Modellierung von Metabolismus, Transkriptom und Zellentwicklung bei *Arabidopsis*, *Listerien* und anderen Organismen

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Roland Schwarz
aus
Kleve

Würzburg, 2008

Eingereicht am: .....

Mitglieder der Promotionskommission:

Vorsitzender: .....

Gutachter: Prof. Dr. Thomas Dandekar

Gutachter: PD Dr. Susanne Berger

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

Zusammenfassung

Im gleichen Maße wie informatisches Wissen mehr und mehr in den wissenschaftlichen Alltag aller Lebenswissenschaften Einzug gehalten hat, hat sich der Schwerpunkt bioinformatischer Forschung in stärker mathematisch und informatisch-orientierte Themengebiete verschoben. Bioinformatik heute ist mehr als die computergestützte Verarbeitung großer Mengen an biologischen Daten, sondern hat einen entscheidenden Fokus auf der *Modellierung* komplexer biologischer Systeme. Zur Anwendung kommen hierbei insbesondere Theorien aus dem Bereich der Stochastik und Statistik, des maschinellen Lernens und der theoretischen Informatik.

In der vorliegenden Dissertation beschreibe ich in Fallstudien die systematische Modellierung biologischer Systeme aus einem informatisch - mathematischen Standpunkt unter Anwendung von Verfahren aus den genannten Teilbereichen und auf unterschiedlichen Ebenen biologischer Abstraktion. Ausgehend von der Sequenzinformation über Transkriptom, Metabolom und deren regulatorischer Interaktion hin zur Modellierung von Populationseffekten werden hierbei aktuelle biologische Fragestellungen mit mathematisch - informatischen Modellen und einer Vielzahl experimenteller Daten kombiniert. Ein besonderer Augenmerk liegt dabei auf dem Vorgang der Modellierung und des Modellbegriffs als solchem im Rahmen moderner bioinformatischer Forschung.

Im Detail umfassen die Projekte (mehrere Publikationen) die Entwicklung eines neuen Ansatzes zur Einbettung und Visualisierung von Multiplen Sequenz- und Sequenz-Strukturalignments, illustriert am Beispiel eines Hemagglutininalignments unterschiedlicher H5N1 Varianten, sowie die Modellierung des Transkriptoms von *A. thaliana*, bei welchem mit Hilfe einer kernelisierten nicht-parametrischen Metaanalyse neue, an der Infektionsabwehr beteiligten, Gene ausfindig gemacht werden konnten. Desweiteren ist uns mit Hilfe unserer Software *YANAsquare* eine detaillierte Untersuchung des Metabolismus von *L. monocytogenes* unter Aktivierung des Transkriptionsfaktors prfA gelungen, dessen Vorhersagen durch experimentelle ^{13}C Isotopologstudien belegt werden konnten. In einem Anschlußprojekt war der Zusammenhang zwischen Regulation des Metabolismus durch Regulation der Genexpression und der Fluxverteilung des metabolischen Steady-State-Netzwerks das Ziel. Die Modellierung eines komplexen organismischen Phänotyps, der Zellgrößenentwicklung der Diatomee *Pseudo-nitzschia delicatissima*, schließt die Untersuchungen ab.

Abstract

In the same way that informatical knowledge has made its way into almost all areas of research in the Life Sciences, the focus of bioinformatical research

has shifted towards topics originating more in the fields of mathematics and theoretical computer science. Bioinformatics today is more than the computer-driven processing of huge amounts of biological data, but it has a special focus on the emphmodelling of complex biological systems. Of special importance hereby are theories from stochastics and statistics, from the field of machine learning and theoretical computer science.

In the following dissertation, I describe the systematic modelling of biological systems from an informatical-mathematical point of view in a case studies approach, applying methods from the aforementioned areas of research and on different levels of biological abstraction. Beginning with the sequence information itself, followed by the transcriptome, metabolome and the interaction of both and finally population effects I show how current biological questions can be tackled with mathematical models and combined with a variety of different experimental datasets. A special focus lies hereby on the procedure of modelling and the concept and notion of a *model* as such in the framework of bioinformatical research.

In more detail, the projects contained the development of a new approach for embedding and visualizing Multiple Sequence and Structure Alignments, which was illustrated using a hemagglutinin alignment from different H5N1 variants as an example. Furthermore we investigated the *A. thaliana* transcriptome by means of a kernelized non-parametric meta-analysis, thus being able to annotate several new genes as pathogen-defense related. Another major part of this work was the modelling of the metabolic network of *L. monocytogenes* under activation of the transcription factor prfA, establishing predictions which were later verified by experimental ^{13}C isotopologue studies. Following this project we investigated the relationship between the regulation of metabolism by changes in the cellular genexpression patterns and the flux distributions of the metabolic steady-state network. Modelling of a complex organismal property, the cell size development of the planktonic diatom *Pseudo-nitzschia delicatissima* concludes this work.

Für Anna

Inhaltsverzeichnis

Teil I

Einführung: Modellierung in der Bioinformatik

Kapitel 1

Der Modellbegriff in der Bioinformatik

1.1 Verschiedene Definitionen

Den Begriff des Modells exakt und präzise zu formulieren fällt schwer angesichts der Unmengen von Ideen und Konzepten, die im allgemeinen Sprachgebrauch und über die Jahrtausende damit in Verbindung gebracht wurden. Ringt man sich schließlich doch zu einer Definition durch, so wird diese im Hinblick auf die Vielfalt, mit der der Begriff heute gebraucht wird, notgedrungen ziemlich allgemein ausfallen, möchte man sich nicht hier schon in Widersprüche verstricken. Denn wer liest schon gerne, dass es zwar viele Modelle gäbe, von denen sich auch viele zurecht so nannten, ausgerechnet das Eigene aber keines sei. Allgemein müsste eine Definition also sein, und aussehen könnte sie etwa so: „Modelle sind vereinfachte Ausschnitte der Wirklichkeit oder Möglichkeit“ und damit je nach Blickwinkel „[...] entweder Vorbild, Abbild, Entwurf oder Ersatz, aber auch Urbild, Muster und Form, Maß, Typ und Exemplar“ (Müller, 1983). Zugegeben, Modelle so allgemein und generell, sagen wir sparsam, zu definieren, dass nahezu alles mit ein wenig Zurechtrückerei des Blickwinkels ein Modell darstellt, mag aus Sicht von Ockhams weisem Satz „Entia non sunt multiplicanda praeter necessitatem“ nicht unvernünftig erscheinen (der im übrigen gar nicht von ihm war, wie wir alle wissen oder zumindest bei Thorburn (1918) nachlesen können). Aber vor allem erleichtert es die Aufgabe, eine Einleitung in eine Dissertation zu schreiben, in der es in erster Linie um *Modelle* geht und die selbst den Begriff sehr großzügig interpretiert, wie wir sehen werden, und auch lose Sammlungen mathematischer Gleichungen schon einmal gerne als *Modell* bezeichnet.

Doch bevor unverhaltene Kritik laut wird: Ganz am Zeitgeist vorbei scheint die Arbeit damit nicht zu zielen. Schon eine Suche bei *Amazon* nach Titeln, die den Begriff *Modell* enthalten, ergibt 3756 Ergebnisse aus allen

Sparten moderner Literatur. Vom „Modell Berlin“ und seiner Schulpolitik zu „Modellhubschrauber. Technik für Fortgeschrittene“, über „Das Modell des standortgerechten Kompostes“ hin zu „Höhere Mathematik: Linearalgebra und Linearmodelle“ oder dem Roman „Das Modell“ ist das *Modell* omnipräsent.

Kein Wunder, sind Modell, Model und Modul, die drei Modellbegriffe die nach Müller stets gemeinsam betrachtet werden müssen doch Ergebnis eines Verschmelzens fünf ursprünglich disjunkter Bedeutungsfelder aus der griechischen bzw. lateinischen Sprache (Müller, 2001):

1. *metron* (lat. *modus / modulus*): Maßstab, Maß, Grenze

2. *typos* (lat. *forma*): Form, Skulptur, Gußform

3. *paradeigma* (lat. *exemplar*): Maler- und Architekturmodell

4. Eine Reihe der eng verwandten griechischen Begriffe *eidos* (Gestalt, Form, Idee, Urbild, Bild), *eidolon* (Abbild, Trugbild) und *eikon* (Bild), die im lateinischen ihre Entsprechungen finden in *idea*, *imago* und *effigies* (Bild, Vorbild, Abbild, Vorstellung), *species* (Aussehen, Bild, Schein, Idee, Musterbild, Art) und *simulacrum* (Abbild, Muster, Puppe, Schatten-, Traum-, Trugbild, Charakterbild)

5. *keroplasto*, ein Begriff aus der plastischen darstellenden Kunst welches erst im 18. bis 19. Jahrundert als *Zero- oder Keroplastik* Einzug in die deutsche Sprache fand. Ursprünglich tauchte das Wort bereits in Platons “Timaios” auf, in welchem der Demiurg den Menschen wie ein Wachsmodellierer (*keroplastes*) erschuf.

Vermutlich aufgrund der vielfältigen Einflüsse und seines kontinuierlichen Wandels im Sprachgebrauch und Laufe der Zeit bleibt der Modellbegriff lange unscharf. Explizite Modellliteratur findet sich erst in der Mitte des letzten Jahrhunderts (vgl. Stachowiak (1973, 1965); Stoff (1969)). Rückblickend kann man heute viele frühere Auffassungen, Philosophien, Systeme oder Theorien als *Modelle* bezeichnen, auch wenn Denker wie Platon und Kopernikus, wie ein Galileo oder Newton, Darwin oder Marx sie selbst kaum so bezeichnet haben dürften (Müller, 2001). In diesem Zusammenhang vertritt Müller die Auffassung, dass eine alleinige Betrachtung des Modellbegriffs, herausgenommen aus seinem Umfeld, unstatthaft sei und zumindest „[...] System-, Analogie- und Funktionsdenken einerseits, Bild-, Symbol- und Abbildtheorie anderseits sowie Ideen-, Zeichen- und Bedeutungslehre [...] gleichgewichtig und in ihrer überaus engen Verzahnung mit der Problematik Modell behandelt werden [sollten].“ Er führt weiter aus, dass eine ernsthafte Beschäftigung mit der Materie „[...] sich dabei um Fragen der Erkenntnistheorie, Hermeneutik und Ontologie bewegen [müsste] und würde sich vom

Nominalismus über den Empirismus und Materialismus bis zur Existenzphilosophie erstrecken. Sie würde hinführen zur Informationstheorie und Kybernetik, zur Linguistik und Semiotik, aber auch zur Philosophy of Science, zur Logik und Metamathematik.“

Er mag Recht haben und uns bleibt nichts anderes als froh zu sein, dass wir dieses Kunststück erst gar nicht versuchen müssen und uns begnügen dürfen mit einem schwächeren und mehr allgemein gehaltenen Modellbegriff und die vorliegende Arbeit nicht umbenennen in „Stochastische Repräsentationen, Vektorraumdrehungen, Analysen mit Kernelmethoden, Gleichungssysteme, konvexe Analysis, stochastische Prozesse und Order Statistics in bioinformatischer Anwendung“. Stattdessen klassifizieren wir all das einfach als *Modelle*.

Zuletzt sollte man vielleicht noch einen Blick werfen auf die Modelliteratur der letzten Jahre. So unternahm beispielsweise Wüstneck einen Versuch den Modellbegriff genauer zu definieren, verlor dabei jedoch nur wenig seiner Allgemeinheit: „Ein Modell ist ein System, das als Repräsentant eines komplizierten Originals auf Grund mit diesem gemeinsamer, für eine bestimmte Aufgabe wesentlicher Eigenschaften von einem dritten System benutzt, ausgewählt oder geschaffen wird, um letzterem die Erfassung oder Beherrschung des Originals zu ermöglichen oder zu erleichtern, beziehungsweise um es zu ersetzen“ (Wüstneck, 1963). Stachowiak verfeinerte 1973 mit seiner *Allgemeinen Modelltheorie* diesen Ansatz und erläuterte den Modellbegriff durch drei wesentliche Charakteristika:

Abbildungsmerkmal: „Modelle sind stets Modelle von etwas, nämlich Abbildungen, Repräsentationen natürlicher oder künstlicher Originale, die selbst wieder Modelle sein können“ (Stachowiak, 1973, S. 131)

Verkürzungsmerkmal: „Modelle erfassen im allgemeinen nicht alle Attribute des durch sie repräsentierten Originals, sondern nur solche, die den jeweiligen Modellerschaffern und/oder Modellbenutzern relevant erscheinen“ (Stachowiak, 1973, S. 132)

Pragmatisches Merkmal: „Modelle sind ihren Originalen nicht per se eindeutig zugeordnet. Sie erfüllen ihre Ersetzungsfunktion a) für bestimmte - erkennende und / oder handelnde, modellbenutzende - Subjekte, b) innerhalb bestimmter Zeitintervalle und c) unter Einschränkung auf bestimmte gedankliche oder tatsächliche Operationen“ (Stachowiak, 1973, S. 132)

Wie schon Wüstneck bemerkte, ist das Ziel der Modellierung dabei meist entweder (i) das Finden eines Ersatzes um nicht mit dem Original arbeiten zu müssen weil es zu teuer, zu gefährlich oder ethisch nicht vertretbar ist, oder (ii) die Überprüfung von Thesen über die Funktionsweise des Originals, die direkt am Original selber nicht oder nur schwer nachvollzogen

werden können. Beispiele für beide Varianten sind zahlreich, denkt man an die Verwendung von Modellorganismen wie der Maus in der experimentalbiologischen Forschung zur Untersuchung der Wirkweise von Substanzen die später auf den menschen Übertragen werden soll (i). Variante (ii) spielt besonders in der mathematischen Modellierung von realweltlichen Problemen und Systemen eine Rolle. Verhält sich das erstellte Modell des realen Systems bei - nach bestimmten Bedingungen - optimal gewählten Parametern ähnlich dem realen Vorbild, so kann versucht werden über einen Analogieschluß Struktureigenschaften des Modells auf das Original zu übertragen.

Der Vollständigkeit halber sei noch angemerkt, dass zwei weitere Ansätze der Definition des Modellbegriffs in der Gegenwartsliteratur zu finden sind, darunter die konstruktionsorientierte nach Schütte und die prozessorientierte nach vom Brocke (Schütte, 2001; vom Brocke, 2003; Wikipedia, 2007b) auf die an dieser Stelle nicht näher eingegangen werden soll.

1.2 Verwendung des Modellbegriffs in der Bioinformatik

Im Bereich bioinformatischer Forschung spielen insbesondere Modelle der zweiten Variante (die Überprüfung von Thesen über die Funktionsweise des Originals) eine besondere Rolle, was unter anderem auch daran liegen mag, dass Bioinformatiker selten in die Verlegenheit kommen mit lebenden Dingen arbeiten zu müssen. Dennoch ist ganz generell der Begriff hier ebenso weit gefaßt, wie in vielen anderen Disziplinen.

Was für eine Rolle spielen Modelle in der bioinformatischen Forschung? Um diese Frage zu beantworten sollte man einen Blick auf die geschichtliche Entwicklung der Bioinformatik werfen. Auch wenn diese junge Disziplin ihren Anfang in den Überschneidungen zahlreicher etablierter Wissenschaften wie der Biologie, Chemie, Mathematik und Informatik fand und ein genauer Ursprung nur schwer festzustellen ist, so gibt es doch einige nennenswerte Meilensteine.

Dazu gehört sicherlich der nach der Erfindung der Proteinsequenziermethoden (Sanger and Tuppy, 1951) und Aufklärung der Proteinsekundärstrukturelemente (α-Helix und β-Faltblatt ,Pauling *et al.* (1951)) von Margaret Dayhoff ins Leben gerufene Atlas der Protein Sequenzen (Dayhoff *et al.*, 1965), deren Nachfolger heute als *Protein Information Ressource*, kurz *PIR*, bekannt ist (Barker *et al.*, 1993; George *et al.*, 1997). Parallel dazu entdeckten Wissenschaftler die Sekundärstrukturelemente der Proteine, . Ebenso nennenswert ist eine der ersten entscheidenden algorithmischen Entdeckungen der Bioinformatik, die Entwicklung des globalen Sequenzalignments (Needleman and Wunsch, 1970). Ein naher Verwandter, das lokale Sequenzalignment, folgte 11 Jahre später durch Smith and Waterman (1981) und legte den Grundstein für moderne Suchalgorithmen auf Sequenzdatenban-

ken. Da die zunehmende Rechenleistung der gängigen Mikroprozessoren — trotz Erfüllung des Mooreschen Gesetzes[1] — der wachsenden Größe jener Sequenzdatenbanken nicht Herr werden konnten, mußten Heuristiken wie das prominente *BLAST* entstehen (Altschul *et al.*, 1990), die quasi-optimale lokale Alignmentsuchen auf Datenbanken nahezu beliebiger Größe erlaubten. Ähnliche Schwierigkeiten ergaben sich beim Versuch das Verfahren von paarweisen Sequenzalignments auf das Alignieren mehrerer Sequenzen gleichzeitig zu übertragen, eine Aufgabe die heute ebenfalls von heuristischen Verfahren gelöst wird wie sie z.B. im bekannten multiplen Alignierer CLUSTALW implementiert sind (Thompson *et al.*, 1994; Larkin *et al.*, 2007).

Ein weiteres Phänomen aus der Sequenzanalyse, welches von Anfang an das Interesse der Forscher weckte, war die Eigenschaft von RNA Molekülen sich in wässriger Lösung zu spezifischen Sekundärstrukturen zu falten. Entscheidende Zellkomponenten, wie die Ribosomen als Proteinbiosynthesemaschinerien, können erst durch ihre charakteristische räumliche Struktur die ihnen zugedachte Funktion ausüben. Es wurden Algorithmen vorgeschlagen, die unter Berechnung der freien Energie aller möglichen Sekundärstrukturen in der Lage waren die wahrscheinlichsten Faltungsmuster alleine aus der RNA Sequenz vorherzusagen (Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981). Die RNA Sekundärstrukturvorhersage spielt bis heute eine wichtige Rolle in der bioinformatischen Arbeit (Wolf *et al.*, 2005; Schultz *et al.*, 2005, 2006; Selig *et al.*, 2007), Sequenzstrukturalignments dienen der exakteren Aufklärung von Verwandtschaftsverhältnissen oder zur Unterscheidung von Arten (Müller *et al.*, 2007).

Heute hat das Feld der Bioinformatik an Reichhaltigkeit hinzugewonnen. Sein Ursprung, die Sequenzanalyse, ist bis heute eines der bedeutendsten Teilbereiche. Hinzugekommen sind die Analyse von Proteinstrukturen und -domänen, von metabolischen und regulatorischen Netzwerken der Zelle, von Gen- und Proteinexpression sowie Dutzende von Modedisziplinen die heute unter dem Schlagwort Omik (engl. Omics) zusammengefasst werden. Sie alle haben die gesamtheitliche holistische Aufklärung (angezeigt durch die Endsilbe -omik) eines Teilbereichs des Organismus oder der Zelle zum Ziel. Von ihnen sind die Transkriptomik (Untersuchung der Gesamtheit aller transkribierten Elemente der Zelle), Proteomik (Aufklärung des vollständigen Proteinbestands der Zelle) und Metabolomik (Aufstellung des gesamten metabolischen Netzwerkes der Zelle) noch einsichtig vernünftig. Wo so manch andere Omiken hingegen bei zartbeseiteten Artgenossen schon leichtes Stirnrunzeln (Fluxomik, Interaktomik) oder gar Kopfschütteln (Gly-komik) verursachen können, wartet der harte Kern noch gespannt auf die erste Publikation aus der Omeomik und was sie uns wohl über all die anderen Omiken Spannendes

[1] Als *Mooresches Gesetz* bezeichnet man die 1965 von Gordon Moore, einem Mitbegründer der Firma Intel, formulierte These dass sich die Schaltkreiskomponenten auf einem Computerchip alle zwei Jahre verdoppeln.

zu berichten weiß (Wikipedia (2007a), weiterführende Informationen findet man auch in der Fachzeitschrift *Omics*).

Schon diese wenigen Ecksteine reichen aus, um einen generellen Trend in der Bioinformatik auszumachen. Umfaßte der ursprüngliche Proteinsequenzatlas von Frau Dayhoff noch eine überschaubare Anzahl an Proteinen (Dayhoff *et al.*, 1965), so kann der Bioinformatiker heute auf eine Vielzahl von Sequenzdatenbanken zurückgreifen mit Größen von mehr als 50 Millionen Sequenzen (Benson *et al.*, 2006). Versuchte man früher noch detailliert Eigenschaften einer einzelnen Sequenz aufzuklären, so analysiert man heute Datensätze mit zehntausenden von Genexpressionswerten auf einmal (Engelmann *et al.*, 2007). War in den Anfängen schon eine erfolgreich aufgeklärte RNA Sekundärstruktur eine Publikation wert (Severini *et al.*, 1996), so sollte es heute besser eine ganze Datenbank sein (Wolf *et al.*, 2005).

Dass eine erfolgreiche Bewältigung dieser anstehenden Datenflut auch Veränderungen innerhalb der Bioinformatik erfordert mag einsichtig erscheinen. Dennoch scheint diese Notwendigkeit nur langsam zur wissenschaftlichen Basis durchzudringen, so dass 2001 Martin Vingron, Direktor des Max-Planck-Instituts für Molekulare Genetik in Berlin, einen entsprechenden Leitartikel in der Zeitschrift *Bioinformatics* verfaßte.

Bioinformatics needs to adopt statistical thinking (Vingron, 2001) war der Titel eines Beitrags der exakt jenen graduellen Umschwung bioinformatischer Fragestellungen aufzeigte und den damit verbundenen notwendigen, nur langsam voranschreitenden, Paradigmenwechsel forderte. Sein Beitrag ist darüberhinaus in mehrerer Hinsicht bemerkenswert. So stellt er die häufig vertretene Annahme in Frage, dass die klassische *hypothesen-gesteuerte* Wissenschaft mehr und mehr ersetzt würde durch eine *hypothesen-freie* Herangehensweise, bei der die reine Sammlung von Information und anschließende holistische Interpretation ohne konkrete Fragestellung im Vordergrund stünde. Er räumte zwar ein, dass die moderne Art bioinformatischer Forschung durchaus in einem starken Kontrast zur traditionellen Molekularbiologie stehe, in welcher eine oder mehrere konkrete Hypothesen durch minutiös geplante Experimente akkurat verifiziert oder falsifiziert wurden, während heute der molekularbiologische Experimentator immer öfter nicht selber in der Lage ist, die Ergebnisse seiner Arbeit zu interpretieren. Dennoch sei es falsch, von einer generellen *Hypothesenfreiheit* zu sprechen. Ganz im Gegenteil: Hypothesen seien nach wie vor forschungsrelevant, allein ihr Eintrittspunkt in das Projekt habe sich zeitlich nach hinten verschoben. An die gesammelten Daten würden heute zunächst allgemeine Fragen gestellt, aus denen sich dann nach und nach echte Hypothesen entwickelten, die dann ebenso konsequent überprüft werden wie es schon immer wissenschaftlicher usus war.

Oder um es anders auszudrücken: In der klassischen Denkweise entstanden aus den komplexen biologischen Systemen in den Gedanken der Molekularbiologen *Modelle*, die ein bestimmtes Phänomen erklären konnten oder

zumindest die notwendigen Abstraktionsschritte zur Erklärung desselbigen beinhalteten (das *Verkürzungsmerkmal* nach Stachowiak), sie bildeten eine erste Abstraktionsebene innerhalb derer dann akkurat und reproduzierbar experimentiert werden konnte.

Heute werden (einen eventuell vorausgehenden Abstraktionsschritt beiseite lassend) zuerst Daten in großem Umfang erhoben, wissentlich der Tatsache, dass damit viele Fragen beantwortet werden könnten, sollte man des Datenumfangs Herr werden. Anschließend erfolgt der Schritt der Modellbildung, oft im Rahmen einer explorativen nicht-parametrischen Datenanalyse, in welchem versucht wird, die Komplexität der Daten zu reduzieren, das Signal-Rausch-Verhältnis zu verbessern, erste wiederkehrende Muster in den Daten zu erkennen und geeignete neue Repräsentationen für die ursprünglichen Daten zu erzeugen. Diese neuen Repräsentationen sind es schließlich, an denen die konkreten Fragen und Hypothesen evaluiert werden, mit aller gebührenden wissenschaftlichen Rigorosität und oftmals den klassischen statistischen parametrischen Tests, welche schon seit Jahrzehnten in der traditionellen Molekularbiologie und experimentellen Medizin ihre Anwendung finden. Es ist allein der Moment der *Modellbildung*, der aufgrund der wachsenden Anforderungen an die statistischen Kenntnisse des Wissenschaftlers heute oftmals in die Hände von Bioinformatikern gelangt, die damit ebenso oft noch überfordert sind (Vingron, 2001). Aber das wird eine der Hauptaufgaben der Bioinformatiker in den kommenden Jahren sein und stellt somit auch die Rolle des Modellbegriffs im Rahmen bioinformatischer Forschung in ein neues Licht. Bioinformatische Modelle sind mittlerweile zu einem Großteil statistischer Natur und um diese Form von Modellen soll es im folgenden auch zuvorderst gehen. Um es mit Herrn Vingrons Worten auszudrücken „The consequence of all this is that we need to get back to school and learn more statistics“.

Teil II

Resultate: Vom Gen zum Organismus über Transkriptom und Metabolom

Kapitel 2

Das Gen: Information aus Sequenz und Struktur

2.1 Anfänge der Sequenzanalyse

Wie einleitend erwähnt liegt der Ursprung bioinformatischer Forschung in der Untersuchung von Protein- und DNA-Sequenzen. Noch vor der Erfindung des ersten optimalen Alignmentalgorithmus Anfang der 70er durch Needleman und Wunsch war es ein Molekularbiologe aus Connecticut, Russell F. Doolittle, der begann, systematische paarweise Sequenzvergleiche zur Detektierung von Homologien zu verwenden. Zusammen mit seinem Sohn war er fasziniert von der Idee, dass die Geschichte allen Lebens auf der Erde eines Tages durch die Analyse von Sequenzen aufgeklärt werden könnte (Jones and Pevzner, 2004). Er verankerte damit das Grundprinzip der Datenbanksuche in der zukünftigen Bioinformatik. Der Schritt vom paarweisen zum multiplen Sequenzvergleich war zumindest gedanklich nicht weit.

Heute, im Hinblick auf die rasant ansteigende Anzahl an Genomprojekten, Metagenomic Projekten (Yooseph *et al.*, 2007) und die exponentiell wachsenden molekularen Sequenzdatenbanken, ist die Sequenzanalyse immer noch eines der zentralen Gebiete bioinformatischer Forschung. Die vorherrschende Datenbasis sind dabei multiple Sequenzalignments (MSAs), welche neben ihrer ursprünglichen Bedeutung als Fundament der Funktionsannotation von Proteinen zahlreiche weitere Anwendungbereiche gefunden haben. Domänenanalysen wie die SMART Datenbank (Schultz *et al.*, 1998; Letunic *et al.*, 2006), PFAM (Sonnhammer *et al.*, 1998; Finn *et al.*, 2006) und InterPro (Apweiler *et al.*, 2000; Mulder *et al.*, 2007), die Vorhersage von Signalpeptiden (SignalP (Nielsen *et al.*, 1997; Bendtsen *et al.*, 2004)), homologiebasierte Proteinstrukturvorhersage, die Erkennung von Transmembranproteinen (TMHMM2 Krogh *et al.* (2001)) und interagierenden Aminosäuren (ipHMMs Friedrich *et al.* (2006)) (mittels Hidden Markov Modellen (HMMs)), sowie sämtliche Methoden zur (molekularen) phylogeneti-

schen Rekonstruktion von Stammbäumen basieren auf multiplen Sequenzalignments (Felsenstein, 2005; Swofford, 2003; Friedrich *et al.*, 2005). Insbesondere im Hinblick auf die phylogenetische Rekonstruktion von Stammbäumen rücken dabei zunehmend auch kombinierte multiple Sequenz-Struktur-Alignments (MSSAs) in den Vordergrund (Hudelot *et al.*, 2003; Gutell *et al.*, 2002). Nachteile sowohl von Sequenz- als auch Sequenzstruktur-Alignments sind ihre schlechte Interpretierbarkeit. Auch wenn sie Ausgangsbasis für eine Vielzahl verschiedener Methoden und Modelle sind, ist eine unmittelbare Analyse aufgrund ihrer Länge und Komplexität mit dem bloßen Auge oder auch nur einfachen Hilfsmitteln nur in Ausnahmefällen durchzuführen. Sie unterscheiden sich darin nicht von den meisten anderen Datenquellen, die heute für bioinformatische Analysen herangezogen werden, seien es Genexpressions- oder Massenspektrometriedaten.

In der klassischen multivariaten Statistik werden explorative Methoden angewendet, um solch komplexe hochdimensionale Daten einer ersten Untersuchung zu unterziehen. Ziel ist dabei nicht die Beantwortung konkreter Fragestellungen; es werden vielmehr erste charakteristische Eigenschaften der Daten detektiert sowie Ausreißer und eventuell damit zusammenhängende methodische Fehler erkannt. Dabei kommen auch dimensionsreduzierende Techniken zur Anwendung, die in der Lage sind hochdimensionale Daten für das menschliche Auge verständlich aufzubereiten. Diese Methoden beinhalten typische Verfahren wie die Hauptkomponentenanalyse (Principal Component Analysis, PCA), Klassische Multidimensionale Skalierung (Classic Multidimensional Scaling, MDS), Korrespondenzanalyse (Correspondence Analysis, CA) oder Kanonische Korrelationsanalyse (Canonical Correlation Analysis, CCA). Die explorative Analyse, also das statistische „Erforschen“ der Daten, wirft dabei oft erste Hypothesen ab, die dann durch weitere Analysen, wie z.B. parametrischen Tests, bewiesen oder widerlegt werden können.

Aufgrund der Komplexität der MS(S)As wäre es naheliegend die Methoden explorativer Statistik auf Multiple Sequenz- und Multiple Sequenz-Struktur-Alignments anzuwenden, doch es bleibt ein entscheidendes Problem — die Einbettung. Die meisten statistischen Verfahren arbeiten — wenig überraschend — mit Zahlen[1], die monotone Aneinanderreihung von Buchstaben wie sie MS(S)As darstellen scheinen bei Mathematikern bisher nur selten Interesse geweckt zu haben, und die Definition einer (sinnvollen) Abbildung aus einem Alphabet auf einen reellwertigen Vektorraum ist nicht trivial.

