll VHDL-A, die Analog Extensions von VHDL, baldmöglichst standardisiert werden.

Während der Jahrestagung des GI-FA 4.5 im September 1995 in Wien wurde in einer Sitzung über eine mögliche Nutzung von VHDL-A zur Modellierung allgemeiner analoger, auch nichtelektrischer Systeme, diskutiert. In der Zeitschrift Eurosim SNE Number 15, November 1995, wurde über die Ergebnisse dieser Sitzung zusammenfassend berichtet.

Alle hier erwähnten aktuellen Informationen, wie das Language Reference Manual, Modelle, Packages, Protokolle der Sitzungen etc. können auch weiterhin auf dem VHDL-A ftp-server: nestor@epfl.ch (128.178.50.20) in dem Unterverzeichnis pub/vhdl/standards/ieee eingesehen und abgeholt werden.

Dr. Ingrid Bausch-Gall

Neue Studienrichtung „Medizinische Informatik" an der Universität Leipzig

An der Universität Leipzig wird ab dem kommenden Wintersemester 1996/97 im Diplomstudiengang Informatik die Studienrichtung „Medizinische Informatik" angeboten. Die zehnsemestrige Ausbildung wird gemeinsam von Instituten der Fakultät für Mathematik und Informatik und der Medizinischen Fakultät getragen.

Die Absolventen dieses interdisziplinär angelegten Studiums sollen in die Lage versetzt werden,

die Methoden der Informatik angemessen für Problemstellungen der Medizin und Biowissenschaften einzusetzen und weiterzuentwikkeln. Die Studienrichtung umfaßt daher eine profunde Informatikausbildung im Grund- und Hauptstudium. Außerdem werden im sechssemestrigen Hauptstudium Themenkreise wie Krankenhaus-Informationssysteme, medizinische Bild- und Signalverarbeitung, Modelle biologischer und kognitiver Systeme, Wissensverarbeitung und deren Anwendung in der Medizin im Rahmen eines Vertiefungsstudiums behandelt. Hinzu kommt ein Nebenfach Medizin, das im Grund- und Hauptstudium belegt werden muß und Grundlagen der Biophysik, Physiologie und Biometrie zum Inhalt hat. Im Hauptstudium ist ein mindestens viermonatiges Berufspraktikum in einer medizinorientierten Einrichtung (z.B. Krankenhaus, Forschungsinstitut, medizinische Verwaltung, pharmazeutische Industrie) abzuleisten.

Dieses für die neuen Bundesländer einzigartige Studienangebot – ähnliche Ausbildungsgänge werden derzeit nur an drei deutschen Universitäten angeboten – bietet gute Berufschancen für Informatikabsolventen. Der Einsatz von Computern und computergestützten Verfahren und Arbeitsweisen in medizinischen Einrichtungen wird in Zukunft auch auf der Basis von gesetzlichen Forderungen stark zunehmen, so daß die Arbeitsmarktchancen von Informatikern mit der Studienrichtung „Medizinische Informatik", die in der Diplomurkunde ausgewiesen wird, als gut eingeschätzt werden dürfen.

Erstmals können sich Studenten zum Wintersemester bis 15.09.1996 für das erste Studienjahr einschreiben. Studienortwechsler der Fachrichtung Informatik können auch in das 2. Studienjahr aufgenommen werden.

Weitere Informationen: Prof. Dr. S. Gerber, Institut für Informatik, Universität Leipzig, Augustusplatz 10, 04103 Leipzig, Tel.: 0341/97-32102, http://www.informatik.uni-leipzig.de

Berichte

GI/ITG-Workshop „Custom Computing" (CCM'96), 19.–21.6.1996

Der vierte Workshop der Fachgruppe „Architekturen für hochintegrierte Schaltungen" (GI-FG 3.1.1 / ITG-FG 4.1.1) vom 19. bis 21. Juni 1996 im IBFI Schloß Dagstuhl hatte als Schwerpunktthema „Custom Computing Machines (CCM)", d.h. Hardware-Architekturen, die durch Strukturprogrammierung an spezielle Anwendungen anpaßbar sind. Diese auch unter dem Namen „Application Specific Instruction Set Processors (ASIP)" bekannten Architekturen bestehen meist aus einem traditionellen Mikroprozessor und anwenderprogrammierbarer Logik (FPL). Aus 32 eingereichten Beiträgen wurden 17 ausgewählt, deren thematisches Spektrum weit gespannt war. Die ausgewählten Beiträge wurden ergänzt durch eingeladene Vorträge von Prof. Hartenstein (Universität Kaiserslautern) über „Szenen und Krisen des High-Performance Computing" und von Toni Gore (OMIMO Brüssel) über „Application Specific Processor Customisation" durch Nutzung der Baustein-Bibliotheken der europäischen Open Microprocessor Initiative (OMI).

