Stefan Braß and Alexander Hinneburg, editors, Proc. 18. Workshop Grundlagen von Datenbanken, Wittenberg, pages 13–23, interner Bericht, Juni 2006. Universit¨at Halle sowie
Abhandlungen der Braunschweigischen Wissenschaftlichen Gesellschaft, volume LVII, pages 9–34. J. Cramer Verlag, Braunschweig, 2007.

# Bioinformatik: Erkenntnisse aus der Datenflut?

Hans-Dieter Ehrich Silke Eckstein Brigitte Mathiak
Andreas Kupfer Claudia Täubner

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.*

*In Deutschland wurden im Jahre 2001 fünf Bioinformatik-Kompetenzzentren mit einer Anschubfinanzierung des BMBF eingerichtet. Eines davon befindet sich in Braunschweig, es hat den Namen* Intergenomics *und soll die Interaktion zwischen Genomen aufklären helfen, insbesondere Infektionsprozesse.*

*In diesem Beitrag werden nach einer Einführung in Probleme und Ansätze der Bioinformatik und des Intergenomics-Kompetenzzentrums Arbeiten in unserem eigenen Teilprojekt vorgestellt. Hier werden z.Z. drei Ansätze verfolgt: (1) Suche nach Bildern in Textdokumenten (PDF) aufgrund der Bildbeschriftungen, (2) diskrete Modellierung und Simulation von Signaltransduktionswegen und (3) Koevolution von Datenbankschemata und Ontologien zur Verbesserung der Datenintegration.*

## 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* [12], 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* [7].

Die Gesamtzahl der Datenbanken ist nirgends erfasst, ebenso wenig wie der Gesamtumfang der darin gespeicherten Daten. Das *European Bioinformatics Institute (EBI)* [11] 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 bereits 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 2006 [18] verzeichnete 858 Datenbanken, 139 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 vielem 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 Network 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 [43]) fördert das BMBF sechs Kompetenzzentren: in Berlin, Braunschweig, Köln, Gatersleben/Halle, Jena und München.

Das Branschweiger Kompetenzzentrum "Intergenomics" [25] 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 im Intergenomics-Projekt gepflegten Datenbanken. 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 [3].

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 geplant.

## 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.

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

**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 |

Abbildung 1: Intergenomics-Datenbanken und -Software-Werkzeuge

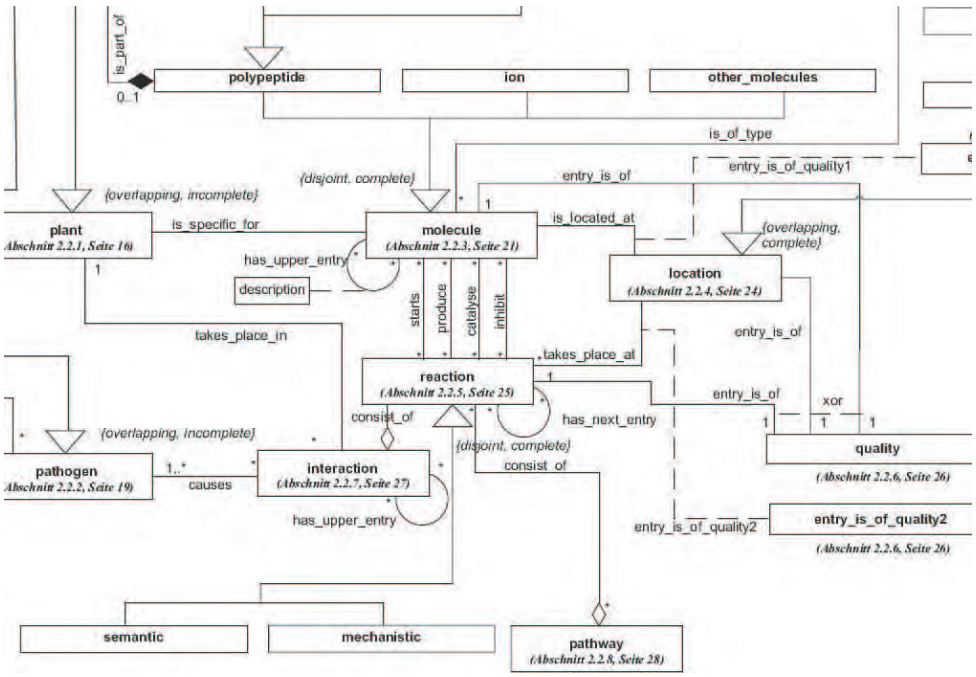



Abbildung 2: Konzeptionelles Schema für die PathoPlant-Datenbank (Teilansicht)

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 Autorenteams) 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 [5, 20, 15]. 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)* [45] erfassten Moleküle.

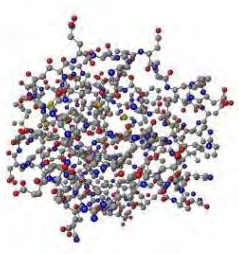
Der BioBrowser stellt eine einfach zu handhabende Benutzerschnittstelle zur Verfügung, die die benötigten Daten über das Protein und eventuelle Selektionen zugänglich macht. Diese Daten werden benutzt, um eine Visualisierung des Proteins zu generieren, mit der sich interaktiv arbeiten lässt: die räumliche 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.

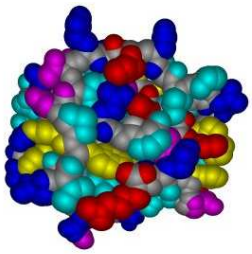
## 3.2 Kontinuierliche Modellierung

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 [2].

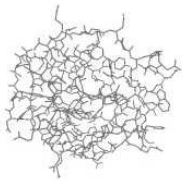
Als Beispiel betrachten wir die Expression von Genen, d.h. das Realisieren der Information, die in der DNA (bei Viren auch RNA) eines Gens gespeichert ist [19]. Im betrachteten Fall wird aufgrund von Erbinformation, die in der DNA gespeichert ist, im Rahmen des Stoffwechsels ein Substrat *(Edukt)* $S$ in ein anderes *(Produkt)* $P$ umgewandelt. Der Modell-Prozess [52] 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) in ihrer Stoffkonzentration nach Reaktionen gebildet, deren zeitlicher Verlauf quantitativ durch Differentialgleichungen beschrieben werden kann. Die Kurvenverläufe der Lösungen sind in Abbildung 4 graphisch dargestellt.



