H.-D. Ehrich: Bioinformatik – Modellbildung als Herausforderung fur die Informatik. In M. Bohnet, H. Hopf, K. Lompe, and H. Oberbeck (Herausgeber), Innovation jenseits von Fachgrenzen, pages 3750, TU Braunschweig, 2004.

Bioinformatik – Modellbildung als Herausforderung für die Informatik

Hans-Dieter Ehrich
Institut für Informationssysteme
Technische Universität Braunschweig
HD.Ehrich@tu-bs.de

Zusammenfassung

Die riesigen Datenmengen, die in der Mikrobiologie anfallen, sind nur mit einem großen Aufwand an Informationsverarbeitung zu bewältigen. Die Bioinformatik soll den Engpass überwinden helfen, der bei der Entwicklung der benötigten Informatik-Methoden entstanden ist. Datenbanktechnik hilft, die Daten abzulegen, wiederzufinden und auf vielfältige Weise miteinander zu verknüpfen. Um die Daten zu Informationen und schliesslich Erkenntnissen zu verdichten, bedient man sich formaler Modelle. Hierbei finden mathematische, zunehmend aber auch informatische Methoden Anwendung. Ziel ist es, biologische Systeme und Prozesse qualitativ und quantitativ immer umfassender darstellen, simulieren, analysieren und prognostizieren zu können – und so besser zu verstehen.

1 Einleitung

Ende 1998 gelang es erstmals, das Genom eines mehrzelligen Lebewesens vollständig zu sequenzieren. Es handelte sich um den kleinen Wurm mit lateinischem Namen *Caenorhabditis elegans* [1], in der Mikrobiologie seit langem bekannt als wichtiger Modellorganismus.

Das Wissen um die richtige Gensequenz muss aus riesigen Mengen von Mess- und Labordaten gewonnen werden. Solche Datenmengen sind nur mit Informatik-Werkzeugen zu bewältigen. Dass man Datenbanktechnik braucht, liegt auf der Hand: die Daten müssen so gespeichert werden, dass sie nach verschiedenen, im voraus oft nicht bekannten Kriterien wiedergefunden werden können. Zudem müssen sie einer Vielzahl von Benutzern gleichzeitig zur Verfügung stehen, auch von ferne über das Internet, sie müssen aus unterschiedlichen Quellen über das Netz zusammengeführt werden können, sie müssen gegen Beschädigungen und Verluste gesichert sein u.s.w.

Dazu dienen biologische Datenbanken, die es mittlerweile in großer und schnell wachsender Anzahl gibt. Unter vielen anderen gibt es eine zentrale Datenbank für *C. elegans* [2].

Die Gesamtzahl der Datenbanken ist nirgends erfasst, ebenso wenig wie der Gesamtumfang der darin gespeicherten Daten. Das *European Bioinformatics Institute (EBI)* [3] hat als gemeinnützige Organisation die Aufgabe, das wachsende Informationsvolumen aus der Molekularbiologie und Genomforschung für die öffentliche Forschung zugänglich zu machen. Dessen *Sequential Retrtieval System (SRS)* verfügte Anfang des Jahres 2004 über 1.5 Terabyte Speicherplatz; dies entspricht dem Umfang von über vier Monaten Spielzeit handelsüblicher Musik-CDs – oder fast zwei

Wochen digitalisierten Spielfilms.

Die Zeitschrift *Nucleic Acids Research* veröffentlich zu Beginn jedes Jahres ihre *Molecular Biology Database Collection*, eine Liste nützlicher öffentlich zugänglicher Datenbanken. Die Liste des Jahres 2004 verzeichnete 548 Datenbanken, 162 mehr als im Vorjahr. Der Zuwachs war grösser als in den Jahren zuvor und setzte damit einen seit Jahren bestehenden Trend fort: der Bestand wächst mit steigenden Zuwachsraten.

Daten allein sind noch keine Erkenntnisse. Um Erkenntnisse zu gewinnen, ist es nötig, die in den Daten steckenden biologischen Strukturen und Prozesse nach verschiedenen Gesichtspunkten darstellen, simulieren, analysieren und prognostizieren zu können.

Dazu dienen u.a. Anwendungsprogramme. In der Folge der Genom-Sequenzierung, d.h. der Erforschung der Abfolge der Basen in einem Nukleinsäuremolekül, entstanden Methoden zum *Alignment*, dem Vergleich von Sequenzen auf größtmögliche Übereinstimmung, und Suchheuristiken nach gegebenen Sequenzen in Datenbanken wie BLAST *(basic local alignment search tool)*. Viele weitere Anwendungen sind im Gebrauch oder in der Entwicklung, etwa zum Aufspüren funktionaler Elemente in DNA-Sequenzen, zum Vergleich von Genomen, zur Ermittlung phylogenetischer Bäume, zur Aufklärung von Molekülstrukturen und vieles mehr.

In einigen Bereichen der Mikrobiologie ist es gelungen, Gesetzmäßigkeiten in mathematische Modelle zu fassen. So gibt es z.B. Modelle metabolischer Prozesse in der Form von Systemen partieller Differentialgleichungen (s.u. Abschnitt 3.2). Diese Modelle erlauben dann Prognosen über das Verhalten im modellierten Ausschnitt der Natur, die am Experiment verifiziert werden können. Oder falsifiziert: wenn die Natur abweichendes Verhalten zeigt, ist das Modell falsch und muss korrigiert werden. Im Auffinden eines Modellfehlers und dessen Korrektur liegt oft eine großer Erkenntnisgewinn. Sind keine Fehler mehr erkennbar, gewinnt das Modell den Status einer wissenschaftlichen *Theorie* im besten Sinne der Tradition der exakten Naturwissenschaften.