Daher arbeiteten die ersten Versuche zur explorative Analyse von MSAs Anfang der 90er Jahre ohne eine solche. Schneider and Stephens führten 1990 Sequenzlogos ein, bei denen die Größe der Buchstaben der Sequenz in

[1] Natürlich dürfen hier Ausnahmen wie das *Non-metric Multidimensional Scaling* nicht vergessen werden

Abhängigkeit vom Konserviertheitsgrad gewählt wurde, so dass konservierte Positionen schnell gefunden werden konnten (Abbildung 2.1).

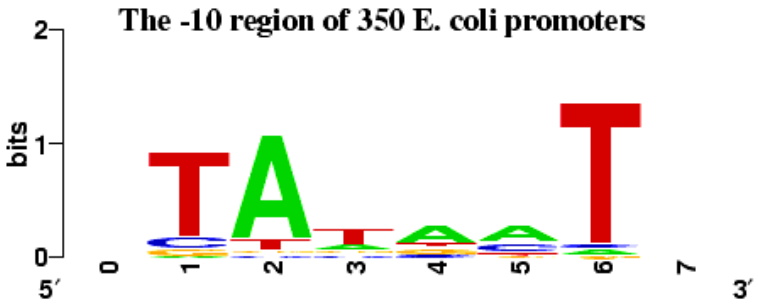

Abbildung 2.1: Sequenzlogo von 350 *E. coli* Promotorregionen (sog. TATA-Box)

Später wurde dieses Prinzip auch auf RNA Sequenz-Struktur Alignments übertragen (Gorodkin *et al.*, 1997). Alternativ stellen viele heutige Alignmenteditoren und Alignmentprogramme wie Seaview (Galtier *et al.*, 1996), ClustalX (Thompson *et al.*, 1997) oder 4SALE (Abbildung 2.1, Seibel *et al.* (2006)) Visualisierungsroutinen zur Verfügung, die beispielsweise Aminosäuren gemäß ihrer biochemischen und physikalischen Eigenschaften einfärben und ein Balkendiagramm an das MSA alignieren, die Auskunft über den Grad der Konserviertheit der Positionen geben.

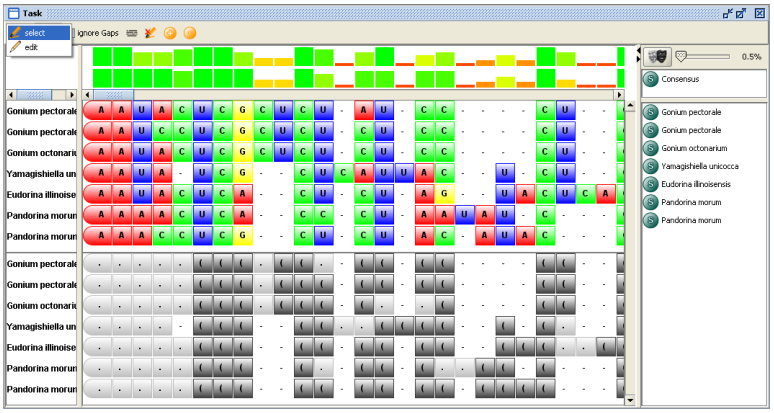

Abbildung 2.2: Screenshot des Sequenz-Strukturalignmentprogramms 4SALE (Seibel *et al.*, 2006) mit Konserviertheitsprofil als Balken darüber.

Parallel entstanden Theorien und Methoden, die aufbauend auf den MSAs deren Leistungsfähigkeit zur Beschreibung von Sequenzähnlichkeiten verbesserten. Profil-HMMs als probabilistische Darstellung von MSAs von

Sequenzfamilien (Churchill, 1992; Haussler *et al.*, 1993; Krogh *et al.*, 1994) wurden verwendet, um uncharakterisierte Sequenzen auf Zugehörigkeit zu Sequenzfamilien zu überprüfen. Konsequenterweise wurden HMM Logos zur Visualisierung eingeführt, welche Entropieterme aus den geschätzten HMM Parametern wie Emissions-, Insertions- und Deletionswahrscheinlichkeiten ableiten und darstellen (Schuster-Böckler *et al.*, 2004).

Obwohl all diese Methoden nützliche Hilfsmittel zur Darstellung positionsspezifischer Informationen wie des Konserviertheitgrades in MSAs sind, ermöglichen sie keine angemessene explorative Analyse. Sie sind nicht in der Lage charakteristische Sequenzcluster zu detektieren, lange Sequenzen adäquat zu repräsentieren oder gar sinnvoll Hypothesen für eine nachfolgende Überprüfung mittels parametrischer Tests vorzugeben. Die größten Schwierigkeiten sind hierbei nach wie vor die Länge der Sequenzen und die Tatsache, dass eine Darstellung der Konserviertheitsgrade der Positionen nur bei im Allgemeinen gut konservierten Sequenzen hilfreich ist. Das Problem der Einbettung bleibt erhalten.

2.2 CAMA: Einbettung und Visualisierung

Im Jahre 1995 stellten Casari *et al.* eine Methode vor um das Dimensionsproblem in MSAs zu lösen, welche dann später in der Software Jalview (Clamp *et al.*, 2004) eine Implementierung fand. Der Algorithmus basiert auf einer Abbildung der Sequenzen auf eine binäre Vektordarstellung (ohne dabei Insertionen & Deletionen (Gaps) im Alignment zu berücksichtigen) und eine anschließende Anwendung einer Hauptkomponentenanalyse. Jaakkola *et al.* verwendeten bereits 1998 zum ersten Mal aus einem HMM abgeleitete *Fisher Scores* zur Einbettung von Sequenzen in einen euklidischen Vektorraum und detektierte auf diese Weise erfolgreich entfernte Sequenzhomologien. Die Fisher Scores fanden später weitere Anwendungen, so in der Klassifikation von G-Protein Coupled Receptors (GPCRs) (Karchin *et al.*, 2002). Die Klassifikation der Proteine basierte ausschließlich auf den Fisher Scores des zugehörigen HMMs. Die beobachtete sehr gute Klassifizierungsqualität erstaunt nicht wirklich, da Fisher Scores eine minimale suffiziente Statistik der HMM-Parameter liefern (Lindgren, 1993). Der erste Schritt der explorativen Analyse von MS(S)As ist die Einbettung von Sequenzen, also Folgen eines Alphabets von Buchstaben, in einen reellwertigen Vektorraum. Die grundlegende Idee ist es, die Wahrscheinlichkeitsinformation, die stochastische generative Modelle wie HMMs und SCFGs über die Verteilung der Sequenzen geben, für diesen Zweck zu verwerten, anstatt eine direkte eins zu eins Zuordnung von Buchstaben des Alphabets zu numerischen Werten vorzunehmen (triviale Einbettung).

Eine alternative Möglichkeit wäre es, direkt die positionsspezifischen Wahrscheinlichkeiten der Sequenzen aus einem HMM zur Einbettung zu

verwenden. Dies hätte jedoch den gravierenden Nachteil, dass zwei völlig verschiedene, jedoch in etwa gleich unwahrscheinliche Sequenzen zu sehr ähnlichen Repräsentationen führen würden. Sie würden entsprechend in einer folgenden Visualisierung irrtümlicherweise nahe beieinander zu liegen kommen. Um dies zu umgehen verwenden wir die von Jaakkola *et al.* eingeführte Einbettung mittels Fisher Scores mit ihren positiven Eigenschaften, die sich in mehrerer Hinsicht gegenüber trivialen alternativ vorgestellten Einbettungen wie der von Casari *et al.* 1995 als überlegen erwiesen hat. Durch Verwendung der Fisher Scores wird sämtliche Information aus dem stochastischen Modell extrahiert, was weit mehr ist als nur die Wahrscheinlichkeiten der einzelnen Sequenzpositionen. Da es sich bei den Fisher Scores um die Ableitung der Log-Likelihood des Modells nach seinen Parametern handelt, fließen hier zusätzliche Informationen darüber ein, wie stark eine Sequenz das Modell beeinflussen würde sowie, welche Teile des Modells für die Emission der gegebenen Sequenz verantwortlich sind. Die Fisher Scores enthalten damit Informationen über die interne Repräsentation der Sequenz innerhalb des Modells.

Weitere entscheidende Vorteile sind die positionsspezifische Einbettung und der Effekt, dass die eingebetteten Sequenzen stets die gleiche Länge besitzen, da diese von dem stochastischen Modell vorgegeben ist. So wird beispielsweise einem 'A' in einer Alignmentspalte, die überwiegend aus 'A's besteht, ein anderer Wert zugewiesen als einem 'A', welches nur einmal in einer Alignmentspalte vorkommt. Da ein Profil-HMM auch positionsabhängige Insertions- und Deletionsereignisse modelliert, so werden diese auch in den Fisher Scores reflektiert.

Die Fisher Score Einbettung wurde von unserer Arbeitsgruppe erfolgreich angewandt und zusammen mit einer Korrespondenzanalyse als Verfahren explorativer Statistik in einer Software implementiert. CAMA — Correspondance Analysis on Multiple Sequence Alignments ermöglicht eine unmittelbare explorative Analyse von MSAs, die Detektion von Sequenzclustern und den dafür verantwortlichen Sequenzpositionen (Abbildung 2.3).

2.3 Ausblick: Visualisierung von SCFGs

HMMs sind neben stochastischen Modellen enge Verwandte von formalen Sprachen. In der Theorie formaler Sprachen wird versucht die syntaktischen, also grammatikalischen Eigenschaften einer Sprache formal zu definieren um durch die so entstehende formale *Grammatik* in der Lage sein zu können syntaktisch korrekte Sätze zu generieren und diese von grammatikalisch falschen zu unterscheiden. In diesem Kontext bildet das HMM die formale Grammatik der Sprache der Aminosäure- oder Nukleotidsequenzen. Seine zugrunde liegende Struktur entspricht genauer einer (stochastischen) *regulären* Grammatik und ist damit in der Lage klassische Sequenzalignments bestehend

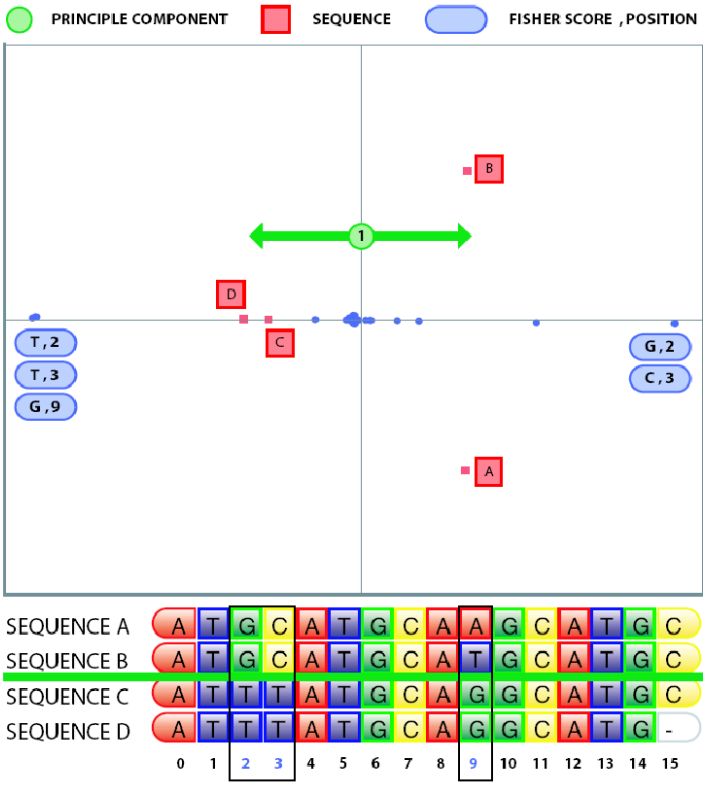

Abbildung 2.3: Beispielhafte zweidimensionale Visualisierung von vier Sequenzen eines artifiziellen multiplen Sequenzalignments. Die X- und Y-Achsen des Scatterplots entsprechen den ersten beiden Hauptachsen der Korrespondenzanalyse. Die Sequenzen $A - D$ sind als rote Quadrate eingezeichnet und es ist deutlich zu sehen, dass die erste Achse (X-Achse) die Sequenzen C und D von den Sequenzen A und B trennt. Die dafür verantwortlichen Sequenzpositionen sind als blaue Ovale in den Plot eingetragen. Hauptunterschiede sind an Position 2 und 3 (G,C gegenüber T,T) und 9 (A bzw. T gegenüber G) zu finden. Die zweite Achse der CA trennt weiterhin Sequenzen A und B gegeneinander auf, die sich an Position 9 voneinander unterscheiden.

aus den Buchstaben eines Alphabets sowie Gaps zu repräsentieren, nicht jedoch horizontale Abhängigkeiten zwischen Sequenzpositionen zu modellieren. Um eine RNA Sequenz wie der ITS2 inklusive ihrer Sekundärstruktur zu generieren muss die zugehörige Grammatik in der Lage sein palindromartige[2] Sprachelemente darstellen zu können. Die dazu nötige und in der

[2]Ein Palindrom (von griechisch παλινδρομος (palíndromos) „rückwärts laufend“) ist eine Zeichenkette, die von vorn und von hinten gelesen gleich bleibt, wie zum Beispiel die

Chomsky Hierarchie nächst höhere Grammatikfamilie sind die kontextfreien Grammatiken. Um das beschriebene Einbettungsverfahren auch auf Multiple Sequenz-Struktur Alignments anwenden zu können, soll in einem Folgeprojekt das Prinzip der Fisher Scores von HMMs auf stochastische kontextfreie Grammatiken (SCFGs) übertragen werden, um sowohl die Sequenz- als auch die Strukturinformation des ursprünglichen Alignments zu erhalten.

Desweitere soll insbesondere die Möglichkeit betrachtet werden *kernelisierte* Versionen der klassischen Methoden (Schölkopf *et al.*, 1998; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004) anzuwenden. Dies hätte zwei konkrete Vorteile: Durch die implizite Projektion der Daten in den sogenannten Feature Space könnten zum einen auch nicht-lineare Abhängigkeiten in den Daten gefunden werden. Zum anderen erwarten wir mit zunehmender Dimension der Daten, d.h. Länge der Alignments, eine deutliche Verbesserung des Laufzeitverhaltens, da für die Berechnung lediglich die Skalarprodukte der Sequenzen herangezogen werden und die Dimension einer Skalarproduktmatrix stets unabhängig von der Länge der Datenvektoren ist.

2.4 Anwendung: Das H5N1 Hemagglutininalignment

In der Untersuchung und Typisierung von Influenza A Viren, eine Aufgabe die gerade im Hinblick auf eine befürchtete Grippeepidemie zunehmend Aufmerksamkeit gewonnen hat, sind insbesondere zwei Glycoproteine von besonderer Bedeutung. Hemagglutinin (HA) sowie die Neuraminidase (NA) sind zwei Oberflächenproteine der Viren deren Hauptaufgabe zum einen die Bindung an die zu infizierenden Zellen sind (HA), zum anderen die Freigabe reproduzierter Viruspartikel aus der Wirtszelle durch die Hydrolasetätigkeit der NA an der Zellmembran. In Wildvögeln und Geflügel wurden weltweit bisher 16 HA und 9 NA Typen registriert, welche in vielfältigen Kombinationen auftreten (z.B. H1N1, H16N3, H5N1) (Olsen *et al.*, 2006). Insbesondere dem Hemagglutinin wird dabei eine besondere Bedeutung zugemessen, da davon ausgegangen wird, dass eine bestimmte Kombination von Mutationen in diesem Gen notwendig wären, um eine effektive Übertragung des Virus von Mensch zu Mensch zu ermöglichen. Damit ist es auch bei der Entwicklung von Impfstoffen in den Fokus der Forscher gerückt (Subbarao and Luke, 2007). Zur Illustration unserer Methode wurden 499 H5 HA Sequenzen von ca 600bp Länge aligniert und anschließend mittels eines HMMs in die benötigte Fisher Score darstellung gebracht. Die 499 Sequenzen stammten insgesamt aus 99 verschiedenen Isolationsorten, davon die meisten aus dem asiatischen Raum, aus einem Zeitraum von 1959 bis 2006 und wurden 34

Wörter ANNA oder auch GNUDUNG (Wikipedia, 2007c).

unterschiedlichen Spezies entnommen. Da noch nicht die kernelisierte Version des Algorithmus zur Anwendung kam, beschränkte sich die Analyse des Datensatzes allerdings zunächst auf ein Subset von 79 Isolaten aus Vietnam. Zusammen mit Annotationen über Isolationsort, Jahr und Species wurden kanonische Korrespondenzanalysen (CCA) durchgeführt um Zusammenhängen zwischen den annotierten Faktoren, Sequenzen und Sequenzpositionen aufzudecken. Zur Illustration der Methode soll die zeitabhängige Evolution der Vietnamisolate betrachtet werden (Abbildung 2.4). Es ist eindeutig zu erkennen, dass die Sequenzen nach den drei Isolationsjahren clustern. Die verantwortlichen Positionen (rote Kreuze) des Alignments entlang der Jahresachsen verweisen auf einzelne Polymorphismen, welche zwar in ihrer Gesamtheit eine Klassifikation nach Jahren erlauben, jedoch nicht charakteristisch für alle Sequenzen des Jahres zu sein scheinen. Es waren keine Sequenzmotive zu erkennen, die als generelles Merkmal für ein bestimmtes Isolationsjahr gelten können. Unter den detektierten SNPs mit besonders hohem Informationsgehalt (Positionen an den oberen und unteren Rändern des Plots) befanden sich der Austausch eines Histidinrests (H) gegen ein Glutamin (Q) an Position 46, eines Threonins (T) gegen ein Lysin (K) an Position 54 sowie der Austausch eines Valins (V) gegen einen Leucinrest (L) an Position 65.

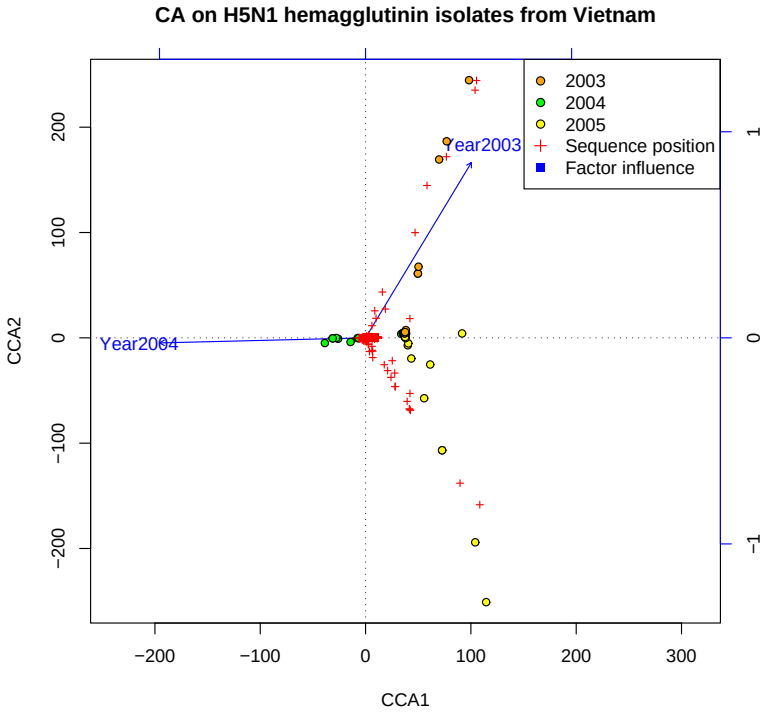

Abbildung 2.4: CCA Plot von 79 H5N1 Isolaten aus Vietnam (Kreise). Die Sequenzen wurden gemäß ihrem Isolationszeitpunkt eingefärbt. Die verantwortlichen Sequenzpositionen sind als rote Kreuze eingezeichnet.

Kapitel 3

Regulation der Gene: Das Transkriptom am Beispiel von *A. thaliana*

3.1 Genexpression und Microarrays

In den letzten Jahren wurden eine große Anzahl von Microarray Experimenten unter einer Vielzahl von experimentellen Bedingungen durchgeführt. Um die Ergebnisse dieser Studien für die Wissenschaft verfügbar zu halten, wurden Datenbanken wie der NCBI Gene Expression Omnibus (Barrett *et al.*, 2007), ArrayExpress (Parkinson *et al.*, 2007) oder NASCArrays (Craigon *et al.*, 2004) zur Archivierung der Daten aufgesetzt. Angesichts des rapiden Preisverfalls bei Microarray Experimenten werden diese Datenbanken vermutlich in naher Zukunft ähnliche explosionsartige Zuwächse erleben wie die Sequenzdatenbanken und damit ebenso sehr nach ausgereiften Verfahren verlangen, um diese Datenmengen angemessen bearbeiten zu können.

Die Verarbeitung dieser heterogenen Datenmengen ist nicht immer einfach, da die Standards der Datenhaltung sowie die Qualitätsanforderungen der Datensätze zwischen den unterschiedlichen Datenbanken stark variieren. Fehlende Rohdaten oder ungenaue Protokollierung der experimentellen Bedingungen sind nur einige der Schwierigkeiten, denen man begegnet. Nichtsdestotrotz bietet die Metaanalyse mehrerer solcher Datensätze Möglichkeiten, Einblicke in die Regulation von Genen und Genclustern zu erhalten, die aus einem Experiment oder einer Serie allein kaum zu erhalten wären. Die Gründe dafür sind zahlreich, oft ist das regulatorische Signal zu schwach oder die Proben werden in einem funktionalen Kontext betrachtet, der von vorneherein bestimmte Signale herausfiltert.

In den letzten Jahren sind eine ganze Reihe von Methoden für die Metaanalyse von Microarray Experimenten vorgeschlagen worden. Die meisten von ihnen verwenden sogenannte Effektgrößenmodelle (*effect size models*)

und betrachten insbesondere die Variation zwischen einzelnen Studien (*interstudy variation*, Choi *et al.* (2003); Conlon *et al.* (2006); Hu *et al.* (2005); Moreau *et al.* (2003)). Damit ähneln sie meist der Vorgehensweise bei der klassischen Genexpressionsanalyse und fügen die Studie selbst als weitere erklärende Variable hinzu. Dabei werden in die Metaanalyse verschiedene Datensätze aus unterschiedlichen Experimenten integriert um so die Anzahl Replikate und dadurch die Mächtigkeit der Tests zu vergrößern. Das bedeutet jedoch auch, dass nur solche Datensätze verwendet werden können, bei denen die Proben gleichen Bedingungen ausgesetzt werden, wodurch derartige Meta-Analysen für gewöhnlich nur eine geringe Anzahl von Studien beinhalten.

Ein zweiter Ansatz der überwachten Metaanalyse von Microarrays ist es direkt biologisches Wissen um die Funktion von Genen in die Analyse mit einfließen zu lassen, um so globale Co-Expression vorherzusagen und funktionale Verwandtschaften zwischen co-regulierten Genen aufzudecken (Huttenhower *et al.*, 2006).

Nichtsdestoweniger basieren alle oben genannten Methoden auf parametrischen Modellen und stützen sich damit auf eine Reihe biologischer und statistischer Annahmen. Der Ansatz der Metaanalyse unserer Arbeitsgruppe hingegen ist im Sinne der eingangs beschriebenen Modellbildung vor der Beantwortung konkreter Fragen und Verifizierung von Hypothesen explorativ und nicht-überwacht. Das Ziel ist es, Einsichten in die biologische Struktur der Daten zu erhalten, Hypothesen vorzugeben und diese Hypothesen schließlich beispielsweise mit den oben beschriebenen parametrischen Verfahren zu überprüfen. Wir verwenden dabei *Affymetrix* ATH-1 Gesamtgenomarrays von *Arabidopsis thaliana* als Ausgangsbasis.

In dieser Studie (Engelmann *et al.*, 2007) zeigen wir, wie man den Herausforderungen, die durch die Heterogenität von Microarraydaten entstehen, mit explorativen Verfahren effektiv begegnen kann. Zuerst wurden Microarray Datensätze aus öffentlichen Datenbanken gesammelt und vorprozessiert, um eine gemeinesame Datenbasis für die folgenden Analysen zu schaffen. Dabei wurden Ausreißer detektiert und entfernt um das Signal-Rausch Verhältnis der Daten zu steigern. Anschließend kamen Ordinationsverfahren wie die kernelisierte Hauptkomponentenanalyse zum Einsatz sowie spektrales und hierarchisches Clustering um Kontraste zu gruppieren und die verantwortlichen Gene identifizieren zu können. Letzteres erfolgte durch Merkmalswahl auf den Loadings der Hauptkomponentenachsen, eine Form der Gendetektion, die unseres Wissens bisher noch nicht im Zusammenhang von Metaanalysen zum Einsatz kam.

Die detektierten Gene spielen eine Rolle entweder in Pflanzenwachstum und Entwicklung (assoziiert mit Auxinbehandlung) oder in der Abwehr von Pathogenen und wurden auf physiologische Prozesse und Funktionen abgebildet, welche durch die Literatur belegt werden konnten. Für nicht vollständig annotierte Gene kann unser Ansatz eine Funktions- und Regu-

lationsannotation vorschlagen, wie hier für die Familie der DUF26 (Domain of Unknown Function) Kinasegene gezeigt.

3.2 Die Metaanalyse - ein Machine-Learning Ansatz

3.2.1 Aufbereitung der *Arabidopsis thaliana* Datensätze

In Vorbereitung der Metaanalyse wurden Microarraydaten von uns aus der Datenbank *GEO* extrahiert (Barrett *et al.*, 2007) und gesammelt. Im Rahmen unserer Analysen bezeichnet ein *Datensatz* einen GEO Eintrag mit seiner zugehörigen eindeutigen GEO Zugriffsnummer. Jeder Datensatz besteht aus mehreren Affymetrix CEL-Dateien, welche die Rohdaten der Microarrayexperimente aus einer Hybridisierung beinhalten, im Weiteren als *Sample* oder *Probe* bezeichnet. Ein *Kontrast* besteht aus mehreren, mindestens aber zwei solcher Wiederholungen von Proben aus der ersten Samplegruppe und mindestens zwei aus der Zweiten. Anstatt also die Gesamheit einzelner Messungen miteinander zu vergleichen wurde jeder Datensatz in Kontraste aufgeteilt und diese bildeten die Datenbasis unserer Analysen. Ein Kontrast ist damit der Unterschied in der Genexpression zwischen zwei experimentellen Bedingungen eines Datensatzes, beispielsweise der Vergleich einer *Arabidopsis thaliana* Mutante mit ihrem Wildtyp. Die meisten Datensätze der GEO Datenbank enthielten mehrere unterschiedliche Kontraste.

Ein solcher Kontrast wurde als logarithmischer Fold-Change aller Gene des ATH-1 Chips dargestellt. Vor Berechnung der Fold-Changes haben wir die rohen Intensitätswerte aller Proben eines Kontrastes mittels des gcRMA Algorithmus normalisiert (*gcrma* Paket (Wu and with contributions from James MacDonald Jeff Gentry, 2005) enthalten im Modul Bioconductor (Gentleman *et al.*, 2004) der Statistiksprache R (R Development Core Team, 2006)).

Die Fold Changes und P-Werte selbst wurden mittels des LIMMA Pakets (Smyth, 2004) berechnet; zur Korrektur der P-Werte gegen die α-Fehler-Inflation kam der Algorithmus von Benjamini and Hochberg zur Anwendung (Benjamini and Hochberg, 2000).

Wir erlegten den in der Datenbank zu findenden Experimenten folgende Bedingungen auf, um für unsere Analyse von Nutzen zu sein: a) Die Affymetrix Rohdaten (CEL-Dateien) mußten verfügbar sein, b) es mussten mindestens zwei Replikate jedes Versuchs enthalten sein, c) bei dem Eintrag handelte es sich nicht um ein Zeitverlaufsexperiment. Im November 2006 erfüllten 20 Datensätze diese Kriterien aus deren Gesamtheit von 424 CEL-Dateien 76 Kontraste gebildet werden konnten.

Die anfängliche zeilen-indizierte Datenmatrix enthielt für jeden der 76 paarweisen Vergleiche Fold Changes von allen 22810 *Arabidopsis* Genen des

ATH-1 Chips. Um eine valide Datenbasis für die integrierte Analyse der unterschiedlichen Experimente zu schaffen, wurden zunächst Ausreißer durch Projektion eines bivariaten Boxplots auf den Varianz-Median Plot der Experimente detektiert (siehe Abbildung 3.2.1, Everitt (2005)).

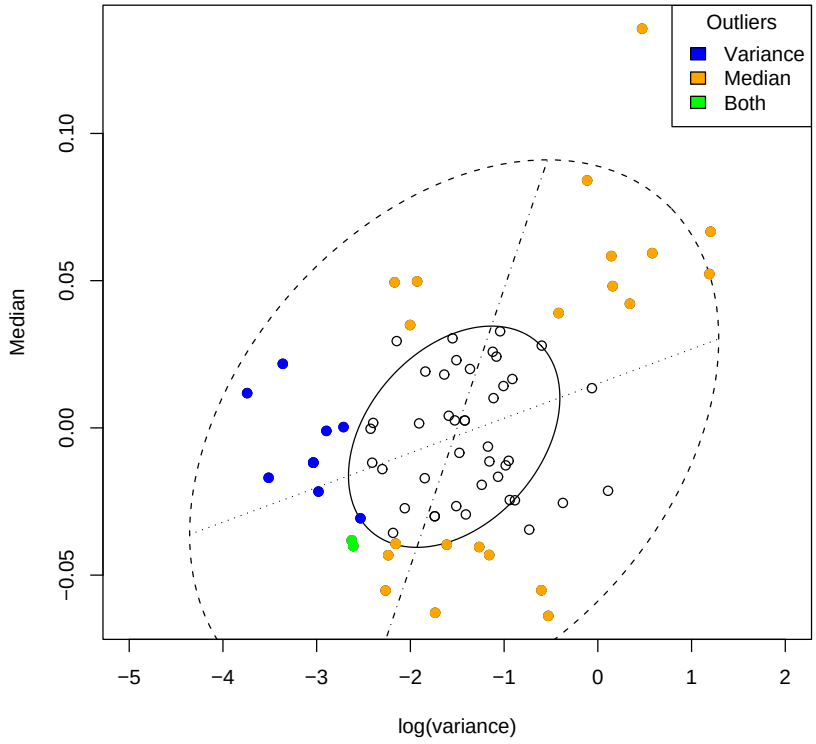

Abbildung 3.1: Median vs. log(Varianz) Plot aller 76 Kontraste sowie der assoziierte Bivariate Boxplot. Die farbliche Markierung charakterisiert die Art des Ausreissers (siehe Legende). Der Bivariate Boxplot ist das zweidimensionale Analogon des bekannten Boxplots univariater Daten. Er besteht aus einem Paar konzentrischer Ellipsen (engl. *hinge* und *fence*, Everitt (2005)) und basiert auf einem robusten Schätzer für Lage, Maß und Korrelation. Nur ungefärbte Kontraste wurden für die weiteren Analysen herangezogen.