Ball and Stick

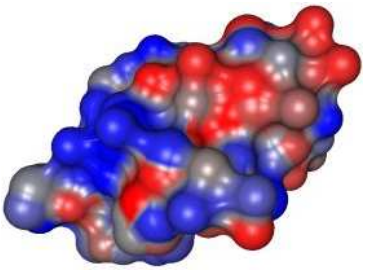


Spacefill



Strichzeichnung

Ribbon-Struktur



Abrollflächen

Abbildung 3: Darstellungsarten für Protein-Moleküle

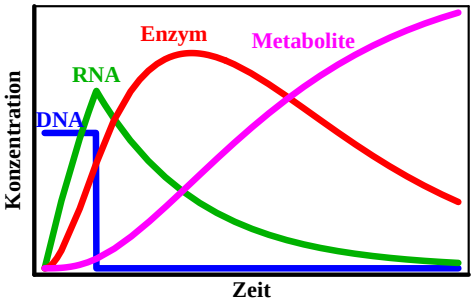



Abbildung 4: Stoffkonzentrationen

Ein solches Modell ist isoliert wenig aussagekräftig, aber wenn eine Vielzahl von Genen und Genprodukten sowie ihre Interaktion betrachtet werden, so können Vorhersagen über mögliche Zustände biologischer Zellen gemacht werden. Problematisch ist, dass die meisten Ausgangsdaten, i.e. Reaktionsraten, 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). Bzgl. eines Beispiels, wie dieses „Rauschen“ stochastisch modelliert werden kann, sei auf [6] verwiesen (es geht um die Expression des GFP *(green fluorescent protein)* bei Bäckerhefe).

### 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)* [28], und der Embryonalentwicklung von *C. elegans* als erweitertes Sequenzdiagramm *(Life Sequence Chart)* [22, 29]. 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 in der Mikrobiologie. 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., [26]). Erst funktionierte es nicht richtig, aber Kam fand den Fehler: es

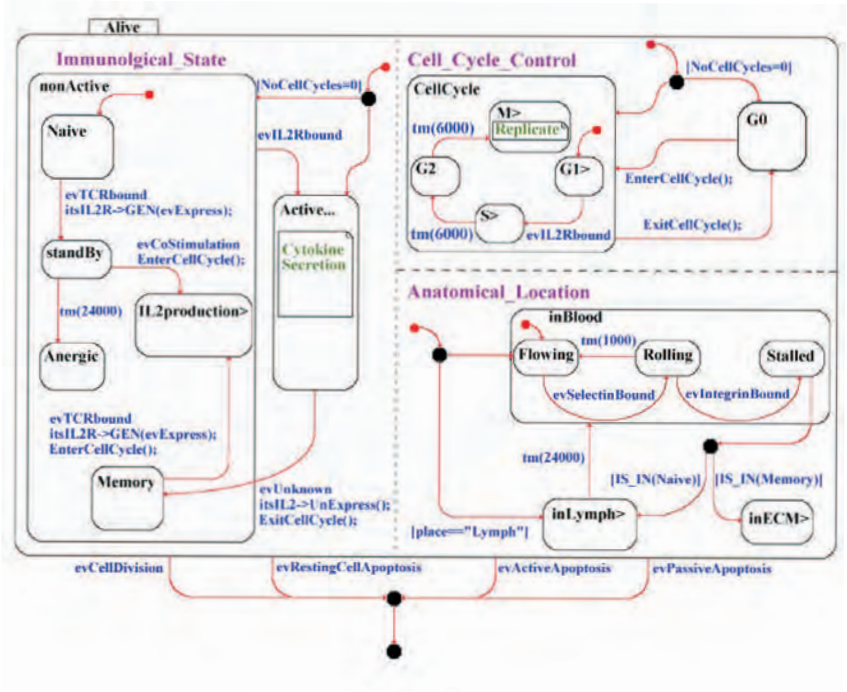



Abbildung 5: Statechart-Modell der T-Zellen-Aktivierung (N. Kam)

fehlte ein Stückchen Information, das in der Literatur gemeinhin übersehen worden war. Nachdem dies korrigiert 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.

# 4 Biologische Daten: Annotationen, Modelle und Ontologien

Die Bioinformatikgruppe des Instituts für Informationssysteme der TU Braunschweig beschäftigt sich im Rahmen des Kompetenzzentrums „Intergenomics“ mit verschiedenen Themen rund um biologische Datenbanken: Zum einen werden Text-Mining-Ansätze verfolgt, um die Biologen bei der Datenannotation – also dem Füllen ihrer Datenbanken mit Informationen aus Publikationen – zu unterstützen. Des weiteren werden Daten über sogenannte Signaltransduktionswege aus den Datenbanken ausgelesen, mit Hilfe von Spezifikationssprachen der Informatik modelliert und die Modelle mit den zugehörigen Werkzeugen simuliert. Und schließlich beschäftigen wir uns mit Integrationsaspekten, um die gemeinsame Verwendung der vielen unterschiedlichen biologischen Datenbanken zu unterstützen. Dazu verfolgen wir einen Ansatz zur Koevolution von Datenbanksschemata und zugehörigen Ontologien. In den folgenden Abschnitten geben wir einen Überblick über die drei Gebiete und verweisen auf weiterführende Publikationen zu den einzelnen Themen.

## 4.1 Suche nach Bildern in Textdokumenten

In den letzten Jahren ist die Anzahl biologischer Datenbanken stark gestiegen, von 548 im Januar 2004 über 719 im Januar 2005 auf 858 im Januar 2006 [17, 18]. Es handelt sich dabei zumeist um recht spezielle Datenbanken, die möglichst alle publizierten Informationen über ein spezielles Thema zusammenstellen wollen. Daher ist eine der aufwändigsten Arbeiten beim Aufsetzen

neuer Datenbanken die Annotation der Literatur. Typischerweise wird die Literaturrecherche in speziellen Literaturdatenbanken wie zum Beispiel PubMed [47] durchgeführt, die ca. 15 Millionen Referenzen und Abstracts aus etwa 4000 biomedizinischen Zeitschriften seit 1950 enthält. Per Stichwortsuche in den Abstracts werden die relevanten Veröffentlichungen gesucht, anschließend von Fachleuten auf ihre Relevanz hin überprüft und ggf. annotiert, d.h. die wissenschaftlichen Ergebnisse in die Datenbank übernommen. Unser Ziel ist es, die Fachleute (Biologen) durch den Einsatz von Text-Miming-Methoden bei der Literaturvorauswahl gezielt zu unterstützen, um so den Annotationsaufwand zu verringern.

Die im Rahmen des Intergenomicsprojekts entwickelte Datenbank PRODORIC [41] enthält Daten über DNA-Bindunsstellen prokariotischer transkriptionaler Regulatoren. Diese Daten sind die Ergebnisse spezieller Experimente wie zum Beispiel DNAse I Footprints oder ElectroMobility gel Shift Assays (EMSA), die typischerweise nicht in den Abstracts der entsprechenden Veröffentlichungen genannt werden. Die Suche mit allgemeineren Stichwörtern, die das Thema beschreiben, wie etwa „gene regulation“, „promoter“ oder „binding site“ ergibt um die 15.000 Treffer, von denen nur etwa 10-20% auch tatsächlich interessierende Daten enthalten. Um diese Artikel allerdings zu finden, ist es bisher trotzdem notwendig, in alle Artikel der Ergebnissmenge hineinzuschauen. Dabei suchen die Fachleute vor allem nach solchen Publikationen, die Abbildungen von DNAse I Footprints oder EMSA assays enthalten, da sie belegen, dass tatsächlich die entsprechenden Experimente vorgenommen wurden.