In der Biologie spielt neben der Mathematik auch die Informatik bei diesem Prozess der Modell- und Theoriebildung eine entscheidende Rolle.

Mathematische Modellrechnungen sind in aller Regel so aufwändig, dass ohne Computer nicht auszukommen ist, und viele erfordern ein anspruchsvolles Instrumentarium an Algorithmik und Datenstrukturen.

Es gibt darüber hinaus Ansätze, dem Instrumentarium mathematischer Modellierung originär informatische Methoden hinzuzufügen. Grundlage bilden die diskreten digitalen Modellvorstellungen für Hard- und Software: der Zeitablauf vollzieht sich nicht kontinuierlich, sondern in sprunghaften Übergängen zwischen stationären Zuständen.

2 Kompetenzzentrum für Bioinformatik

Mitte der achtziger Jahre begann in Deutschland eine einschlägige Konferenzserie, die jetzt GCB *(German Conference on Bioinformatics)* heißt. Als sie begann, war es weltweit die erste Konferenzserie zu diesem Thema. In der Folge wurden wissenschaftliche Förderprogramme des Bundesministeriums für Bildung und Forschung (BMBF) und der Deutschen Forschungsgemeinschaft (DFG) aufgelegt, die zunehmend zu einer Kooperation von Biologen und Informatikern führten.

Im Rahmen des NGFN *(National Genome Research Network)* fördert das BMBF Projekte im Bereich der Genomforschung mit ihren medizinischen Anwendungen, die zunehmend auf Arbeiten in der Bioinformatik Bezug nehmen. Im Rahmen des HNB *(Helmholtz Net-*

work Bioinformatics) fördert das BMBF die Entwicklung einer Plattform zur Integration eines weiten Spektrums von Dienstleistungen und Werkzeugen, um sie den Forschergruppen in der Biologie leichter zugänglich zu machen.

Im Jahre 2000 begann das BMBF eine intensive Förderung der Bioinformatik mit dem Ziel, die Bioinformatik-Aktivitäten in Deutschland zu bündeln sowie untereinander und mit anderen Disziplinen zu vernetzen. Um dem erkennbaren Engpass an Fachpersonal in der Bioinformatik zu begegnen, sollten zudem in enger Abstimmung mit den Landesregierungen entsprechende Studien- und Ausbildungsmöglichkeiten geschaffen werden.

Im Rahmen des NBCC (Network of Bioinformatics Competence Centers [4]) fördert das BMBF sechs Kompetenzzentren: in Berlin, Braunschweig, Köln, Gatersleben/Halle, Jena und München.

Das Branschweiger Kompetenzzentrum "Intergenomics" [5] hat das Ziel, bioinformatische Werkzeuge bereitzustellen, mit denen interaktive genomgesteuerte Prozesse während der Infektion von Säuger- oder Pflanzenorganismen modelliert werden können. Zu diesem Zweck soll eine integrierte Infrastruktur geschaffen werden, die die in der Region vorhandenen Wissensbasen, Werkzeuge und Dienste ebenso umfasst wie diejenigen, die im Projekt neu entwickelt werden. Partner im Projekt sind die Gesellschaft für biotechnologische Forschung (GBF) in Braunschweig, die Technische Universität Braunschweig, die Fachhochschule Braunschweig-Wolfenbüttel, das Universitätsklinikum Göttingen und die Firma BIOBASE GmbH in Wolfenbüttel.

Der kommerzielle Partner kann mögliche Verwertungen und professionelle Bedürfnisse an Bioinformatikentwicklungen artikulieren. Insofern ist das Projekt auf Nachhaltigkeit angelegt, mit vielversprechenden Verwertungsperspektiven für die zu entwickelnden Werkzeuge.

Abbildung 1 zeigt die zu Beginn des Jahres 2004 im Intergenomics-Projekt gepflegten Datenbanken und Software-Werkzeuge. Der Informatiker ist hier weniger an der inhaltlichen Bedeutung der Daten orientiert als an deren Strukturierung und anforderungsgerechten Verwaltung. Dazu ist ein sorgfältiger Datenbankentwurf nötig, an dem Anwender und Informatiker zusammenwirken müssen. Dies ist ein komplexer Vorgang, der von der Erhebung der Benutzeranforderungen bis zur Implementierung eine Reihe von Entwurfsphasen durchläuft. Abbildung 2 zeigt eine Teilansicht eines konzeptionellen Datenmodells für die PathoPlant-Datenbank (ca. 20% des Gesamtmodells). Aus diesem wurde ein logisches Datenbankschema entwickelt, mittels dessen schließlich die Datenbank implementiert wurde [6].

Ein wichtiges Ziel ist der Aufbau von Studien- und Ausbildungsangeboten im Fach Bioinformatik, die den Anforderungen an die Interdisziplinarität des Faches Rechnung tragen. An der TU Braunschweig werden bereits Veranstaltungen in diesem Bereich angeboten. Die Einführung eines Studiengangs für Bioinformatik ist in Vorbereitung.

3 Modellierung

Daten in Datenbanken sind nicht das Ziel, sondern der Ausgangspunkt wissenschaftlicher Erkenntnis. Anwendungs- und Verarbeitungsprogramme (einige wurden oben erwähnt) verarbeiten diese Daten und liefern aggregierte Daten, die wieder weiter verarbeitet werden, u.s.w. So entsteht eine Methodik der Darstellung, Simulation, Analyse und Prognose biologischer Strukturen und Prozesse. Aggregierte Daten und Programme, die solche Verarbeitungen ermöglichen, lassen sich als Computer-Modelle biologischer Sachverhalte auffassen.