Auf besonderes Interesse stießen verschiedene Ansätze zur Unterstützung des Entwurfs und zur Automatisierung der Implementierung von CCMs (insbesondere von H. Högl, Mannheim, M. Wannenmacher, Hagen, F. Mayer-Lindenberg, Hamburg-Harburg, und von M. Weinhardt, Karlsruhe), Berichte über Anwendungen von CCMs im Consumerbereich (T. Friedrich, Philips), zur Simulation Neuronaler Netze (S. Jones, Loughborough) und Zellularer Automaten (T. Hahn et al., Darmstadt), Entwurfsmethodiken zur Optimierung von Befehlssätzen (G. Markwardt, Dresden, M. Schutti, Linz) und ein Übersichtsvortrag mit einer systematischen Klassifizierung verschiedener Arten anwendungsspezifischer Hardware (J. Becker, Kaiserslautern). Ein für die Anwender immer wichtiger werdendes Thema wurde allerdings kaum behandelt, nämlich die Qualitätssicherung bzw. Methoden für den Nachweis der Korrektheit von CCMs.

Der von D. Monjau herausgegebene Tagungsband ist als Chemnitzer Informatik Bericht CSR-96-05 erschienen (Fakultät für Informatik, TU Chemnitz-Zwickau, Straße der Nationen 62, 09111 Chemnitz).

Veranstaltungen

Workshop Beherrschbarkeit informationsverarbeitender Systeme, 30.–31.10.1996 in München

Der Workshop wird vom Fachbereich 3 der GI veranstaltet. Ziel des Workshops soll sein, zwischen Teilnehmern aus der Industrie, den Hochschulen und den Forschungseinrichtungen das Profil und die Struktur eines Forschungsprogramms zu diskutieren, das der Fachbereich 3 in Zukunft initiieren und längerfristig weiterverfolgen möchte. Grundlagenarbeiten und industrielle Anwendungen der Ergebnisse sollen deshalb mit der Absicht der gegenseitigen Anregung im Vordergrund stehen.

Zum Thema:

Die fatale Abhängigkeit unserer Gesellschaft von informationsverarbeitenden Systemen und die steigende Tendenz dieser Abhängigkeit machen es dringend erforderlich, darüber nachzudenken, wie Konstruktion, Herstellung und Betrieb solcher hochkomplexer verteilter Systeme besser beherrscht werden können. Trotz der großen gesellschaftlichen und wirtschaftlichen Bedeutung dieser Systeme

bestehen beispielsweise nach wie vor noch erhebliche Defizite in der Spezifikation, im Entwurf, in der Modellierung sowie in der Verhaltensanalyse solcher großer heterogener Systeme. Die Dynamik des Systemverhaltens ist unter Aspekten wie Zuverlässigkeit, Sicherheit, Leistung oder Datenschutz aus globaler Sicht oft nicht überschaubar und infolgedessen auch nicht so zu regeln, daß ein System beispielsweise bei Fehlfunktionen stets nur sichere Zustände einnimmt und dennoch – falls notwendig – einen weiteren Betrieb ermöglicht.

Die Initiative des FB 3 möchte Arbeiten sowohl in der Forschung als auch in der industriell orientierten Entwicklung in die Wege leiten, die diesen Zustand langfristig verbessern, wenn nicht gar überwinden. Aspekte der Systemzuverlässigkeit und der Systemsicherheit werden dabei in Anbetracht zahlreicher Beispiele von spektakulären Auswirkungen fehlerhafter Großsysteme im Vordergrund stehen müssen.

Um das geplante Forschungsprojekt möglichst effizient und wirkungsvoll zu machen, erscheint zu Beginn eine Konzentration auf bestimmte Einsatzbereiche komplexer informationsverarbeitender Systeme sinnvoll. Gedacht ist hier insbesondere an Verkehrsleitsysteme und Medizinische Informationssysteme.

Willkommen sind Beiträge aus dem Teilnehmerkreis (in Form von ca. 15minütigen Kurzvorträgen), in denen bevorzugt über Probleme, Forschungsbedarf und Erfahrungen aus der Praxis berichtet werden.

Weitere Informationen zu Verkehrsleitsystemen: Prof. Dr. Detlef Schmid, Uni Karlsruhe, Inst. f. Rechnerentwurf und Fehlertoleranz, 76128 Karlsruhe, Tel.: 0721 608-3960, Email: schmid@ira.uka.de

Zu Medizinischen Informationssystemen: Dr. Elmar Holler, Forschungszentrum Karlsruhe, Inst. f. Angewandte Informatik, 76344 Eggenstein-Leopoldshafen, Tel.: 07247 825757, Email: holler@iai.fzk.de

Arbeitstagung HL7

22.–23.10.1996, Göttingen

„Implementierungserfahrungen – Neuere Entwicklungen bei Kommunikationsstandards" ist das Thema der nächsten Arbeitstagung. Am Vortag, dem 21.10.1996, wird ein spezieller Trainingskurs für HL7-Implementierungen angeboten.

Weitere Informationen: Geschäftsstelle der HL7-Benutzergruppe in Deutschland, Heinrich-Buff-Ring 44, 35392 Gießen, Tel. 0641/702-4501, Fax 0641/78788.

GMDS/GI Workshop SoftKis '97

20.–21.2.1997, Universität Dortmund

Der Workshop „Erfolgsfaktor Softwaretechnik für die Entwicklung von Krankenhaus-Informationssystemen (SoftKis '97)" richtet sich an Forscher, Entwickler sowie an EDV-Leiter in Kliniken und einschlägigen Softwareunternehmen. Ziel der Tagung ist es, Theorie und Praxis miteinander ins Gespräch zu bringen.