Das Problem bei Veröffentlichungen über diese speziellen experimentellen Daten ist, dass die Suche nach entsprechenden Stichwörtern in Abstracts oder auch im kompletten Text nicht wirklich weiterhilft: In den Abstracts werden die entsprechenden Experimente nicht erwähnt und in den Veröffentlichungen selbst werden im Literaturüberblick oft diverse experimentelle Methoden diskutiert, die das gesamte Gebiet betreffen. Die gesuchten Stichwörter sind somit oft auch in den Veröffentlichungen zu finden, die ganz andere Experimente beschreiben und sich von anderen Ansätzen abgrenzen.

Es gibt allerdings Stellen in den Publikationen, an denen die gesuchten Experimentbezeichnungen zuverlässig und mit einer geringen Fehlerrate zu finden sind: die Bildunterschriften zu den erwähnten Abbildungen der Experimente. Auf dieser Idee basiert die von uns entwickelte Suchmaschine *CaptionSearch*, deren Oberfläche in Abbildung 6 zu sehen ist.

Ein wissenschaftliches Dokument, das heutzutage meist im PDF-Format vorliegt, ist schwieriger zu analysieren, als es im ersten Moment scheint. Leser können die Struktur und die Bedeutung der verschiedenen Schriftarten und Bilder leicht erkennen. Diese Information ist aber nicht ohne weiteres zugänglich, wenn man das PDF-Dokument – ohne menschlichen Leser – analysieren will. Während beispielsweise ein HTML-Dokument, aus dem alle Tags herausgelöscht wurden, einen lesbaren Text ergibt, sieht die Sache bei PDF-Dokumenten deutlich komplizierter aus: Im Extremfall ist in einem PDF-Dokument für jeden Buchstaben und für jedes Bild einzeln festgelegt, wo auf der Seite sie positioniert werden. Tatsächlich sind es meistens Textstücke und nicht einzelne Buchstaben, die über gemeinsame Positionsangaben verfügen, wobei Anfang und Ende dieser Textstücke aber keine semantische Bedeutung haben, also nicht mit z.B. Absatzanfang und -Ende zusammenfallen. Die meisten PDF-Konverter emulieren diese Stück-für Stück-Positionierung in ASCII oder HTML. Informationen über die Lesereihenfolge oder semantische Zusammenhänge von Textabschnitten gehen dabei aber verloren und müssen wiederhergestellt werden.

Im Folgenden geben wir einen groben Überblick darüber, wie unsere Bildsuchmaschine arbeitet: Abbildung 7 stellt alle Schritte vom Herunterladen der Paper bis zur Anfrageausführung dar.

Im ersten Schritt wird zunächst die komplette infrage kommende Literatur heruntergeladen. Der zweite Schritt gliedert sich in drei Teile. Teil A funktioniert im Prinzip wie jeder beliebige PDF-zu-Text-Konverter und basiert auf dem in Java implementierten Konverter PDFBox [37], den wir so angepasst haben, dass die Positions- und Layoutinformationen des Textes erhalten

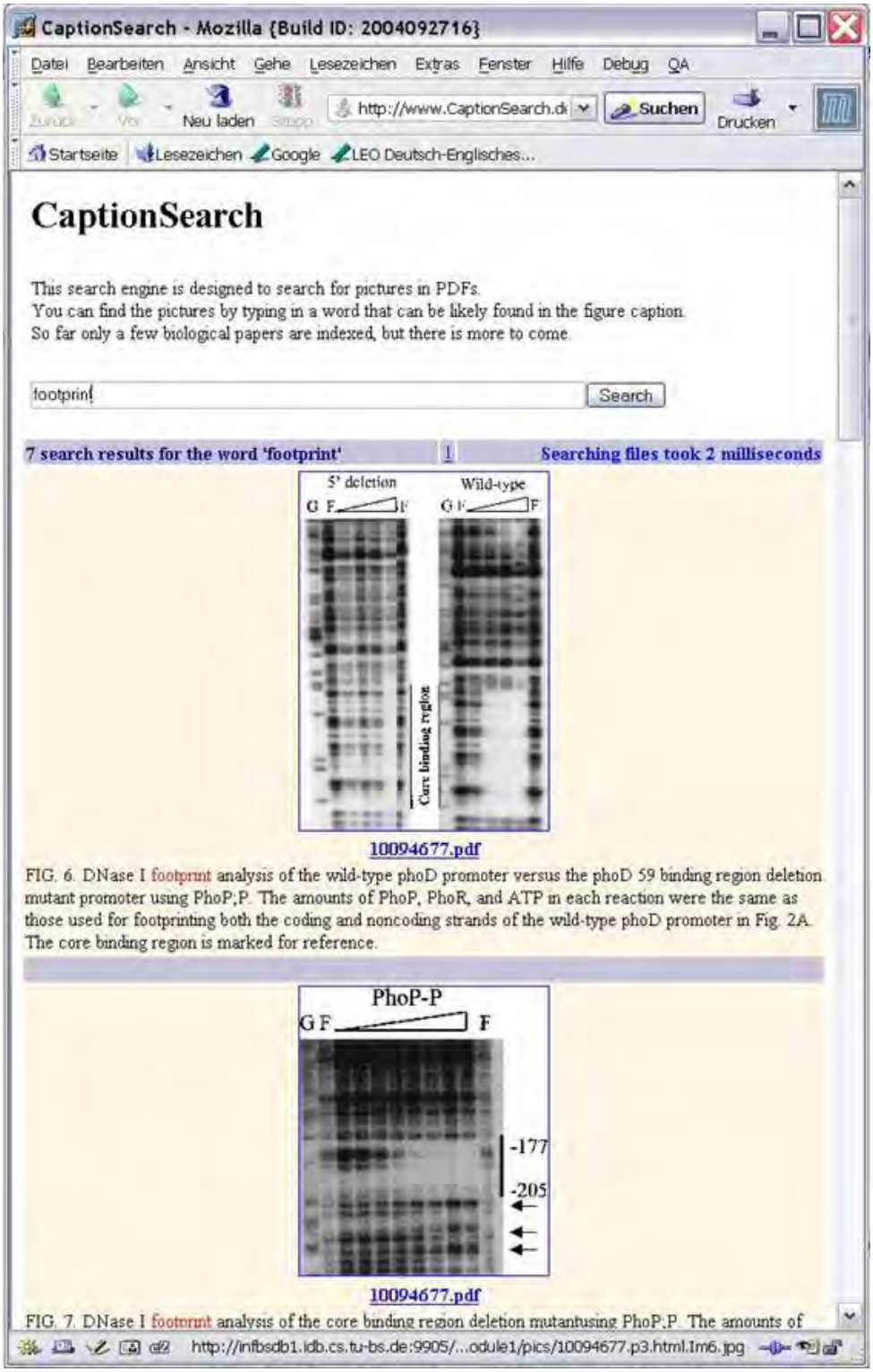



Abbildung 6: CaptionSearch

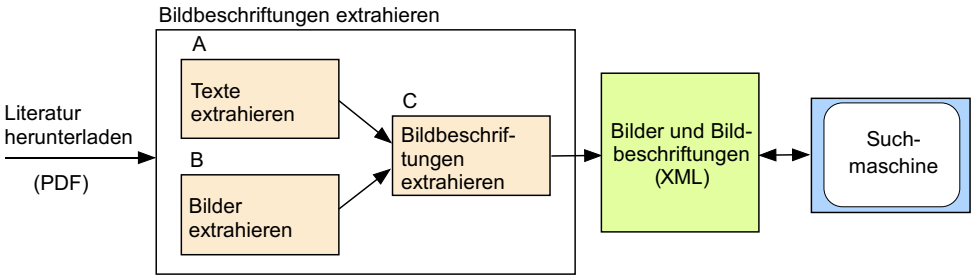



Abbildung 7: Datenfluss

bleiben. In Teil B werden die Bilder extrahiert und in Dateien abgespeichert. In den Text werden anstelle der Bilder Textblöcke integriert, die Informationen über die Position des Bildes und den Namen der neuen Bilddatei enthalten. Teil C schließlich besteht aus einem Algorithmus, der in jeder Datei die zusammengehörigen Paare von Bildern und Bildunterschriften sucht.