Um Missverständnisse zu vermeiden: "Mo-

BIOBASE (nur Eukaryoten)

TRANSFAC	Transkriptionsfaktoren und deren Bindungsstellen
TRANSCompel	kombinatorische Elemente in Promotoren
PathoDB	pathologisch relevante veränderte Transkriptionsfaktoren und deren Bindungsstellen
TRANSPRO	humane und murine Promotor-Sequenzen
S/MARt DB	DNA-Anheftungsstellen am nukleären Gerüst
CYTOMER	hierarchische Klassifizierung von Zellen, Strukturen, Organen
TRANSPATH	Signaltransduktionswege

TU Braunschweig

PRODORIC	Transkriptionsfaktoren und Signaltransduktion bei Prokaryoten
PathoPlant	Pflanze-Pathogen-Wechselwirkung/Signaltransduktions-Komponenten
AthaMap	genomweite Karte von vorhergesagten Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom
JVirGel	Tool zur Vorhersage des Laufverhaltens von Proteinen im 2-dimensionalen Gel (Auftrennung nach Masse und IP)

Abbildung 1: Intergenomics-Datenbanken und -Software-Werkzeuge

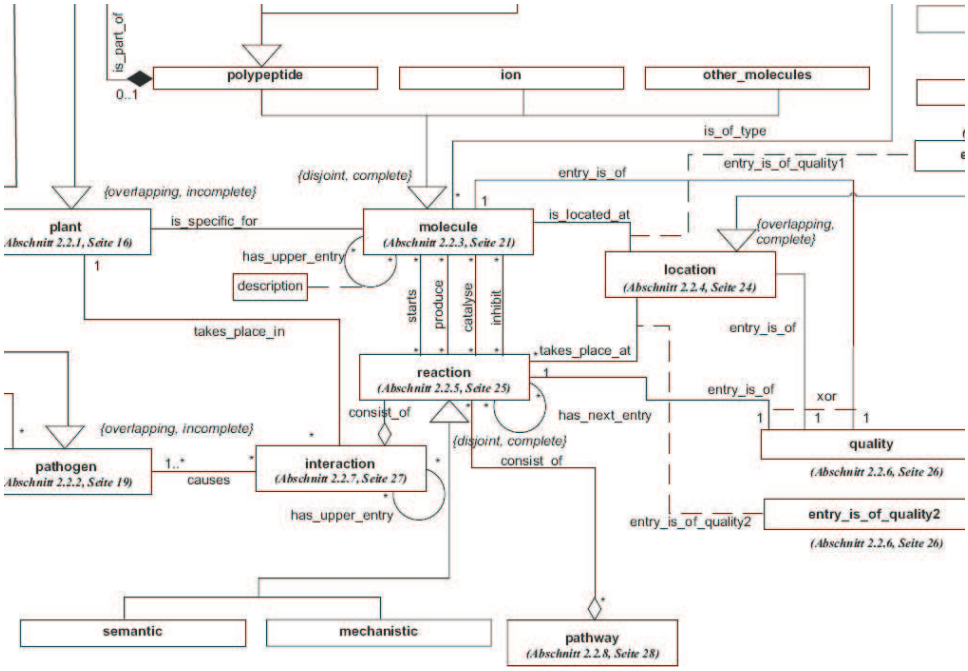


Abbildung 2: Konzeptionelles Schema für die PathoPlant-Datenbank (Teilansicht)

dell" bedeutet hier etwas anders als bei dem Begriff des "Modellorganismus", wie er in der Biologie geläufig ist und oben in der Einleitung für *C. elegans* verwendet wurde: ein Modellorganismus wird erforscht in der begründeten Hoffnung, Erkenntnisse auf andere Organismen übertragen zu können. Das ist im folgenden nicht gemeint.

Gemeint ist eine symbolische Darstellung der Sachverhalte, auf die sich exakte Analysen und Prognosen gründen lassen. Seit Jahrhunderten hat die Mathematik die Grundlagen und Werkzeuge hierfür geliefert, sehr erfolgreich etwa für die Physik zur Formulierung ihrer Naturgesetze. Computer-Modelle erfüllen im Prinzip denselben Zweck. Sie erweitern den Anwendungsbereich für exakte Modellierungsmethoden, denn sie können oft dort verwendet werden, wo eine mathematische Modellierung nicht möglich oder nicht zweckmäßig ist. Insofern sind sie die Fortsetzung der mathematischen Modellierung mit anderen Mitteln.

In der Biologie spielen mathematische Modelle ebenfalls eine große Rolle, zum Beispiel bei der Erfassung der Gesetzmäßigkeiten des Metabolismus, aber auch in anderen Bereichen wie der Genexpression (s.u. Abschnitte 3.2 und 3.3).

Konzepte aus der Informatik ergänzen jedoch das Modellierungsinstrumentarium. Offensichtlich wird Informatik bei der algorithmischen und rechentechnischen Durchführung von Berechnungen auf der Grundlage mathematischer Modelle benötigt, z.B. bei der numerischen Lösung partieller Differentalgleichungen. Aber das ist nicht alles: hinzu kommen originär informatische Modellierungskonzepte, etwa zur Visualisierung biologischer Strukturen und Prozesse mit Methoden der Computergraphik. Im Abschnitt 3.1 wird ein Beispiel zur Darstellung und Erkundung der Struktur biologischer Makromoleküle beschrieben. In neueren Ansätzen werden diskrete Modelle aus der Digitaltechnik (Hardware und Software) zur Modellierung biologischer Sachverhalte verwendet; dazu zwei Beispiele im Abschnitt 3.4.