Weitere Informationen: Dr. W. Hasselbrink, Universität Dortmund, Informatik 10, Baroperstr. 301, Tel. 0231/7554712, Fax 0231/7552061, Email: willi@ls10.informatik.uni-dortmund.de

3. ITG/GI/GMM-Workshop Beschreibungssprachen und Modellierung von Schaltungen und Systemen

27.–28.2.1997 in Holzhau/Erzgebirge

Veranstalter: ITG/GI/GMM-Fachgruppe, TU Chemnitz-Zwikkau

Weitere Informationen: Prof. Dr.-Ing. Dieter Monjau, TU Chemnitz-Zwickau, Fakultät für Informatik, Email: monjau@informatik.tu-chemnitz.de, http://www.informatik. th-darmstadt.de/ISS/deegener/itg.htlm

GI-Fachtagung BTW 97

5.–7.3.1997, Universität Ulm

Die GI-Fachtagung „Datenbanken im Büro, Technik und Wissenschaft" (BTW '97) ist ein Repräsentations- und Diskussionsforum für moderne Anforderungen an Datenbanktechnologien. Die BTW '97 hat sich insbesondere zum Ziel gesetzt, den Kontakt zwischen Anwendern, Herstellern und Forschern zu intensivieren.

Weitere Informationen: P. Dadam, Universität Ulm, Fakultät für Informatik, 89069 Ulm, Tel. 0731/502-4130, Fax 0731/502-4134, Email: Dadam@informatik.uni-ulm.de

4. Internationaler Workshop

Fuzzy-Neuro-Systeme '97 – Computational Intelligence

12.–14.3.1997, GH Paderborn

Veranstalter des Workshops ist der Fachausschuß 1.2 „Inferenzsysteme" der Gesellschaft für Informatik e. V. (GI) in Zusammenarbeit mit dem Forschungsschwerpunkt „Sensorik/Aktorik" des Landes Nordrhein-Westfalen an der Universität-GH Paderborn, Abteilung Soest. Das wissenschaftliche Programm umfaßt eingeladene Vorträge, die Präsentation eingereichter Arbeiten in Form von Vorträgen sowie Poster-Veranstaltungen. Es ist das Ziel, einen Überblick zum gegenwärtigen Stand der Forschung und Entwicklung von Fuzzy-Systemen und Neuronalen Netzen zu geben. Kreative Diskussionen sollen es ermöglichen, eine innovative Brücke zwischen Theorie und Praxis zu schlagen.

Weitere Informationen: Prof. Dr. A. Grauel, Universität-GH Paderborn, Abt. Soest, FB 16, Steingraben 21, D-59494 Soest, Email: fns97@uni-paderborn.de, WWW: http://www.uni-paderborn.de/~fns97/.

8. E.I.S.-Workshop

Entwurf Integrierter Schaltungen (E.I.S.)

8.–9.4.1997 in Hamburg

Veranstalter: GMD, Universität Hamburg, GMM, GI, ITG

Weitere Informationen: Frau I. Reinhardt, GMD-SE, Schloss Birlinghoven, 53754 Sankt Augustin, Tel.: 02241/142873, Fax: 02241/142035, Email: reinhardt@gmd.de, http://set.gmd.de/SET/ws/eis8.htlm

Fachtagung „Krankenhausinformationssysteme"

10.–11.4.1997, Heidelberg

Krankenhausinformationssysteme werden zunehmend auch in mittleren und kleineren Krankenhäusern eingesetzt. Die Arbeitsgruppe „Krankenhausinformationssysteme" hat es sich zum Ziel gesetzt, mit speziellen Arbeitstagungen insbesondere Praktiker aus Krankenhäusern zu erreichen. Eine erste, sehr erfolgreiche Tagung wurde im Mai 1996 in Göttingen durchgeführt.

Weitere Informationen: PD Dr. K. Kuhn, Medizinische Klinik Ulm, Tel. 0731/502-4348, FAX 0731/502-4731, E-Mail: klaus.kuhn@informatik.uni-ulm.de

GI/ITG Workshop: Zielarchitekturen Eingebetteter Systeme

11.9.1997, Rostock

Der Workshop findet im Rahmen der Tagung Architektur von Rechensystemen (ARS'97) statt. Veranstalter ist die gemeinsame Fachgruppe „Architekturen für hochintegrierte Schaltungen" der Gesellschaft für Informatik (GI) und der Informationstechnischen Gesellschaft (ITG).

Weitere Informationen: Prof. Dr. K. Waldschmidt/Dr. B. Klauer, J.W. Goethe-Universität Frankfurt, Technische Informatik, Postfach 111932, 60054 Frankfurt am Main, Tel.: (069) 79828248, Fax: (069) 79822351

Veranstaltungen

Symposien, Kongresse, Workshops, Tagungen

14. bis 18.10.1996 – Tübingen

International Conference on **Theory and Practice of Geometric Modeling**
Veranstalter: Universität Tübingen
ⓘ Prof. Dr.-Ing. W. Straßer, Eberhard-Karls-Universität Tübingen, Fakultät für Informatik, Auf der Morgenstelle 10, C9, 72076 Tübingen, Tel. 07071/296356, Email: strasser@gris.informatik.uni-tuebingen.de

21. bis 22.10.1996 – München

Informatik-Kongreß **Markt und Wettbewerb im Internet**
Veranstalter: GI
ⓘ Gesellschaft für Informatik e.V. (GI), Wissenschaftszentrum, Ahrstr. 45, 53175 Bonn, Email: gi-bonn@gmd.de