Die Schwierigkeit bei der Zuordnung von Bildern und Bildunterschriften besteht – wie oben bereits angedeutet – darin, dass in PDF-Dokumenten nur die Positionen von Bildern und Textpassagen, die auch bis zu einzelnen Buchstaben kurz sein können, gespeichert wird, nicht aber Informationen wie z.B. die Lesereihenfolge oder weitere Angaben zur Struktur. Es müssen also Textblöcke in der Nähe des jeweiligen Bildes gefunden werden. Dabei gibt es, anders als der Name suggeriert, verschiedene Möglichkeiten, wo eine Bildunterschrift positioniert sein kann, nämlich über, unter oder neben dem Bild. Weitere Schwierigkeiten kommen bei zweispaltigem Satz oder nebeneinander stehenden Abbildungen hinzu. Der Algorithmus wird ausführlich in [39] erläutert.

Anschließend werden die bisher zusammengetragenen Informationen in je eine XML-Datei pro PDF-Datei geschrieben. Die XML-Dateien werden indiziert sodass sie mit unserer Suchmaschine abgefragt werden können.

In [38] stellen wir Evaluationsergebnisse vor, die zeigen, dass unser Ansatz durchaus vielversprechend ist. Hinzu kommt, dass den Fachleuten oft schon ein Blick auf das Miniaturbild genügt, um die Qualität der präsentierten Experimente einschätzen zu können. Für sie ist die Arbeit mit unserer Suchmaschine allein schon deshalb nützlich, weil sie nun schneller größere Mengen an Literatur durchforsten können.

Der hier verfolgte Ansatz der Layoutanalyse könnte außer zur Bildsuche auch zur Tabellensuche verwendet werden. Dabei können semi-automatisch Daten direkt aus der Veröffentlichung in eine Datenbank übernommen werden.

## 4.2 Diskrete Modellierung und Simulation von Signaltransduktionswegen

Signaltransduktionswege beschreiben die Reaktion von Zellen auf extrazelluläre Signale. Beispielsweise können solche extrazellulären Signale durch Rezeptoren in der Zellmembran aufgenommen und durch Signalkaskaden in Form von Protein-Protein-Interaktionen innerhalb der Zelle weitergeleitet werden. Sie führen dann z.B. zu Soffwechselregulationen oder Genexpressionen. Die beteiligten Proteine werden auch Signalmoleküle genannt. Man kann sie sich als molekulare Schalter vorstellen, die über mindestens zwei Zustände verfügen.

Die Datenbank TRANSPATH der Firma Biobase [51] stellt Informationen über Signaltransduktionswege, die beteiligeten Moleküle sowie die chemischen Reaktionen zur Verfügung. Dabei handelt es sich hauptsächlich um Signaltransduktionswege in Säugetieren. In der zur Zeit aktuellen Version 7.1 sind Informationen über 50949 Moleküle, 21570 Gene, 91923 Reaktionen sowie 57 Signaltransduktionswege abgelegt. Hinzukommen 29779 Verweise auf die zugrunde liegenden Publikationen.

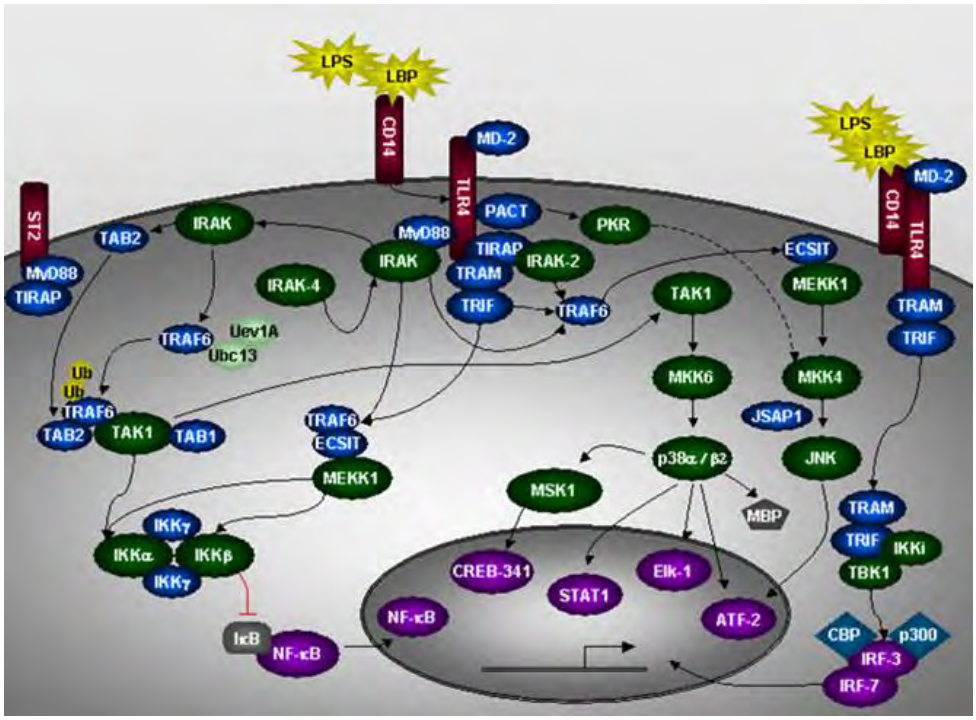



Abbildung 8: TLR4

Signaltransduktionswege lassen sich aus unterschiedlichen Blickwinkeln betrachten, je nachdem ob man einen ersten Einblick in den Ablauf, die Beteiligung von Molekülen und deren Interaktionsorte bekommen möchte oder für ein tieferes Verständnis des Pathways an den zugrunde liegenden chemischen Reaktionen interessiert ist. Ersteres wird auch als semantische und letzteres auch als mechanistische Sichtweise bezeichnet.

Bei der semantischen Sichtweise handelt es sich um qualitative Betrachtungen, die in TRANSPATH durch abstrakte, handgemachte und sensitive Maps unterstützt werden. Abbildung 8 zeigt den TLR4-Pathway, der die Immunreaktion der Lungenepitheliumzelle auf einen Angriff des Bakteriums *Pseudomonas aeroginosa* beschreibt, in einer solchen Darstellung. Sensitiv ist die Map in dem Sinne, dass bei einem Klick auf die einzelnen Komponenten die entsprechenden Einträge in der Datenbank angezeigt werden.