Alle Datensätze außerhalb der 15% und 85% Quantile wurden entfernt. Das Ergebnis war eine neue Datenmatrix X mit 41 verbleibenden Kontrasten. Die 35 herausgefilterten Kontraste wurden stichprobenartig auf Fehler untersucht und das Ergebnis bestätigte unsere Vorgehensweise. Die meisten Datensätze waren fehlerhaft, einige sogar derart entartet, dass sämtliche

P-Werte der Genexpressionsanalysen einem Wert von 1 nahe kamen.

Bei der Untersuchung von experimentellen Datensätzen aus unterschiedlichen Labors und experimentellen Bedingungen sind effiziente Datentransformationsmethoden notwendig um ein vertretbares Maß an Vergleichbarkeit zwischen den Datensätzen herzustellen. Bei der Betrachtung von Fold Changes aus Microarrayexperimenten muss zusätzlich noch in Betracht gezogen werden, dass diese implizit durch ihr algebraisches Vorzeichen eine *Richtung* der differentiellen Expression angeben. Diese verliert jedoch mehr und mehr an Interpretierbarkeit, je mehr unterschiedliche experimentelle Bedingungen in die Studie mit einfließen. Daher wurde hier nur der Betrag der Fold Changes für die weiteren Analysen verwendet, jeder der 41 Kontraste x wurde individuell mittels einer *Box-Cox-Transformation* (Box and Cox, 1964)

$$x' = \begin{cases} (x^p - 1)/p & \text{if } p \neq 0 \\ log(x) & \text{if } p = 0 \end{cases} \tag{3.1}$$

nahe an eine Normalverteilungsform gebracht und anschließend auf Mittelwert 0 und Varianz 1 standardisiert. Die Power-Koeffizienten p der *Box-Cox-Transformation* wurden mittels Maximum-Likelihood aus den Daten geschätzt und das transformierte Ergebnis durch QQ-Plots überprüft. Die durchschnittlichen errechneten Parameter p lagen um 0.13, was in etwa einer logarithmischen Transformation entspricht.

3.2.2 Dimensionsreduktion durch eine Kernel-PCA

Die Hauptkomponentenanalyse (PCA) gehört zu den Ordinationsverfahren und hat das Ziel hochdimensionale Daten in eine niederdimensionale Darstellung zu bringen. Dies geschieht durch Errechnung eines neuen Satzes von Basisvektoren aus der Eigenwertzerlegung der zugehörigen Kovarianzmatrix und Projektion der Daten in diesen neuen *linearen* Unterraum.

Das Verfahren stößt an seine Grenzen da, wo die beteiligten Datenpunkte nicht mehr linear zu trennen sind. Zu diesem Zweck wurde im Bereich des maschinellen Lernens die sogenannte Kernel PCA eingeführt (Schölkopf *et al.*, 1998). Sie ist eine nicht-lineare Erweiterung der klassischen PCA die vor der Dimensionsreduktion die Daten aus dem Eingaberaum I implizit in einen — möglicherweise noch höherdimensionalen — Merkmalsraum F, den sogenannten *Feature Space*, projiziert. Implizit, da die Abbildung durch Austauschen des Standardskalarprodukts mit einer positiv-semidefiniten symmetrischen Bilinearform, der Kernel Funktion κ erfolgt (Gleichung 3.2). Der PCA Algorithmus wird derart umgeformt, dass er seine sogenannte *Dualform* erhält und sämtliche Berechnungen nur noch mittels der Matrix XX' (Shawe-Taylor and Cristianini, 2004) der paarweisen Skalarprodukte der Datenpunkte ausführt, welche auch die Bezeichnungen Gram- oder Kernelmatrix K erhält (Gleichung 3.3). Etwas präziser formuliert, für eine zeilen-

indizierte Datematrix X und eine Abbildung $\phi : I \to F$, $x \mapsto \phi(x)$ sind die Kernelfunktion κ und die zugehörige Kernelmatrix K definiert wie folgt:

$$\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (3.2)$$

$$K_{ij} = \kappa(x_i, x_j). \quad (3.3)$$

Durch die mögliche Nichtlinearität der Kernelfunktion und die entsprechende „Krümmung“ des Merkmalsraums ist die Kernel PCA in der Lage auch nichtlineare Muster in den Daten zu erkennen, die die klassische PCA nicht oder nur unvollständig berücksichtigt hätte.

Für unsere Analysen kam die Implementierung des Kernel PCA Algorithmus im Paket *kernlab* zum Einsatz (Karatzoglou *et al.*, 2004). Als Kernelfunktion κ wählten wir einen polynomiellen Kernel

$$\kappa(x_i, x_j) = (s \langle x_i, x_j \rangle + k)^d$$

vom Grad $d = 2$, Skalierung $s = 1$ und Offset $k = 0$.

Der Algorithmus war in der Lage den gesamten Informationsgehalt der Matrix mittels 38 Hauptkomponenten wiederzugeben. 22810 Gene von 41 Experimenten konnten somit in Form einer 41×38 Datenmatrix dargestellt werden ohne nennenswerten Informationsverlust. Wenn man darüberhinaus annimmt, dass etwa 20% der Varianz in den Daten Rauschen ist — eine sicherlich nicht zu pessimistische Schätzung angesichts der Heterogenität der Daten —, so reichen die ersten 25 Hauptkomponenten aus, um eine Rauschreduzierte Datenmatrix abzubilden die immer noch für 80.585% der Varianz der Daten aufkommt. Für einen detaillierten Überblick über die Informationsverteilung auf die einzelnen Achsen siehe auch Tabelle 3.1.

3.2.3 Drei Gruppen von Kontrasten

Eine graphische Inspektion der PCA Plots (Abbildung 3.2.3) zeigte unmittelbar drei Hauptcluster von Experimenten auf, dazu noch mehrere kleinere.

Im Gegensatz zu typischen anderen Metaanalysen wurden die Cluster unmittelbar durch die nicht-überwachte Analyse detektiert, während sie sonst oft als a-priori definierte Covariate in die Analysen einfließen. Die drei sichtbaren Cluster wurden anschließend durch Anwendung eines Spektralclusterings (Karatzoglou *et al.*, 2004; Ng *et al.*, 2001) statistisch verifiziert. Ausgehend von der Annotation der Datensätze aus der GEO Datenbank korrespondierten die drei gefundenen Experimentcluster mit drei grundlegenden experimentellen Bedingungen, denen die Pflanzen während der Experimente ausgesetzt waren:

1. Auxin Behandlung oder Inhibierung (Dreiecke)

2. Aktivierung der pflanzlichen Pathogenabwehr (gefüllte Kreise)

	PC1	PC2	PC3	PC4	PC5
PV	0.10035	0.05383	0.05003	0.04640	0.03887
CP	0.10035	0.15418	0.20422	0.25062	0.28949

	PC6	PC7	PC8	PC9	PC10
PV	0.03725	0.03250	0.03226	0.03142	0.02973
CP	0.32674	0.35925	0.39151	0.42293	0.45267

	PC11	PC12	PC13	PC14	PC15
PV	0.02793	0.02699	0.02647	0.02606	0.02470
CP	0.48061	0.50761	0.53409	0.56016	0.58486

Tabelle 3.1: Varianzverteilung der ersten 15 Hauptkomponenten auf die 41×22810 Datenmatrix der *Arabidopsis thaliana* Microarray Daten. Die 15 Achsen erklären beinahe 60% der Gesamtvarianz. Abkürzungen: PV = Varianzanteil / Proportion of Variance, CP = Kumulativer Varianzanteil / Cumulative Proportion of variance.

3. Experimente die nicht einer der beiden Kategorien zugeordnet werden konnten, im folgenden als "andere" bezeichnet.

Für eine detaillierte biologische Interpretation der Cluster, siehe 3.2.5.

Um weitere strukturelle Einsichten in die Experiment-Kontrast Beziehungen zu gewinnen, kam ein hierarchischer Clustering Algorithmus nach Ward's Verfahren der minimalen Varianz zur Anwendung (Ward, 1963), welcher versucht kompakte sphärische Cluster mittels einer euklidischen Metrik zu bestimmen. Um dabei Achsen mit hohem Informationsgehalt aus der vorhergehenden PCA Analyse ein höheres Gewicht geben zu können wurden die Datenpunkte mit der Wurzel der Eigenwerte der Kernelmatrix K skaliert, d.h.:

$$\widetilde{X} = S\Lambda^{1/2}. \tag{3.4}$$

Die Robustheit der Gruppen wurde anschließend durch ein 1000-faches Multiscale Bootstrapping (*pvclust* Paket, (Suzuki and Shimodaira, 2006)) verifiziert (siehe Abbildung 3.2.3). In Übereinstimmung mit den Ergebnissen des Spektralclusterings zuvor und der graphischen Inspektion der paarweisen Scatterplots wurden die drei hauptsächlichen experimentellen Cluster auch in diesem Dendrogramm mit großer Unterstützung durch die Bootstrapanalyse (AU Werte ≥ 89) wiedergegeben.

Da die drei Hauptcluster in erster Linie durch die X-Achse des kPCA Scatterplots getrennt wurden, postulierten wir dass die erste Achse ausreichen könnte um Gene zu selektieren deren Koregulationsmuster eindeu-

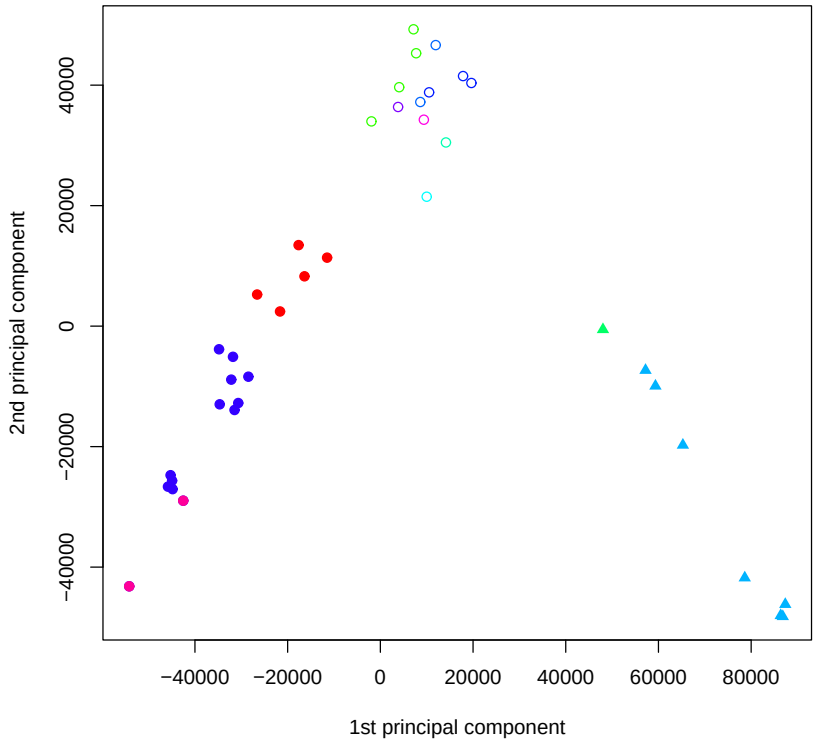

Abbildung 3.2: Scatterplot der Datenpunkte auf den ersten beiden KPCA Hauptachsen. Deutlich zu erkennen sind auch hier die drei Hauptcluster, die den Hauptgruppen Pathogen (Dreiecke), IAA (gefüllte Kreise) und Andere entsprechen.

tig zwischen auxinabhängigen, pathogenabhängigen und den übrigen Kontrasten unterscheiden könnten.

3.2.4 Eine neue Methode der Genauswahl

Um eine effektive Merkmalsidentifikation zu erreichen, d.h. um jene Gene zu finden, die zur Einteilung in die einzelnen Cluster maßgeblich beitragen, wurden bereits eine Vielzahl an Methoden vorgeschlagen. Selbstorganisierende Karten (SOMs) (Tamayo *et al.*, 1999), Maximal Margin Linear Programming (MAMA) (Antonov *et al.*, 2004), Korrelationsbasierte Merkmalseliminierung (CFS) (Hall, 1999) oder Rekursive Merkmalseliminierung (RFE) mittels Support Vektor Maschinen (SVMs) Guyon *et al.* (2002); Zhang *et al.* (2006). In konsequenter Fortführung unseres Ansatzes der ex-

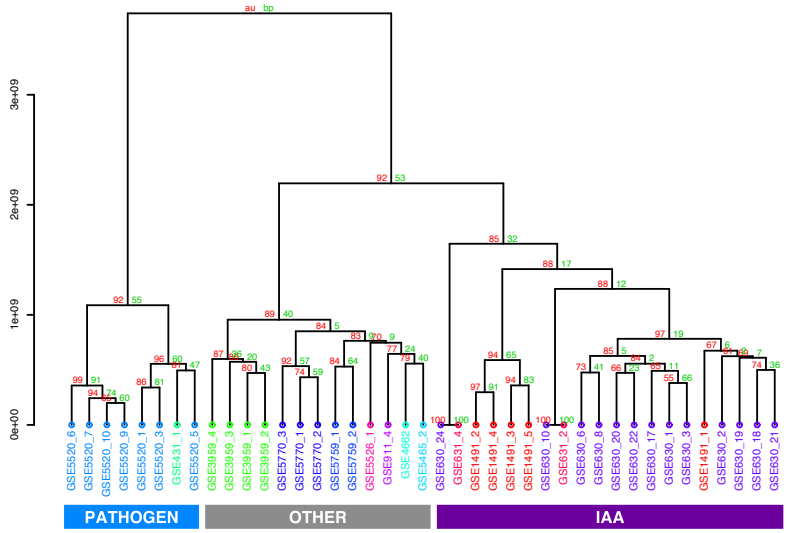

Abbildung 3.3: Hierarchischer Clusterbaum der 41 in der Metaanalyse verwendeten Kontraste. Die drei Hauptgruppen Pathogen, IAA und Andere sind deutlich zu erkennen.

plorativen Analyse haben wir Gene ausgewählt, die eine starke Assoziation mit der ersten kPCA Hauptkomponentenachse zeigten, d.h. wir haben den Beitrag (die sog. Loadings) der Gene zu den Achsen berechnet und darauf basierend unsere Auswahl getroffen. Um dies im Hinblick auf den kerneldefinierten Merkmalsraum zu erreichen, projizierten wir einzelne künstlich erzeugte Kontraste mit jeweils nur einem deregulierten Gen auf das neue Koordinatensystem. Jedes der 22810 künstlichen Experimente wurde derart gewählt, dass es einen großen absoluten Fold Change in dem interessanten, einen Fold Change gleich Null hingegen in allen anderen Genen besaß. Aus der resultierenden 22810×38 Matrix mit den Loadings der Gene auf die Achsen wählten wir die 500 stärksten Einflüsse aus, für sowohl positive (auxinabhängige) wie auch negative (pathogenabhängige) Extrema. Um die Genauigkeit des Auswahlverfahrens zu überprüfen wiederholten wir den vorangegangenen kPCA Schritt auf den ausgewählten Genen und inspizierten die paarweise Scatterplots der ersten 20 Hauptachsen für sowohl auxinannotierte als auch pathogenannotierte Gene. Alle kPCA Plots des IAA (Auxin) Datensatzes zeigten eine weite Verteilung der IAA-annotierten Kontraste über die Achsen, während alle anderen Kontraste auf einen kompakten lokalen Cluster projiziert wurden. Ein nahezu identisches Ergebnis wurde bei den pathogenannotierten Datensätzen festgestellt und zeigt, dass Expres-

sionsmuster, die nicht zu einem der beiden genannten Kontraste gehören, effektiv durch das Auswahlverfahren entfernt wurden.

3.2.5 Biologische Interpretation der Cluster

Das hierarchische Clustern auf allen kPCA Achsen in Abbildung 3.2.3 offenbarte drei hauptsächliche Cluster von Kontrasten: Solche die die Pathogenabwehr untersuchen (blau), solche die Auxin / IAA Effekte untersuchen (violett) und "andere" (grau). Die Bezeichnungen an den Knoten verbinden die GEO Zugriffsnummer mit einem Index der die Kontrastnummer angibt. Für eine detaillierte Beschreibung der einzelnen Kontraste siehe Tabelle 3.2.

	Sample Group 1		Sample Group 2		
Contrast	Genotype	Treatment	Genotype	Treatment	Cluster
GSE1491_1	WT Col-0	IAA	WT Col-0	non	IAA
GSE1491_2	WT Col-0	IAA inhibitor A	WT Col-0	non	IAA
GSE1491_3	WT Col-0	IAA inhibitor B	WT Col-0	non	IAA
GSE1491_4	WT Col-0	IAA/IAA inhibitor A	WT Col-0	non	IAA
GSE1491_5	WT Col-0	IAA/IAA inhibitor B	WT Col-0	non	IAA
GSE3959_1	MU LEC2GR	1h LEC2 induction	MU LEC2GR	no LEC2 induct.	other
GSE3959_2	MU LEC2GR	4h LEC2 induction	MU LEC2GR	no LEC2 induct.	other
GSE3959_3	MU LEC2GR	1h LEC2 induction	WT WS-0	4h LEC2 induct.	other
GSE3959_4	MU LEC2GR	4h LEC2 induction	WT WS-0	NA	other
GSE431_1	pmr4-1 MU	non	pmr4-1 MU	powdery mildew	pathogen
GSE4662_1	MU STA1	non	WT	NA	other
GSE5465_2	MU OETOP6B	non	WT	NA	other
GSE5520_1	WT Col-0	DC1318 Cor 10e6	MU STA1	non	pathogen
GSE5520_10	WT Col-0	EcTUV86-2 fliC 10e8	WT Col-0	non	pathogen
GSE5520_3	WT Col-0	DC3000 10e6	WT Col-0	non	pathogen
GSE5520_5	WT Col-0	DC1318 Cor 5x10e7	WT Col-0	non	pathogen
GSE5520_6	WT Col-0	DC3000 hrpA-fliC 10e8	WT Col-0	non	pathogen
GSE5520_7	WT Col-0	DC3000 hrpA 10e8	WT Col-0	non	pathogen
GSE5520_9	WT Col-0	EcO157H7 10e8	WT Col-0	non	pathogen
GSE5526_1	WT?	non	WT?	non	other
GSE5759_1	WT Col-0	dark + lincomycin	WT Col-0	dark	other
GSE5759_2	WT Col-0	red light + lincomycin	WT Col-0	red light	other
GSE5770_1	WT Col-0	lincomycin	WT Col-0	non	other
GSE5770_2	abi4-102 MU	lincomycin	abi4-102 MU	non	other
GSE5770_3	gun1-1 MU	lincomycin	gun1-1 MU	non	other
GSE630_1	WT Col-0	IAA (2h 5μM)	WT Col-0	EtOH (2h)	IAA
GSE630_10	MU arf2-6	IAA (2h 5μM)	MU arf2-6	EtOH (2h)	IAA
GSE630_17	MU IAA17-6	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_18	MU arx3-1	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_19	MU i5i6i19	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_2	MU nph4-1	IAA (2h 5μM)	MU nph4-1	EtOH (2h)	IAA
GSE630_20	MU IAA17-6	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_21	MU arx3-1	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_22	MU i5i6i19	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_24	MU arf2-6	IAA (2h 5μM)	WT Col-0 A2	IAA (2h 5μM)	IAA
GSE630_3	MU arf19-1	IAA (2h 5μM)	MU arf19-1	EtOH (2h)	IAA
GSE630_6	MU IAA17-6	IAA (2h 5μM)	MU IAA17-6	EtOH (2h)	IAA
GSE630_8	MU i5i6i19	IAA (2h 5μM)	MU i5i6i19	EtOH (2h)	IAA
GSE631_2	MU arf2-6	IAA (2h 5μM)	MU arf2-6	non	IAA
GSE631_4	MU arf2-6	IAA (2h 5μM)	WT Col-0	IAA (2h 5μM)	IAA
GSE911_4	35S::LFY	non	WT ler	35S::LFY	other

Tabelle 3.2: Jeder Kontrast besteht aus zwei Gruppen, die durch ihren genetischen Hintergrund und die Behandlung beschrieben sind. Die letzte Spalte "Cluster" stammt aus der Spektralclusterung der kPCA Analyse. Die Kontrastebezeichnungen setzen sich aus der GEO Zugriffsnummer und einem fortlaufenden Kontrastindex zusammen.

Jeder Kontrast besteht aus einem Vergleich zweier Proben (Samples), wovon bei jedem der genetische Hintergrund (bspw. Wildtyp oder Mutante) sowie die Behandlung (Bestrahlung, Toxine, Pathogene, Hormone) auf-

geführt ist.

Bei genauerer Betrachtung des IAA Clusters ließ sich dieser weiter in zwei Gruppen auftrennen. Die Studie GSE1491, die die Wirkweise eines IAA Inhibitors untersuchte, war klar getrennt von den verbleibenden Kontrasten die hautpsächlich die Auswirkung von IAA auf verschiedene Mutanten untersuchten, besonders auf solche mit Defekten in der IAA Biosynthese oder assoziierter Signalübertragungswege. Der Cluster der „übrigen“ Kontraste bestand aus Studien aus einer ganzen Reihe von Effekten, beispielsweise der Zugabe von Lycomycin, einem Inhibitor der Proteinstranslation in Plastiden, Regulationsänderungen eines Transkriptionsfaktors zur Embryogenese oder Untersuchungen von stresstoleranten Mutanten. Natürlich kamen in diesem Cluster von unterschiedlichsten Kontrasten solche aus einer Studie nahe beieinander zu liegen. Die Architektur des hierarchischen Clusterdendrogramms zeigt, dass die initiale Aufbereitung der Daten zusammen mit der Kernel PCA in der Lage waren dafür zu sorgen, dass Kontraste aus biologisch ähnlichen Experimenten in der Tat als näher zueinander verwandt angesehen wurden als solche aus unterschiedlichen. Damit sind wir mit unserer Analyse in der Lage eine Vergleichbarkeit von Microarray Datensätzen aus unterschiedlichen Laboratorien und mit unterschiedlichen experimentellen Bedingungen herzustellen — eine nichttriviale Aufgabe, wenn man die Vielzahl der Datenquellen betrachtet die in dieser Analyse zum Einsatz kamen.

3.2.6 Regulation von *Arabidopsis thaliana* Genen durch Indol-3-essigsäure (IAA)

Zur Verifikation der vorhergesagten Gene, die sich für den IAA Cluster als repräsentativ erwiesen hatten, kam das Programm MapMan (Usadel *et al.*, 2005) zum Einsatz. Es zeigte sich, dass die von uns als IAA-abhängig annotierten Gene auch in der MapMan Darstellung zum Großteil in der MapMan-eigenen IAA Kategorie zu liegen kamen, die dort 215 Gene umfaßt. Von den 500 augewählten Genen waren 43 auch in der MapMan Gruppe enthalten, so dass 2% der Gesamtgene bereits 20% der als IAA-abhängigen MapMan Gene ausmachten.

In der *Hormon*-Untergruppe *Ethylen* und in der Kategorie *Transkriptionsfaktor* sind viele Gene unter IAA Behandlung reguliert, während eine kleinere Anzahl von Genen auf die MapMan Gruppen *Cytochrome P450* und *Cell Wall* entfällt.

Regulierte Gene aus der *Ethylen* Untergruppe sind entweder in Ethylensynthese oder Signaltransduktion involviert. Ethylen ist Bestandteil der Regulation einer Reihe von Entwicklungsprozessen, oft im Zusammenspiel mit anderen pflanzlichen Signalhormonen. So kann z.B. Auxin eine Ethylensynthese anregen, wobei Ethylen wiederum eine Erhöhung des Auxinspiegels auslösen kann. Einige Prozesse, wie z.B. Wurzelwachstum, das differentiel-

le Wachstum im Hypokotyl oder die Bildung von Wurzelhaaren werden in *Arabidopsis thaliana* sowohl von Auxin als auch Ethylen reguliert (Stepanova *et al.*, 2005). Bei allen GEO Datensätzen, die wir als IAA-assoziiert annotiert haben, stammt die extrahierte RNA aus Setzlingen, wo derartige Prozesse sehr wahrscheinlich unter IAA Einfluss reguliert werden.

Cytochrome P450 Monooxygenasen sind in zahlreiche biosynthetische Reaktionen verwickelt, z.B. die Synthese von Pflanzenhormonen oder Stoffe des Immunsystems. Die Regulation von Zellwandgenen ist ebenfalls zu erwarten, da Auxin die Verlängerung von Zellen durch ein Dehnen der Zellwand auslöst, welches Restrukturierungsprozesse der Zellwand in Gang setzt.

Zusammenfassend läßt sich sagen, dass das Genauswahlverfahren unserer nicht-überwachten Metaanalyse eine Vielzahl von Genen ausgewählt hat, die sich im Nachhinein als IAA-reguliert herausgestellt haben, sei es durch Annotation oder unabhängige Validierung.

3.2.7 Regulation von Genen durch Pathogenexposition

Unser Genauswahlverfahren für den pathogen-assoziierten Cluster lieferte eine große Anzahl regulierter Gene aus den folgenden MapMan Kategorien (Usadel *et al.*, 2005): *Biotischer Stress*, *Rezeptorkinasen*, *Photosynthese* (Lichtreaktionen), *Alkaloid-ähliche Proteine* aus *Sekundärmetabolismus*, *Nitrilasen*, *Zellwandgene* und *WRKY Transkriptionsfaktoren* (Abbildungen 3.2.7 und 3.2.7).

Von all diesen funktionalen Kategorien wurde bereits berichtet, dass sie durch Pathogenangriffe reguliert werden und eine Rolle in der pflanzlichen Immunabwehr spielen. Im folgenden soll nur eine kurze Beschreibung der Funktionen der einzelnen Kategorien gegeben werden:

Biotischer Stress: Diese Kategorie beschreibt eine Reihe von Genen die als relevant für die Pathogenabwehr annotiert sind.

Rezeptorkinasen: Rezeptorkinasen können Signaltransduktionskaskaden auslösen und werden im nächsten Abschnitt detaillierter beschrieben

Photosynthese: Eine Veränderung des Kohlenhydratstoffwechsels wurde nach einem Angriff durch *Pseudomonas synrinae* und *Botrytis cinerea* von Berger *et al.* festgestellt (Berger *et al.*, 2004). Die Autoren berichteten von einer Koregulation von Genen der Immunabwehr sowie der Photosynthese als Antwort auf die Infektion mit besagten Mikroorganismen.

Alkaloide: Alkaloide sind Sekundärmetabolite welche im Allgemeinen nicht essentiell sind für die grundlegenden metabolischen Prozesse der Pflanze. Sie spielen jedoch eine wichtige Rolle in der pflanzlichen Pathogenabwehr (Dixon, 2001) und werden von der Pflanze produziert um

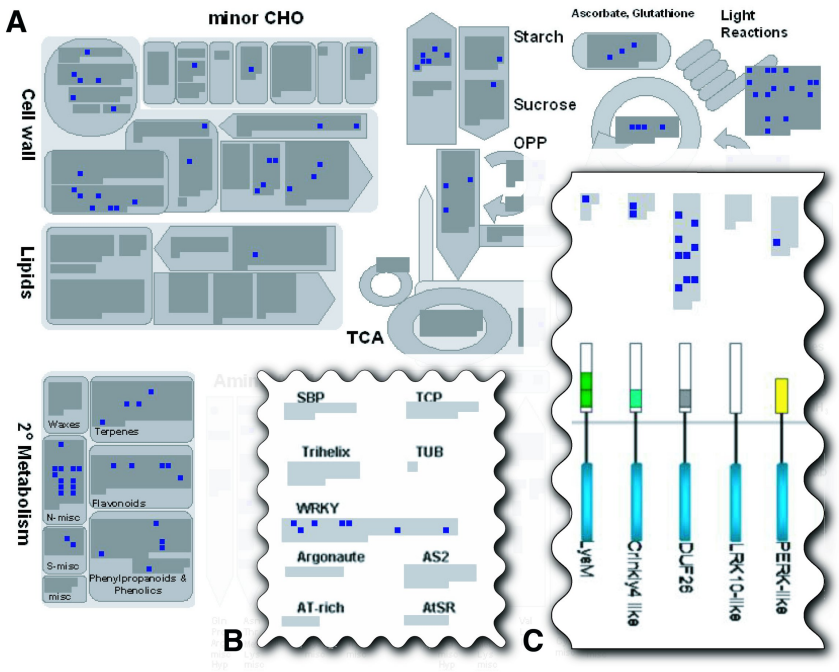

Abbildung 3.4: Überblick über Gene die als pathogenabwehrassoziiert vorhergesagt wurden. Ein blauer Punkt entspricht einem von unserer Genselektion ausgewählten Gen. Dunkelgraue Punkte zeigen die übrigen Gene. A) Überblick über den Stoffwechsel. Eine Regulation von Zellwandgenen (oben links), Alkaloide, die in die Kategorie N-misc. des Sekundärmetabolismus fallen, sowie Lichtreaktionen der Photosynthese ist zu erkennen. B) Teil der Transkriptionskarte, die die Regulation der WRKY Transkriptionsfaktoren zeigt. C) Ausschnitt aus der Karte der Rezeptorkinasen, die die Regulation der DUF26 Gene aufzeigt. Die Abbildung basiert auf Karten des Programms *MapMan* zur Analyse von Stoffwechselwegen (Usadel *et al.*, 2005).

die Energieversorgung der Erreger zu erschweren. Die Anhäufung von antimikrobischen Substanzen ist oft reguliert von Signaltransduktionswegen die die Erkennung des Erregers durch Pflanzenrezeptoren erfordern, welche von spezifischen Wirtsresistenzgenen codiert werden (Dangl and Jones, 2001; Piroux *et al.*, 2007). Die Regulation der DUF26 Gene die von unserem Genauswahl ebenfalls betroffen waren und von uns als pathogen-abhängig annotiert wurden könnte daher mit dieser Erkennungsfunktion zu tun haben.