21. bis 25.10.1996 – Montreal, Kanada

2nd International Conference on Engineering of Complex Computer Systems **ICECCS**
Veranstalter: IEEE, Technical Segment Committee on ECCS
ⓘ Prof. Dr.-Ing. Bernd Krämer, FernUniversität, 58084 Hagen, Email: bernd.kraemer@fernuni-hagen.de

22. bis 23.10.1996 – München

Tagung **Informationsverarbeitung in der Konstruktion**
Veranstalter: VDE-EKV; VDMA; VDA; GI-FG 4.2.1 (Rechnerunterstütztes Entwerfen und Konstruieren (CAD))
ⓘ VDI-Gesellschaft Entwicklung Konstruktion Vertrieb (VDI-EKV), Postfach 10 11 39, 40002 Düsseldorf, Tel. 0211/6214-271, Fax 0211/6214-575

23. bis 25.10.1996 – Wien, Österreich

15th International Conference on Computer Safety, Reliability and Security **SAFE COMP '96**
Veranstalter: EWICS TC 7, Austrian Research Centre Seibersdorf, Federal Research and Testing Centre Arsenal
ⓘ E. Schoitsch, Austrian Research Centre Seibersdorf, A-2444 Seibersdorf, Tel. +432254 7803117

24. bis 25.10.1996 – Dresden

Workshop **Prozeßorientierte Dokumentation im Betrieblichen Umweltinformationssystem**
Veranstalter: GI-FA 4.6 (Informatik im Umweltschutz), GI-Arbeitskreis „Betriebliche Umweltinformationssysteme" der GI-FG 5.4.2 (Informationssysteme in der Personalwirtschaft), TU Dresden
ⓘ Prof. Dr. Eric Schoop, TU Dresden, Fakultät Wirtschaftswissensschaften, 01062 Dresden, Tel. 0351/463-2198

28.10.1996 – Hamburg

6. Treffen der Arbeitsgruppe „Petrinetze und Informationssysteme in der Praxis" **Simulation, Analyse und Optimierung von Abläufen**
(siehe GI-Mitteilungen Band 19, Heft 3)
Veranstalter: GI-FG 0.0.1 (Petrinetze), GI-FG 2.5.2 (Entwicklungsmethoden für Informationssysteme und deren Anwendung EMISA), INNOBIS GmbH
ⓘ Peter Langner, INNOBIS GmbH, Holsteiner Chaussee 183b, 22457 Hamburg, Tel. 040/5598760, Fax 040/55987699

9.11.1996 – Tübingen

Workshop
Computernetze in Afrika – das Projekt Congo-NIC
Veranstalter: GI-FG 8.2.2 (Informatik und Dritte Welt) und FIFF e.V. (Forum InformatikerInnen für den Frieden und gesellschaftliche Verantwortung e.V.)
ⓘ Hans Rauschmayer, software design & management GmbH & Co. KG, Thomas-Dehler-Str. 27, 81737 München, Tel. 089/63812-248, Email: Hans.Rauschmayer@sdm.de

18. bis 19.11.1996 – Erlangen

Workshop **3D-Bildanalyse und -synthese**
Veranstalter: Graduiertenkolleg „3D-Bildanalyse und -synthese" bei FAU Erlangen-Nürnberg, GI-FG 4.1.2 (Imaging und Visualisierungstechniken)
ⓘ Prof. Dr. Heinrich Müller, Universität Dortmund, Informatik VII, Otto-Hahn-Str. 16, 44221 Dortmund, Tel. 0231/7556324, Email: mueller@ls7.informatik.uni-dortmund.de

28. bis 29.11.1996 – Boppard

Workshop über Realzeitsysteme **PEARL '96**
Veranstalter: GI-FG 4.4.2 (Echtzeitprogrammierung PEARL); GI-FG 4.4.1 (Echtzeitsysteme)
ⓘ Dr. H. Windhauer, Werum GmbH, Erbstorfer Landstr. 14, 21337 Lüneburg, Tel. 04131/8900-66, Email: windauer@werum.de

1. bis 4.12.1996 – Austin, Texas

2nd World Conference on **Integrated Design and Process Technology**
Veranstalter: Society for Design and Process Science
ⓘ Prof. Dr.-Ing. Bernd Krämer, FernUniversität, 58084 Hagen, Email: bernd.kraemer@fernuni-hagen.de

10. bis 12.12.1996 – Frankfurt/Main

2nd Working Conference on **Personal Wireless Communications** (Wireless local Access)
Veranstalter: IFIP TC 6 – Task Group Wireless Communication
ⓘ Prof. Dr. Oswald Drobnik, Fachbereich Informatik (Telematik), Johann Wolfgang Goethe-Universität, D-60054 Frankfurt/Main, Tel. +49-69-79828362, Fax +49-69-798 23340, Email: pcw96@tm.informatik.uni-frankfurt.de, www: http//www.tm.informatik.uni-franfurt.de/pcw96/

27. bis 28.1.1997 – Dortmund

Workshop **Hypermedia-Systeme im Umweltbereich – Konzepte, Organisation und Realisierung**
Veranstalter: Universität Dortmund, Lehrstuhl für Software-Technologie
ⓘ Dr. Stefan Dißmann, Universität Dortmund, Tel. 0231/7552482, Email: hsu97@ls10.informatik.uni-dortmund.de