Der mechanistischen Sichtweise liegen quantitative Betrachtungen zugrunde, d.h. es interessiert hier, wieviele Moleküle welcher Art an einer bestimmten Reaktion beteiligt sind bzw. als Ergebnis derselben entstehen. Solche Betrachtungen werden in TRANSPATH durch den sogenannten PathwayBuilder unterstützt, der per Mausklick eine Übersicht automatisch aus den zugrundeliegenden chemischen Reaktionen generiert (vgl. Abb. 9).

Beide Sichtweisen beschreiben aus Sicht der Informatik Interobjektverhalten. Einen weiteren Blickwinkel stellt das Intraobjektverhalten dar. Übertragen auf die hier betrachteten biologischen Phänomene wäre das beispielsweise die Sicht in spezielle Molekülkomplexe hinein. Dabei stehen dann nicht mehr die kompletten Abläufe im Pathway im Vordergrund, sondern es interessieren im Detail die chemischen Vorgänge in einem Molekülkomplex. Auch hier kann, je nach Abstraktionsgrad, wieder zwischen semantischer und mechanistischer Sichtweise unterschieden werden.

Die Visualisierungstools der TRANSPATH helfen zwar dabei, einen Überblick über die Signaltransduktionswege zu erlangen, sie sind aber letztlich doch nur statische Darstellungen dynamischer Vorgänge. Hier bietet es sich an, Modellierungs- und Simulationswerkzeuge der Informatik einzusetzen, um dynamische Vorgänge auch dynamisch darzustellen. Je nach Blickwinkel

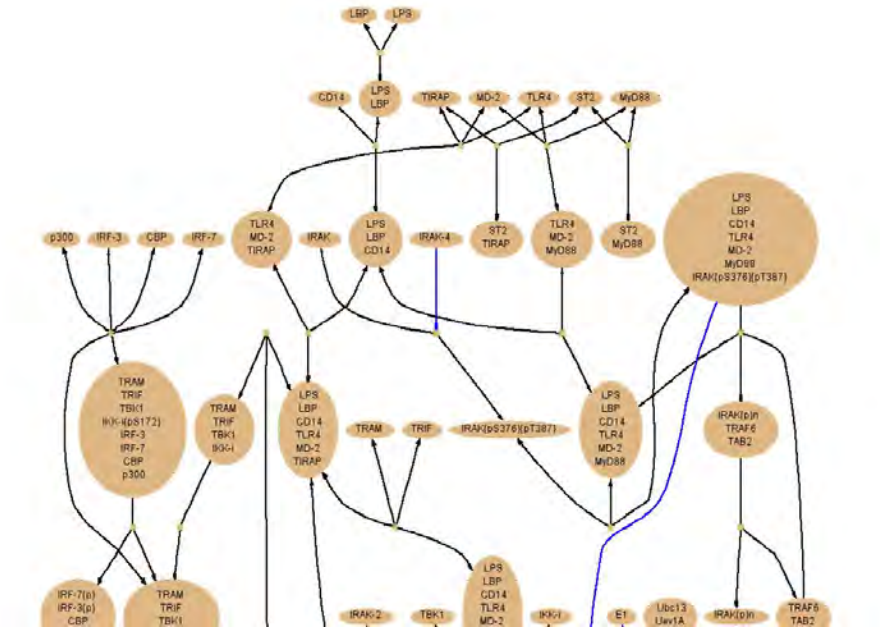



Abbildung 9: TLR4 – vom PathwayBuilder erzeugte Darstellung

eignen sich dabei bestimmte Modellierungssprachen besser als andere.

Das Interobjektverhalten in der semantischen Sichtweise kann mit Life Sequence Charts (LSC) [9] dargestellt und mit der Play-Engine [22] simuliert werden. Vereinfacht ausgedrückt stellt ein LSC die Elemente dar, die an einer Interaktion teilnehmen. In einem solchen Diagramm steht die Abfolge von Nachrichten im Vordergrund. Die beteiligten Moleküle (Objekte) werden oben im Diagramm angeordnet und enthalten eine senkrechte Lebenslinie an der entlang die Nachrichten ausgetauscht werden. Abbildung 10 zeigt einen Ausschnitt aus einer LSC-Darstellung des TLR4-Pathways [35]. Zusätzlich bietet die Play-Engine die Möglichkeit, graphische Oberflächen einzubinden, sodass z.B. animierbare Versionen der oben erwähnten Maps erstellt werden können.

Zur Darstellung des Interobjektverhaltens in der mechanistischen Sichtweise bieten sich Petri-Netze an [23]. Petri-Netze [46, 42] sind bipartite, gerichtete Graphen, die aus Stellen und Transitionen bestehen. Stellen visualisieren beispielsweise die Anzahl der beteiligten Moleküle eines Typs, Transitionen beschreiben die Reaktionen. Es gibt eine Vielzahl verschiedener Petri-Netz-Arten. Wir verwenden gefärbte Petri-Netze (Coulerd Petri Nets, CPNs), weil durch die Farben unterschiedliche Moleküle visualisiert werden können [16]. Als Simulationstool kommt Design/CPN [8] zum Einsatz. Eine CPN-Darstellung des TLR4-Pathways ist in Abbildung 11 zu sehen.

Das Intraobjektverhalten stellen wir mit Statecharts [21] und dem Rhapsodytool [27] dar. Ein Statechart ist ein Zustandsdiagramm, das die dynamischen Aktivitäten eines Objekts, in unserem Fall also eines Moleküls bzw. Molekülkomplexes, beschreibt und dessen Zustände, Aktivitäten, Reaktionen und Ereignisse während seiner Lebenszeit visualisiert. Abbildung 12 zeigt ein Statechart-Diagramm aus dem TLR4-Pathway [31].

Zur Zeit werden mit den genannten Sprachen und Werkzeugen Signaltransduktionswege von Hand modelliert und simuliert. Ziel ist es aber, die biologischen Daten aus der TRANSPATH und später auch aus anderen Datenbanken automatisch in die Werkzeuge zu importieren und anschließend ihr Verhalten zu simulieren und zu validieren.