Nitrilasen: Nitrilasen sind an der IAA Biosynthese beteiligt und katalysieren die Umwandlung von Indol-3-acetonitril in IAA. Die Induktion von vier *Arabidopsis thaliana* Nitrilasen durch das Pathogen *Pseodo-*

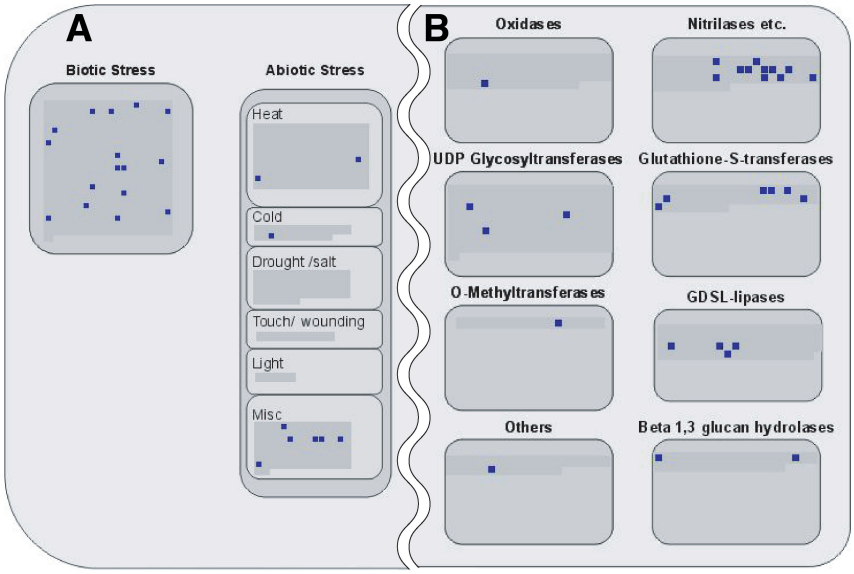

Abbildung 3.5: Überblick über Stressgene (A) und Gene großer Enzymfamilien (B) aus pathogen-assoziierten Kontrasten. Regulierte Gene sind als blaue Punkte, nicht regulierte Gene in grau angezeigt. Die Kategorien *Biotischer Stress* und *Nitrilasen* enthalten eine große Anzahl Gene repräsentativ für pathogeninduzierte Expressionsänderungen. Zur Erstellung der Karten kam das Programm *MapMan* zum Einsatz (Usadel *et al.*, 2005).

monas syringae wurde von Bartel and Fink gezeigt (Bartel and Fink, 1994).

Zellwandgene: Da die Zellwand eine natürliche Barriere für Pflanzenpathogene darstellt umfaßt die Immunantwort der Pflanze Modifikationen der Zellwand ebenso wie Induktion von Biosynthesewegen zur Verstärkung der Zellwände und zusätzlichen Erschwerung weiterer Angriffe (Cheong *et al.*, 2002).

WRKY TFs: Die Regulation der WRKY Transkriptionsfaktoren wird auch von der Begleitpublikation zum Datensatz GSE5520 beschrieben Thilmony *et al.* (2006). Unsere Beobachtungen stützen die Vermutung der Autoren, dass diese Transkriptionsfaktoren die Immunantwort der Pflanzen mitregulieren.

Insgesamt war das Genauswahlverfahren der Metaanalyse in der Lage biologisch wichtige Gene herauszufiltern, von denen viele experimentell als im Zusammenhang mit der Pathogenabwehr der Pflanzen stehend verifiziert werden konnten. Die Funktion der übrigen Gene bleibt zu klären, es ist

jedoch nicht unwahrscheinlich sie im Umfeld der pflanzlichen Immunantwort zu finden.

3.2.8 Serine-threonine Kinasen und die Immunabwehr

Wie in Abbildung 3.2.7 gezeigt umfasst der Satz an pathogenassoziierten Genen auch eine Reihe von Rezeptorkinasen, wovon viele zur Familie der Serine / Threonine Kinasen, der DUF26 Unterfamilie gehören. Sie alle teilen die selbe Domänenkomposition und -reihenfolge und bestehen aus einem Signalpeptid, einer extrazellulären Region die zwei Domänen unbekannter Funktion enthält (DUF26 und PF01657) und eine cytosolische Serin- / Threoninkinase Domäne (pkinase, PF00069). Laut der SMART Datenbank (Letunic *et al.*, 2006) kommen Proteine dieser Familie ausschließlich in *Streptophyta* vor. Die 9 potentiellen Rezeptorkinasen haben eine hohe Ähnlichkeit in Domänenzusammensetzung und Nukleotidsequenz mit der rezeptorartigen Kinase 4 von *Arabidopsis thaliana* (Swiss-Prot-ID Q9C5T0). Dieses Enzym ist bekannt als Mitglied eines systemisch aquirierten Resistenzweges in höheren Pflanzen. Seine Expression kann von einem regulatorischen Protein aktiviert werden, aktiviert von einer Interaktion zwischen Salicylsäure und dem Pathogen (Du and Chen, 2000). Salicylsäure wiederum ist ein Signalmolekül welches eine systemisch erworbene Resistenz in der Wirtspflanze auslösen kann (Ryals *et al.*, 1996). Diese Ergebnisse unterstützen die Hypothese einer Funktion der potentiellen Rezeptorkinasen in Prozessen der Pathogenabwehr.

Zwei der DUF26 Gene (At4g21400 und At4g21410) waren auch in den Kontrasten des Datensatzes GSE3959 sowie in einem Kontrast des Datensatzes GSE5770 differentiell exprimiert. Ersterer hatte eine Untersuchung des B3 Domänen-Proteins LEAFY COTYLEDON2 (LEC2) zum Ziel, einem Transkriptionsfaktor, der für mehrere Bereiche der Embryogenese benötigt wird. In letzterem Datensatz wurden *abi4* Mutanten mit Lincomycin behandelt und mit unbehandelten Mutanten verglichen. Daraus könnte man schließen, dass die beiden DUF26 Kinasegene entweder in mehr als einem Signaltransduktionsweg eine Rolle spielen oder aber, dass der gleiche Signalweg mehrere zelluläre Funktionen reguliert. In jedem Fall könnte dies ein interessanter Ausgangspunkt für eine genauere Untersuchung dieser Signalwege sein.

Wie aus Abbildung 3.2.8 zu sehen ist waren die DUF26 Gene nicht in jedem Experiment zur Pathogenabwehr differentiell exprimiert. Auch dafür kommen mehrere Gründe in Frage. So könnte z.B. die Varianz in den einzelnen Microarray Experimenten so hoch gewesen sein, dass die differentielle Expression nicht detektiert werden konnte, oder biologische Effekte sorgten dafür dass der Fold Change zu niedrig war um als signifikant bewertet zu werden. Auch diese Erkenntnis könnte ein interessanter Startpunkt zur weiteren Untersuchung der DUF26 Kinasegene sein.

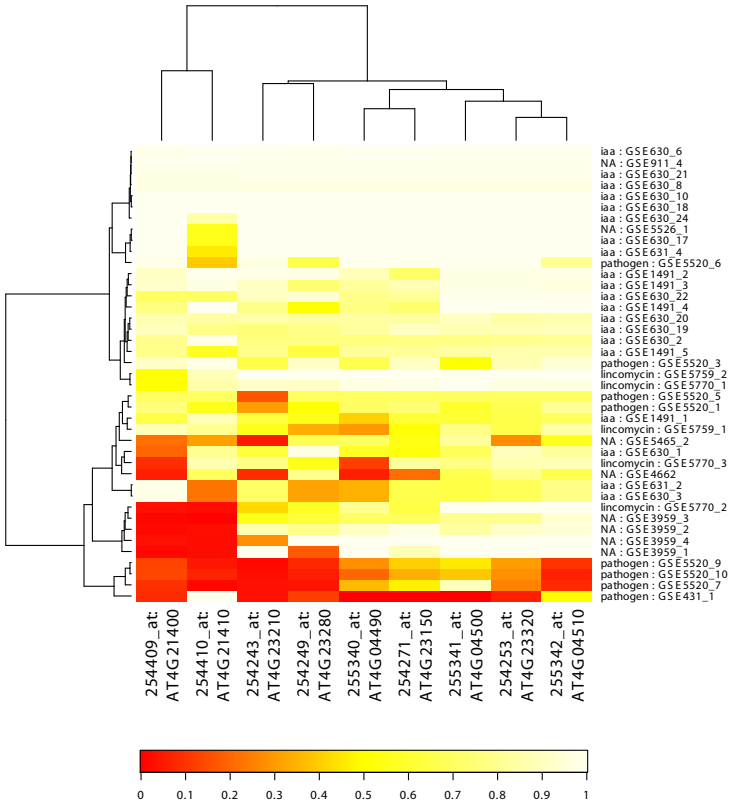

Abbildung 3.6: Heatmap der P-Werte der DUF26 Kinasegene. Rote Felder weisen auf niedrige P-Werte, gelbe Felder auf hohe P-Werte hin. Die DUF26 Kinasegene sind in den vier pathogen-assoziierten Kontrasten stark differentiell exprimiert.

3.3 Diskussion

Öffentliche Microarray Datenbanken häufen seit langem große Mengen von Daten an die bisher nur spärlich für Metaanalysen eingesetzt wurden. Durch ihre großen Informationsmengen können zusätzliche Schlüsse über die Funktion und Regulation von Genen gezogen werden die nicht aus einzelnen Microarray Datensätzen ableitbar sind. In dieser Studie haben wir einen neuartigen Ansatz einer nicht-überwachten Metaanalyse vorgeschlagen. Durch eine konservative Vorprozessierung der Daten in Kombination mit einer kernelisierten Hauptkomponentenanalyse und mehreren Clusteringverfahren waren wir in der Lage robuste und biologisch signifikante Cluster aufzuklären.

Um die für die Cluster verantwortlichen Gene zu finden wurde ein neuer Ansatz der Merkmalsauswahl definiert, der auf den Loadings der Haupt-

komponentenachsen basiert. Ein derartiger Ansatz hat bisher im Kontext der Metaanalyse unseres Wissens noch keine Anwendung gefunden, und hat gegenüber konkurrierenden Verfahren einen entscheidenden Vorteil: Es wird nach Linearkombinationen von Merkmalen (Genen) gesucht, die die Experimentklassen möglichst gut voneinander trennen. Dies erweist sich im Allgemeinen als schwierig, da der abzusuchende Raum aller möglichen Linearkombinationen üblicherweise zu groß ist um vollständig durchsucht zu werden. Durchdachte Heuristiken und Optimierungsverfahren müssen zu diesem Zwecke zur Anwendung kommen die sehr wahrscheinlich unterschiedliche Ergebnisse liefern (Zhang *et al.*, 2006). Eine Methode wie die hier vorgeschlagene umgeht dieses Problem effizient indem sie direkt auf den Loadings der PCA Analyse arbeitet. Die Eigen- oder Singulärwertzerlegung der Kernelmatrix ist deterministisch und damit auch die Ergebnisse des Genauswahlverfahrens. Die von uns gefundenen Gene haben sich als repräsentativ für eine Gruppe von Kontrasten herausgestellt und konnten — zumindest in Teilen — experimentell verifiziert werden. Auch erwiesen sich die Gene als robust gegenüber einem künstlichen Hinzufügen von Rauschen zur Datenmatrix, eine weitere positive Eigenschaft der von uns vorgeschlagenen Methode.

Es ist das Genauswahlverfahren, das in erster Linie von einer Metaanalyse profitiert. Schwache regulatorische Signale können in einem einzelnen Datensatz leicht übersehen werden, d.h. den Genen wird wahrscheinlich aufgrund des niedrigen Fold-Changes und einer hohen Gesamtvarianz ein zu hoher P-Wert attestiert. Diese Situation verschlimmert sich noch zusätzlich nachdem das Gesamtniveau der P-Werte durch eine Multiple-Testing-Korrektur weiter angehoben wurde, eine effektive Methode zum Herausfiltern dieser subtilen Signale. In einer Metaanalyse die viele Datensätze berücksichtigt kann ein Signal welches zwar schwach aber konsistent über mehrere Kontraste sichtbar ist dennoch berücksichtigt werden. Um diesen Effekt zu erhalten und einen frühzeitigen Verlust an Information zu verhindern, wurden in dieser Studie Fold-Changes und nicht die P-Werte der ursprünglichen Datensätze verwendet. Wir haben desweiteren die Analyse auf den Absolutwerten der Fold Changes durchgeführt, da das algebraische Vorzeichen bei unterschiedlichen Experimenten an Bedeutung verliert, ja sogar irreführend sein kann. Wenn z.B. ein Kontrast eines Experiments den Überschuss eines Faktors mit einer Kontrollprobe vergleicht und ein weiterer Kontrast das Fehlen desselben Faktors gegen die Kontrolle, dann erwarten wir Fold Changes mit gegensätzlichen Vorzeichen, möchten aber dennoch, dass die Experimente nahe beieinander zu liegen kommen da der gleiche Faktor in beiden untersucht wurde. In einigen Fällen war die „Richtung“ des Experimentaufbaus noch nicht einmal aus der Annotation der Datensätze ersichtlich.

Um sicherzustellen, dass diese Resultate nicht durch einfachere Modelle auch hätten erlangt werden können, und um ein Overfitting der Daten zu

verhindern, haben wir die Ergebnisse mit denen der herkömmlichen linearen PCA verglichen. Obwohl diese prinzipiell in der Lage war, einige der Hautpcluster von Experimenten wiederzufinden, blieb ihre Genauigkeit im hierarchischen Clustering sowie im Genauswahlverfahren weit hinter der der kernelisierten Version zurück. Zudem sollte bemerkt werden, dass die kPCA gegenüber der linearen Variante enorme Geschwindigkeitsvorteile mit sich bringt, da die Kernelmatrix nur einmal berechnet werden muss und ihre Dimension unabhängig von der Anzahl der Gene in der ursprünglichen Datenmatrix ist (22810 im Falle der ATH-1 Arrays).

Wir zeigen hier für einen großen *Arabidopsis thaliana* Datensatz, wie ein Genauswahlverfahren basierend auf den Hauptachsen der Kernel PCA Gene vorschlägt, die für IAA- oder Pathogenabwehrassoziierte Kontraste charakteristisch sind. Es konnte im Nachhinein durch Literaturrecherche und frühere experimentelle Studien gezeigt werden, dass diese Gene tatsächlich zu IAA-Effekten in Bezug stehen oder an der pflanzlichen Reaktion auf Pathogenangriffe teilhaben. Wir haben durch unsere Ergebnisse weiterhin postuliert, dass die Gene der DUF26 Kinase Familie ebenfalls eine Rolle in der pflanzlichen Immunabwehr spielen. Natürlich müssen weitere Experimente diese Hypothese überprüfen, aber es ist ein schönes Beispiel dafür, wie eine explorative Analyse weitere konkrete biologische Fragestellungen aufbringen kann.

Ganz allgemein können Metaanalysen, die mehrere höchst divergente Experimente in einer Analyse zusammenfassen, neuartige Annotationen zur Funktion oder Regulation von Genen aufwerfen, indem sie das Verhalten der Gene in den unterschiedlichen Bedingungen gemeinsam betrachten. Bemerkenswert ist dabei, dass diese Analysen nicht nur Experimentgrenzen sprengen und heterogene Daten analysieren können, sondern dass sie ganz erheblich von der Heterogenität der Daten profitieren.

Natürlich ist eine Metaanalyse wie die vorliegende auch spekulativ, aber so wie es in der klassischen Statistik gang und gäbe ist vor parametrischen Tests mittels explorativen Analysen die Integrität und Qualität der Daten zu überprüfen und erste Muster in den Daten zu erkennen, so möchten wir vorschlagen, dass es auch im Feld der Metaanalysen so gehalten wird. Hypothesen aus nicht-überwachten Verfahren können dann mittels parametrischer Methoden und biologischer Experimente verifiziert werden.

Wir haben hier gezeigt dass es möglich ist Datensätze aus unterschiedlichsten Datenquellen und mit unterschiedlichsten Fragestellungen in einer kohärenten Studie zu vereinen. Diese Analyse wurde angewandt, um repräsentative Gene von Clustern zu finden. Von Expressionsveränderungen zwischen solchen Clustern konnte auf die Funktion und Regulation einzelner Gene geschlossen werden. Die Affymetrix ATH-1 Plattform kam hier zum Einsatz, aber unser Ansatz kann auf jeder Plattform und an jeden Organismus zum Einsatz kommen, der die Berechnung eines logarithmischen Fold Changes erlaubt. Um einen einfachen Zugang zu unseren Daten zu erhalten,

ist eine Datenbank in Planung, der neue Datensätze einfach hinzugefügt und mit unseren bestehenden Datensätzen von *Arabidopsis thaliana* verglichen werden können.

Kapitel 4

Metabolismus: Netzplan der Zelle

4.1 Der Stoffwechsel

Nach unserer Betrachtung globaler Genexpressionsmuster in Form einer Metaanalyse soll der Blick von der Regulation der Gene zunächst weg und hin zu einer konkreten Klasse von Proteinen gewendet werden, den metabolisch aktiven Enzymen des Zellstoffwechsels. Mit ihrer Katalysatorfunktion im Stoffwechsel bilden diese Proteine große Netzwerke aus, deren holistische Beschreibung ohne Unterstützung von Algorithmen und Softwaresystemen kaum möglich ist. Insbesondere im Hinblick auf die Entwicklung neuer Pharmaka sind detaillierte Analysen metabolischer Netzwerke und ihrer Regulation wünschenswert, um effektive Ansatzpunkte zur medikamentösen Behandlung von Krankheiten aufzudecken und mögliche Gefahren und Nebenwirkungen auszuschließen.

In biochemischen Lehrbüchern werden Stoffwechselwege üblicherweise aus didaktischen Gründen als lineare Abfolge enzymatischer Reaktionen mit einigen wenigen Verzweigungen beschrieben, man denke beispielsweise an das Zusammenspiel aus dem Pentose Phosphatweg (PPP) und der Glykolyse. Die detaillierte Analyse zeigt aber, dass reale metabolische Netzwerke aus einer verblüffenden Vielfalt von Wegen, Subnetzen und Zyklen bestehen. So ergeben sich stark unterschiedliche metabolische Situationen in Abhängigkeit vom Bedarf der Zelle. Beispielsweise kann eine Zelle zu einem Zeitpunkt hauptsächlich Zucker (Glukose) für den Energiestoffwechsel benötigen, während zu einem anderen Zeitpunkt vor allem das Zellwachstum einschließlich Nukleotidproduktion eine Rolle spielt. Um diese unterschiedlichen Nutzungen des metabolischen Netzwerks aus Glykolyse, PPP und weiteren Enzymen zu beschreiben, sind mathematische Verfahren und Modelle notwendig die die Gesamtheit des Stoffwechsels auf der einen Seite, seine Modularität und internen Abhängigkeiten auf der anderen Seite,

präzise beschreiben zu können.

4.1.1 Möglichkeiten der Beschreibung: Dynamische vs. topologische Analyse

Unterschiedlichste Verfahren wurden bisher vorgeschlagen, um metabolische Netzwerke im Hinblick auf ihre Topologie, den zeitabhängigen Stoffumsatz, ganz allegemein ihre Kinetik, mögliche Flaschenhälse oder zentrale regulierende Enzyme zu untersuchen. Im Wesentlichen kann man jedoch zwei Möglichkeiten der Analyse unterscheiden — dynamische, also zeitabhängige und topologische oder statische Analysen. Die Wahl des Verfahrens ist stark abhängig von der zu beantwortenden Fragestellung, je nach gewähltem Verfahren trät das resultierende Netzwerkmodell unterschiedliche Verkürzungs- und pragmatische Merkmale. Dynamische Modelle fokussieren meist auf die Kinetik der beteiligten Enzyme. Zu beantwortende Fragestellungen sind oftmals, wie schnell bestimmte Reaktionsfolgen ablaufen, wieviel eines bestimmten Produktes in einer bestimmten Zeit produziert werden kann oder in welchem zeitlichen Rahmen Regulation stattfindet. Die Modelle enthalten oft nur eine begrenzte Anzahl von Enzymen und Reaktionen, da ausführliche Informationen über die Kinetik der Enzyme bei unterschiedlichen Umgebungsbedingungen benötigt werden. Das Modell selbst präsentiert sich oft als System von Differentialgleichungen, die die Metabolitkonzentrationen in Abhängigkeit von der Zeit beschreiben. Die benötigten Parameter müssen aus experimentellen Daten geschätzt werden, wobei sich durch die Natur der Differentialgleichungen kleine Fehler immer weiter aufsummieren, ein weiterer Grund warum derartige Modelle oft auf wenige Enzyme beschränkt bleiben. Prominente Beispiele dynamischer Modellierungssoftware beinhalten GEPASI (Mendes, 1993, 1997) und PLAS (siehe auch Abbildung 4.1.1 und vgl. Voit (2000)).

Im Gegensatz dazu verfolgt die topologische Modellierung einen anderen Ansatz. Verfahren dieser Kategorie werden auch als Steady-State Verfahren (Schilling *et al.*, 1999) bezeichnet, da sie ein Quasi-Fließgleichgewicht im Netzwerk annehmen, und sind auch für genomweite Netzwerke geeignet. Die Grundidee ist hier nicht die Beschreibung der Metabolitkonzentrationen über die Zeit, sondern allein der Stöchiometrie des Netzwerkes. Das System wird dazu in zwei Klassen von Metaboliten unterteilt, gepufferte, *externe* Metabolite und ungepufferte, *interne*. Die Steady-State Regel besagt nun, dass interne Metabolite sich nicht anhäufen dürfen, d.h. dass die stöchiometrischen Koeffizienten der Zuflüsse zu einem Metabolit gleich den Abflüssen sein müssen. Unter dieser Bedingung schränkt sich die Anzahl der Möglichkeiten des Stoffflusses durch das Netzwerk ein, Steady-State Analyse haben dann meist die vollständige Aufzählung aller stöchiometrisch ganzzahligen Wege durch das Netzwerk zum Ziel oder aber das Finden von Basisvektoren, deren Linearkombination alle erlaubten Zustände des Netzwerks beschreibt.

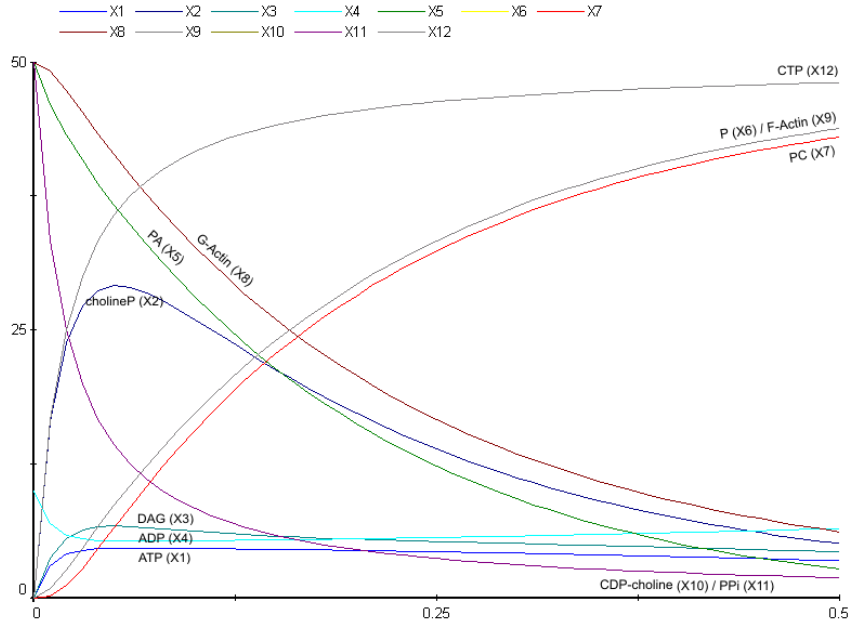

Abbildung 4.1: Dynamische Modellierung einer Phospholipidkaskade aus dem murinen Phagosom. Phospholipidkonvertierungen induzieren die Polymerisierung von granulärem G-Actin zu filamentösem F-Actin an der phagosomalen Membran. Bild erstellt mittels *PLAS - Power Law Analysis and Simulation* (Voit, 2000)

Derartige topologische Verfahren wie die Elementarmodenanalyse (EMA) Schuster and Hilgetag (1994); Schuster *et al.* (1999), die assoziierte Theorie der *Extreme Pathways* Papin *et al.* (2004) oder die *Flux Balance Analyse* (FBA) Kauffman *et al.* (2003) haben sich als besonders effektiv zur Modellierung von metabolischen Netzwerken erwiesen. Interessante Fragestellungen hierbei sind die Bestimmung sogenannter *Minimal Cutsets* Klamt and Gilles (2004), minimal zu blockierender Sätze von Reaktionen, die die Produktion eines bestimmten Stoffes verhindern, die Identifikation potentieller neuer pharmakologischer Targets Becker *et al.* (2006); Trawick and Schilling (2006), die Analyse von Netzwerkrobustheitsfaktoren Edwards and Palsson (2000); Wilhelm *et al.* (2004) oder die Bestimmung maximal möglicher Metabolitausbeuten Schuster *et al.* (2000).

4.1.2 Die Elementarmodenanalyse und verwandte Verfahren

Um eine ganzheitliche Netzwerkanalyse vorzunehmen, müssen, wie bereits eingangs erwähnt, sämtliche denkbaren Stoffwechselwege auf ihre thermodynamische und stöchiometrische Realisierbarkeit überprüft werden. Dazu

wird zunächst das Stoffwechselnetz in Form einer stöchiometrischen $m \times r$ Matrix S ausgedrückt welche die stöchiometrischen Koeffizienten der Metabolite (Anzahl Metabolite = m) in den einzelnen Reaktionen (Anzahl Reaktionen = r) darstellt. Ausgehend von der vereinfachten Annahme, dass die Konzentrationen der *internen* Metabolite sich zwar kurzfristig ändern aber integriert über die Zeit konstant bleiben, kann dann die Steady-State Bedingung über

$$S * x = 0 \tag{4.1}$$

ausgedrückt werden, wobei x der Vektor mit den Flußkoeffizienten der Reaktionen ist. Biochemisch rechtfertigen lässt sich dies mit dem Prinzip der Masseerhaltung und der Tatsache, dass hohe Metabolitkonzentrationen über einen längeren Zeitraum die Zelle schädigen würden, da der osmotische Druck Zellkomponenten beschädigen könnte und das ungünstige thermodynamische Reaktionsgleichgewicht der beteiligten Reaktionen den Stoffwechsel an dieser Stelle zum Erliegen bringen müßte. Bei dieser Modellannahme geht jedoch der Informationsgehalt der kurzfristigen Metabolitkonzentrationen der Zelle völlig verloren, so dass Signaltransduktionswege, in denen oftmals zeitabhängige Veränderungen von Transmitterkonzentrationen eine entscheidende Rolle spielen, dann mit diesen Verfahren nicht mehr adäquat modelliert werden können. Zusatzlich zur stöchiometrischen Matrix des Systems und dem Status der Metabolite (intern oder extern) müssen zu den Reaktionen noch Reversibilitätseigenschaften angegeben werden um die Einschränkungen des Koeffizientenvektors x festzulegen. Das resultierende homogene Gleichungssystem ist fast immer unterbestimmt, da die Anzahl der Reaktionen die Anzahl der Metabolite meist bei weitem übertrifft, und wird mit Verfahren aus der Konvexen Analysis (vgl. Rockafellar (1970) gelöst. Durch die Nichtnegativitätsbedingungen ist der resultierende Lösungsraum ein gerichteter konvexer vielflächiger Kegel (Abbildung 4.1.2), dessen Basisvektoren die sogenannten *Extreme Pathways* darstellen. Ein jeder unter der Steady-Sate Bedingung und der zusätzlichen Irreversibilitätsbedingungen zulässige Zustand des Systems, also jeder zulässige Vektor x, läßt sich dann als Linearkombination dieser Basisvektoren darstellen, wobei auch hier die Irreversibilitätsbedingungen gelten, d.h. die Linearkombinationskoeffizienten müssen ≥ 0 sein wenn nicht jede beteiligte Reaktion des Extreme Pathways reversibel ist. Die Basisvektoren werden für weitere Berechnungen als Spaltenvektoren in einer $r \times e$ Matrix E zusammengefaßt, wobei e die Anzahl der Basisvektoren beschreibt.

Dieses Grundprinzip der Steady-State-Analysen liefert eine neuartige Beschreibung des metabolischen Netzwerks. Der Satz von Extreme Pathways ist alleine ausreichend um das Netzwerk vollständig zu beschreiben und Ausgangsbasis für alle weiteren Untersuchungen. Die Elementarmodenanalyse liefert darüber hinaus noch mittels eines Verfahrens der ganzzahligen linearen Optimierung alle weiteren möglichen stöchiometrisch ganzzahligen

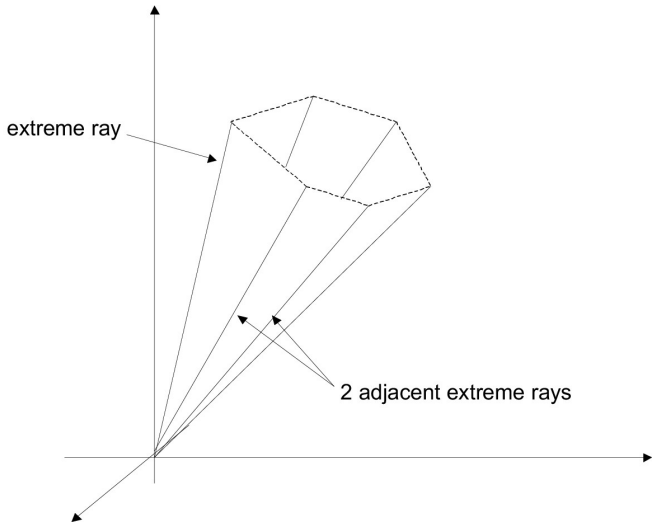

Abbildung 4.2: Der Raum der zulässigen Flüsse hat die Form eines gerichteten konvexen vielflächigen Kegels. Die konvexe Basis oder Extreme Pathways / Rays liegen an den Rändern des Kegels. Ein jeder Zustand des Systems kann durch eine positive Linearkombination der Extreme Pathways erreicht werden.