Dies sind nur einige Beispiele, ein Anspruch auf Vollständigkeit wird nicht erhoben. Auch andere Arbeitsrichtungen in der Biologie sind für exakte Methoden zugänglich, und es gibt weitere Ansätze zur mathematischen oder informatischen Modellierung als die hier gezeigten. Motiv für die hier getroffene Auswahl war (neben den Vorlieben des Autors) das Bestreben, eine Bandbreite von Ansätzen zu zeigen, die auf unterschiedliche Anwendungsszenarien anwendbar sind und ahnen lassen, wie breit das Feld biologischer Phänomene ist, die exakten Methoden zugänglich sind.

3.1 Graphische Modellierung

Als Beispiel soll der *BioBrowser* vorgestellt werden, ein innovatives Computergraphik-System zur interaktiven Visualisierung hochkomplexer Protein-Moleküle [7, 8, 9]. Es wurde in einer von der Deutschen Forschungsgemeinschaft geförderten Zusammenarbeit des Instituts für Computergraphik der TU Braunschweig (Prof. Fellner) und der Abteilung Strukturbiologie der GBF Braunschweig (Prof. Heinz) entwickelt.

Das Werkzeug bietet die Möglichkeit, die 3D-Struktur auch sehr großer und komplexer Protein-Moleküle interaktiv auf Standard-Rechnern zu visualisieren. Dies ist u.a. Voraussetzung für den gezielten Entwurf von Arzneimitteln, da die Funktion eines solchen Moleküls eng mit seiner 3D-Struktur zusammenhängt. Die Bedeutung eines solchen Werkzeugs wächst mit der Anzahl und Größe der in der RCSB PDB *(Research Collaboratory for Structural Bioinformatics Protein Data Bank)* [10] erfassten Moleküle.

Der BioBrowser stellt eine einfach zu handhabende Benutzerschnittstelle zur Verfügung, die die benötigten Daten über das Protein und

eventuelle Selektionen zugänglich macht. Diese Daten werden benutzt, um eine Visualisierung des Proteins zu generieren, mit der sich interaktiv arbeiten lässt: die räumliche Darstellung kann nach Belieben gedreht und geschoben werden.

Es werden die gebräuchlichen Arten der Visualisierung unterstützt (s. Abbildung 3). Dazu gehören zunächst die in der Chemie üblichen Ball-and-Stick- und Spacefill-Darstellungen, die die Positionen der Atome genau wiedergeben, jedoch bei großen Molekülen wesentliche Teile des Bildes verdeckt halten. Ebenfalls unübersichtlich sind die Strichzeichnungen, die die im Molekül vorkommenden Bindungen darstellen. Bewährt haben sich 'Ribbon'-Darstellungen der Hauptkette bzw. bestimmer Strukturelemente sowie die Darstellung als molekulare Oberfläche.

Es ist beabsichtigt, weitere Darstellungen wie zum Beispiel eine textbasierte Ausgabe der Primärstruktur eines Proteins in den Bio-Browser zu integrieren. Geplant ist dann der Ausbau zu einem Werkzeug, das es dem Forscher ermöglicht, zusätzliche Informationen in die 3D-Struktur einzubetten oder auch abzurufen. Diese Informationen sind i.a. textbasiert, aber es sollen auch Hyperlinks möglich sein, wie sie von WebBrowsern her bekannt sind.

3.2 Kontinuierliche Modellierung

Dies ist die klassische mathematische Modellierung mittels infinitesimaler Methoden, meist Differentialgleichungen. Dieser Ansatz ist immer dann geeignet, wenn wir eine (fast) beliebig teilbare „glatte" Materie vor uns haben, deren Teile bzgl. der kontinuierlich veränderlichen Messgrößen (Temperatur, Druck, Konzentration eines bestimmten Stoffes u.s.w.) gleiche Eigenschaften haben. Solche Modelle sind in der Physik, aber z.B. auch in der Chemie, der Verfahrenstechnik und auch in der Simulation biologischer Prozesse Standard [11].

Als Beispiel betrachten wir die Expression von Genen, d.h. das Realisieren der Information, die in der DNA (bei Viren auch RNA) eines Gens gespeichert ist [12]. Im betrachteten Fall wird aufgrund von Erbinformation, die in der DNA gespeichert ist, im Rahmen des Stoffwechsels ein Substrat *(Edukt)* S in ein anderes *(Produkt)* P umgewandelt. Der Modell-Prozess [13] beginnt mit einer kurzzeitigen Aktivierung der DNA als Modell-Eingabe. Daraufhin werden RNA, Enzym (das ist ein spezielles Protein) und Produkt (das *Metabolit*, z.B. Glukose oder Fruktose) nach folgenden Reaktionen gebildet.

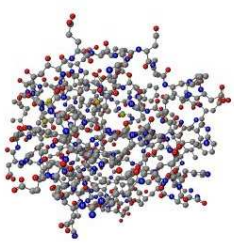
$$\begin{aligned}
\mathrm{DNA} &\xrightarrow{k_1} \mathrm{DNA} + \mathrm{RNA} \\
\mathrm{RNA} &\xrightarrow{k_2} \mathrm{RNA} + \mathrm{Enzym} \\
\mathrm{RNA} &\xrightarrow{k_3} \textit{abgebaut} \\
\mathrm{Enzym} &\xrightarrow{k_4} \textit{abgebaut} \\
\mathrm{S} + \mathrm{Enzym} &\xrightarrow{k_5} \mathrm{P} + \mathrm{Enzym}
\end{aligned}$$

Der zeitliche Verlauf der Stoffkonzentrationen werden quantitativ durch folgende Differentialgleichungen beschrieben.