10. bis 14.2.1997 – Plzen/Prag, Tschechische Republik

5th International Conference on Computer Graphics and Visualization 97 **WSCG '97**
Veranstalter: IFIP WG 5.10
ⓘ Vaclav Skala, Computer Sci.Dept., University of West Bohemia, Univerzitni 22, Box 314, Plzen, Czech Republic, Email: wscg97@kiv.zcu.cz

17. bis 22.2.1997 – Braunschweig

Fachtagung Kommunikation in Verteilten Systemen **KiVS '97**
Veranstalter: GI; ITG; GI/ITG-FG 3.3.1 (Kommunikation und Verteilte Systeme)
ⓘ Professor Dr. M. Zitterbart, KiVS '97, IBR, TU Braunschweig, Bültenweg 74/75, 38106 Braunschweig, Email: zit@ibr.cs.tu-bs.de

20. bis 21.2.1997 -Dortmund

Workshop Erfolgsfaktor Softwaretechnik für die Entwicklung von Krankenhausinformationssystemen **SoftKIS '97**
Veranstalter: GMDS/GI-Arbeitskreis „Methoden und Werkzeuge für das Management von Krankenhausinformationssystemen" der GI-FG 4.7.1 (Medizinische Informatik)

ⓘ Dr. W. Hasselbring, Universität Dortmund, Informatik 10 (Software-Technologie), 44221 Dortmund, Tel. 0231/755-4712, Email: willi@ls10.informatik.uni-dortmund.de

26. bis 28.2.1997 – Kiel

1st International Workshop Cooperative Information Agents – DAI meets Database Systems **CIA-97**
Veranstalter: Universität Kiel, GI-FG 1.1.6 (Verteilte Künstliche Intelligenz), GI-FG 2.5.1 (Datenbanksysteme)
ⓘ Matthias Klusch, Christian-Albrechts-Universität Kiel, Institut für Informatik, Olshausenstr. 40, 24118 Kiel, Tel. 0431/880-4474, Email: mkl@informatik.uni-kiel.d400.de

26.2. bis 1.3.1997 – Weimar

Internationales Kolloquium über Anwendungen der Informatik und der Mathematik in Architektur und Bauwesen **IKM 97**
Veranstalter: Universität Weimar
ⓘ HAB Weimar-Universität-, IKM 97, 99421 Weimar, Tel. 03643/584251, Email: ikm@informatik.hab-weimar.de

27. bis 28.2.1997 – Holzhau

3. Workshop **Beschreibungssprachen und Modellierung von Schaltungen und Systemen**
Veranstalter: GI-FG 3.5.7 (Hardware-Beschreibungssprachen und Modellierungsparadigmen)
ⓘ Prof. Dr.-Ing. Dieter Monjau, TU Chemnitz-Zwickau, Fakultät für Informatik, 09107 Chemnitz, Tel. 0371/531-1467, Email: monjau@informatik.tu-chemnitz.de

27. bis 28.2.1997 – Rostock

5. Workshop **SEUH**
Software Engineering im Unterricht an Hochschulen (siehe GI-Mitteilungen Band 19, Heft 3)
Veranstalter: GI-FG 2.1.1 (Software Engineering), GI-FG 2.1.6 (Requirements Engineering), ACM German Chapter
ⓘ Professor Dr. Peter Forbrig, Universität Rostock, FB Informatik, 18051 Rostock, Email: pforbrig@informatik.uni-rostock.de

3. bis 6.3.1997 – Dresden

Software-Ergonomie '97
(siehe GI-Mitteilungen Band 19, Heft 3)
Veranstalter: GI-FA 2.3 (Ergonomie in der Informatik); German Chapter of ACM; TU Dresden
ⓘ Prof. Dr.-Ing. Liskowsky, TU Dresden, Fak. Informatik, 01062 Dresden, Tel. 0351/4575-389, Email: liskowsk@is2163.inf.tu-dresden.de

5. bis 7.3.1997 – Bad Honnef

XPS-97
Veranstalter: GI-FA 1.5 (Expertensysteme)
ⓘ Dr. Hans Voss, GMD-FIT.KI, Schloß Birlinghoven, 53754 Sankt Augustin, Tel.: 02241/14-2532, Email: hans.voss@gmd.de

5. bis 7.3.1997 – Ulm

Fachtagung Datenbanksysteme in Büro, Technik und Wissenschaft **BTW 97**
(siehe GI-Mitteilungen Band 19, Heft 4)
Universität Zürich, Institut für Informatik, Winterthurer Str. 190, CH-8057 Zürich, Tel. +41-1-2574312, Fax +41-1-3630035, Email: {dittrich,geppert}@ifi.unizh.ch
Veranstalter: GI-FA 2.5 (Rechnergestützte Informationssysteme)
ⓘ Prof. Dr. Peter Dadam, Universität Ulm, Fakultät für Informatik (DBIS), 89069 Ulm, Tel. 0731/502-4130, Email: dadam@informatik.uni-ulm.de

12. bis 14.3.1997 – Soest

4. Internationaler Workshop **Fuzzy-Neuro-Systeme '97**
(siehe GI-Mitteilungen Band 19, Heft 3)
Veranstalter: GI-FA 1.2 (Inferenzsysteme), Universität-GH-Paderborn (Abteilung Soest)
ⓘ Prof. Dr. Adolf Grauel, Universitität-GH-Paderborn, Abteilung Soest FB 16, Steingraben 21, 59494 Soest, Tel. 02921/378162