Hierzu bedienen wir uns der Technik des Model Driven Engineerings (MDE) – der modellgetriebenen Softwareentwicklung, die es auf Basis von Metamodellen ermöglicht, Modell-

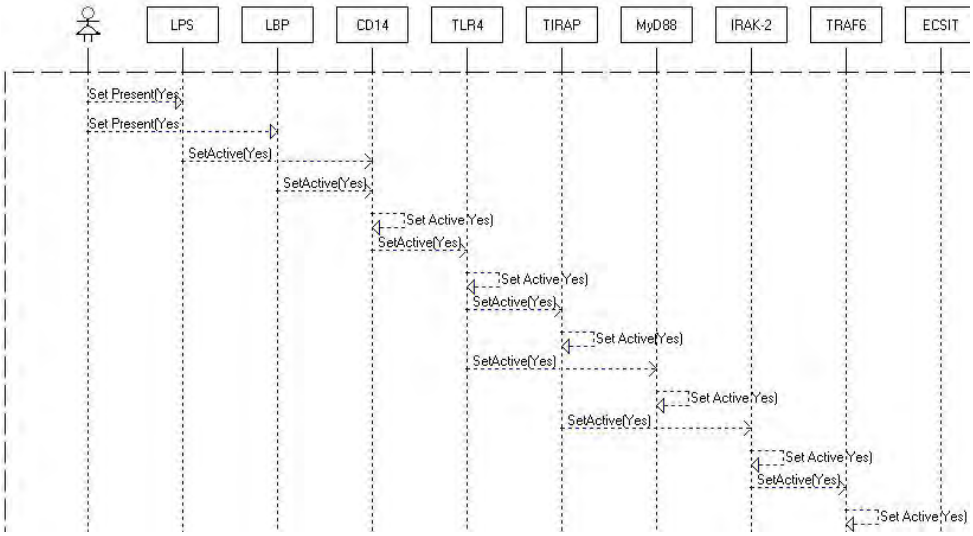



Abbildung 10: TLR4 – Ausschnitt aus einem LSC-Diagramm

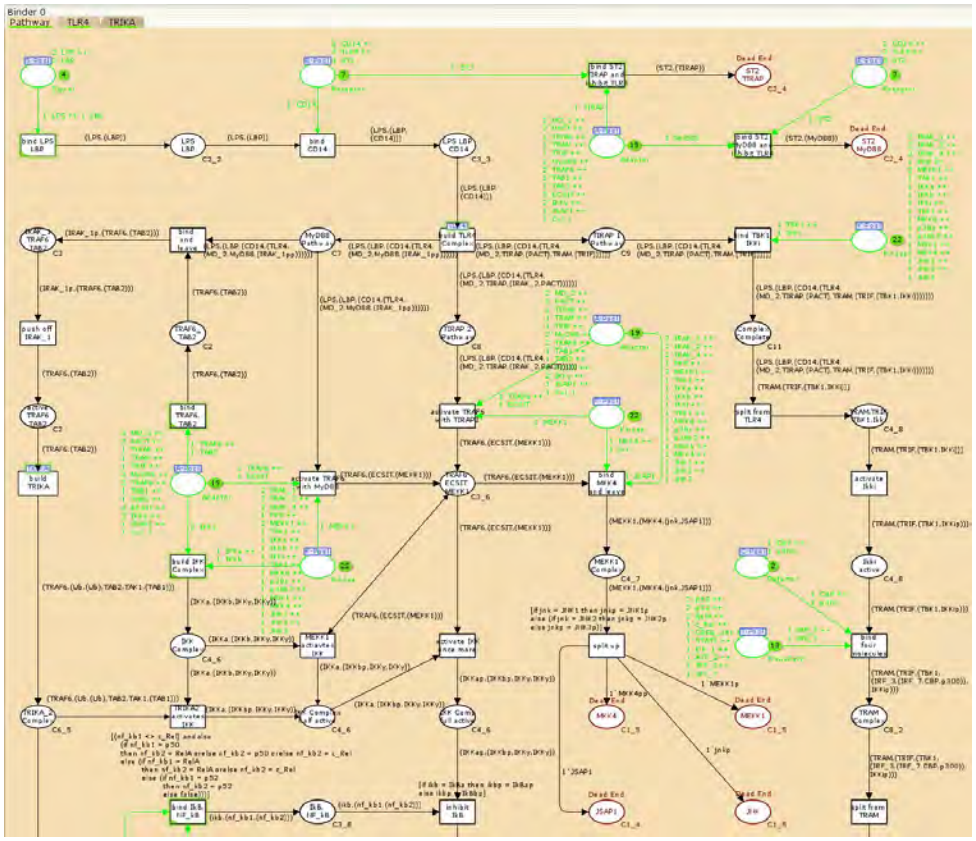


Abbildung 11: TLR4 – Ausschnitt aus der Petri-Netz-Darstellung

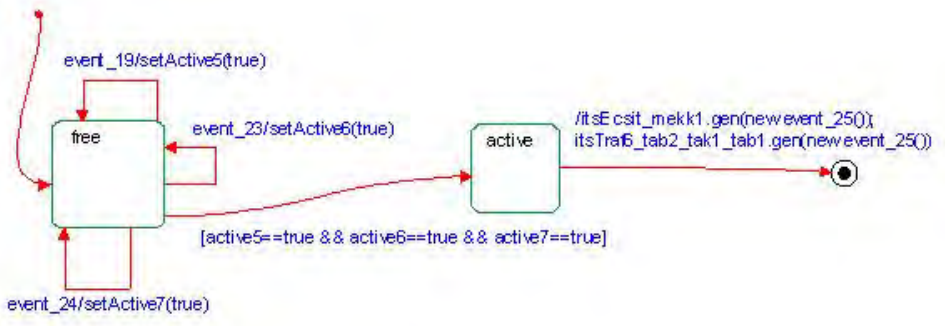



Abbildung 12: TLR4 – Statechart-Diagramm

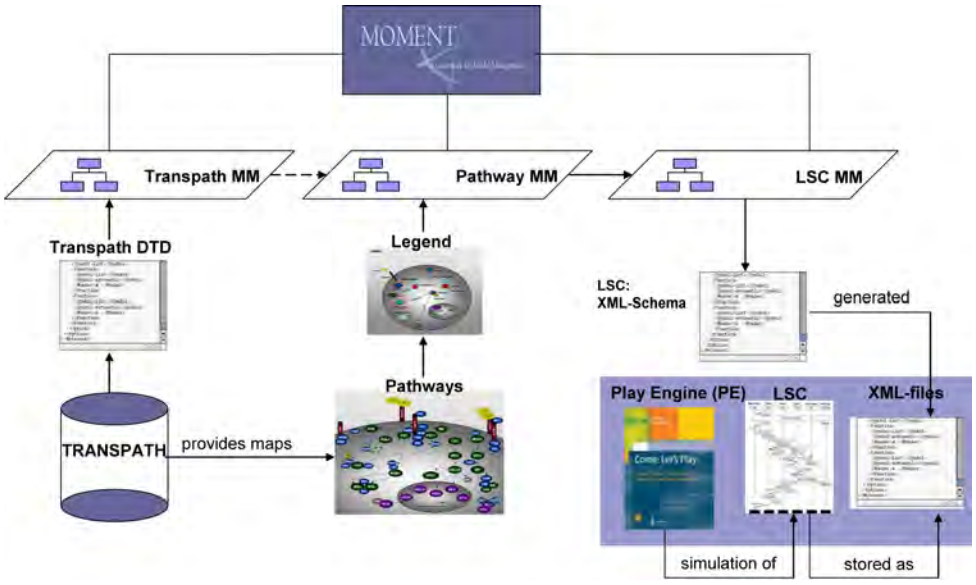



Abbildung 13: Erzeugung von LSC- aus Pathway-Modellen

Transformationen durchzuführen. Zum Einsatz kommt dabei das MDE-Werkzeug MOMENT [40], in dem – in unserem Fall – zum einen ein Metamodell für Pathways, zum anderen Metamodelle für die Play-Engine, Rhapsody und Design/CPN sowie Abbildungsregeln zwischen den Metamodellen abgelegt werden. Als Eingabe erhält MOMENT dann Pathway-Modelle und erzeugt als Ausgabe Modelle, die wiederum als Eingabe für die oben genannten Simulationswerkzeuge dienen. Abbildung 13 zeigt diesen Ansatz für die Erzeugung von LSC-Modellen. Die nächste Aufgabe ist es daher, eine Software zu entwickeln, die die TRANSPATH-Datenbank mit dem MOMENT-Tool verbindet. Sie muss eine graphische Oberfläche zur Verfügung stellen, auf der der Benutzer Pathways aus der Datenbank auswählen und festlegen kann, für welche der Simulationstools MOMENT Modelle daraus erzeugen soll.