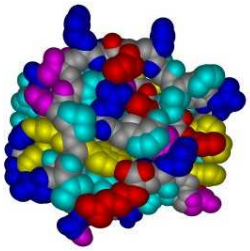
$$\begin{aligned}
\frac{d\ [\mathrm{RNA}]}{d\ t} &= k_1[\mathrm{DNA}] - k_3[\mathrm{RNA}] \\
\frac{d\ [\mathrm{Enzym}]}{d\ t} &= \frac{k_2[\mathrm{RNA}]}{K_{m2} + [\mathrm{RNA}]} - k_4[\mathrm{Enzym}] \\
\frac{d\ [\mathrm{S}]}{d\ t} &= -\frac{k_5[\mathrm{Enzym}]\ [S]}{K_{m5} + [\mathrm{S}]} \\
\frac{d\ [\mathrm{P}]}{d\ t} &= \frac{k_5[\mathrm{Enzym}]\ [S]}{K_{m5} + [\mathrm{S}]}
\end{aligned}$$

Die Kurvenverläufe der Lösungen sind in Abbildung 4 graphisch dargestellt.

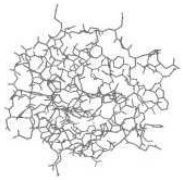
Ein solches Modell ist isoliert wenig aussagekräftig, aber wenn eine Vielzahl von Genen und Genprodukten sowie ihre Interaktion betrachtet werden, so können Vorhersagen



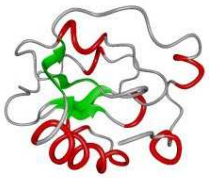
Ball and Stick



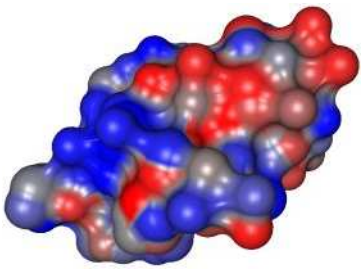
Spacefill



Strichzeichnung



Ribbon-Struktur



Abrollflächen

Abbildung 3: Darstellungsarten für Protein-Moleküle

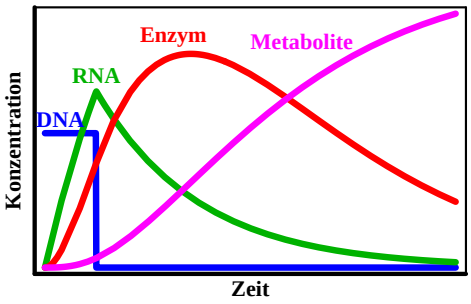


Abbildung 4: Stoffkonzentrationen

über mögliche Zustände biologischer Zellen gemacht werden. Problematisch ist, dass die meisten Reaktionsraten k_i nicht bekannt und z.T. auch nicht messbar sind.

Derartige Anwendungen könnte man eher der *Biomathematik* zurechnen, aber die Informatik ist auf zweierlei Art mit betroffen. Zum einen bedarf die Lösung großer Gleichungssysteme des massiven Rechnereinsatzes; dies ist die Domäne des *Wissenschaftlichen Rechnens*, eines interdisziplinären Gebiets zwischen Mathematik und Informatik. Zum anderen müssen für eine breite Erforschung der so modellierbaren Phänomene sehr viele Modelle generiert und untersucht werden, mehr als von Hand zu schaffen ist. Nötig wäre die automatische Generierung von Modellen aus Datenbanken! Dies könnte z.B. im Rahmen einer rechnergestützten interaktiven Werkbank für Biologen geschehen, in die sowohl Datenbanken als auch die einschlägigen mathematischen Methoden als Werkzeuge eingebettet sind. Konzeption und Implementierung solcher Arbeitsumgebungen ist eine typische Gemeinschaftsaufgabe der Informatik und des Anwendungsgebiets, in diesem Fall der Bioinformatik.

3.3 Stochastische Modellierung

An vielen biologischen Prozessen sind nur wenige Moleküle beteiligt, und diese sind dazu von unterschiedlicher Art und Funktion. Eine Betrachtungsweise als Kontinuum wäre nicht

angemessen. Auch ist das Ergebnis nicht selten keine kontinuierliche Messgröße, sondern einer von mehreren möglichen diskreten Zuständen, die zufällig auftreten. Wenn es möglich ist, viele Realisierungen eines solchen Prozesses zu beobachten, kann man stochastische Aussagen über die Verteilung der Zustände machen.

Ein Beispiel für einen derartigen biologischen Prozess ist die Genexpression bei Eukarioten: sie ist inhärent stochastischer Natur (was erklärt, warum clonale eukariotische Populationen recht heterogen sein können). Blake et al geben ein Beispiel, wie dies „Rauschen" stochastisch modelliert werden kann [14]. Es geht um die Expression des GFP *(green fluorescent protein)* bei Bäckerhefe.

Abbildung 5(a) zeigt zwei Gene. Codierende Bereiche sind grau, regulierende schwarz umrandet gezeichnet. TetR, exprimiert von P_{GAL10^*}, unterdrückt die Expression von yEGFP. ATc *(Anhydrotetracyclin)* und GAL *(Galactose)* sind erforderlich, um die Expression von yEGFP zu induzieren.

Abbildung 5(b) zeigt Dosis-Response-Kurven von P_{GAL10^*} beim Exprimieren von yEGFP, und zwar die mittlere Fluoreszenz bei Veränderung eines Stoffes bei jeweils festgehaltenem anderen (Kurven A und B: Details können der Originalarbeit entnommen werden).

Abbildung 5(c) lässt die inhärent stochastische Natur des Prozesses erkennen: gezeigt wird der Verlauf der Reaktion von Zellen gemäß A und B (s. Abbildung 5(b)). Die Histogramme zeigen die Anzahlen der Zellen über dem Fluorescenzgrad zu bestimmten Zeitpunkten, von links nach rechts in der Zeit fortschreitend. Während fast alle Verteilungen einer Normalverteilung gleichkommen, zeigt die zweite Verteilung im B-Bild einen ungewöhnlichen Verlauf: ein Effekt, der nur bei stochastischer Modellierung sichtbar wird. Bzgl. der weiteren Diskussion sei auf die Originalarbeit verwiesen.