19. bis 21.3.1997 – Linz, Österreich

Joint Modular Languages Conference '97
Veranstalter: Johannes Kepler Universität Linz, GI, OCG, SI, BCS
ⓘ Prof. Dr. Hanspeter Mössenböck, Johannes Kepler Universität, A-4040 Linz, Tel. +43732/2468-7130, Email: moessenboeck@ssw.uni-linz-ac.at

20. bis 21.3.1997 – Würzburg

13th European Workshop on Computational Geometry **CG '97**
Veranstalter: Universität Würzburg, GI-FG 0.1.2 (Algorithmische Geometrie)
ⓘ Prof. Dr. Hartmut Noltemeier, Univ. Würzburg, Institut f. Informatik I, Am Hubland, 97074 Würzburg, Tel. 0931/8885054, Email: cg97@informatik.uni-wuerzburg.de

9. bis 11.4.1997 – Berlin

3rd International Symposium on Autonomous Decentralized Systems **ISADS 97**
Veranstalter: IEEE-CS, IPSJ, SICE, GI-FA 3.1 (Systemarchitekturen)
ⓘ Barbara Intelmann, GMD-Fokus, Hardenbergplatz 2, 10623 Berlin, Tel. 030/25499-209/200, Email: isads97@fokus.gmd.de

21. bis 24.4.1997 – München

International Conference on Acoustics, Speech and Signal Processing **ICASSP-97**
Veranstalter: IEEE, GI-FG 4.0.3 (Physik, Informatik, Informationstechnik)
ⓘ Prof. Dr. M. Lang, TU München, Lehrstuhl für Mensch-Maschine-Kommunikation, Arcisstr. 21, 80333 München

28.4. bis 2.5.1997 – Whistler Resort, Kanada

International Symposium on Environmental Software Systems 1997 **ISESS 1997**
(siehe GI-Mitteilungen Band 19, Heft 2)
Veranstalter: University of Guelph; Austrian Research Centre Seibersdorf; IFIP WG 5.11; GI-FA 4.6 (Informatik im Umweltschutz)
ⓘ Prof. Dr. Ralf Denzer, Hochschule für Technik und Wirtschaft des Saarlandes, Goebenstr. 40, 66117 Saarbrücken

12. bis 16.5.1997 – San Diego, Ca.

International Symposium on Integrated Network Management **ISINM '97**
Veranstalter: IFIP WG 6.6, IEEE-CS
ⓘ Roberto Saracco, CSELT, Via Reiss Romoli 274, I-10148 Torino, Email: roberto.saracco@cselt.stet.it

24. bis 27.5.1997 – Bonn

Conference on Women, Work, and Computerization **WWC '97**
Termine: beiträge erbeten bis 1.10.96
Veranstalter: GI-FA 8.1.1 (Frauenarbeit und Informatik), IFIP, GMD, Universität Hamburg
ⓘ Doris Köhler, Rechenzentrum der Universität Hamburg, Schlüterstr. 70, 20146 Hamburg, Fax 040/4123-6270, Email: IFIP-WWC97@rrz.uni-hamburg.de

10. bis 12.9.1997 – Kiel

7th International Conference on Computer Analysis of Images and Patterns **CAIP '97**
Termine: Beiträge erbeten bis 1.2.97
Veranstalter: IAPR
ⓘ Dr. Kostas Daniilidis, Christian-Albrechts-Universität Kiel, Inst. f. Informatik, Preusserstr. 1-9, 24105 Kiel, Tel. 0431/560473, Email: caip97@informatik.uni-kiel.de

22. bis 25.9.1997 – Zürich, Schweiz

6th European Software Engineering Conference **ESEC 1997**
Veranstalter: CEPIS Member Societies
ⓘ Prof. Dr. Helmut Schauer, Universität Zürich-Irchel, Institut für Informatik, Winterthurerstr. 190, CH-8057 Zürich

24. bis 26.9.1997 – Aachen

Informatik '97 Jahrestagung der GI: Informatik als Innovations-Motor
Veranstalter: GI
ⓘ Prof. Dr. Matthias Jarke, RWTH Aachen, Informatik V, Ahornstr. 55, 52056 Aachen, Tel. 0241/8021501, Email: gi-97@informatik.rwth-aachen.de

29.9. bis 2.10.1997 – Dortmund

Tagung Hypertext-Information Retrieval-Multimedia 97 **HIM 97**
Veranstalter: Universität Dortmund
ⓘ Prof. Dr. Gisbert Dittrich, Email: dittrich@ls1.informatik.uni-dortmund.de

Keine Gewähr für die Richtigkeit der Angaben.

übrigens ... Globale Herausforderungen in der Informationstechnologie

Die zentrale Herausforderung beim Übergang in ein neues Jahrtausend heißt nachhaltige Entwicklung. Die Erde ist heute bedroht durch eine immer rascher wachsende Weltbevölkerung, den ungebremsten Verbrauch von Ressourcen, die zunehmende Erzeugung von Umweltbelastungen und schließlich die immer raschere Beschleunigung von Innovationsprozessen, die letztlich zu einer Unregierbarkeit unserer Gesellschaften führen können. Die Hoffnung, daß der technische Fortschritt, z.B. in Form einer zunehmenden Dematerialisierung, die resultierenden Probleme lösen wird, hat sich bis heute nicht erfüllt. Das ist u.a. eine Folge des sogenannten Rebound-Effekts, der im Kern dazu führt, das Einsparungen, die aus technischen Fortschritten resultieren könnten, sofort in vermehrte menschliche Aktivitäten umgesetzt werden.