Stoffwechselwege durch das Netzwerk, eine vollständige Aufzählung der metabolischen Kapazitäten des Netzwerks. Die ursprüngliche Implementierung existiert in Form des Programms METATOOL (Pfeiffer *et al.*, 1999; von Kamp and Schuster, 2006), eines kommandozeilenbasierten Programms für Windows und UNIX, wurde aber in modifizierter Forma auch in andere Softwarepakete aufgenommen, wie GEPASI (Mendes, 1993) or SNA (Urbanczik, 2006).

4.2 Die Entwicklung von YANA und YANAsquare

Trotz der Effizienz des implementierten Algorithmus fehlt METATOOL als kommandolzeilenbasiertem Programm die Benutzerfreundlichkeit und Bedienbarkeit einer graphischen Benutzeroberfläche. Aus diesem Grund und um die einfache Implementierung von Zusatzalgorithmen und Modulen zu ermöglichen, haben wir das Softwarepaket YANA (Schwarz *et al.*, 2005, 2007) und seine Erweiterung YANAsquare (im folgenden einfach YANA genannt) entworfen. YANA bietet eine integrierte Modellierungsumgebung für metabolische Netzwerke mit Ein- und Ausgabe in standardisierten Datenaustauschformaten wie der XML basierten Systems Biology Markup Language (SBML, Hucka *et al.* (2003); Finney and Hucka (2003)). Es integriert Algorithmen zur Analyse metabolischer Netzwerke, zu denen neben den ob-

ligatorischen Steady-State Analysen (EMA und Berechnung der Extreme Pathways) Algorithmen zur Zerlegung großer Netzwerke, die Robusheitsanalyse und die Berechnung von Flussverteilungen aus Pathwayaktivitäten und vice versa gehören. Dazu integriert das Softwarepaket eine Visualisierung metabolischer Netzwerke in Graphenform inklusive entsprechender etablierter Layoutroutinen wie dem Force Directed Layout und ist in der Lage durch eine direkte Abfrage der KEGG Datenbank in kurzer Zeit genomweite metabolische Netzwerke aufzustellen.

4.2.1 Der Steady-State und die kombinatorische Explosion

Um die Steady-State Analysen durchzuführen agiert YANA entweder als Frontend zu METATOOL und überläßt diesem die Berechnung der Elementarmoden oder verwendet seine eigene Implementierung des Algorithmus nach Schuster *et al.* (1999). Dieser berechnet die EMs durch eine schrittweise Erfüllung der Steady-State-Bedingung für jeden Metabolit. Die Originalversion des Algorithmus wurde nach Klamt and Gilles (Klamt and Gilles, 2004) verbessert, indem die EMs während der Berechnung durch Bitmuster an Stelle ihrer Flusskoeffizienten dargestellt werden. Dies ist nur aufgrund der Existenz einer direkten Abbildung der Menge der Reaktionen eines EM zu den Flusskoeffizienten dieser Reaktionen möglich. Die am häufigsten aufgerufene Funktion bei der Berechnung der Elementarmoden, der Test auf Nicht-Zerlegbarkeit, wird damit auf eine einfache Bitoperation zurückgeführt, was das Laufzeitverhalten des Algorithmus drastisch verbessert. Diese Implementierung wird auch von einem Softwarepaket verwendet, welches die Berechnung von chemischen Organisationseinheiten in chemischen Reaktionsnetzen zum Ziel hat (Dittrich and di Fenizio, 2007).

Nach ausgeführter Analyse können die errechneten Pathways in tabellarischer und graphischer Form angezeigt werden (vgl. Abschnitt 4.2.5) inklusive Angaben zu den beteiligten Reaktionen, ihren Reaktionsgleichungen und dem Nettostoffumsatz des Stoffwechselwegs.

Um die kombinatorische Explosion der Elementarmoden zu verhindern, die oft mit großen Netzwerken einhergehen, kommen in YANA zwei gegensätzliche Strategien zur Anwendung. Beide ändern systematisch den intern / extern Status der Metabolite in Abhängigkeit von der Konnektivität[1] selbiger. Die erste Strategie (Schuster *et al.*, 2002b) zerlegt das Netzwerk in Teilnetze indem sie Metabolite mit einer Konnektivität *oberhalb* eines bestimmten Schwellwerts als extern, also gepuffert, setzt. Dies führt dazu, dass zentrale Knoten aufgelöst werden und Elementarmoden, die vorher durch den Knoten liefen ihn jetzt als Start- bzw. Endpunkt verwenden können, was die Anzahl der Moden im System drastisch reduziert.

[1] Die *Konnektivität* eines Metabolits bezeichnet hier ein Maß für den Vernetztheitsgrad des Metabolits, beispielsweise die Anzahl der Reaktionen an denen er beteiligt ist, graphentheoretisch also der Grad des Knotens.

Alternativ bietet YANA die Option Metabolite mit Konnektivitäten *unterhalb* eines bestimmten Schwellwerts als extern zu markieren. Im Gegensatz zur oben genannten Methode werden dadurch in großen Netzwerken die schwach vernetzten Aussenbereiche ausgeblendet, und die Analyse konzentriert sich auf den zentralen hochvernetzten Kern des Systems, was in vielen Fällen ebenfalls zu einer Reduzierung der Anzahl an Elementarmoden führt. Die verbleibenden Stoffwechselwege beinhalten immer noch die wichtigsten EMs, lediglich verkürzt und fokussiert auf die zentralen Knotenpunkte des Netzes (sog. Hub-Metabolite, Schmidt *et al.* (2003)).

4.2.2 Von EM Aktivitäten zu Flußverteilungen...

Wie bereits in Abschnitt 4.1.1 beschrieben liefert die Steady-State Analyse einen Satz von Basisvektoren in Form der $r \times e$ Matrix E zur Beschreibung des gültigen Flussraums des Netzwerks und ein jeder Punkt y dieses Raums kann durch positive Linearkombination $y = Ew$ der Elementarmoden erreicht werden, wobei w die Linearkombinationskoeffizienten der Basisvektoren und damit die Aktivitäten der einzelnen Extreme Pathways sind. Jeder derart berechenbare Punkt erfüllt dabei stets die Steady-State-Bedingung, d.h. $Sy = 0 \; \forall y = Ew, w \geq 0$. Der Vektor von Linearkombinationskoeffizienten w kann in YANA direkt eingestellt werden, indem jedem Pathway ein Wert von 0 bis 1 (0% bis 100%) zugewiesen werden kann. w wird anschließend durch Normierung mit der 2-norm auf den Einheitskreis projiziert, um vergleichbare Ergebnisse zu erhalten und das Zurückschätzen der Pathwayaktivitäten zu erleichtern. Der wesentliche Informationsgehalt des Vektors y, der jeder Reaktion des Systems einen Fluxkoeffizienten zuweist (die sog. Flussverteilung, engl. flux distribution), liegt in der relativen Höhe der Koeffizienten zueinander und nicht in ihrem Absolutwert, so dass diese Normierung keine weiteren Nachteile mit sich bringt. Für ein konkretes w kann die Flussverteilung von YANA dann tabellarisch oder in Diagrammform dargestellt werden (Abbildung 4.2.2).

Die derartige Berechnung von Flussverteilungen aus einem gegebenen theoretischen Satz von Elementarmodenaktivitäten ist wichtig um z.B. die Bedeutung von Enzymen im Netzwerk abschätzen zu können (Gagneur and Klamt, 2004). Dennoch wäre es wünschenswert, wenn man durch experimentelle Daten bestimmte Flussverteilungen zur Rückbestimmung der EM Aktivitäten nutzen könnte, um festzustellen, wie gut sie mit dem Netzwerkmodell übereinstimmen und um gegebenenfalls die Experimente oder die Modellierung zu verbessern (Poolman *et al.*, 2004).

4.2.3 ...und zurück

Um Flussverteilungen lebender Zellen zu erhalten, könnte man entweder direkt die Metabolitflüsse messen oder Flüsse aus Proteinquantifizierungen

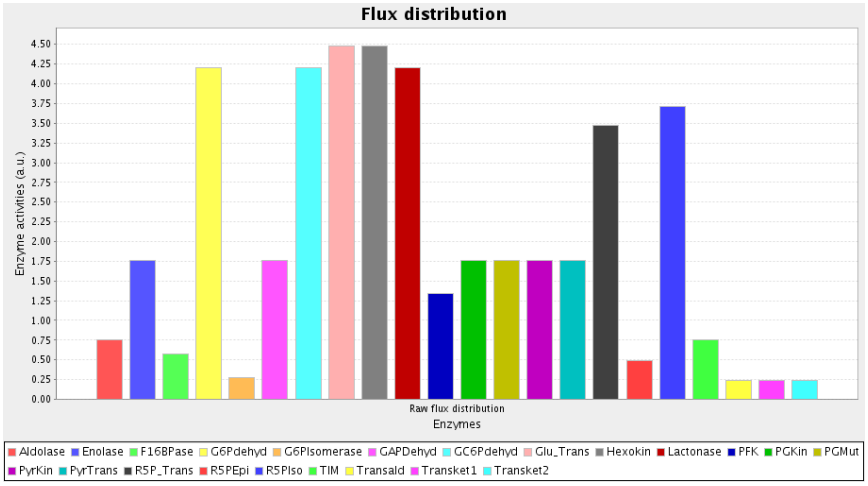

Abbildung 4.3: Flussverteilung eines Systems aus Glykolyse und Pentose Phosphatweg (PPP). Nur ein kleiner Teil der aufgenommenen Glukose (Glukosetransporter *Glu_trans* bei ca 4.5) wird zu Pyruvat umgesetzt (Pyruvattransporter *Pyr_trans* nur bei 1.75, trotz des stöchiometrischen 1:2 Verhältnisses zwischen Glukose und Pyruvat). Ein Großteil geht in den PPP ein und dient der Produktion von Ribose (Ribose-5-Phosphat Transporter *R5P_trans* bei 3.5.

und Angaben zur Enzymkinetik abschätzen. Proteinmengen werden wiederum üblicherweise durch Genexpressionsanalysen oder Proteomuntersuchungen wie Massenspektroskopie bestimmt. Beispielsweise wurde in einer RNA und Proteinexpressionsstudie in Hefe gezeigt, dass im Durchschnitt pro mRNA Kopie ca 4000 Proteinmoleküle synthetisiert werden (Ghaemmaghami *et al.*, 2003) mit individuellen Unterschieden durch mRNA Stabilität, translatorische Regulation und Promotoraktivitäten. Die Bestimmung all dieser unterschiedlichen Faktoren, die die Expressionslevel und schließlich auch die Enzymaktivität selbst beeinflussen können, ist nicht trivial. Erst das komplexe Zusammenspiel aus mRNA Expressionslevel, Proteinexpressionslevel, Niveau der Enzymaktivität und der resultierenden metabolischen Flüsse erlaubt es dem Organismus, sich optimal auf wechselnde Umweltbedingungen einzustellen. Daher sollte in Betracht gezogen werden, dass im Allgemeinen jede dieser Messungen nur bedingt zum Schätzen der Metabolitflüsse dienen kann. Dennoch reicht in der Praxis meist eine grobe Abschätzung der Aktivitätsunterschiede der modellierten Enzyme aus, die in Zusammenhang mit den wichtigsten die Expressionslevel beeinflussenden regulatorischen Signalen (bspw. sollten besonders instabile Proteine oder mRNA Moleküle nicht vernachlässigt werden) zum Schätzen von Flussverteilungen genutzt werden können.

Ist erst eine valide Flussverteilung gefunden bietet YANA verschiedene Algorithmen um die zugehörigen Pathwayaktivitäten zu schätzen. In jedem Fall versucht die Routine den quadratischen Abstand zwischen gefitteter $\hat{z}$ und experimenteller z Flussverteilung so klein wie möglich zu halten. Es gilt also

$$\| Ew - z \|^2$$

zu minimieren. Gesucht wird demnach

$$\underset{w}{\operatorname{argmin}} \langle Ew - z, Ew - z \rangle , \tag{4.2}$$

wobei $\langle ., . \rangle$ das Standardskalarprodukt in einem euklidischen Raum bezeichnet. Andere Metriken wären jedoch ebenso denkbar.

Einfach Umformulierung des Problems ergibt als alternative Form der Zielfunktion

$$\underset{w}{\operatorname{argmin}}(w'E'Ew - 2w'E'z + z'z), \tag{4.3}$$

wobei " ' " die transponierte eines Vektors oder der Matrix bezeichnet. Ohne die angesprochenen Irreversibilitätsbedingungen könnte diese Funktion durch einfachs Differenzieren minimiert werden und würde zu der aus der Regressionsanalyse bekannten Normalengleichung

$$w = (E'E)^{-1}E'z \tag{4.4}$$

führen, in der $()^{-1}$ die Inverse oder Pseudo-/ Moore-Penrose-Inverse (Penrose, 1954; Ben-Israel, 2002) einer Matrix bezeichnet. Die Berechnung von Pathwayaktivitäten mittels einer Pseuodinversen und anschließender Korrektur zur Einhaltung der Nichtnegativitätsbedingungen wurde zuerst von Poolman *et al.* vorgeschlagen (Poolman *et al.*, 2004). Das Verfahren hat den Nachteil, dass die für die Pseudoinverse nötige Singulärwertzerlegung (SVD) recht rechenaufwändig ist und für große Netzwerke zu Laufzeitproblemen führen kann. Eine direkte Berechnung von $(E'E)^{-1}$ ist dagegen nicht möglich, da die Matrix E vollen Spaltenrang besitzen muss, eine Voraussetzung, die Datenmatrizen in Regressionsanalysen üblicherweise erfüllen, die Elementarmodenmatrizen hingegen nicht. Die dazu nötige lineare Unabhängikeit der Spaltenvektoren ist nicht gegeben da es sich nur um Basisvektoren im Sinne den konvexen Analysis handelt, deren lineare Unabhängigkeit nur im Falle positiver Linearkombinationskoeffizienten erhalten bleibt.

Ein Optimierungsproblem wie in Gleichung (4.2) mit linearen Ungleichungsbedingungen ($w \geq 0$ für alle irreversiblen Pathways) ist in der Optimierungstheorie als Quadratisches Programm bekannt (Schwartz and Kanehisa, 2005). Genauer handelt es sich sogar um ein konvexes quadratisches Programm, da die zugrunde liegende räumliche Struktur eine konvexe Menge darstellt und die Zielfunktion eine affine Funktion ist mit einer Hesse-Matrix

= 0. Was diese Art von Optimierungsverfahren attraktiv macht ist die Tatsache, dass jedes lokale Optimum auch ein globales Optimum ist. Letzteres ist zudem eindeutig wenn die Funktion strikt konvex ist. Dies bedeutet in letzter Instanz, dass alle Algorithmen die ein Minimum der Zielfunktion 4.2 finden, immer das gleiche Optimum finden und dieses ist dann auch das globale.

Zur direkten Optimierung dieser Zielfunktion kam daher in YANA als alternativer Optimierungsalgorithmus ein Gradientenabstiegsverfahren mit Backtracking Line Search zum Einsatz, welches sich gegenüber der SVD durch ein deutlich verbessertes Laufzeitverhalten auszeichnete. Als dritte Option wurde noch ein evolutionärer Algorithmus hinzugefügt, der jedoch ein um ein Vielfaches schlechteres Laufzeitverhalten besitzt und daher nur selten zum Einsatz kommt. Seine Vorteile liegen in der einfachen Implementierung auch komplexer Fitnessfunktionen oder der einfachen Veränderung der Zielfunktionen, die nicht auf lineare oder konvexe Probleme beschränkt sind (Schwarz *et al.*, 2005).

4.2.4 YANAsquare und der KGB

Eine erfolgreiche Analyse metabolischer Netzwerke, von der erwartet wird, dass sie biologisch wertvolle Resultate liefert, hängt in erster Linie von der akkuraten Rekonstruktion des Netzwerks ab. Der vollständige Satz von beteiligten Metaboliten und Enzymen muss erfaßt und Systemgrenzen sorgfältig definiert werden. Jedes Enzym muss auf seine An- oder Abwesenheit in dem jeweiligen Organismus hin überprüft werden, durch Auswerten von Genannotationen, Homologiesuchen oder Literaturrecherchen. Diese Aufgabe wird noch zusätzlich durch die fehlende oder uneinheitliche Bezeichnung der Enzyme in den unterschiedlichchen Quellen erschwert. Öffentliche Datenbanken wie Brenda Barthelmes *et al.* (2007), Enzyme Bairoch (2000) oder die KEGG Enzyklopädie Kanehisa *et al.* (2004) haben diesen Vorgang mittlerweile merklich erleichtert, indem sie Algorithmen und Schnittstellen zur Verfügung stellen, um den korrekten Satz an Enzymen und ihren zugehörigen Stoffwechselnetzen zu finden und sie auf Vorkommen in dem zu untersuchenden Organismus zu überprüfen. Desweiteren versuchen sie die verschiedensten Nomenklaturen der Enzyme und ihrer Funktionen nach und nach zu vereinheitlichen oder zumindest eindeutige Abbildungen zwischen diesen herzustellen. Dennoch musste der eigentliche Modellierungsschritt in der Software bisher von Hand vorgenommen werden.

Um das anfängliche Aufsetzen des Netzwerks zu beschleunigen und dem Benutzer ein Werkzeug zum schnellen Erstellen eines ersten metabolischen Netzes an die Hand zu geben, haben wir das Kegg Browser (KGB) Modul für YANA implementiert. Es ist in der Lage sich direkt mit der KEGG Datenbank zu verbinden und somit online die Stoffwechselkarten, Reaktionslisten und Metabolite zu durchsuchen und die gewünschten Reaktionen nach YA-

NA zu importieren. Dabei können Organismen oder der Referenzpathway ausgewählt werden, um nur die Reaktionen zu übertragen, die im ausgewählten Organismus als vorhanden annotiert sind. Alle zugehörigen Metabolite werden automatisch gesammelt und mitübertragen; dabei können besonders hoch vernetzte Metabolite wie H^-, CO^2, H_2O, ADP oder ATP herausgefiltert werden, um die Netzwerkkomplexität für die EMA zu reduzieren.

Ein Problem bei der Behandlung der KEGG Datenbankeinträge ist die Tatsache, dass Metabolit- und Reaktionsbezeichnungen oft Sonderzeichen enthalten. Aus Gründen der Abwärtskompatibilität zu den METATOOL Eingabedateien und zur Erfüllung der Vorgaben des SBML2 Standards müssen daher die Enzym- und Metabolitnamen standardisiert werden. Wir haben eine automatische Abkürzungsroutine integriert, welche auf eine standardisierte Weise die Enzym- und Metabolitnamen in eine Kurzdarstellung bringt, die mit den METATOOL und SBML2 Standards konform ist. Die direkte Abfrage von KEGG über eine Internetverbindung mit geringer Bandbreite ist zeitaufwändig, weshalb wir einen semilokalen Zugriff mittels eines lokalen Zwischenspeichers implementiert haben, der die Reaktionsdefinitionen enthält. Allerdings muss dieser Zwischenspeicher regelmässig über eine integrierte Aktualisierungsroutine auf den neuesten Stand gebracht werden.

Zur Anbindung der KEGG Datenbank wurde version 1.3 der Apache Axis Bibliothek als SOAP Implementierung verwendet zusammen mit der Web Services Description Language (WSDL) auf dem HTTP Protokoll. Die Benutzeroberfläche wurde mittels der SWING Bibliothek erstellt (siehe Abbildung 4.2.4). Die Pathways und Reaktionen werden vom Benutzer spezifiziert und in einer tabellarischen Darstellung gesammelt. Anschließend erlaubt ein intelligenter Editor das schnelle Editieren und Anpassen der Daten an die eigenen Bedürfnisse bevor sie in YANA importiert werden. Die Software kann relativ mühelos an eine neue Datenbank angepasst werden, so beispielsweise eine hauseigene Metabolit und Reaktionsdatenbank im KEGG Format.

4.2.5 Visualisierung von metabolischen Netzwerken

Mit zunehmender Netzwerkkomplexität nimmt auch die Anzahl der resultierenden Elementarmoden rapide zu. Trotz verschiedener Verfahren zur Einschränkung der kombinatorischen Explosion der EMs (vgl. Abschnitt 4.2.1) ist die Interpretation großer Netzwerke dennoch extrem schwierig, und mehrere tausend Elementarmoden sind keine Seltenheit. Typische Fragen hierbei sind, wie die Flussverteilungen eines bestimmten Satzes von Enzymen sich im Netzwerk darstellt, welche Teilbereiche des Netzes von bestimmten Extreme Pathways betroffen sind oder wo ganz allgemein der grö"ste Massefluss hinzeigt. Um diese Fragen einfach beantworten zu können haben wir eine Visualisierungsroutine in YANA implementiert die in der Lage ist das metabolische Netzwerk in Form eines bipartiten Graphen darzustellen,

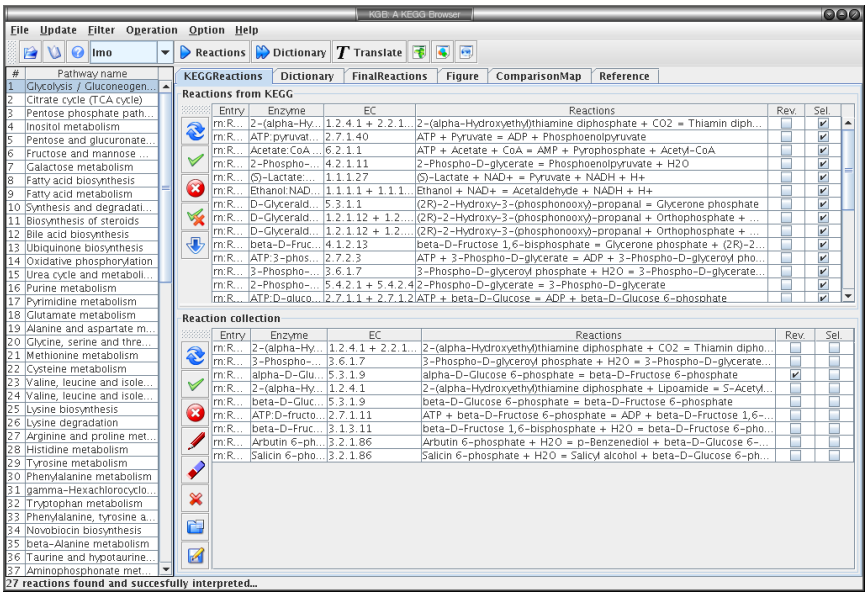

Abbildung 4.4: Screenshot des KeggBrowser Moduls (KGB) in YANAsquare. Über eine komfortable Benutzeroberfläche können die Einträge der KEGG Datenbank online gelesen und importiert werden.

in dem Metabolite und Enzyme die beiden Knotenklassen darstellen. Mehrere felderprobte Layoutalgorithmen (z.B. Spring-embedded, Radial Tree oder Sugiyama) wurden hinzugefügt um eine automatische übersichtliche Darstellung des Netzwerks zu ermöglichen. Während schon das Erstellen des Netzwerkes durch diese Erweiterung sehr vereinfacht wird, ist sie insbesondere nach erfolgter Steady-State-Analyse interessant, um die aktuelle Flussverteilung in Form von unterschiedlich skalierten Pfeilen in den Graphen einzuzeichnen und es so dem Anwender zu ermöglichen unmittelbar die Funktion des Netzwerks graphisch zu erfassen (vgl. Abbildung 4.2.5). Auch einzelne Elementarmoden können auf diese Art und Weise graphisch dargestellt werden.

4.2.6 Ein Robustheitsmaß

Experimente in der Pharmakologie, der Infektionsbiologie oder der Genetik haben oft die Untersuchung der Robustheit von Netzwerken gegen Enzymdeletionen zum Ziel. Durch spezifische Knock-outs werden bestimmte metabolische Prozesse aus dem System entfernt und die anschließenden Fähigkeiten des Netzwerks die ursprünglichen Stoffe zu synthetisieren oder abzubauen untersucht. Um YANA's Netzwerkanalysefähigkeiten noch zu erweitern haben wir in Erweiterung früherer Arbeiten (Wilhelm *et al.*, 2004) einen Al-

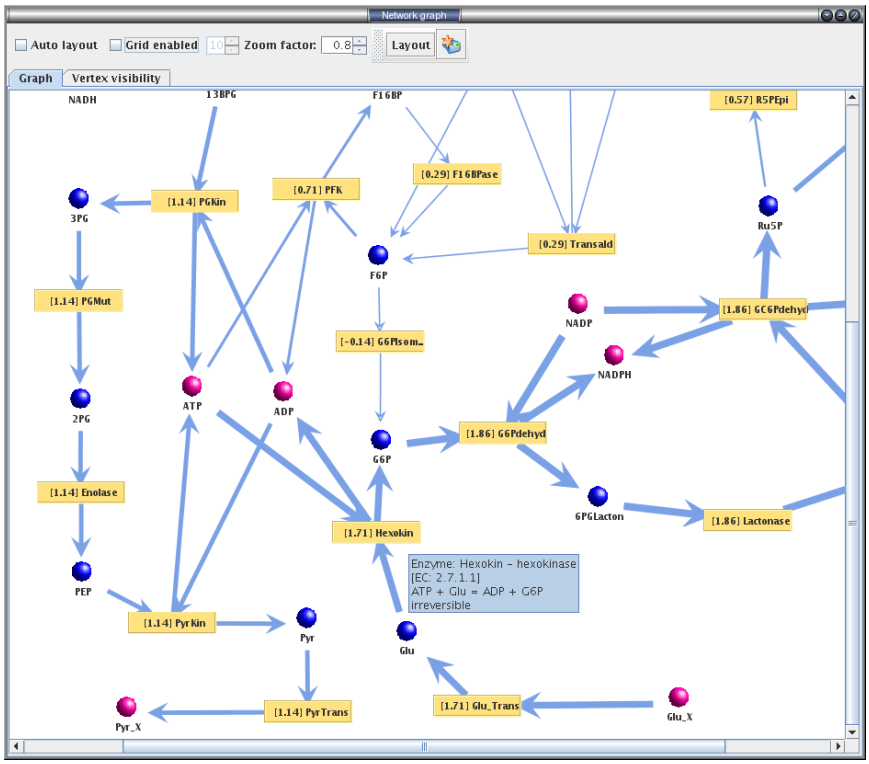

Abbildung 4.5: Visualisierung eines metabolischen Netzwerks mittels eines bipartiten Graphen durch YANAsquare. Die Kugeln stellen die Metabolite da (violett für externe, blau für interne Metabolite), die Enzyme sind als gelbe Rechtecke angezeigt.

gorithmus zur Robustheitsanalyse implementiert. Für eine stöchiometrische $m \times n$ Matrix S mit m Metaboliten und n Enzymen, eine Elementarmodenmatrix E $(n \times e)$ aus e konvexen Basisvektoren und einen Vektor v von Linearkombinationskoeffizienten oder Pathwayaktivitäten läßt sich der Vektor

$$p = SEv$$

des Metabolitumsatzes errechnen. Der Algorithmus iteriert dann über die Enzyme, deaktiviert alle Stoffwechselwege, welche das Enzym verwenden (setzt den entsprechenden Koeffizienten von v auf 0) und simuliert auf diese Weise einen Gen-Knockout. Das Ergebnis sind n unterschiedliche Vektoren mit EM Aktivitäten v_i, $i \in 1..n$. Nach jedem Schritt berechnet das System die Anzahl der noch produzierbaren Metabolite und faßt sie in einem Vektor

der durchschnittlichen Gesamtproduktion

$$p^{(avg)} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}(SEv_i)$$

zusammen, wobei $\mathcal{H}(x)$ die Heaviside Funktion

$$\mathcal{H}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

beschreibt und hier komponentenweise definiert ist. Auf die gleiche Art definieren wir den Vektor der gesamten Metabolitproduktion

$$p^{(tot)} = \mathcal{H}(SE\mathbf{1})$$

des unlimitierten Netzwerks ($\mathbf{1}$ ist der Einheitsvektor), welches Auskunft gibt über die Möglichkeiten des Netzwerks falls kein Enzym inhibiert ist. Der Vektor $p^{(avg)}$ liefert die durchschnittliche Produktion nach Ausfall eines Enzyms. Durch Aufsummieren der Vektorkomponenten erhalten wir die (skalaren) Gesamt- und Durchschnittssummen der Produktion $s^{(tot)}$ und $s^{(avg)}$ sowie den Verlust l

$$\begin{aligned} s^{(avg)} &= \sum_{j=1}^{m} p_j^{(avg)} \\ s^{(tot)} &= \sum_{j=1}^{m} p_j^{(tot)} \\ l &= s^{(tot)} - s^{(avg)}. \end{aligned}$$

Das Verhältnis

$$r = \frac{s^{(avg)}}{s^{(tot)}}$$

ausgedrückt in Prozent wird schließlich als Gesamtrobustheitsmaß des Netzwerks gegen einzelne Enzymdeletionen verwendet. Zusätzlich berechnen wir aus der durchschnittlichen EM Aktivität

$$v^{(avg)} = \frac{1}{n} \sum_{i=1}^{n} v_i$$

die durchschnittliche Anzahl noch aktiver EMs nach einer einzelnen Enzymdeletion

$$t = \sum_{i=1}^{e} v_i^{(avg)}.$$

Der durschnittliche EM Verlust m ist dann durch

$$m = e - t$$

gegeben und wird ebenfalls als ein Ergebnis der Robustheitsanalyse zurückgegeben.