### 4.3 Koevolution von Datenbankschemata und Ontologien

Ontologien sind eine der Schlüsseltechniken für Datenintegration und Metadatenbanken, da sie Datenbanken auf semantischer Ebene verknüpfen können. Ein arbeitsintensives Problem sind Änderungen im Datenbankschema, da hierdurch bestimmte Teile der Ontologie manuell angepasst werden müssen. In diesem Teilprojekt wird ein Ansatz verfolgt, der dem Datenbankschema und der zugehörigen Ontologie ermöglicht, sich zu ändern und weiterzuentwickeln, ohne

die Verbindung zueinander zu verlieren. Dazu wird eine intitiale Ontologie zu einem gegebenen Datenbankschema generiert, Annotationstechniken für die Ontologie untersucht und ein halbautomatischer Prozess implementiert, der auftretende Änderungen auf die Ontologie möglichst verlustfrei übertragen kann.

Datenbankschemata in relationalen Datenbanken enthalten unter anderem die Struktur der gespeicherten Daten. Hiermit können, über Angaben wie etwa Tabellen- oder Spaltennamen, die gewünschte Informationen selektiert werden. Diese Namen sind aber nur einfache Zeichenketten und besitzen keine definierte semantische Bedeutung. Die Namen können im Normalfall auch nicht über mehrere Datenbanken hinweg verwendet werden, da es keine gemeinsamen Namenskonventionen gibt. Bei Data-Warehouses oder im Bereich der Bioinformatik sind häufig eine Vielzahl von Datenbanken involviert. Wenn nun bestimmte Informationen in diesen Datenbanken oder einfach nur die passende Datenbank für ein gegebenes Thema gesucht werden, sind semantische Annotationen hilfreich [48].

Ontologien beschreiben die Konzepte und Relationen einer bestimmten Domäne. Sie bieten ein erheblich reichhaltigeres Datenmodell und werden eingesetzt, um Semantik sowohl maschinen- als auch menschenlesbar darzustellen [30]. Es wurde gezeigt, dass Ontologien Strukturen von XML-Dokumenten [13], wie auch von Datenbanken abbilden können. Momentan werden Ontologien hauptsächlich in Form eines kontollierten Vokabulars, wie etwa GeneOntology [1], eingesetzt, um die Begriffe für die Annotation zu vereinheitlichen [49].

Ein Ziel dieses Teilprojektes ist es, biologische Datenbanken mit Hilfe von Ontologien zu annotieren. Sie alle enthalten spezifische Forschungsdaten, die sich nutzbringend kombinieren lassen [34]. Mit der Datenbankontologie kann die Domäne der Datenbank beschrieben und die notwendigen Informationen für den Datenbankzugriff gespeichert werden. Sie liefern beispielsweise einem Programm zur Integration die notwendigen Daten, ohne dass dieses auf unterschiedliche Bezeichner, etwa bei Tabellennamen, Rücksicht nehmen muss. Ebenso läßt sich damit unterscheiden, zu welchem Organismus die in einer Datenbank gespeicherten Moleküle gehören, etwa zum Mensch, zur Maus oder zur Hefe. Die Annotation kann somit als Grundlage für andere Anwendungen, wie Datenintegration oder Anfragebearbeitung über heterogene Datenbanken, eingesetzt werden [44, 50].

Obwohl Ontologien große Vorteile bieten, werden sie bislang selten zur Beschreibung von Datenbankschemata eingesetzt. Dies hat zwei Gründe. Zunächst unterstützen die meisten Datenbank-Managementsysteme keine Ontologien zur semantischen Annotation. Daraus ergibt sich, dass die Ontologie extern gespeichert werden muss und Änderungen im Datenbankschema manuell auf die Ontologie übertragen werden müssen. Denn die Ontologie kann nur solange verwendet werden, wie sie auch zum Datenbankschema passt. Die hier betrachteten biologischen Forschungsdatenbanken ändern sich häufig genug, so dass ein beträchtlicher Teil der Arbeitszeit hierfür investiert werden müsste. Im Bereich der Datenintegration gibt es dafür bereits verschiedene Ansätze. Diese verwenden allerdings spezialisierte Datenmodelle, wie etwa das Hypergraph Datenmodell in [14], und nicht allgemein nutzbare Modelle, wie Ontologien.

Die im laufenden Betrieb entstehende Datenbankschemaevolution und die wünschenswerte Ontologieevolution scheinen sich gegenseitig zu behindern. Die Ontologie darf nicht bei jeder Änderung neu generiert werden, da sonst die bisher erfolgte Annotation gelöscht wird. Hier wird daher ein Synchronisationsmechanismus zwischen den beiden entwickelt, der die auftretende Arbeitslast für den Anwender reduziert.

Die Koevolution von Datenbankschemata und Ontologien beginnt mit der Abbildung eines Datenbankschemas. Diese Abbildung fungiert als Bindeglied zwischen der Datenbank und der annotierten Semantik, so dass Verknüpfungen zu den einzelnen Elementen des Datenbankschemas bereit gestellt werden. Die Generierung ist ein automatischer Prozess, der als Ausgangspunkt für die Annotation dient. Für die Ontologien verwenden wir die Web Ontology Language (OWL) [4], dies ermöglicht eine Verwendung der Annotationen in verschiedenen OWL-Anwendungen. OWL ist eine Empfehlung des World Wide Web Consortiums (W3C) und unterstüzt besonders das Zu-

sammenspiel mehrerer Ontologien in dieser Sprache. Diese Möglichkeit fehlte einigen der älteren Ontologie-Sprachen. Im hier beschriebenen Ansatz wird kein globales Schema erzeugt, sondern jede Datenbankontologie enthält Instanzen, die das Datenbankschema einer bestimmten Datenbank beschreiben und die wiederum miteinander leicht kombiniert werden können. beschrieben [10], hier wird stattdessen mit OWL Lite nur die kleinste Menge an Sprachmitteln verwendet. Dadurch wird die Weiterverarbeitung der generierten Ontologien erheblich vereinfacht.