Solche vermehrten Aktivitäten führen – in einer historischen Perspektive – zu einer wachsenden Bevölkerung, mehr Konsum, mehr Mobilität und einer ständig höheren Umweltbelastung. Als Folge der zunehmenden Globalisierung stehen dabei kurzfristig gewaltige zusätzliche Umweltbelastungen durch das hohe wirtschaftliche Wachstum in den Schwellenländern und damit zusammenhängend – als neues Phänomen – ein rasanter Abfluß von Arbeit aus den reichen Industrieländern mit wachsender Arbeitslosigkeit und Bedrohung unserer Sozialsysteme an. Bei Fortsetzung der bisherigen Trends drohen einerseits erhebliche soziale Konflikte, andererseits ein Klimakollaps, und es ist absolut unklar, wie wir diese Situation konkret bewältigen sollen.

Offensichtlich ist, daß eine friedliche Bewältigung dieser Herausforderungen nur im Rahmen weltweiter Lösungen erfolgen kann, also im Rahmen von Vereinbarungen zwischen Nord und Süd, Ost und West, die allen Menschen auf diesem Globus eine positive Perspektive für die Zukunft versprechen. Dies erfordert das graduelle Schließen der heute unerträglich großen Differenz zwischen Reich und Arm, aber ebenso die weltweite Durchsetzung – und Mitfinanzierung – von Umwelt- und Sozialstandards. Entsprechende Mechanismen der Zusammenarbeit (z.B. Umweltzertifikate, weltweite Sozialsysteme, Maßnahmen des Joint Implementation zwischen Nord und Süd) würden den Aufbau von globalen Infrastrukturen ermöglichen und den Weg in eine nachhaltige Entwicklung marktwirtschaftlich absichern. Zugleich würden sie zu wirklich zukunftssicheren Arbeitsplätzen führen und damit auch unsere Sozialsysteme zu stabilisieren erlauben. Geeignete globale Rahmenbedingungen sind dann auch die Voraussetzung dafür, daß regionale Initiativen in zielführender Weise möglich werden, gemäß der Leitidee „Think globally, act locally".

Informations- und Kommunikationstechnologie ist für die beschriebenen Prozesse der Globalisierung ein ganz wesentlicher Faktor. Zum einen wirkt IT empowernd, erlaubt weltweit Menschen, sich effizient in den Wirtschaftsprozeß einzubringen, ist ein wesentlicher treibender Faktor für eine preiswerte weltweite Organisation von Wertschöpfungsketten, damit indirekt eine wichtige Ursache für den Abfluß von Arbeit aus den Industriestaaten. IT ist andererseits Teil der Lösung, denn Informations- und Kommunikationstechnik ermöglicht besonders weitgehende Effekte der Dematerialisierung durch Technik, und bei Vermeidung von Rebound-Effekten durch geeignete gesellschaftliche Rahmenbedingungen eröffnet dies gute Chancen für langfristig tragfähige Lösungen. Noch nie war es so preiswert und umweltverträglich möglich, Menschen überall auf der Welt in gleichberechtigter Weise in die weitere Entwicklung einzubeziehen.

Die Bewältigung der Zukunft wird im wesentlichen in einer geeigneten Austarierung des Spannungsverhältnisses zwischen Wirtschaft, sozialen Anforderungen und der Umwelt bestehen. Aufgrund der Globalisierung des Wirtschaftens wird dieses Austarieren auf Dauer allerdings nicht mehr national oder regional, sondern nur noch global zu bewältigen sein. Die entscheidenden Fragen sind insofern Fragen hinsichtlich der weltweiten Durchsetzung sozialer und ökologischer Mindeststandards, die eine Ausrichtung des Wirtschaftens hin zu einer nachhaltigen Entwicklung, aber auch zu einem sozialen Miteinander – und damit zu einer weitergehenden Verwirklichung der Menschenrechte – bringen werden. Natürlich erfolgen solche Standards partiell zu Lasten des insgesamt erreichbaren Produktionsumfangs, verbessern dafür aber die Lebensqualität, den Grad an sozialer Gerechtigkeit, die ökologische Situation und insgesamt die Durchsetzung der Menschenrechte. Offensichtlich sind Lösungen der angedeuteten Art nur denkbar,

wenn sie auch weltweit und fair finanziert werden, z.B. über Mechanismen der Zusammenarbeit wie Umweltzertifikate, Ausbildungshilfen, Maßnahmen des Joint Implementation zwischen Nord und Süd. Eine gedeihliche Zukunft ist nur im Rahmen weltweiter Lösungen, im Rahmen von Vereinbarungen zwischen Nord und Süd, Ost und West erreichbar, und diese werden letztlich allen Menschen auf diesem Globus eine positive Perspektive versprechen müssen. Der mögliche Beitrag der Informations- und Kommunikationstechnik zur Erreichung dieser Ziele ist eine zentrale Frage.