4.3 Diskussion

In den vergangenen Jahren wurde eine Reihe von Werkzeugen zur Netzwerkanalyse vorgestellt. Die bekanntesten davon beinhalten *METATOOL* von Kamp and Schuster (2006), *FluxAnalyzer* Klamt and Stelling (2003), *Jarnac*, ein Modul für die Systems Biology Workbench Sauro *et al.* (2003), *Gepasi* Mendes (1997), *ScrumPy* Poolman *et al.* (2004) und *COPASI* Hoops *et al.* (2006). Dazu kommt noch das kürzlich von Urbanczik eingeführte *SNA*, ein Analysepaket für die Mathematica-Umgebung Urbanczik (2006). All diese Programme beinhalten effiziente Implementierungen eines Steady-State-Algorithmus oder binden sich direkt an *METATOOL*. Einige, wie der *FluxAnalyzer*, stellen sogar eine einfache graphische Darstellung des Netzwerks zur Verfügung, brauchen jedoch oft, wie *SNA*, eine gültige Matlablizenz.

YANA und sein Nachfolger YANAsquare bieten eine integrierte Modellierungsumgebung für metabolische Netzwerke die neben der obligatorischen Steady-State-Analyse eine durchdachte Visualisierung des Netzwerks anbietet. Hinzu kommen zahlreiche weitere Analysemodule die einige der wichtigsten Fragen der Analyse großer Netzwerke angehen: Zwei Strategien zur Reduzierung der kombinatorischen Explosion der Elementarmoden des Systems (Schwarz *et al.*, 2005; Schuster *et al.*, 2002a), die Möglichkeit zu Berechnung der Flussverteilung und ihrer graphischen Darstellung im Netzwerk, die Rückrechnung von Pathwayaktivitäten aus besagten Flussverteilungen (vgl. auch Poolman *et al.* (2004); Stelling *et al.* (2002)), sowie eine Routine zur Bestimmung der Robustheit des Netzwerks gegenüber Enzymdeletionen. Zusammen mit dem KGB Modul für das schnelle Erstellen von Netzwerkmodellen aus der bekannten KEGG Datenbank leistet das Programm einen wichtigen Beitrag zur Analyse metabolischer Netzwerke, es ist darüberhinaus plattformunabhängig und wird als Open-Source Software vertrieben.

Trotz seiner einfachen Benutzbarkeit muss natürlich betont werden, dass der Import von Stoffwechselwegen aus der KEGG Datenbank mit angemessener Vorsicht geschehen muss. Obwohl die Datenbank ein wertvolles Werkzeug darstellt und die Erstellung von Netzwerken drastisch vereinfacht müssen die Modelle im nachhinein überprüft werden auf Korrektheit und Konsistenz. Die Daten aus KEGG sind schon aufgrund der Größe der Datenbank nicht immer verläßlich und können Netzwerk- oder stöchiometrische Inkonsistenzen enthalten oder falsch annotiert sein. Zusätzliche Literaturre-

cherche ist zwingend erforderlich um beispielsweise zusätzliche Informationen über Pathwayvarianten in bestimmten Organismen zu erhalten. Aber das KGB Modul war auch nicht als Ersatz für ein sorgfältig von Hand aufgestelltes metabolisches Netzwerk gedacht, sondern soll dem Anwender die Möglichkeit geben, sich schnell eine vernünftige Datenbasis zusammenzustellen, von der aus das Netzwerk dann verfeinert und akkurat aufgestellt werden kann. Neben den angesprochenen Punkten gehört dazu auch das sorgfältige Definieren des Intern- / Externstatus der Metabolite und das Hinzufügen von Transportprozessen (z.B. aus der TransportDB, Ren *et al.* (2004, 2007)) die noch nicht in KEGG enthalten sind.

YANA ist des Weiteren auch in der Lage verschiedene Zellkompartimente durch Unternetzwerke zu definieren, indem Transportreaktionen zwischen den Kompartimenten eingeführt werden (z.B. Mitchondrium $\leftrightarrow$ Zytoplasma) und ansosten darauf geachtet wird, dass Reaktionen nur innerhalb eines Kompartiments ablaufen.

Die implementierte Robustheitsroutine erlaubt einen schnellen Einblick in den Effekt von Enzymdeletionen auf das Netzwerk. Dabei ist die Frage der statistischen Signifikanz unterschiedlicher Robustheitsscores nur schwer zu beantworten. Die aus der Steady-State-Analyse resultierende Matrix der Basisvektoren ist deterministisch, so dass stochastische Variation nur in Form der stöchiometrischen Matrix vorliegt, ein Raum auf dem es nur sehr schwer möglich ist, ein vernünftiges Nullmodell zu definieren. Wir glauben allerdings, dass die sich aus einer biologisch sinnvollen Definition der stöchiometrischen Matrix ergebenen Robustheit ebenfalls eine biologische Signifikanz hat, die gegebenenfalls jedoch experimentell überprüft werden muss.

4.4 Zusammenfassung

Für jede spezifische Frage aus dem Bereich der metabolischen Netzwerkmodellierung sind unterschiedlichste Lösungen gedacht und implementiert worden (Mendes, 1997; Sauro *et al.*, 2003; Urbanczik, 2006; Klamt and Stelling, 2003; von Kamp and Schuster, 2006; Poolman *et al.*, 2004). YANA besticht dabei insbesondere durch die Möglichkeit auch große Netzwerke analysieren zu können. Als modulare Open-Source Anwendung ist es für alle akademischen Benutzer frei verfügbar und kann mit relativ geringem Aufwand erweitert und den eigenen Wünschen angepasst werden.

Wir haben YANA u.a. verwendet um das komplexe Phospholipidnetzwerk im murinen Phagosom zu untersuchen (Schwarz *et al.*, 2007). Obwohl das verwendete Netzwerk eine deutliche Abstraktion des wirklich Signalnetzes der Zelle ist, verhielten sich die meisten Lipide experimentell so wie vorhergesagt, sie inhibierten oder aktivierten die Nukleierung von Aktin an der phagosomalen Membran. Desweiteren haben wir mit der Software die genomweiten metabolischen Netzwerke von 5 *Staphylococci* Spezies vergli-

chen (Schwarz *et al.*, 2007) und damit gezeigt, dass YANA nicht nur in der Lage ist große Netzwerke effizient aufzustellen und zu modellieren sondern auch Netzwerkvergleiche dieser Größe durchzuführen. Zukünftige Versionen beinhalten eventuell zusätzliche Möglichkeiten der Analyse von Zellkompartimenten, erweiterte Layoutalgorithmen für die Graphendarstellung, sowie die Möglichkeit zum direkten Editieren des Netzwerks durch die Graphendarstellung.

Kapitel 5

Das Transkriptom als Regulator des Metabolismus bei *L. monocytogenes*

5.1 Regulation des Metabolismus

Obwohl viele Studien gezeigt haben, dass Steady-State-Analysen metabolischer Netzwerke im Allgemeinen und die EMA im Speziellen hervorragend geeignet sind, um die Topologie von Netzwerken aufzudecken und mögliche Pfade durch das Netz offenzulegen, wurden sie von jeher zu Recht dafür kritisiert, keinerlei Aussage über den tatsächlichen regulatorischen Zustand des Stoffwechsels einer lebenden Zelle machen zu können. Sie zeigen zwar auf, welche Möglichkeiten zur Produktion das Netzwerk hat, welche davon aber tatsächlich aktiv sind und welche regulatorisch abgeschaltet werden, bleibt offen. Um dieses Manko auszugleichen wurden Versuche unternommen experimentell erhobene biologische Daten auf das System aus Elementarmoden oder Extreme Pathways abzubilden, um so einen Eindruck vom regulatorischen Zustand des Netzes zu erhalten oder zumindest die vorhergesagte Topologie experimentell verifizieren zu können. In den meisten Fällen wurden dabei experimentell erhaltene Flussverteilungen oder Metabolitproduktionsraten mit den Ergebnissen aus FBAs verglichen, und diese — oftmals dünnbesetzten — Datensätze konnten weitestgehend erfolgreich *in-silico* vorhergesagt werden (Edwards *et al.*, 2001). Desweiteren wurde gezeigt, dass aus einer EMA geschätzte Flussverteilungen unter bestimmten Bedingungen sehr gute Prediktoren für zelluläre Regulation sein können, zumindest im Hinblick auf die Menge an transkribierter mRNA (Stelling *et al.*, 2002). Auch innerhalb des Netzes wurde gezeigt, dass die vorhergesagten strukturellen Eigenschaften und bevorzugten Stoffwechselwege aus einer EMA in guter Übereinstimmung mit den durch ^{1}3C Isotopologstudien bestimmten Metabolitflüssen sind (Eisenreich *et al.*, 2006). Dennoch wa-

ren die bisherigen Korrelationen und Vorhersagen von und zu Steady-State-Modellen metabolischer Netzwerke meist eindimensional ohne akkurate statistische Überprüfung oder Crossvalidierung durch weitere experimentelle Datensätze.

Mit Genexpressionsmessungen als Indikator für den regulatorischen Zustand der Zelle und präzisen Messungen der Metabolitproduktion soll gezeigt werden, dass die konvexe Basis des Steady-State-Cones ein gültiges Modell für das metabolische Potential eines Netzwerks ist. Die aus den gemessenen Metabolitkonzentrationen vorhergesagten Pathwayaktivitäten implizieren eine Flussverteilung, die mit unter gleichen experimentellen Bedingungen erhobenen Genexpressionsdaten korreliert wird. Wir reformulieren dazu den auf Nebenbedingungen basierenden Ansatz der Flux Balance Analyse (FBA), indem unmittelbar innerhalb des Steady-State-Cones und damit im Bereich zulässiger Flussverteilungen gesucht wird. Unser Ansatz ist in der Lage, sowohl Metabolitproduktions- oder Metabolitflussmessungen als auch Reaktionsflüsse als Zielfunktion zu verwenden und zieht dabei die Reversibilität der Reaktionen sowie etwaiger weiterer Einschränkungen in Betracht. Die Methode, alle Schritte und Resultate, wird am zentralen Kohlenstoff- und Aminosäurestoffwechsel von *Listeria monocytogenes* (Glaser *et al.*, 2001) demonstriert und erläutert.

5.2 Das Zusammenspiel von Stoffwechsel und Genexpression

Aufgrund der Verfügbarkeit von zwei unterschiedlichen Datensätzen besteht unser Ansatz einer Korrelationsstudie aus vier Teilbereichen: i) Aufstellen des Netzwerks, ii) Extraktion der Struktureigenschaften, iii) Trainieren des Netzes um die gemessene Metabolitproduktion wiedergeben zu können und iv) Vorhersage von Enzymaktivitäten und Korrelation selbiger mit den Genexpressionsdaten.

5.2.1 Aufstellen des Netzwerks

Das Netzwerkmodell, welches in der Analyse verwendet wurde, basierte aus dem von uns etablierten Modell welches schon bei einer früheren Untersuchung an *Listerien* zum Einsatz kam (Eisenreich *et al.*, 2006). Es wurde ursprünglich mittels Daten der KEGG Datenbank erstellt (Kanehisa *et al.*, 2004), welche anschließend von Hand nach nachannotiert wurden. Zu den enthaltenen Pathways gehören wesentliche Bestandteile des zentralen Kohlenstoffstoffwechsels sowie der Aminosäuresynthsewege, namentlich die Glykolyse, der Pentose Phosphatweg, der hufeisenförmige TCA inklusive dem listerienspezifischen Carboxylase Bypass (Glaser *et al.*, 2001) sowie der Aminosäuresynthesewege von und zu allen in der Analyse verwendeten Ami-

noräuren (Ala, Arg, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Lys, Phe, Pro, Ser, Thr, Trp, Tyr, Val). Das finale System besteht aus 127 Metaboliten (von denen 91 interne Metabolite sind) und 118 Reaktionen.

Zusätzlich zur Stöchiometrie ist der Intern- / Extern-Status der beteiligten Metabolite eine weitere bedeutende und oft vernachlässigte Netzwerkeigenschaft (siehe auch Abschnitt 4.1.1 und folgende oder Papin *et al.* (2004)). Während man bei der Frage nach der Reversibiliät der Reaktionen in den meisten Fällen auf der sicheren Seite ist, wenn man eine Reaktion als reversibel annimmt, ist die Beantwortung der Frage nach dem Metabolitstatus deutlich schwieriger, da die Wahl die Anzahl der Elementarmoden und die Topologie des Netzwerks drastisch beeinflußt. Während Algorithmen vorgeschlagen wurden, die diese Eigenschaft willentlich ausnutzen, um die kombinatorische Explosion der Elementarmoden zu verhindern (Dandekar *et al.*, 2003; Schuster *et al.*, 2002a; Schwarz *et al.*, 2005), ist für überschaubare Netzwerke eher eine biologisch motivierte Entscheidung sinnvoll.

Die Theorie besagt, dass externe Metabolite solche sind die an den Rändern des Systems sitzen, also die Metabolite, die von der Zelle aufgenommen bzw. abgegeben werden; sie sind die Anfangs- und Endpunkte der Extreme Pathways bzw. Elementarmoden. Für die internen Metabolite hingegen gilt die Steady-State-Bedingung, sie werden unmittelbar weiterverarbeitet. Diese Grenzen werden jedoch zunehmend unscharf, wenn man nur einen Teilbereich des Netzwerks modelliert und nur einen Ausschnitt aller Reaktionen der Zelle in Betracht zieht. Die Frage die sich dann stellt ist, wie man mit Metaboliten verfahren soll, die zwar hoch vernetzt und zentral im modellierten Teil des Netzwerks stehen und damit in der Mehrheit aller EMs eine Rolle spielen werden, die aber ebenso auch in zahlreichen anderen Teilen des wahren Netzwerks in großem Maße produziert und verbraucht werden. Hier davon auszugehen, dass sie intern, also ungepuffert sind, ist sicherlich zu restriktiv angesichts der vielen anderen möglichen Zuflüsse. Solche Metabolite auf extern zu setzen ist ebenfalls ungünstig, da sie damit effektiv aus dem Netz entfernt werden, also nicht nur an anderer Stelle verbraucht sondern auch produziert werden dürfen — vollständig gepuffert eben. Pyruvat beispielsweise ist einer der zellulären Kohlenstoffträger mit den höchsten Konnektivitäten überhaupt (Schmidt *et al.*, 2003; Wagner and Fell, 2001). Dass jedes verbrauchte Molekül Pyruvat an anderer Stelle produziert werden muss macht für ein Modell, welches den zentralen Kohlenstoffwechsel betrachtet, sicherlich Sinn. Dass jedes produzierte Molekül an anderer Stelle auch sofort weiterverarbeitet werden muss hingegen nicht, dies könnte auch problemlos außerhalb des modellierten Teil des Netzwerks geschehen. Diese Überlegungen werden umso wichtiger je größer der Unterschied zwischen der Konnektivität des Metabolits im Netzwerk und der Konnektivität des Metabolits im Gesamtnetz der Zelle ist. Im Fall des *Listerien* Netzwerks haben wir die Metabolitkonnektivitäten des Netzwerks mit denen des KEGG Referenzpathways verglichen um eine Abschätzung für diese Diskrepanz zu

erhalten (siehe Tabelle 5.1).

Name	*Original*	*Modell*	*KEGG*	*Beschreibung*
NAD	int	11	640	nicotinamide dinucleotide
NADH	int	11	640	reduced NAD
NADP	int	12	640	NAD phosphate
NADPH	int	12	640	reduced NADP
ATP	ext	20	463	adenosine triphosphate
CoA	int	6	357	coenzyme A
ADP	ext	16	332	adenosine diphosphate
NH3	int	9	288	ammonia
Pyr	int	8	152	Pyruvate
A_Ketoglutarat	int	8	146	α-ketoglutarate
AcetylCoA	int	6	133	acetyl coA
Glu	int	18	123	glutamate

Tabelle 5.1: Metabolitkonnektivitäten hochvernetzter Metabolite im Netzwerkmodell von *Listeria* („Modell“), die zugehörigen KEGG Referenzpathwaykonnektivitäten („KEGG“) und ihr Zustand im Basisnetzwerk vor Modifikation („Original“).

Wir möchten hier zwei Möglichkeiten vorschlagen um mit dieser Diskrepanz umzugehen. Eine wäre es, all jene Metabolite auf extern zu setzen, deren Unterschied zwischen modellierter und tatsächlicher Vernetztheit jenseits eines bestimmten Schwellwerts sind. Aber das würde an entsprechender Stelle auch einen Substratfluss in das Netz implizieren, was aufgrund von dadurch verschwindender struktureller Abhängikeiten zwischen Teilen des Netzwerks nicht immer wünschenswert ist. Im Falle des *Listerien*netzwerks wurde ATP aus diesem Grund stets als extern definiert, u.a. auch da der Fokus der Analyse auf der Modellierung des Kohlenstoffwechsels lag, nicht auf der Energiebereitstellung. Ein zweiter Ansatz um mit hoch vernetzten Metaboliten umzugehen, ist das Einführen zusätzlicher Abfluss- oder Zuflussreaktionen, eine Art künstlicher Kompartimentierung die den modellierten Teil des metabolischen Netzes vom Rest abtrennt. Welche der beiden Verfahren zur Anwendung kommt hängt stark von den Annahmen und dem Zweck des Modells ab. In unserem Falle haben wir das Orignalmodell derart geändert, dass α-ketoglutarate und NH3 extern wurden und haben eine Abflussreaktion für Glutamat eingeführt. Pyruvat kann über die Pyruvatcarboxylase durch den TCA in externes Succinat umgewandelt werden, so dass hier keine Abflussreaktion erforderlich war. ADP und ATP wurden extern gehalten und die NAD Gruppe wurde als eine der hauptsächlichen Verbindungen zwischen Glykolyse, PPP und der Aminosäuresynthese komplett intern gehalten. Zusätzliche Abflussreaktionen im Falle der NAD Gruppe brachten

keine Änderungen der konvexen Basis, was zeigt, dass die Regulation der NADH und NADPH Produktion durch spezifische Pathwayaktivierung allein präzise genug ist und keine überschüssigen Reduktionsäquivalente verbleiben. Das gleiche scheint für CoA / AcetylCoA der Fall zu sein, welche auch für den Rest der Analyse als intern geführt wurden.

5.2.2 Steady-State-Analyse

Nach dem Aufstellen des Netzwerks erfolgte die Berechnung der konvexen Basis des Systems mit Hilfe des Softwarepakets YANA (siehe Abschnitt 4.1.1 und folgende, Schwarz *et al.* (2005, 2007)). Das *Listerien* System bestand aus 75 konvexen Basisvektoren bei 6587 EMs. Die Multiplikation der stöchiometrischen Matrix S mit der Matrix der konvexen Basisvektoren E liefert den Metabolitumsatz pro EM dargestellt in der Matrix M. Dabei gilt zu beachten, dass die Einträge dieser Matrix Null sind für alle internen Metabolite, da die Extreme Pathways natürlich die Steady-State-Bedingung erfüllen.

5.2.3 Trainieren des Modells

Ausgehend von dem in Abschnitt 4.2.3 beschriebenen Verfahren zum Rückschätzen von Pathwayaktiviäten aus experimentell gemessenen Flußverteilungen wurde der Ansatz auf das Lernen aus Metabolitmessungen übertragen. Anstatt Gleichung 4.2 zu minimieren wurde dann versucht den Metabolitumsatz M der Extreme Pathways über

$$\underset{w}{\operatorname{argmin}}\langle Mw - z, Mw - z\rangle = \underset{w}{\operatorname{argmin}}\langle SEw - z, SEw - z\rangle, \tag{5.1}$$

an die Metabolitproduktionsdaten aus der ^{1}3C Isotopologmessung anzupassen. Dies sollte im Vergleich zu Genexpressionsdaten präzisere Resultate liefern, da erstere doch allgemein als verrauscht und fehlerbehaftet gelten und da zelluläre Regulation nicht durch Änderungen der Genexpression allein erreicht wird.

Im Falle des *Listerien*netzwerks verwendeten wir Messungen der Konzentration von 14 Aminosäuren im Medium um den optimalen Vektor w zu finden. Dabei wurden Leu, Iso und Val aus dem Trainingsprozess entfernt da sie nicht oder nur in geringem Maße überhaupt *de-novo* synthetisiert worden waren. Da sie jedoch ursprünglich auch im Medium vorlagen hätte ein Wert von Null auch verhindert, dass sie hätten aufgenommen werden können. Zur Überprüfung des Trainingsprozesses kam ein Regressionsmodell zum Einsatz welches mit einem R^2 nahe eins bestätigte, dass das Modell in der Lage war die experimentell bestimmten Metabolitkonzentrationen zu lernen, eine Grundvorausetzung an das Modell, um später prädiktiv sein zu können (Abbildung 5.1).

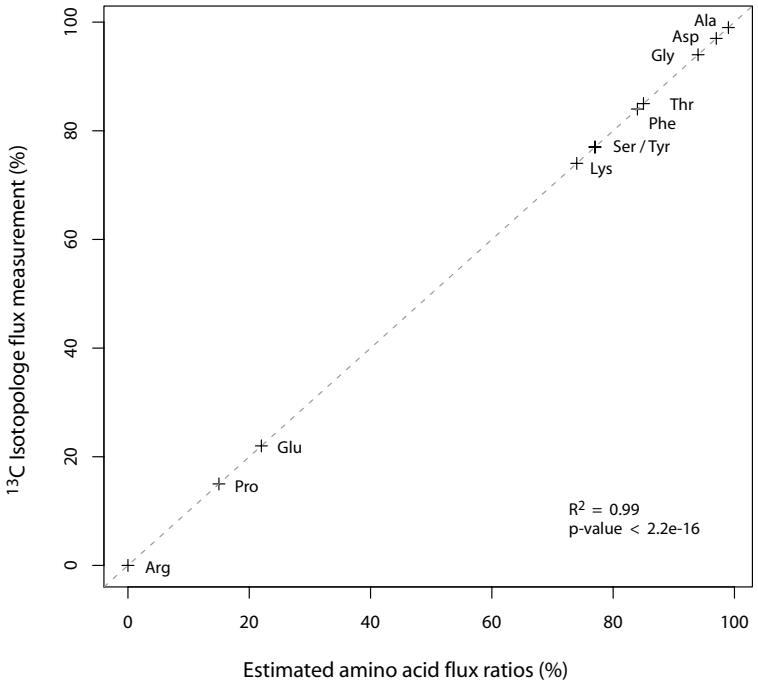

Abbildung 5.1: Scatter plot of estimated metabolite flux against flux measurements from ^{13}C isotopolog studies. Fitting of a regression model resulted in an adjusted R^2 of 0.99 (p-value $< 2.2e-16$).

5.2.4 Vorhersage und Vergleich mit der Genexpression

Falls auf lange Sicht die zelluläre Regulation tatsächlich in erster Linie durch die Genexpression erfolgt, sollten Flussverteilungen aus einer EMA unter stabilen Bedingungen mit entsprechenden Genexpressionsdaten korrelieren. Zur Bestimmung der Flussverteilung wurde dafür der Betrag der Elementarmodenmatrix $|E|w$ verwendet. Die resultierende Flussverteilung wurde mit den 90 zugeordneten Genen des Netzwerks verglichen (Abbildung 5.1). Für 18 Gene des zentralen Kohlenstoffwechsels (Glykolyse und PPP) war die Vorhersage akkurat. Der Grad der Korrelation wurde durch ein Regressionsmodell statistisch validiert. Der korrigierte R^2 Wert lag bei 0.8541 (multiples R^2 von 0.8623) mit einem P-Wert von $9.768e-09$. Das Einhalten der Modellannahmen wurde durch diagnostische Plots verifiziert und mit zufallsgenerierten Flussverteilungen verglichen um stochastische Effekte auszuschliessen.

Für die restlichen Gene war die geschätzte Flussverteilung zwar kein

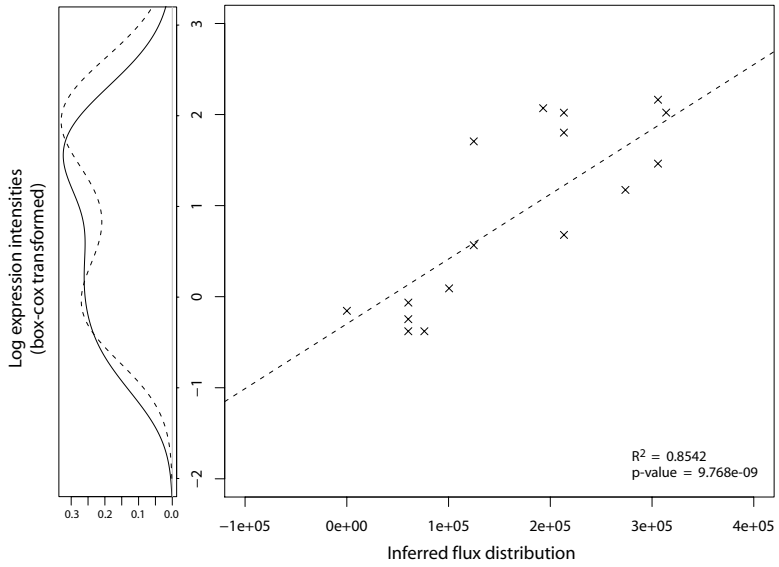

Abbildung 5.2: Scatterplot der vorhergesagten Genexpression gegen experimentell bestimmte (log-transformiert) für 18 Gene aus Glykolyse und PPP. Die geschätzte Flussverteilung gibt die Messungen adäquat wieder ($R^2 = 0.8542$, P-Wert < 0.05). Dichteschätzungen (links) sowohl des Gesamtdatensatzes (durchgezogen) als auch des hier gezeigten (getrichelt) zeigen, dass die Korrelation nicht auf einen Ausschnitt der Gesamtdaten beschränkt ist.

allgemein gültiger Prediktor, markierte aber eine untere Schranke der Genexpression (blaue Markierung, Abbildung 5.2.4). Desweiteren sollte bemerkt werden, dass die festgestellte Verbindung zwischen Metabolitproduktion und Genexpressionsmustern reversibel ist. Ein Versuch bei dem die Genexpressionsdaten als Trainingsdatensatz verwendet wurden und damit die Metabolitproduktion vorhergesagt wurde verlief erfolgreich. Die Metabolitmessungen konnten bei einem R^2 Wert von 0.81 signifikant reproduziert werden.

5.3 Diskussion

Es wurden bisher einige Methoden vorgeschlagen, um biologische Datensätze auf metabolische Netzwerke abzubilden, mit unterschiedlichem Erfolg. In diesem Zusammenhang sind im wesentlichen drei verschiedene Herangehensweisen zu beobachten. i) die zeitabhängige dynamische Beschreibung von kurzen Enzymkaskaden mittels Differentialgleichungen und entsprechenden Vergleichen mit echten metabolischen Flussmessungen (Ferreira *et al.*, 2003;

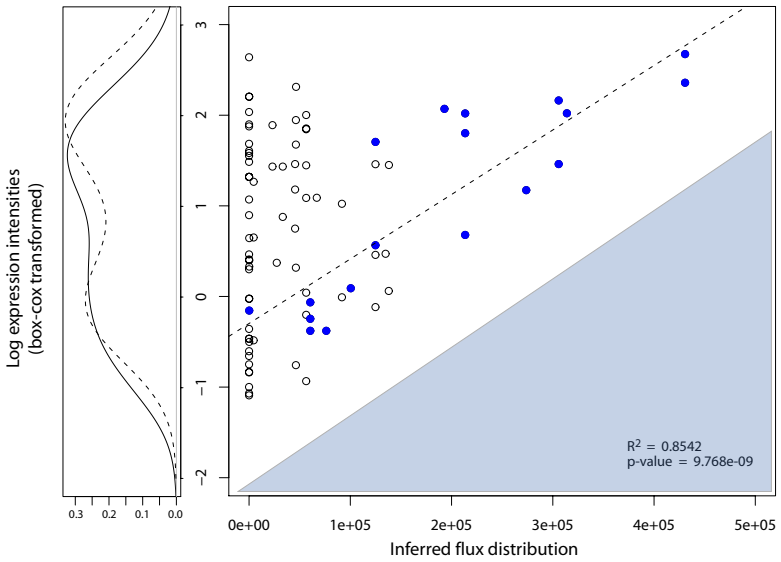

Abbildung 5.3: Scatterplot der vorhergesagten Genexpression gegen experimentell bestimmte (log-transformiert) für alle 90 Gene des Netzwerks. Es ist deutlich zu sehen, dass die geschätzte Flussverteilung eine untere Schranke für die Genexpression angibt. Blau markierte Punkte sind stellen die Gene aus Glykolyse und PPP dar. Dichteschätzungen (links) sowohl des Gesamtdatensatzes (durchgezogen) als auch des hier gezeigten (getrichelt) zeigen, dass die Korrelation nicht auf einen Ausschnitt der Gesamtdaten beschränkt ist.

Bettenbrock *et al.*, 2006; Voit, 2002), ii) der Vergleich der Struktur metabolischer Netzwerke mit Flussmessungen (Herrgard *et al.*, 2006) und kinetischen Daten (Schwartz and Kanehisa, 2006) sowie iii) die Korrelation der Resultate aus Steady-State-Analysen mit Genexpressionswerten (Covert *et al.*, 2001; Covert and Palsson, 2003) mit oder ohne biologischen Kovariaten wie Transkriptionsfaktorbindestellen (TFBS, Yeang and Vingron (2006)) oder Schätzungen der Biomasseproduktion (Stelling *et al.*, 2002).

All diese Verfahren bearbeiten dabei meist einzelne biologische Phänomene und entsprechende Messungen. Sie enthalten ausgeklügelte Verfahren um diese Daten mit den *in-silico* Vorhersagen zu vergleichen, beinhalten demgegenüber aber selten weitere experimentelle Datensätze um die gemachten Aussagen zu validieren. Wir zeigen hier, wie der biologische Zusammenhang zwischen zwei wichtigen organismischen Eigenschaften — der Metabolitproduktion und Regulation des Stoffwechsels durch Änderungen

der Genexpression — in ein kohärentes Modell eingebettet werden können. Es ist nicht nur in der Lage die gelernten Daten akkurat wiederzugeben, sondern hat sich in seiner Vorhersage zumindest gegenüber der Expression von Genen aus der Glykolyse und dem PPP als zuverlässig erwiesen. Auch wenn die Prädiktivität sich nicht für alle Gene des Metabolismus bestätigte, stellt die vorhergesagte Genexpression dennoch eine untere Schranke für die tatsächlichen experimentellen Werte dar (Abbildung 5.2.4), d.h. es wurden keine Enzyme gefunden mit hohem vorhergesagtem Flusskoeffizient aber niedriger tatsächlicher Genexpression. Einige Gene hatten zwar höhere Expressionswerte als vorhergesagt, angesichts der Tatsache, dass das modellierte metabolische Netz nur ein kleiner Ausschnitt aus dem tatsächlichen Gesamtnetzwerk ist, ist das allerdings wenig überraschend. Dies gilt insbesondere für die hochvernetzten Bereiche im Stoffffluss unterhalb von Pyruvat.