3.4 Diskrete Modellierung

Bei biologischen Prozessen, an denen wenige heterogene Moleküle beteiligt sind und bei denen die Zustände und Zustandsübergänge eher diskret betrachtet werden, ist es nicht immer möglich oder sinnvoll, stochastische Aussagen über große Anzahlen zu machen. Für einige derartige Prozesse gibt es Ansätze, sie mit informatischen Konzepten zu modellieren, wie sie für den Entwurf digitaler Systeme (Hard- und Software) entwickelt wurden. Zu den Ansätzen, die zunehmend auch Anwendungen in der Mikrobiologie finden, gehören Zustands– und Sequenzdiagramme in verschiedenen Varianten und Erweiterungen.

Erste Beispiele sind die Modellierung des Prozesses der Aktivierung von T-Zellen im Immunsystem als erweitertes Zustandsdiagramm *(Statechart)* [15], und der Embryonalentwicklung von *C. elegans* als erweitertes Sequenzdiagramm *(Life Sequence Chart)* [16, 17]. Die Modelle lassen Simulationen zu, die bereits zu Erkenntnissen in den Anwendungsbereichen beigetragen haben.

Abbildung 6 zeigt das Statechart-Modell der Aktivierung von T-Zellen im Immunsystem. Die Funktionsweise kann hier nicht erklärt werden, vielmehr soll der Nutzen einer solchen Modellierung kommentiert werden: sie gab erstmals ein umfassendes und übersichtliches Bild eines sehr komplexen Vorgangs. Die Fachwissenschaftler hatten über den Prozess eine Unmenge von Daten, waren von einem Verständnis aber ein gutes Stück entfernt. Kam konstruierte das Modell nach intensivem Literaturstudium, als er sicher war, alle bekannten Fakten über den Prozess berücksichtigt zu haben. Er testete das Modell dann mit dem Software-Werkzeug *Rhapsody* (I-Logix, Inc., [18]). Erst funktionierte es nicht richtig, aber Kam fand den Fehler: es fehlte ein Stückchen Information, das in der Literatur gemeinhin übersehen worden war. Nachdem dies kor-

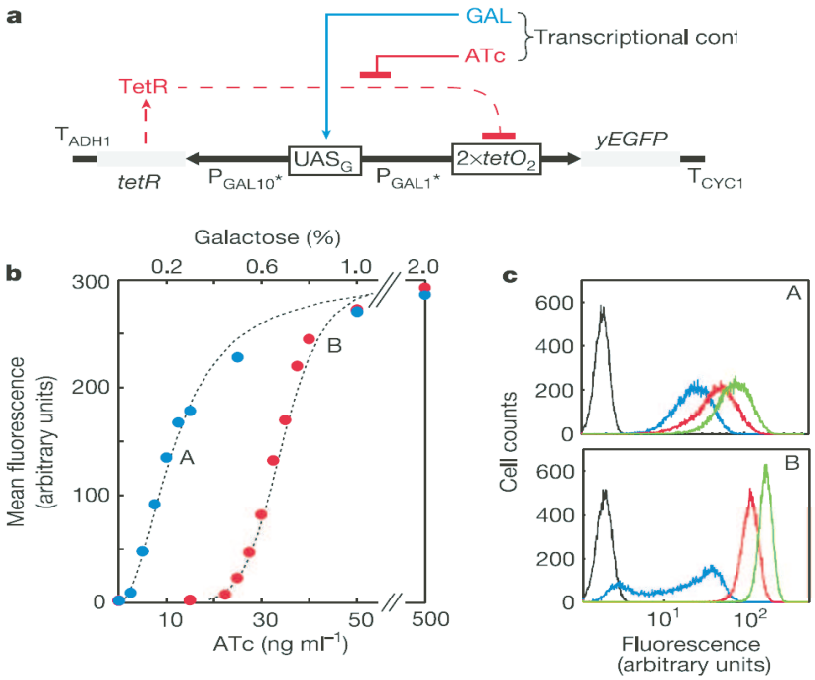


Abbildung 5: Genexpression

rigiert war, funktionierte das Modell einwandfrei und gab den Biologen erstmals einen nachvollziehbaren Überblick über den Prozess.

Die Geschichte zeigt in einer Nussschale den zweifachen Nutzen der Modellierung: den Erkenntnisgewinn durch die Korrektur eines fehlerhaften Modells und den Nutzen für die Fachwissenschaftler, einen besseren Überblick zu gewinnen.

Entsprechende Modellierungsarbeiten haben im Intergenomics-Projekt begonnen, zunächst am Beispiel des Ethylen Pathway bei *Arabidopsis thaliana* [19].

Beim zweiten Beispiel geht es um Sequenzdiagramme: sie sind in der Hard- und Softwaretechnik gebräuchlich, um Szenarien des Zusammenspiels digitaler Funktionseinheiten darzustellen. In der Arbeit [20] wurden Sequenzdiagramme zu LSCs *(Live Sequence Charts)*, einem universellen Darstellungsmittel für kommunizierende Prozesse, weiterentwickelt. Sogleich wurde diese Theorie auf die Modellierung biologischer Prozesse angewandt [17]. In dem Beispiel handelt es sich um die Analyse von Szenarien in der Embryonalentwicklung von *C. elegans*, die zum programmierten Zelltod führen. Abildung 7 zeigt ein Photo mit einigen solchen Zellen [21]. Abbildung 8 zeigt einen Schnappschuss des animierten LSC-Modells [22] zu dem Zeitpunkt, da Zelle P7 den Zelltod erleidet: das LSC oben links zeigt, dass es ein Szenario gibt, das zu diesem Zustand führt, und das LSC unten zeigt den Weg dorthin. Durch Ausblenden bereits gefundener Szenarien und Wiederholung der Simulation konnten bis dahin unbekannte Wege in den Zelltod aufgedeckt werden. Eingeblendet ist neben den LSC-Szenarien eine Darstellung des Prozesses, wie sie Biologen geläufig ist, die auf diesem Gebiet arbeiten.