Das Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW) in Ulm hat in diesem Zusammenhang für die Europäische Kommission in Form der Koordinierung einer Expertengruppe in 1995 eine Studie zum Thema der Wechselwirkung zwischen den beiden Leitideen *Informationsgesellschaft* und *nachhaltige Entwicklung* erarbeitet. Es ist dies ein diffiziles Thema. In der Diskussion ist klar geworden, daß zum einen das beschriebene Dreieck von Anforderungen im wirtschaftlichen, sozialen und ökologischen Bereich auszutarieren ist und daß zum anderen die beiden Leitideen nicht automatisch konvergieren. Konkret kann man sich zwar bei der heutigen Ausgangssituation kaum eine nachhaltige Welt vorstellen, die nicht wesentlich auf Informationstechnologien aufbaut, aber man kann sich sehr wohl Gesellschaften vorstellen, die auf Informationstechnologien aufbauen und nicht nachhaltig ausgerichtet sind.

Die beschriebene Studie hat in Form einer Präambel gewisse Leitprinzipien für die weitere Entwicklung herausgearbeitet, die hier erwähnt werden sollten. Hierzu gehört

(1) die hohe Relevanz des Themas der Nachhaltigkeit (die in ihrer Wichtigkeit vergleichbar ist mit den Menschenrechten, Demokratie und dem Anspruch auf Arbeit),

(2) die Feststellung, daß Nachhaltigkeit immer aus einer globalen wie aus einer lokalen Perspektive betrachtet werden muß,

(3) die Erkenntnis, daß mit dem Ziel einer nachhaltigen Entwicklung fast unauflösbar die Notwendigkeit verbunden ist, vergleichbare Lebensbedingungen für Menschen überall auf diesem Globus herbeizuführen,

(4) die Berücksichtigung der Interessen zukünftiger Generationen und

(5) die Feststellung, daß die Informations- und Kommunikationstechnologie ein großes Potential besitzt, um einen Beitrag zur Erreichung dieser Ziele zu leisten. Allerdings erschließt die Informations- und Kommunikationstechnologie diese Chancen nur dann, wenn Rebound-Effekte vermieden werden können. Dies erfordert

(6) neue gesellschaftliche Rahmenbedingungen, die derartige Effekte verhindern. Ein Denkmodell ist die Mobilisierung der Marktkräfte in Form einer ökologisch und sozial ausgerichteten, globalen Marktwirtschaft. Hierfür sind die Randbedingungen des Marktes geeignet zu definieren. In einer bestimmten ökonomischen Interpretation geht es dabei um die Internalisierung von externen Kosten (sozialer und ökologischer Art). Schließlich werden

(7) entsprechende weltweite, leistungsfähige und integrierte Infrastrukturen benötigt, die am besten über marktgetriebene Prozesse unter geeigneten gesellschaftli-chen Rahmenbedingungen entstehen. Wenn dies alles in der richtigen Weise angegangen wird, dann bestehen gute Aussichten, daß man aus Sustainability einen Business Case machen kann und sich die Leitidee „Think globally – act locally" umsetzen läßt. Tatsächlich würden in diesem Rahmen die neuen zukunftssicheren Arbeitsplätze entstehen, und wahrscheinlich ließen sich entlang dieser Idee in geeigneten Übergangs- und Anpassungsprozessen auch die nationalen Sozialstaatmodelle langfristig absichern.

Wie sieht es nun mit der aktuellen Umsetzung aus? Nach der Etablierung der Initiative zum Aufbau einer globalen Informationsinfrastruktur als Folge der Al Gore-Kampagne im Rahmen der Zusammenarbeit der G7-Staaten haben die Entwicklungsländer gefordert, daß sie in diesen Prozeß, der aus ihrer Sicht enorme Chancen beinhaltet, der aber erneut auch zu einer weiteren Vertiefung der Kluft zwischen Nord und Süd führen kann, adäquat eingebunden werden. Die ISAD-Konferenz, die in Südafrika auf Einladung von Präsident Mandela zum Thema *Information Society and Development* im Zeitraum 13.–15. Mai 1996 stattfand, hatte genau dieses Thema zum Gegenstand; die Europäische Union hat die Organisation dieser Veranstaltung wesentlich unterstützt. Die ISAD-Konferenz in Südafrika hat die große Lücke, die besteht, deutlich gemacht, aber auch erste Ansatzpunkte zur Überwindung der bestehenden Probleme und zur besseren Nutzung der bestehenden Chancen aufgezeigt. Dies betrifft nicht zuletzt das Potential moderner Informations- und Kommunikationstechnik. Anerkennung findet dabei immer auch das europäische Gesellschaftsmodell, vor allem die soziale und ökologische Orientierung. Deutschland ist in all diesen Themen besonders gefordert, denn für unser Land bietet sich die Chance, als Gastgeber der EXPO 2000, die unter dem Motto Mensch, Natur, Technik steht, in einem äußerst sensiblen Moment (einer Jahrtausendwende) Antwort auf die brennenden Fragen der Menschheit zu geben. Das FAW hat in diesem Kontext mit anderen wissenschaftlichen Partnern für die Expo 2000 Gesellschaft in einem weltweit ausgerichteten thematischen Prozeß in Form von 22 Thesen ein Schlüsseldokument erstellt, das neben anderen FAW-Exponaten ebenfalls auf der ISAD-Konferenz verfügbar gemacht wurde.

Prof. Dr. Dr. F. J. Radermacher
Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW), Ulm