Im Vergleich zu anderen Methoden der Flussvorhersage, z.B. der FBA, gibt es einige größere Unterschiede. FBA maximiert das Skalarprodukt zwischen den Vektoren der experimentell bestimmten und vorhergesagten Flussverteilungen und nimmt dabei die Steady-State-Bedingung als Nebenbedingung auf. Im Gegensatz dazu beginnt die Suche nach der optimalen Flussverteilung in unserem Ansatz bereits im Raum der erlaubten Flüsse. Durch die Beschränkung auf die (positive) Linearkombination der konvexen Basisvektoren ist jede gefundene Flussverteilung automatisch eine erlaubte und die Nebenbedingung kann aus der Optimierung entfernt werden. Aus Gründen der Konvexheit des Optimierungsproblems (siehe Abschnitt 4.2.3) ist das Optimum das gleiche wie von Poolman *et al.* vorgeschlagen (Poolman *et al.*, 2004). Allerdings ist die dort vorgeschlagene Berechnung der Pseudoinversen mit einer Singulärwertzerlegung oder direkt verwandten Verfahren verbunden, was bei großen Matrizen zu erheblichen Laufzeitproblemen führt. Unser Optimierungsansatz verhält sich demgegenüber deutlich performanter und ist zudem in der Lage ohne Reformulierung des Problems sowohl Metabolit- also auch Enzymdaten zu optimieren.

In der von Herrgard *et al.* vorgeschlagenen Lösung wird das Problem als sogenanntes *Bilevel Optimization Problems* formuliert (Herrgard *et al.*, 2006). Es wird versucht durch systematisches Entfernen von Reaktionen und anschließender FBA das System an die experimentellen Messungen anzupassen. Ein Entfernen von Reaktionen aus dem System entspricht einer Transformation des konvexen Kegels der erlaubten Flüsse dahingehend, dass der resultierende neue Kegel ein echter konvexer Unterraum, ein Unterkegel, des Urbildes ist. Dies ist äquivalent zu einer Linearkombination der Basisvektoren des Urbildes indem die entsprechenden Extreme Pathways die das fragliche Enzym enthalten, einen Linearkombinationskoeffizienten von Null erhalten. Somit ist eine Modifikation des Netzwerks und anschließendes erneutes Ausführen der Steady-State-Analysen vollständig überflüssig. Unser Ansatz umfaßt den alternativ vorgeschlagenen in seiner Gänze ohne die zuvor handannotierte und biologisch verifizierte Netzwerkstruktur zerstören zu

müssen.

Dabei ist zu bemerken, dass eine präzise und rigorose Aufstellung des Netzwerkes notwendige Ausgangsbasis für eine solche Analyse ist. Die umfaßt wie oben beschrieben eine exakte Untersuchung der einzelnen Enzyme auf An- oder Abwesenheit in dem jeweiligen Organismus genauso wie detaillierte Informationen über die Konnektivität der modellierten Metabolite im Modell und der Zelle. Desweiteren muss man bedenken, dass der Zusammenhang zwischen metabolischer Regulation und Genexpression bei höheren Eukaryoten sicherlich nicht so eindeutig zu bestimmen sein wird wie es hier der Fall war, wo der allosterischen Regulation durch gezielte Aktivierung und Inhibierung von Enzymaktivitäten eine größere Rolle zukommt. Die schließt ein hohes Level von posttranskriptionaler Regulation mit ein, beispielsweise durch micro-RNAs und RNA Response Elemente, und wird es somit nötig machen weitere Kovariate in das Modell mit aufzunehmen. Weitere Herausforderungen umfassen das Abbilden von modellierten Reaktionen auf konkrete Enzyme, welches durch multifunktionale Proteine oder Splicevarianten in eukaryotischen Systemen erschwert wird. Auch werden möglichst genaue Messungen sowohl der Metabolitproduktion oder -flüsse als auch der Genexpression benötigt um ein ausreichend hohes Signal-Rausch-Verhältnis zu erreichen.

Es ist die bijektive Modellierung der Abhängigkeit zwischen zwei unabhängig erhobenen aber biologisch abhängigen Datensätzen die unseren Ansatz auszeichnet. Dieser demonstriert und stärkt die Position der Steady-State-Analysen als Rückgrat der Analyse von metabolischen Netzen und ihrer Regulation und liefert sowohl eine biologische als auch statistische Verifikation. Zukünftige Arbeiten werden eukaryotische Systeme betrachten, in welchen die zunehmende organismische Komplexität durch das Einbeziehen regulatorische Netzwerke oder zuätzlicher Kovariater (z.B. TFBS, Yeang and Vingron (2006)) angegangen werden soll.

Kapitel 6

Ein probabilistisches Modell der Zellgröße bei *Pseudo-nitzschia delicatissima*

Als Modell auf Organismusebene möchte ich zuletzt das Augenmerk auf eine Art von Kieselalgen, *Pseudo-nitzschia delicatissima* richten, deren Lebenszyklus einige interessante Besonderheiten aufweist.

6.1 Der Lebenszyklus von Diatomeen

Pseudo-nitzschia delicatissima ist eine kettenformende federartige planktonische Diatomee, welche oft im Überfluss in wohltemperierten Gewässern der Weltmeere zu finden ist. Umgeben von einer starren Silikathülle durchwandert die Alge im Rahmen der vegetativen Zellteilung eine Reihe sukzessiver Verkleinerungen. Schließlich wechselt sie in ein Stadium der sexuellen Reproduktion, wirft die hinderliche Schale von sich und stellt durch Verschmelzung mit einer weiteren andersgeschlechtlichen Diatomee in etwa ihre originale Ausgangsgröße wieder her (Mann, 1993; Round *et al.*, 2007). 2005 führten Amato *et al.* Paarungsexperimente an drei unterschiedlichen *Pseudo-nitzschia delicatissima* Proben aus dem Golf von Neapel durch. Nachkommen der ursprünglichen Proben wurden in monoklonaler Kultur aufgezogen, um eine sexuelle Reproduktion zu verhindern. Dann wurde die kontinuierliche Reduktion der Zellgröße der Algen über einen Zeitraum von 265 Tagen mit Hilfe eines Lichtmikroskops vermessen (Amato *et al.*, 2005). Die beobachteten Zellgrößen reichen von $80\mu m$ apikale Achsenlänge bis hin zu $18\mu m$ nach 265 Tagen bis die Zellen schließlich abstarben (zur Illustration siehe Abbildung 6.1).

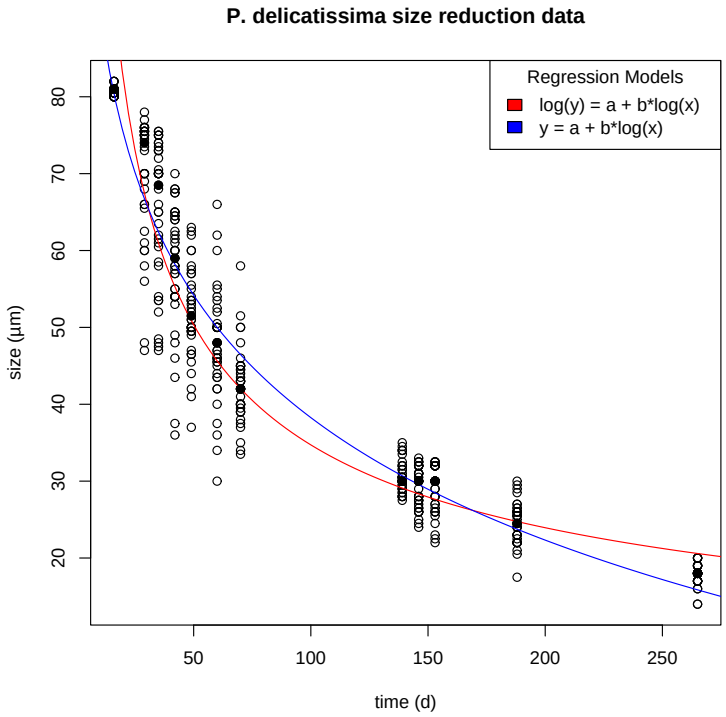

Abbildung 6.1: Originaldatensatz der Zellgrößen im Verlauf der Zeit. Die blauen und roten Linien sind lineare Regressionsmodelle die mittels Logtransformation der erklärenden Variablen (blau) oder der erklärenden und Zielvariablen (rot) gefittet wurden. Die Inverse der Regressionsfunktion des Modells 1 (rot) wurde später zum Abschätzen der Haltezeiten der einzelnen Zustände verwendet. Beide Modelle zeigten eine gute Performance mit einem R^2 Wert von 0.92 und P-Werten $<$ 2.2e-16

Neben des interessanten Lebenswandels der Diatomeen an sich spielt die Größenreduktion auch in anderen Bereichen, wie der Gemeinschaftsanalyse (Snoeijs *et al.*, 2002), eine wichtige Rolle. Dennoch haben sich die meisten mathematischen Analysen bisher auf Regressionsmodelle unter Einbeziehung unterschiedlicher Kovariater beschränkt (Mizuno, 1991) oder waren damit zufrieden einzelne Charakteristika wie durchschnittliche Teilungsraten oder maximale Größenreduktionsraten zu schätzen (Amato *et al.*, 2005; Fehling *et al.*, 2005; Jewson, 1992). Ohne Zweifel haben diese Studien zu wichtigen Einsichten in den kryptischen Lebenswandel der Diatomeen geführt, schon allein aufgrund der massiven Datensätze die durch sie erhoben wurden. Dennoch ist bisher nach unserem Wissen noch kein Modell beschrieben worden, welches in der Lage war, die Größenreduktion einzelner Zellen

stochastisch zu modellieren und damit die immanenten Abläufe hinter den Größenverteilungen einer Diatomeenpopulation aufzudecken.

Um dies zu erreichen wählen wir hier (Schwarz *et al.*, 2008) einen Ansatz der Modellierung mittels einer homogenen Markov Kette (MC). Markov Ketten sind stochastische Prozesse, die eng mit den Petri-Netzen verwandt sind. Bestehend aus Zuständen und Zustandsübergängen wandert eine Marke gemäß der den Übergängen zugeordneten Übergangswahrscheinlichkeiten von Zustand zu Zustand. Eine der wichtigsten Eigenschaften dabei ist es, dass die Wahrscheinlichkeit in den nächsten Zustand zu wechseln einzig und allein vom Zustand abhängt, in dem die Marke sich zu dem Zeitpunkt befindet, nicht von den Zuständen zuvor. Dies ist die sogenannte *Markov-Eigenschaft*, sie garantiert dass die Kette erinnerungslos ist, das Verhalten also nicht von ihrer Vergangenheit abhängt. Im Spezialfall der homogenen Markov-Ketten verändern sich die Übergangswahrscheinlichkeiten zudem nicht mit der Zeit (Grimmett and Stirzaker, 2001). Zufallsprozesse wie diese hier wurden schon zuvor erfolgreich zur Modellierung bestimmter Vorgänge eingesetzt. Einer davon ist die stochastische Beschreibung der Polymerase Kettenraktion (PCR) durch Weiss and von Haeseler. Dort dient ein sogenannte *Bifurcating Tree* oder auch *Branching Process* der Beschreibung der schrittweisen Verdopplung der DNA Moleküle inklusive bestimmter Fehlerraten. Um zudem auch die Mutationen zu modellieren, die während des fehlerbehafteten Replikationsprozesses auftreten können, haben die Autoren dem Branching Prozess einen zusätzlichen Poisson Prozess zur Modellierung seltener Ereignisse überlagert. Dieser Ansatz wurde erfolgreich angewandt und erweitert (Weiss and von Haeseler, 1997) und wurde später auch von anderen Arbeitsgruppen aufgenommen (Saha *et al.*, 2004, 2007).

Wir zeigen hier, wie eine individuelle Modellierung der Zellgrößen innerhalb einer Diatomeenpopulation mit Hilfe einer Markovkette weitaus zufriedenstellendere Resultate liefern kann als die bisher vorherrschende Herangehensweise mittels linearer oder polynomieller Regression. Unser Modell ist in der Lage die Größenverteilungen dreier unabhängig entnommener Populationen wiederzugeben. Wir beschreiben und validieren die Zuverlässigkeit unserer Modellierung durch Rückschätzen des Populationsalters aus den erhobenen Daten und formalisieren die Integration zusätzlicher Modelleigeneschaften wie das Einnehmen eines Sporenstadiums oder den Übergang zum sexuellen Reproduktionszyklus. Insbesondere letzteres kann durch konventionelle Methoden üblicherweise nicht modelliert werden (Schwarz *et al.*, 2008).

6.2 Größenreduktion als Markov Kette

Aus der origniären Arbeit von Amato *et al.* erhielten wir drei unabhängig entnommene Diatomeenpopulationen, genannt F1-5 (Abb. 6.1), F1-13 und

F1-14. Die Population F1-5 wurde im Weiteren als Trainingsdatensatz verwendet, während die anderen beiden Populationen nur zur Validierung der Ergebnisse zum Einsatz kamen.

Wie eingangs erwähnt kam zur Modellierung der Größenreduktion von *Pseudo-nitzschia* eine zustandsdiskrete zeitkontinuierliche homogene Markovkette zur Anwendung bestehend aus einer festen Anzahl von Zuständen. Jeder Zustand entspricht dabei einem bestimmten Größenbereich der Zellen. Die Populationen entwickeln sich kontinuierlich über die Zeit, und daher kann eine Zellteilung und die damit verbundene Größenreduktion auch jederzeit eintreten (Zeitkontinutität). Die Wahrscheinlichkeit eines Zustandsübergangs hingegen, also einer oder eine Reihe von Zellteilungen, verändert sich nicht in Abhängigkeit des Alters der Population sondern nur in Abhängigkeit der Größe der Einzelzellen (Homogenität). Wir nehmen des Weiteren an, dass die Zellen sterben, sobald sie die niedrigste Größenklasse erreicht haben und schließen für unser initiales Modell keinerlei Kovariate mit ein. Da die natürliche Größenreduktion pro Teilungsschritt im Allgemeinen sehr klein ist ($< 1.5 \mu m * gen^{-1}$, Amato *et al.* (2005)) und die Zellen im Gegensatz dazu über eine große Varianz ihrer Größe von 18 bis 80 μm verfügen, hätte dies zu mehr als 40 unterschiedlichen Zuständen geführt. Das Trainieren des Modells hätte nicht nur unverhältnismäßig viel Zeit in Anspruch genommen, die Auflösung des Datensatzes hätte auch für eine derart feine Modellierung nicht ausgereicht, insbesondere da zwischen den Tagen 70 und 139 keine weiteren Messungen stattgefunden haben.

6.2.1 Diskretisierung der Trainingsdaten

Um eine angemessene Anzahl an Zuständen zu finden, die einfach zu handhaben sind aber die Größenreduktion noch adäquat wiedergeben können, kam ein Gaussian Mixture Model mehrerer Normalverteilungen zum Einsatz (Abb. 6.2) um die wahrscheinlichste Anzahl unterscheidbarer Größenkategorien und ihrer Grenzen zu detektieren (Tabelle 6.1).

State 1	**State 2**	**State 3**	**State 4**	**State 5**	**State 6**	**State 7**
80.0	74.5	57.0	48.0	37.5	21.0	14.0

Tabelle 6.1: Untergrenzen der Größenklassen die sich aus dem ML-Trainieren des Mixture Models ergaben (in μm).

Um ein Overfitting zu vermeiden evaluierten wir das Bayesian Information Criterion (BIC) der Modelle unterschiedlicher Gruppenanzahl jeweils für eine Variante mit konstanter und mit variabler Varianz aller Verteilungen. (Abb. 6.3).

Es ergab sich, dass sieben Gruppen variabler Varianz am besten geeignet waren, um unseren Trainingsdatensatz abzubilden; wir wiesen die Daten-

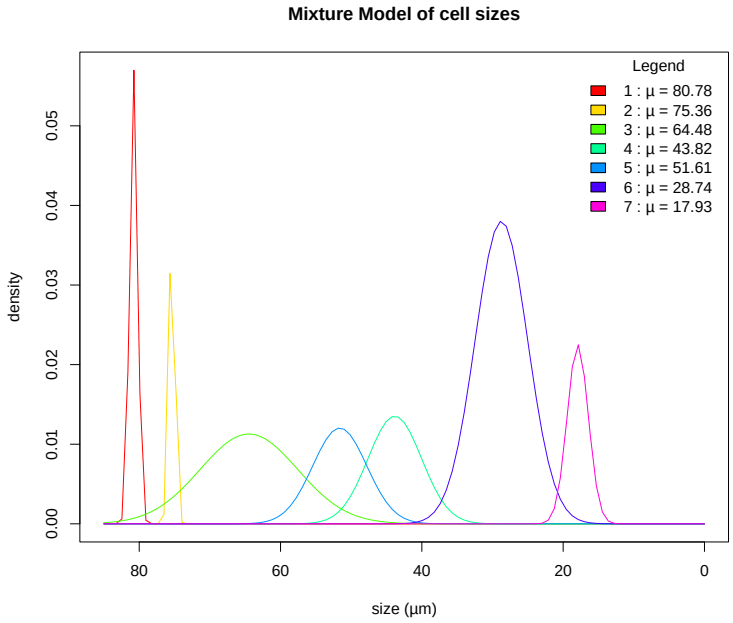

Abbildung 6.2: Kombinierte Dichteverteilung der sieben skalierten Normalverteilungen (für die Mittelwerte siehe Legende). Diese sieben Verteilungen wurden verwendet, um die originalen Größendaten in sieben eindeutige Gruppen zu unterteilen und deren Grenzen exakt angeben zu können. Jede dieser Gruppen wurde anschließend durch einen Zustand der Markovkette dargestellt.

punkte entsprechend der Vorhersagen des Mixture Models einer der sieben Kategorien zu, von denen jede einen Größenbereich zwischen 5.5 und 17.5 μm abdeckt. Für die diskretisierten Daten siehe Tabelle 6.2.

6.2.2 Schätzen einer initialen Ratenmatrix

Neben der Definition der Zustände und einer anfänglichen Populationsverteilung π_0 ist eine Markov Kette vollständig durch ihre Ratenmatrix Q beschrieben. Um die für unsere Trainingspopulation optimale Ratenmatrix zu finden, entschieden wir uns zunächst, eine initiale Matrix anhand der Aufenthaltszeiten der Zellen in den einzelnen Zuständen zu lernen. Die geschah durch Anwendung eines Regressionsmodells auf dem Originaldatensatz (Abb. 6.1). Die Matrix wurde anschließend durch numerische Auswertung des Maximum Likelihood Schätzers der MC weiter optimiert.

In der Theorie der Markovketten ist bekannt, dass die Aufenthaltszeiten (sog. Holding Times) der Kette in einem Zustand exponentialverteilt sind

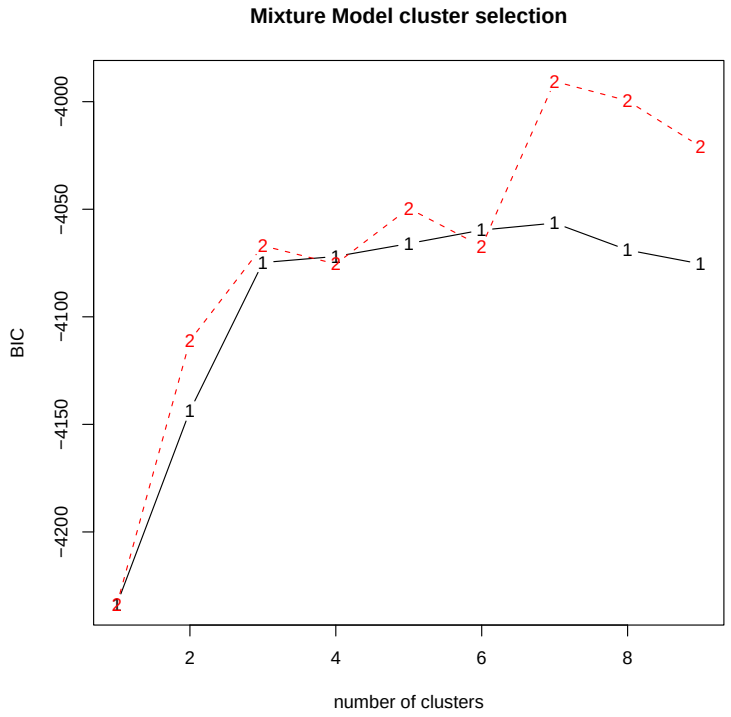

Abbildung 6.3: Ergebnisse der Anwendung des Mixture Models. Die Anzahl der gewählten Gruppen wurde gegen das Bayesian Information Criterion (BIC) evaluiert um ein Overfitting der Trainingsdaten zu vermeiden. Sowohl Modelle mit variabler (rot) und konstanter Varianz (schwarz) kamen dabei zum Einsatz. In beiden Fällen waren sieben Gruppen am besten in der Lage die Originalverteilung wiederzugeben.

mit einem Mittelwert gleich $1/q_{ii}$ (der Hauptdiagonalen der Ratenmatrix Q, (Grimmett and Stirzaker, 2001, S. 259)). Wir haben unsere Ratenmatrix weiter dahingehend eingeschränkt, dass wir den Prozess als einen reinen Birth-Death-Prozess mit einer $n \times n$ Ratenmatrix

$$Q = \begin{pmatrix} q_{11} & q_{12} & & 0 \\ & \ddots & \ddots & \\ & & q_{kk} & q_{kn} \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{6.1}$$

$(k = n - 1)$ modellieren und damit nur Zustandsübergänge von einem Zustand in den nächst kleineren erlauben. Dies entspricht der biologischen Intuition, dass Zellteilungen in erster Linie nacheinander erfolgen. Zum Schätzen

days	**16**	**29**	**35**	**42**	**49**	**60**	**70**	**139**
State 1	1	0	0	0	0	0	0	0
State 2	0	0.487	0.128	0	0	0	0	0
State 3	0	0.436	0.667	0.641	0.231	0.077	0.026	0
State 4	0	0.051	0.154	0.256	0.564	0.436	0.103	0
State 5	0	0.026	0.051	0.077	0.179	0.41	0.769	0
State 6	0	0	0	0.026	0.026	0.077	0.103	1
State 7	0	0	0	0	0	0	0	0

days	**146**	**153**	**188**	**265**
State 1	0	0	0	0
State 2	0	0	0	0
State 3	0	0	0	0
State 4	0	0	0	0
State 5	0	0	0	0
State 6	1	1	0.949	0
State 7	0	0	0.051	1

Tabelle 6.2: Relative Zustandshäufigkeiten des Trainingsdatensatzes nach der Diskretisierung.

der Aufenthaltszeiten kam Modell 1 (die rote Regressionslinie in Abb. 6.1) zum Einsatz, welches von den beiden Alternativen mit einem R^2 Wert von 0.9216 und einem P-Wert $< 9.22 * 10^{-16}$ die bessere Passung zeigte. Da das Regressionsmodell 1 sowohl die erklärende Variable wie auch die Zielvariable logtransformiert betrachtete, wurde die Inverse Funktion

$$x = y^{\frac{1}{b}} * e^{-\frac{a}{b}} \tag{6.2}$$

verwendet, um die Aufenthaltszeiten durch die vorher bestimmten Grenzen der Größenklassen zu bestimmen. Die resultierenden Aufenthaltszeiten reichten von 2.99 Tagen (Zustand 2) bis zu 168.7 Tagen (Zustand 6). Vergleiche auch Tabelle 6.3.

State 1	**State 2**	**State 3**	**State 4**	**State 5**	**State 6**
5.077451	2.994306	15.589484	14.984200	31.949934	168.739609

Tabelle 6.3: Aufenthaltszeiten (in Tagen) der Zustände der Markovkette wie sie durch das Mixture Model und die Regression vorgegeben wurden. Der letzte Zustand gilt in unserem Modell zusätzlich als Zustand des Tods der Zellen, so dass ihm eine unendliche Aufenthaltszeit zukommt.

Diese Werte brachten uns unmittelbar zur initialen Ratenmatrix Q durch

Verwendung des Kehrbruchs und Multiplikation mit -1.

6.2.3 Auf der Suche nach dem optimalen Q

Nach dem Aufsetzen einer ersten Ratenmatrix optimierten wir die Übergangswahrscheinlichkeiten durch numerische Evaluation des MLE, um die Parameter zu finden, die unseren Trainingsdatensatz am ehesten generieren konnten. Im folgenden soll K die Anzahl der Zustände der MC, L die Anzahl der Datensätze (d.h. die Anzahl der Zeitpunkte, an denen Messungen erfolgten) und N die (konstante) Anzahl der Messungen pro Zeitpunkt angeben. Wenn des Weiteren

$$\pi(t_j) = \pi_0 * e^{t_j Q}$$

die Zustandsverteilung zum Zeitpunkt t_j angibt und $x_{ij} \in \{1 \dots K\}, i \in \{1 \dots N\}, j \in \{1 \dots L\}$ der beobachtete Zustand des Individuums i zum Zeitpunkt j ist, dann ist die Likelihood der Parameter t und Q gegeben die Daten

$$\begin{aligned} L(Q;t) &= \prod_{i=1}^{N} \prod_{j=1}^{L} \left[\pi(t_j)\right]_{x_{ij}} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{L} \left[\pi_0 * e^{t_j Q}\right]_{x_{ij}} . \end{aligned}$$

Der Einfachheit halber definieren wir mit

$$n_{ij} := \#(x_j = i)$$

die Anzahl der Individuen in Zustand i zum Zeitpunkt j. Durch Ersetzen von x durch n in der Likelihood und Logtransformation erhalten wir

$$\mathcal{L}(Q;t) = \sum_{j=1}^{L} \sum_{i=1}^{K} n_{ij} \ \log\left(\pi_0 \ e^{t_j Q}\right)_i . \tag{6.3}$$

Das gesuchte $\operatorname{argmax}_Q \mathcal{L}(Q;t)$ wurde im Rahmen unserer Analyse dann letzlich numerisch gefunden durch Anwendung der quasi-Newton BFGS Variante L-BFGS-B (Byrd *et al.*, 1995) in ihrer Implementierung im R Softwarepaket (R Development Core Team, 2006).

6.2.4 Konfidenzintervalle

Nach Berechnung einer optimalen Ratenmatrix ist es nun möglich, mit Hilfe der *Fisher Information* der Markovkette die Varianz des Maximumlikelihoodschätzers zu bestimmen und damit Konfidenzintervalle des Zeitschätzers anzugeben. Grundlage ist die asymptotische Konvergenz des MLEs in Ver-

teilung gegen eine Normalverteilung, also

$$\frac{(\hat{\theta}_n - \theta)}{se} \rightsquigarrow N(0,1),$$

wobei der Standardfehler se durch die inverse Fisherinformation

$$se \approx \sqrt{1/I_n(\theta)}$$

angegeben ist (Wasserman, 2005, S.129 u. 135f.). Ausgehend vom *Score* einer Beobachtung

$$s(X;\theta) = \frac{\partial}{\partial\theta} \log f(X;\theta) \tag{6.4}$$

ist die *Fisher Information* über einer Probe aus n Beobachtungen definiert als

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\sum_{i=1}^{n} s(X_i;\theta) \right) \\ &= \sum_{i=1}^{n} \mathbb{V}_\theta(s(X_i;\theta)). \end{aligned} \tag{6.5}$$

Da gezeigt werden kann, dass der Erwartungswert $\mathbb{E}_\theta(s(X;\theta)) = 0$, folgt daraus dass

$$\mathbb{V}_\theta(s(X;\theta)) = \mathbb{E}_\theta(s^2(X;\theta)).$$

Desweiteren ergibt sich als Korollar aus der Additivität der Information (Gleichung 6.5) bei unabhängigen Ereignissen, dass

$$I_n(\theta) = nI_1(\theta) = nI(\theta). \tag{6.6}$$

Alternativ kann dann die Information auch als

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2} \right) \\ &= -\int \left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2} \right) f(x;\theta)dx \end{aligned} \tag{6.7}$$

beschrieben werden (für Beweise zu Gleichungen 6.6 und 6.7 siehe z.B. Lindgren (1993); Wasserman (2005)).

Im konkreten Fall unserer Markovkette haben wir eine nach t parametrisierten Familie von Verteilungen

$$\pi_t(i) = \left(\pi_0 \; e^{tQ}\right)_i,$$

die die Wahrscheinlichkeit angibt zum Zeitpunkt t in Zustand i zu sein. Die Log-Likelihood *einer* Beobachtung und damit die Scorefunktion ergibt sich

somit als

$$\begin{aligned} s(i;t) &= \frac{\partial}{\partial t} \log f(i;t); \\ &= \frac{\partial}{\partial t} \log \left(\pi_0 \; e^{tQ}\right)_i \\ &= \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i}{\left(\pi_0 \; e^{tQ}\right)_i}. \end{aligned}$$

Entsprechend ist nach Gleichung 6.7 die Information *einer Beobachtung* gegeben durch

$$\begin{aligned} I_1(t) &= I(t) \\ &= -\mathbb{E}\left[\frac{\partial}{\partial t} \; \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i}{\left(\pi_0 \; e^{tQ}\right)_i}\right] \\ &= -\mathbb{E}\left[\left(\pi_0 \; Q^2 \; e^{tQ}\right)_i * \left(\pi_0 \; e^{tQ}\right)_i^{-1} + \left(\pi_0 \; Q \; e^{tQ}\right)_i * - \left(\pi_0 \; e^{tQ}\right)_i^{-2} * \left(\pi_0 \; Q \; e^{tQ}\right)_i\right] \\ &= -\mathbb{E}\left[\frac{\left(\pi_0 \; Q^2 \; e^{tQ}\right)_i}{\left(\pi_0 \; e^{tQ}\right)_i} - \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i^2}{\left(\pi_0 \; e^{tQ}\right)_i^2}\right] \\ &= -\sum_{i=1}^{K}\left[\left(\frac{\left(\pi_0 \; Q^2 \; e^{tQ}\right)_i}{\left(\pi_0 \; e^{tQ}\right)_i} - \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i^2}{\left(\pi_0 \; e^{tQ}\right)_i^2}\right) * \left(\pi_0 \; e^{tQ}\right)_i\right] \\ &= -\sum_{i=1}^{K}\left(\left(\pi_0 \; Q^2 \; e^{tQ}\right)_i - \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i^2}{\left(\pi_0 \; e^{tQ}\right)_i^2}\right). \end{aligned} \tag{6.8}$$

Damit ist die Gesamtinformation

$$I_n(t) = \sum_{i=1}^{K} N_{ij}\left(\left(\pi_0 \; Q^2 \; e^{tQ}\right)_i - \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i^2}{\left(\pi_0 \; e^{tQ}\right)_i^2}\right), \tag{6.9}$$

wobei N_{ij} die Anzahl der Beobachtungen im Zustand i zum Zeitpunkt j ist.