Es wurde zunächst eine abstrakte Datenbankontologie mit den Konzepten *Datenbank*, *Relation* und *Attribut* entworfen und ein Programm implementiert, die ein gegebenes Datenschema in Instanzdaten dieser Ontologie umwandelt [32]. Die resultierende Datenbankontologie ist hierarchisch über die Objekteigenschaft *besteht-aus* aufgebaut, so dass eine Datenbank aus mehreren Relationen besteht, die wiederum aus ihren Attributen besteht. Die Datentypen aus dem Schema werden auf XML-Schema abgebildet.

Im nächsten Schritt wurden die auftretenden Änderungen am Datenbankschema in Form von Änderungsprimitiven klassifiziert und Strategien implementiert, um diese Primitive auf Ontologien anzuwenden [33]. Als Beschreibung der Datenbankänderung wird SQL verwendet und drei Klassen von Änderungsprimitiven unterschieden: *Erzeugen*, *Löschen* und *Ändern*. Dabei werden nur Schemaänderungen betrachtet, die Einfluss auf die Ontologie haben, wie etwa das Erstellen einer neuen Relation oder die Umbenennung eines Attributs.

Aktuell wird untersucht, wie die Schemaänderungen protokolliert werden können. Außerdem ist noch zu klären, wie der Annotationsprozess der Ontologie unterstützt werden kann und welche Evolutionsschritte der Ontologie von der aktuellen Implementierung unterstützt werden können. Daraus werden dann Erweiterungen für die einzelnen Teilbereiche entwickelt.

## 5 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 [36]:

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

Drittmittelgeber wie die DFG und das BMBF 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 einigen Jahren 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 danken wir C. Choi, D. Fellner, R. Hehl, D. Jahn, R. Münch und K. Neumann. Unser ganz besonderer Dank gilt J. Weimar, von dem wir viel über Bioinformatik erfahren haben, und der wesentliche Materialien zu einer früheren Version dieser Ausarbeitung beigesteuert hat. Für alle verbliebenen Fehler, Auslassungen, Verzerrungen und Ungenauigkeiten sind jedoch allein die Autorinnen und Autoren verantwortlich.*

## Literatur

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

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

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

[4] S. Bechhofer, F. v. Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, and L.A. Stein. OWL web ontology language reference, February 2004. World Wide Web Consortium.

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

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

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

[8] CPN group at the University of Aarhus. Design/cpn. http://www.daimi.au.dk/designCPN/, 2006.

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

[10] C.P. de Laborda and S. Conrad. Relational.OWL - A Data and Schema Representation Format Based on OWL. In *Conceptual Modelling 2005 (APCCM05)*, pages 89–96. Australian Computer Society, 2005.

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

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

[13] M. Erdmann and R. Studer. How to structure and access XML documents with ontologies. *Data Knowl. Eng.*, 36(3):317–335, 2001.

[14] H. Fan and A. Poulovassilis. Schema evolution in data warehousing environments - a schema transformation-based approach. In *ER 2004, 23rd Int. Conf. on Conceptual Modeling, Proc.*, volume 3288 of *LNCS*, pages 639–653. Springer, 2004.

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

[16] N. Fleischer. Modellierung und Simulation der P. aeruginosa Infektion mit Petri Netzen. Diplomarbeit, Technische Universität Braunschweig, 2005.

[17] M.Y. Galperin. The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Research*, 33(Database-Issue):5–24, 2005.

[18] M.Y. Galperin. The Molecular Biology Database Collection: 2006 update. *Nucl. Acids Res.*, 34(suppl_1):D3–5, 2006.

[19] 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.

[20] 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.

[21] D. Harel. Statecharts: A visual Formalism for complex systems. *The Science of Computer Programming*, (8):231–274, 1987.

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

[23] M. Heiner and I. Koch. Petri Net Based Model Validation in Systems Biology. *ICATPN*, pages 216–237, 2004.

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

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

[26] www.ilogix.com/.

[27] Ilogix. Rhapsody. http://www.ilogix.com/homepage.aspx, 2006.

[28] 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.

[29] 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.

[30] V. Kashyap and A. Sheth. Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In *Cooperative Information Systems*, pages 139–178. Academic Press, San Diego, 1998.

[31] B. Kass. Modellierung und Simulation der Epithelzellen-Infektion durch Pseudomonas aeruginosa mit Statecharts. Studienarbeit, Technische Universität Braunschweig, 2005.

[32] A. Kupfer and S. Eckstein. Coevolution of database schemas and associated ontologies in biological context. In *22nd British National Conference on Databases*, volume 2, pages 45–50. University of Sunderland Press, 2005.

[33] A. Kupfer, S. Eckstein, K. Neumann, and B. Mathiak. Keeping track of changes in database schemas and related ontologies. In *7th Int. Baltic Conference on Databases and Information Systems*, 2006. Wird veröffentlicht.

[34] Z. Lacroix and T. Critchlow. *Bioinformatics – Managing Scientific Data*. Morgan Kaufmann, 2003.

[35] H. Langhorst. Entwurf und Realisierung einer GUI-Anwendung zum Datenaustausch zwischen TRANSPATH und der Play-Engine. Diplomarbeit, Technische Universität Braunschweig, 2006.

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

[37] B. Litchfield. PDFBox. http://www.pdfbox.org/ or http://sourceforge.net/, 2004.

[38] B. Mathiak, A. Kupfer, R. Münch, C. Täubner, and S. Eckstein. Analysing layout information: searching pdf documents for pictures. In *Lernen, Wissensentdeckung und Adaptivität (LWA) 2005*, pages 190–195. DFKI, 2005.

[39] B. Mathiak, A. Kupfer, R. Münch, C. Täubner, and S. Eckstein. Improving literature preselection by searching for images. In *Knowledge Discovery in Life Science Literature, PAKDD 2006*, volume 3886 of *LNCS*, pages 18–28. Springer, 2006.

[40] Moment. http://moment.dsic.upv.es/, 2005.

[41] R. Münch, K. Hiller, H. Barg, H. Heldt, S. Linz, E. Wingender, and D. Jahn. Prodoric: prokaryotic database of gene regulation. *Nucleic Acids Research*, 31(1):266–269, 2003.

[42] T. Murata. Petri-Nets: Properties, Analysis and Applications. *Proceeding of the IEEE*, 77(4), 1989.

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

[44] C. Ben Necib and J.C. Freytag. Query Processing Using Ontologies. In *17th Int. Conf. CAiSE 2005*, volume 3520 of *LNCS*, pages 167–186. Springer, 2005.

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

[46] C. A. Petri. Kommunikation mit Automaten; Dissertation. Rhein.-Westf. Inst. für Instr. Mathematik an der Universität Bonn, 1962.

[47] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/, 2004.

[48] P. Spyns, R. Meersman, and M. Jarrar. Data modelling versus ontology engineering. *SIGMOD Rec.*, 31(4):12–17, 2002.

[49] R. Stevens, C. Wroe, P. Lord, and C. Goble. Ontologies in Bioinformatics. Int. Handbooks on Information Systems, pages 635–657. Springer Verlag, Heidelberg, 2004.

[50] H. Stuckenschmidt and F. v. Harmelen. *Information Sharing on the Semantic Web*. Advanced Information and Knowledge Processing. Springer Verlag, 2005.

[51] http://www.biobase.de/pages/index.php?id=63.

[52] J. Weimar, pers. Mitteilung.