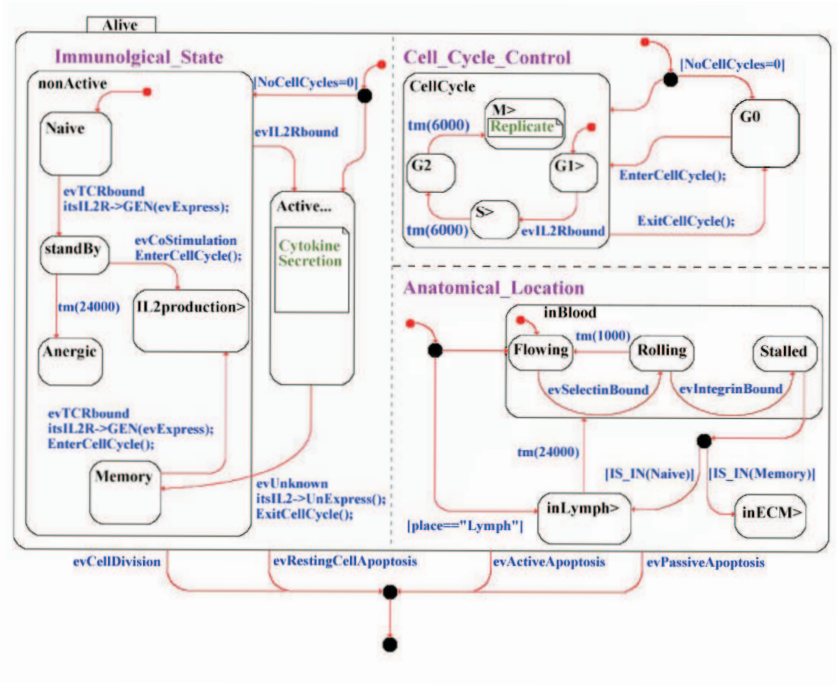


Abbildung 6: Statechart-Modell der T-Zellen-Aktivierung (N. Kam)

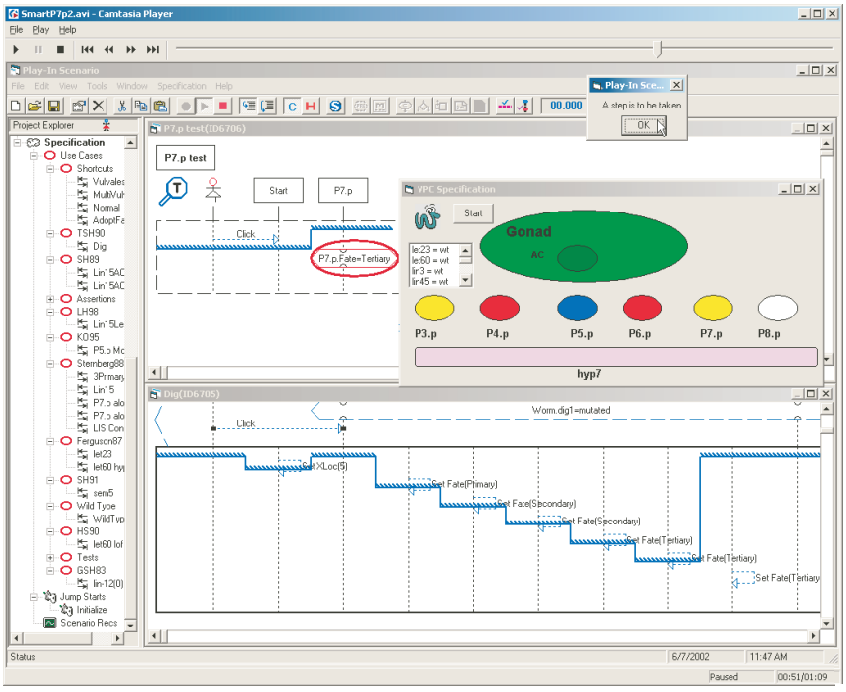


Abbildung 8: LSC-Modellierung der Embryonalentwicklung von *C. elegans* (Schnappschuss)

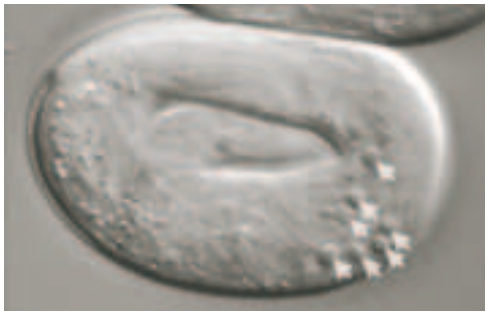


Abbildung 7: Zelltod in der Embryonalentwicklung von *C. elegans*

4 Schlussbemerkungen

Informationsverarbeitung ist ein entscheidender Engpass moderner Biologie geworden. Bioinformatik ist die Anwort auf diese Herausforderung. Thomas Lengauer, Direktor am Max-Planck-Institut für Informatik und Leiter der dortigen Arbeitgruppe *Computational Biology and Applied Algorithmics*, definiert das Gebiet so [23]:

> Bioinformatiker bedienen sich der Methodik der Informatik, um neue Softwarewerkzeuge zu schaffen, mit denen man moderne Biologie treiben kann.
>
> Sie bearbeiten beide Disziplinen gleichrangig.
>
> Dies schafft eine Fülle neuer Möglichkeiten, Türen in beiden beteiligten Disziplinen aufzustoßen.

Drittmittelgeber wie die DFG und das BMFT unterstützen die Bioinformatik als eigenständige Disziplin zwischen der Biologie und der Informatik. Nachdem die ersten Initiativen von Seiten der Biologie kamen, ist die Informatik dabei, das Thema auf breiterer Front für sich zu entdecken. So gibt es seit dem vergangenen Jahr eine eigene Unterreihe der sehr erfolgreihen *Lecture Notes in Computer Science* des Springer-Verlags, die *Lecture Notes in Bioinformatics (LNBI)*.

An mehr als zehn Universitäten sowie mehreren Fachhochschulen wird ein eigenständiger Bioinformatik-Abschluss angeboten. Es gibt einen Studien- und Forschungsführer, der Informationen zu den Studiengängen sowie zu den Forschungsinitiativen in der Industrie und Großforschungseinrichtungen zusammenfasst [24].

In der Region um Braunschweig gibt es einschlägige Kompetenzen in sehr günstiger Zusammenstellung. Eine Zusammenführung lokaler Arbeitsgruppen hat im Rahmen des Intergenomics Kompetenzzentrums begonnen. Es ist zu hoffen, dass hieraus eine stabile Grundlage für innovative Forschung und für ein fundiertes Lehrangebot in diesem Gebiet entstehen.

Danksagungen

Für Hilfen, Anregungen und Unterstützung in vielfacher Hinsicht danke ich S. Eckstein, D. Fellner, R. Hehl, D. Jahn, T. Mack, M. Pirsch und C. Täubner. Mein ganz besonderer Dank gilt J. Weimar, von dem ich viel über Bioinformatik dazugelernt habe, und der wesentliche Materialien zu dieser Ausarbeitung beigesteuert hat. Spezieller Dank gilt J. Weimar und D. Fellner für Korrekturen und Hinweise zu einer Vorversion dieser Arbeit. Für alle verbliebenen Fehler, Auslassungen, Verzerrungen und Ungenauigkeiten bin ich jedoch allein verantwortlich.

Literatur & Links

[1] The C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, 282:2012–2018, 1998.

[2] C. elegans database: www.wormbase.-org/about/about_celegans.html.

[3] www.ebi.ac.uk/.

[4] www.cubic.uni-koeln.de/nbcc/-intro3.htm.

[5] www.intergenomics.de/new/home.php.

[6] M. Balkenhol. Entwicklung einer biologischen Datenbank zur Darstellung von Pflanze-Pathogen-Wechselwirkungen unter Verwendung der UML. *Diplomarbeit, Inst. f. Informationssysteme, TU Braunschweig*, 2003.

[7] BioBrowser: www.cg.cs.tu-bs.de/-research/projects/BioBrowser.

[8] A. Halm, L. Offen, and D. Fellner. Visualization of Complex Molecular Ribbon Structures at Interactive Rates. In *Proc. Information Visualization 2004*, London, July 2004. to appear.

[9] D. Fellner and A. Halm and L. Offen. BioBrowser: Concepts for Fast Protein Visualization. 2004. Wird veröffentlicht.

[10] www.rcsb.org/pdb/.

[11] P. W. Atkins. *Physical Chemistry*. Oxford Univ. Press, 1994.

[12] M.A. Gibson and E. Mjolsness. Modeling the activity of single genes. In J.M. Bower and H. Bolouri, editors, *Computational Methods in Molecular and Cellular Biology: from Genotype to Phenotype*, pages 1–48, Boston, 2001. MIT Press.

[13] J. Weimar, pers. Mitteilung.

[14] W.J. Blake, M. Kærn, C.R. Cantor, and J.J. Collins. Noise in eukaryotic gene expression. *Nature*, 422:633–637, 2003.

[15] N. Kam, I. R. Cohen, and D. Harel. The Immune System as a Reactive System: Modeling T Cell Activation with Statecharts (extended abstract). In *Proc. Visual Languages and Formal Methods (VLFM01), part of IEEE Symposia on Human-Centric Computing Languages and Environments (HCC01)*, pages 15–22, 2001.

[16] D. Harel and R. Marelly. *Come, Let's Play: Scenario-Based Programming Using LSCs and the Play-Engine*. Springer, Berlin, 2003.

[17] N. Kam, D. Harel, R. Marelly, A. Pnueli, E.J.A. Hubbard, and M.J. Stern. Formal Modeling of C. elegans development: A Scenario-Based Approach. In *Proc. Int. Workshop on Computational Methods in Systems Biology (CMSB 2003)*. Kluwer, 2003.

[18] www.ilogix.com/.

[19] C. Täubner. Modellierung des Ethylen-Pathways mit UML-Statecharts. Informatik-Bericht 2003-2, Institut für Informationssysteme, Technische Universität Braunschweig, März 2003.

[20] W. Damm and D. Harel. LSCs: Breathing life into message sequence charts. *Formal Methods in System Design*, 19(1), 2001.

[21] C. elegans Zelltod: www.bio.unc.edu/-faculty/goldstein/lab/movies.htm.

[22] www.wisdom.weizmann.ac.il/-mathusers/kam/celegansmodel/-demos.htm.

[23] www.informatik2003.de/personen/-referenten/keynotes/lengauer.htm.

[24] R. Hofestädt, R. und Schnee. *Studien- und Forschungsführer Bioinformatik*. Spectrum Verlag, Heidelberg, 2002.