Das 95% Konfidenzintervall ergibt sich damit aus

$$\hat{t} \pm 2\sqrt{1/I_n(t)}. \tag{6.10}$$

Für die Konfidenzintervalle des Zeitschätzers unserer Modellvalidierung siehe Tabelle 6.4.

6.3 Diskussion

Das hier vorgeschlagenene stochastische Modell der Größenreduktion in *Pseudo-nitzschia delicatissima* hat gegenüber den bisherigen Ansätzen mittels klassischer linearer Regression entscheidende Vorteile (vgl. Mizuno (1991); Amato

real	0	13	19	26	33	44
estimate	0.500	10.522	16.852	22.507	31.255	41.400
low	0.042	7.783	12.918	17.544	24.781	33.267
high	0.958	13.260	20.785	27.470	37.728	49.533
real	54	123	130	137	172	249
estimate	53.082	136.284	136.284	136.284	144.350	300.000
low	43.058	105.290	105.289	105.289	110.204	197.434
high	63.107	167.279	167.279	167.279	178.495	402.566

Tabelle 6.4: Konfidenzintervalle (*low* und *high*) der Zeitschätzer (*estimate*) der F15 Population.

et al. (2005); Fehling *et al.* (2005); Jewson (1992)). Erst durch Beschreibung des Teilungsverhaltens mittels eines stochastischen Prozesses besteht die Möglichkeit, die Rücküberführung der Zellen in ihre originale Größe durch Übergang in den sexuellen Reproduktionszyklus adäquat abzubilden. Insbesondere die bisherigen linearen Modelle stoßen dabei an ihre Grenzen.

Mittels konsequenter Validierung durch zwei unabhängige Datensätze und Rückschätzung der Erhebungszeitpunkte der Proben inklusive Berechnung der 95%-Konfidenzintervalle konnten wir eine solide Abschätzung der Genauigkeit unseres Modells vornehmen. Wir sind davon überzeugt, dass es unter Einfluß von mehr und präziseren Trainingsdaten noch weiter an Genauigkeit gewinnen wird.

Ein weiterer interessanter Aspekt dabei ist die Tatsache, dass ein Birth-Death Prozess wie von uns beschrieben in äquivalenter Form als Power-Law dargestellt werden kann (Karev *et al.*, 2002, 2003) . Angesichts der vielen biologischen Größen, deren Zusammenhang einer Power-Law Gesetzmäßigkeit folgt, wäre dies auch bei den Diatomeen nicht abwegig, bedarf natürlich aber weiterer Verifikation.

In einem nächsten Schritt soll das Modell jetzt erweitert werden, um den sexuellen Reproduktionszyklus sowie für andere Diatomeenspezies eventuelle Sporenzustände zu integrieren. Dann können unmittelbar aus der Umwelt entnommene Proben mit unserem Modell verglichen werden um Populationsalter zu schätzen und Abweichungen von einer normalen Populationsentwicklung frühzeitig zu erkennen und vorherzusagen. Desweiteren könnte eine zweiter assoziierter Prozess nach dem Vorbild von Weiss and von Haeseler 1995; 1997 verwendet werden, um nicht nur die relativen Häufigkeiten der Größenklassen, sondern auch die Größe der Population selbst zu modellieren.

Teil III

Abschließende Diskussion und Fazit

6.4 Allgemeine Diskussion

Ähnlich rasant wie die Entwicklung der molekularbiologischen Verfahren, der Sequenziermethoden und Verfahren zur Messung der Genexpression, hat sich die Bioinformatik in den vergangenen Jahren weiterentwickelt. Immer mehr treten systembiologische Fragestellungen in den Vordergrund, also solche, die über die Betrachtung einzelner Phänomene und biologischer Mechanismen hinaus Organelle, Zellen oder Organismen als Ganzes zu betrachten versuchen, also als Ziel haben ein biologisches System möglichst vollständig zu verstehen. Auch im Rahmen der hier vorliegenden Arbeit, deren ursprünglicher Fokus auf der Modellierung metabolischer Netzwerke lag, sind im Laufe der drei Jahre zahlreiche Publikationen entstanden aus Bereichen weit jenseits der Modellierung metabolischer Netze. Sie beschreiben den ganzen Weg vom Gen (Kapitel 2, unser DFG Antrag in Evaluationsphase) über das Transkriptom (Kapitel 3, Engelmann *et al.* (2007)) über das Metabolom (Kapitel 4, Schwarz *et al.* (2005, 2007); Eisenreich *et al.* (2006)) hin zur Beschreibung organismischer Entwicklung (Kapitel 6, Schwarz *et al.* (2006)). Hinzu kamen Projekte die die Einzelbereiche im Sinne der Systembiologie miteinander verbunden haben, so die Untersuchung des Zusammenhangs zwischen Genexpression und Metabolitproduktion mittels Daten aus ^{13}C Isotopologstudien und Steady-State-Modellen metabolischer Netzwerke, (Kapitel 5, Manuskript in Vorbereitung). Neben einer fundierten theoretischen Grundlage legten die Projekte stets Wert auf anwendungsnahe bioinformatische Forschung an Modellorganismen (Kapitel 3, *A. thaliana*), medizinisch bedeutenden Pathogenen (Kapitel 4 und 5, *L. monocytogenes* und *S. aureus*) oder allgemeinbiologisch interessanten Organismen (Kapitel 6, *P. delicatissima*). Hinzu kamen noch zahlreiche kleinere Nebenprojekte wie beispielsweise die Verteilung der Ameisenpatrilinien in südamerikanischen Ameisenhügeln (unveröffentlicht), die Verteilung von Genlängen auf Genomen (Levin *et al.*, 2005), das Phospholipidnetzwerk im Phagosom (Schwarz *et al.*, 2007) und die Vermeidung von Antibiotikaresistenzen in Bakterien (Ziebuhr *et al.*, 2004; Becker and Palsson, 2005).

Neben der direkten biologischen und medizinischen Relevanz zeigt gerade diese Doktorarbeit, dass moderne statistische Modelle und Verfahren unumgänglich sind um all diese Probleme effektiv anzugehen. Von Methoden der klassischen linearen Algebra und konvexen Analysis (Kapitel 4 und 5) über Verfahren aus dem Bereich der KI, statistischen Lerntheorie, Signal- und Informationsverarbeitung (Kapitel 2 und 3) bis hin zur Schätztheorie und stochastischen Prozessen (Kapitel 6) deckt die vorliegende Arbeit ein breites Spektrum mathematisch relevanter Verfahren der heutigen bioinformatischen Forschung ab. Für eine detaillierte Diskussion der einzelnen Projekte siehe die entsprechenden Kapitel des Resultateteils.

Neben den hier beschriebenen Resultaten werden die Ergebnisse der letzten drei Jahre meine zukünftigen Arbeiten in der Bioinformatik auch weiter-

hin beeinflussen. Im anstehenden Projekt zusammen mit der AG Schön des Instituts für Hygiene und Infektionsbiologie der Universitätsklinik Würzburg werden die erworbenen Fertigkeiten und entwickelten Verfahren an Meningokokken (*N. meningitidis*) zum Einsatz kommen und zusammen mit Transkriptomdaten und phylogenetischer Rekonstruktion der evolutionären Stammbäume wertvolle Einblicke in das Leben von Neisserien ermöglichen. Ziel ist u.a. die Beantwortung der Frage nach dem exakten Alter der verschiedenen Meningokokken und einer möglichen Erklärung für die immer wieder überraschend und plötzlich auftretenden Erkrankungen.

6.5 Fazit

Zusammenfassend bleibt noch zu bemerken, dass der Schwerpunkt der vorliegenden Arbeit neben dem angesprochenen systembiologischen Gesamtkontext nicht immer nur und primär auf dem Problem selbst lag, sondern vor allem auf seiner Modellierung, seiner Abstraktion, also der Strukturbildung und damit letztlich der Mathematik. Denn was sonst ist Inhalt der Mathematik als in konkreten Problemen Muster zu sehen, Strukturen zu erkennen, sie so präzise wie möglich zu formulieren, dabei die wesentlichen Einflußgrößen zu extrahieren und andere außen vor zu lassen, also *abzubilden*, zu abstrahieren oder zu *verkürzen*, kurz zu Modellieren. Die zu beantwortende Frage ist dabei das pragmatische Merkmal des Stachowiak (Stachowiak, 1973), *die gedankliche oder tatsächliche Operation* für die das Modell entwickelt wurde, mit dem Bioinformatiker, also mir selbst, als *handelndem und modellbenutzendem Subjekt*.

Damit drängt sich gerade die Frage auf, was denn eigentlich im Vordergrund stehen sollte bei einer Dissertation in der Bioinformatik, die biologische Fragestellung oder der informatisch-mathematische Hintergrund? Dies allgemeingültig zu beantworten ist sicherlich schwierig. Zumindest muss man sich dafür die Entwicklung dieser Wissenschaft in den letzten Jahren noch etwas genauer vor Augen führen. Entstanden aus einem Zusammenspiel von Biologie und Informatik und damit auch der Mathematik hat das interdisziplinäre Feld von jeher Wissenschaftler jeder Couleur angezogen, Biologen und Biochemiker ebenso wie Mathematiker, Informatiker und Physiker. So heterogen wie das Forscherfeld so unterschiedlich waren auch die Ziele und Bezeichnungen der neuen Wissenschaft, von denen sich *Bioinformatics* und *Computational Biology* bis heute durchsetzten. Die Unterschiede beider Bereiche die heute meist einfach unter *Bioinformatik* zusammengefaßt werden sind deutlich, die Übergänge jedoch fließend. Während in der *Bioinformatics*-Community der Schwerpunkt meistens deutlich auf der Algorithmik und Mathematik liegt, also auf der Entwicklung neuer Methoden für die biologische Anwendung, ausgehend vom aktuellen Stand der Forschung der Informatik und Mathematik, so steht demgegenüber der Bereich der *Computatio-*

nal Biology mit einem mehr biologischen Schwerpunkt, also der klassischen Bearbeitung oftmals großer Mengen biologischer Daten mit informatischen Methoden und Programmen zur Beantwortung konkreter biologischer Fragestellungen.

Doch es ist gerade letzterer Bereich, der in den vergangenen Jahren im Wandel begriffen war. Mit der Verfügbarkeit großer Rechenleistung auf Basis normaler Arbeitsplatzrechner und der durch die fortschreitende Technisierung immer höhere Ausbildungsstand an grundlegenden informatischen Tätigkeiten, die neben Alltagsarbeiten am PC auch bereits die Programmierung kleinerer Programme beinhalten, ist dieser Bereich der bioinformatischen Forschung mehr und mehr auch von anderen Fachbereichen aufgenommen worden. Es sind nicht mehr spezialisierte Informatiker nötig, um größere Datenmengen zu analysieren oder kleinere Programme in Skriptform zu entwerfen. Die Bioinformatik hat Einzug gehalten in jeden Bereich der Life Sciences, moderne Laborarbeit wäre ohne schnelles Suchen in Onlinedatenbanken kaum denkbar. Von informatischer Seite sind ausgeklügelte Programme entwickelt worden, um den komplexen Ablauf des Analysierens von Daten oder molekularen Prozessen zu vereinfachen und sie dem Wissenschaftler ohne Informatikhintergrund zu ermöglichen. Es ist nicht mehr die *Computational Biology* als solche die biologische Daten mit informatischen Methoden bearbeitet, die Biologie selbst ist mittlerweile *computational* geworden, also informatisch und rechenintensiv, und diese Entwicklung wird sich in den nächsten Jahren fortsetzen. Das, was vor einiger Zeit noch als bioinformatisches Spezialwissen galt, wird oder hat bereits Einzug gehalten in das biologische Grundstudium und wird in Zukunft für jeden angehenden Biologen vorausgesetzt.

Was bleibt den Bioinformatikern? Konzentration auf den stärker mathematik- und informatiklastigen Bereich der *Bioinformatics*, also der Modellierung biologischer Probleme und Verbesserung von Algorithmen zu Analyse biologischer Daten. Was heißt das in Konsequenz? Die Anforderungen an die Bioinformatikgemeinschaft werden größer. Bioinformatiker ohne mathematischen Hintergrund sind mehr und mehr gezwungen, sich auf die Formalsprache ihrer mathematischer orientierten Kollegen einzulassen, während jene sich ihrer Position in einem interdisziplinären Feld an der Schnittstelle zwischen Life Sciences und Mathematik stärker denn je bewußt werden müssen. Bioinformatiker werden auch weiterhin dort zum Einsatz kommen, wo große Datenmengen zu bewältigen sind, nur dass diese Grenzen sich in dem Maße verschieben, in dem der Rest der Life Sciences Community bioinformatisches Wissen erwirbt. Diese Datenmengen rufen damit umso mehr nach fundierten Methoden, damit aus dem Hintergrundrauschen experimenteller Erhebungen verwertbare biologische Signale und in deren Interpretation Erkenntnis über biologische Zusammenhänge gewonnen werden können. Kandidaten dafür sind unter anderem Methoden der Mustererkennung, des maschinellen Lernens, unterstützt von Resultaten der statistischen Lern-

theorie, also Methoden der Statistik. Dies wird eine der Herausforderungen sein für die kommende Generation von Bioinformatikern und die vorliegende Arbeit war ein Weg der Erkenntnis, von biologischen Fragestellungen im Zusammenhang mit metabolischen Netzwerken zur allgemeineren mathematischen, insbesondere statistischen, Modellierung biologischer Probleme. All dies hat der Arbeit im Laufe der Jahre eine interessante Wendung gegeben. Um es im Sinne eines bekannten Bioinformatikers frei noch einmal zu formulieren: „The consequence of all this was that I went back to school and learned more statistics“ (Vingron, 2001).

Material und Methoden

Zur Erstellung der Dissertation und der beteiligten Projekte kamen ein handelsüblicher FujitsuSiemens[1] Desktop PCs mit einem 3.2 GHz HT Prozessor sowie ein Laptop der Firma Dell[2] zum Einsatz. Rechenintensive Anwendungen wurden auf dem 24 Node Cluster des Instituts für Bioinformatik der Universiät Würzburg ausgeführt. Zur exakten Datenbanksuche mittels des Smith-Waterman Algorithmus (Smith and Waterman, 1981) kam ein Genematcher 2 der Firma Paracel zum Einsatz. Bei der Erstellung der im Rahmen der Dissertation angefertigten Programme wurde das Java Software Development Kit (SDK) der Firma SUN in Versionen 1.4, 1.5 und 1.6 verwendet[3]. Als integrierte Entwicklungsumgebung (IDE) diente dabei der JBuilder 2005 der Firma Borland[4]. Zum Setzen der wissenschaftlichen Artikel sowie des vorliegenden Dokuments diente die LaTeX 2_ε Distribution TeTeX in Kombination mit der Entwicklungsumgebung AucTeX für Emacs, beide in Entwicklung der Free Software Foundation[5]. Im Rahmen der statistischen Modellierung und Berechnungen kam die statistische Programmiersprache R zum Einsatz[6] (R Development Core Team, 2006). Für spezifische biologische Fragestellungen war insbesondere das Paket Bioconductor[7] (Gentleman *et al.*, 2004) von Interesse. Als Entwicklungsumgebeung diente hierbei die Emacs Erweiterung ESS[8] (Heiberger, 2001). Für rechenintensivere numerische Berechnungen wurde das Matlabsystem der Firma Mathworks[9] verwendet, als Computeralgebrasystem kam wxMaxima[10] zum Einsatz.

Genaue Angaben zu den verwendeten Daten, Verfahren und eingesetzten Werkzeugen der Teilprojekte entnehmen Sie bitte den einzelnen Detailkapiteln sowie den zugehörigen wissenschaftlichen Publikationen und Konferenzbeiträgen.

[1] http://www.fujitsu-siemens.de
[2] http://www.dell.de
[3] http://www.java.com
[4] http://www.borland.com
[5] http://www.gnu.org, http://www.fsf.org
[6] http://www.r-project.org
[7] http://www.bioconductor.org
[8] http://ess.r-project.org
[9] http://www.mathworks.de
[10] http://wxmaxima.sourceforge.net

Abkürzungsverzeichnis

BIC Bayesian Information Criterion

CA Correspondence Analysis

CAMA Correspondance Analysis on Multiple Sequence Alignments

CCA Canonical Correspondance Analysis

CCA Canonical Correspondence Analysis

CFS Correlation-based Feature Elimination

DUF Domain of Unknown Function

EMA Elementary Mode Analysis

FBA Flux Balance Analysis

GPCR G-Protein Coupled Receptor

HMM Hidden Markov Model

IDE Integrated Development Environment

ipHMM Interaction-Profile HMM

KGB Kegg Browser

MAMA Maximal Margin Linear Programming

MC Markov Chain

MDS Multi-Dimensional Scaling

MSA Multiples Sequenzalignment

MSSA Multiples Sequenzstrukturalignment

PCA Principal Component Analysis

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

RFE Recursive Feature Elimination

SBML Systems Biology Markup Language

SCFG Stochastic Contextfree Grammar

SDK Software Development Kit

SOAP Simple Object Access Protocol

SOM Self-Organizing Maps

SVD Singular Value Decomposition

SVM Support Vector Machine

TFBS Transcription Factor Binding Site

TMHMM Trans-Membrane HMM

WSDL Web Services Description Language

Literaturverzeichnis

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990): Basic local alignment search tool. *J Mol Biol*, **215** (3): 403–410, doi: 10.1006/jmbi.1990.9999.

Amato A, Orsini L, D'Alelio D and Montresor M (2005): Life cycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom Pseudo-nitzschia delicatissima (Bacillariophyceae). *Journal Of Phycology*, **41** (3): 542–556.

Antonov AV, Tetko IV, Mader MT, Budczies J and Mewes HW (2004): Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20** (5): 644–652, doi:10.1093/bioinformatics/btg462.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P *et al.* (2000): InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16** (12): 1145–1150.

Bairoch A (2000): The ENZYME database in 2000. *Nucleic Acids Res*, **28** (1): 304–305.

Barker WC, George DG, Mewes HW, Pfeiffer F and Tsugita A (1993): The PIR-International databases. *Nucleic Acids Res*, **21** (13): 3089–3092.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF *et al.* (2007): NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, **35** (Database issue): D760–D765, doi:10.1093/nar/gkl887.

Bartel B and Fink GR (1994): Differential regulation of an auxin-producing nitrilase gene family in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, **91** (14): 6649–6653.

Barthelmes J, Ebeling C, Chang A, Schomburg I and Schomburg D (2007): BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res*, **35** (Database issue): D511–D514, doi: 10.1093/nar/gkl972.

Becker D, Selbach M, Rollenhagen C, Ballmaier M, Meyer TF, Mann M and Bumann D (2006): Robust Salmonella metabolism limits possibilities for new antimicrobials. *Nature*, **440** (7082): 303–307, doi: 10.1038/nature04616.

Becker SA and Palsson B (2005): Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol*, **5** (1): 8, doi:10.1186/1471-2180-5-8.

Ben-Israel A (2002): The Moore of the Moore-Penrose Inverse. *J Linear Algebra*, **9**: 150–157.

Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004): Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340** (4): 783–795, doi:10.1016/j.jmb.2004.05.028.

Benjamini Y and Hochberg Y (2000): On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**: 60–83.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (2006): GenBank. *Nucleic Acids Res*, **34** (Database issue): D16–D20, doi: 10.1093/nar/gkj157.

Berger S, Papadopoulos M, Schreiber U, Kaiser W and Roitsch T (2004): Complex regulation of gene expression, photosynthesis and sugar levels by pathogen infection in tomato. *Physiologia Plantarum*, **122**: 419?–428.

Bettenbrock K, Fischer S, Kremling A, Jahreis K, Sauter T and Gilles ED (2006): A quantitative approach to catabolite repression in Escherichia coli. *J Biol Chem*, **281** (5): 2578–2584, doi:10.1074/jbc.M508090200.

Box GEP and Cox DR (1964): An Analysis Of Transformations. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, **26** (2): 211–252.

vom Brocke J (2003): *Referenzmodellierung - Gestaltung und Verteilung von Konstruktionsprozessen*. Ph.D. thesis, Westfalische Wilhelms-Universitat Munster, Institut fur Wirtschaftsinformatik.

Byrd RH, Lu P, Nocedal J and Zhu CY (1995): A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, **16** (6): 1190–1208.

Casari G, Sander C and Valencia A (1995): A method to predict functional residues in proteins. *Nat Struct Biol*, **2** (2): 171–8.

Cheong YH, Chang HS, Gupta R, Wang X, Zhu T and Luan S (2002): Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis. *Plant Physiol*, **129** (2): 661–677, doi:10.1104/pp.002857.

Choi JK, Yu U, Kim S and Yoo OJ (2003): Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**: 84–90.

Churchill GA (1992): Hidden Markov-Chains And The Analysis Of Genome Structure. *Computers & Chemistry*, **16** (2): 107–115.

Clamp M, Cuff J, Searle SM and Barton GJ (2004): The Jalview Java alignment editor. *Bioinformatics*, **20** (3): 426–427, doi:10.1093/bioinformatics/btg430.

Conlon EM, Song JJ and Liu JS (2006): Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, **7**: 247, doi:10.1186/1471-2105-7-247.

Covert MW and Palsson BO (2003): Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J Theor Biol*, **221** (3): 309–325.

Covert MW, Schilling CH and Palsson B (2001): Regulation of gene expression in flux balance models of metabolism. *J Theor Biol*, **213** (1): 73–88, doi:10.1006/jtbi.2001.2405.

Craigon DJ, James N, Okyere J, Higgins J, Jotham J and May S (2004): NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*, **32** (Database issue): D575–D577, doi:10.1093/nar/gkh133.

Dandekar T, Moldenhauer F, Bulik S, Bertram H and Schuster S (2003): A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems*, **70** (3): 255–270.

Dangl JL and Jones JD (2001): Plant pathogens and integrated defence responses to infection. *Nature*, **411** (6839): 826–833, doi:10.1038/35081161.

Dayhoff MO, Eck, Chang and Sochard (1965): *Atlas of Protein Sequence and Structure*. .

Dittrich P and di Fenizio PS (2007): Chemical Organization Theory. *Bulletin of Mathematical Biology*, **69** (4): 1199–1231, doi:10.1007/s11538-006-9130-8, in print.

Dixon RA (2001): Natural products and plant disease resistance. *Nature*, **411** (6839): 843–847, doi:10.1038/35081178.

Du L and Chen Z (2000): Identification of genes encoding receptor-like protein kinases as possible targets of pathogen- and salicylic acid-induced WRKY DNA-binding proteins in Arabidopsis. *Plant J*, **24** (6): 837–847.

Edwards JS, Ibarra RU and Palsson BO (2001): In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, **19** (2): 125–130, doi:10.1038/84379.

Edwards JS and Palsson BO (2000): Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics*, **1**: 1.

Eisenreich W, Slaghuis J, Laupitz R, Bussemer J, Stritzker J, Schwarz C, Schwarz R *et al.* (2006): 13C isotopologue perturbation studies of Listeria monocytogenes carbon metabolism and its modulation by the virulence regulator PrfA. *Proc Natl Acad Sci U S A*, **103** (7): 2040–2045, doi:10.1073/pnas.0507580103.

Engelmann JC, Schwarz R, Blenk S, Friedrich T, Seibel PN, Dandekar T and Müller T (2007): Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation. *Bioinformatics and Biology Insights (in submission)*.

Everitt B (2005): *An R and S-PLUS Companion to Multivariate Analysis*. Springer-Verlag London Limited.

Fehling J, Davidson K and Bates SS (2005): Growth dynamics of non-toxic Pseudo-nitzschia delicatissima and toxic P. seriata (Bacillariophyceae) under simulated spring and summer photoperiods. *Harmful Algae*, **4** (4): 763–769.

Felsenstein J (2005): PHYLIP (Phylogeny Inference Package) version 3.6, URL `http://evolution.gs.washington.edu/phylip.html`, distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferreira AEN, Freire AMJP and Voit EO (2003): A quantitative model of the generation of N(epsilon)-(carboxymethyl)lysine in the Maillard reaction between collagen and glucose. *Biochem J*, **376** (Pt 1): 109–121, doi:10.1042/BJ20030496.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S *et al.* (2006): Pfam: clans, web tools and services. *Nucleic Acids Res*, **34** (Database issue): D247–D251, doi:10.1093/nar/gkj149.

Finney A and Hucka M (2003): Systems biology markup language: Level 2 and beyond. *Biochem Soc Trans*, **31** (Pt 6): 1472–1473, doi:10.1042/.

Friedrich J, Dandekar T, Wolf M and Müller T (2005): ProfDist: a tool for the construction of large phylogenetic trees based on profile distances. *Bioinformatics*, **21** (9): 2108–2109, doi:10.1093/bioinformatics/bti289.

Friedrich T, Pils B, Dandekar T, Schultz J and Müller T (2006): Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics*, **22** (23): 2851–2857, doi: 10.1093/bioinformatics/btl486.

Gagneur J and Klamt S (2004): Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, **5**: 175, doi:10.1186/1471-2105-5-175.

Galtier N, Gouy M and Gautier C (1996): SEAVIEW and PHYLO-WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*, **12** (6): 543–8.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B *et al.* (2004): Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5** (10): R80, doi: 10.1186/gb-2004-5-10-r80.

George DG, Dodson RJ, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC *et al.* (1997): The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Res*, **25** (1): 24–28.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK *et al.* (2003): Global analysis of protein expression in yeast. *Nature*, **425** (6959): 737–741, doi:10.1038/nature02046.

Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P *et al.* (2001): Comparative genomics of Listeria species. *Science*, **294** (5543): 849–852, doi:10.1126/science.1063447.

Gorodkin J, Heyer LJ, Brunak S and Stormo GD (1997): Displaying the information contents of structural RNA alignments: the structure logos. *Comput Appl Biosci*, **13** (6): 583–586.

Grimmett G and Stirzaker D (2001): *Probability and Random Processes*. Oxford University Press.

Gutell RR, Lee JC and Cannone JJ (2002): The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, **12** (3): 301–310.

Guyon I, Weston J, Barnhill S and Vapnik V (2002): Gene selection for cancer classification using support vector machines. *Machine Learning*, **46** (1-3): 389–422.

Hall M (1999): *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, Hamilton NZ: Waikato University, Department of Computer Science.

Haussler D, Krogh A, Mian IS and Sjolander K (1993): Protein Modeling using Hidden Markov Models: Analysis of Globins. *In: Proceedings of the 26th Hawaii International Conference on System Sciences*, pp. 792–802, IEEE Computer Society Press, Honolulu.

Heiberger RM (2001): Emacs Speaks Statistics: One Interface — Many Programs. *DSC 2001 Proceedings of the 2nd International Workshop on Distributed Statistical Computing, Vienna*.

Herrgard MJ, Fong SS and Palsson B (2006): Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol*, **2** (7): e72, doi:10.1371/journal.pcbi.0020072.

Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M *et al.* (2006): COPASI–a COmplex PAthway SImulator. *Bioinformatics*, **22** (24): 3067–3074, doi:10.1093/bioinformatics/btl485.

Hu P, Greenwood CMT and Beyene J (2005): Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, **6**: 128, doi:10.1186/1471-2105-6-128.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP *et al.* (2003): The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19** (4): 524–531.

Hudelot C, Gowri-Shankar V, Jow H, Rattray M and Higgs PG (2003): RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol*, **28** (2): 241–252.

Huttenhower C, Hibbs M, Myers C and Troyanskaya OG (2006): A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22** (23): 2890–2897, doi: 10.1093/bioinformatics/btl492.

Jaakkola T, Diekhans M and Haussler D (2000): A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, **7** (1-2): 95–114.

Jewson DH (1992): Life-Cycle Of A Stephanodiscus Sp (Bacillariophyta). *Journal Of Phycology*, **28** (6): 856–866.

Jones NC and Pevzner PA (2004): *An Introduction to Bioinformatics Algorithms*. The MIT Press, Cambridge, Masachusetts.

von Kamp A and Schuster S (2006): Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22** (15): 1930–1931, doi: 10.1093/bioinformatics/btl267.

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004): The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32** (Database issue): D277–D280, doi:10.1093/nar/gkh063.

Karatzoglou A, Smola A, Hornik K and Zeileis A (2004): kernlab - An S4 package for kernel methods in R. *Research Report Series / Department of Statistics and Mathematics*, **9. August 2004**.

Karchin R, Karplus K and Haussler D (2002): Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18** (1): 147–159.

Karev GP, Wolf YI and Koonin EV (2003): Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, **19** (15): 1889–1900.

Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS and Koonin EV (2002): Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol*, **2** (1): 18.

Kauffman KJ, Prakash P and Edwards JS (2003): Advances in flux balance analysis. *Curr Opin Biotechnol*, **14** (5): 491–496.

Klamt S and Gilles ED (2004): Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20** (2): 226–234.

Klamt S and Stelling J (2003): Two approaches for metabolic pathway analysis? *Trends Biotechnol*, **21** (2): 64–69.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994): Hidden Markov-Models in Computational Biology, Applications To Protein Modeling. *Journal of Molecular Biology*, **235** (5): 1501–1531.

Krogh A, Larsson B, von Heijne G and Sonnhammer EL (2001): Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305** (3): 567–580, doi: 10.1006/jmbi.2000.4315.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F *et al.* (2007): ClustalW and ClustalX version 2.0. *Bioinformatics*, doi:10.1093/bioinformatics/btm404.

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J and Bork P (2006): SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, **34** (Database issue): D257–D260, doi:10.1093/nar/gkj079.

Levin AM, Ghosh D, Cho KR and Kardia SLR (2005): A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, **21** (12): 2867–2874, doi: 10.1093/bioinformatics/bti417.

Lindgren BW (1993): *Statistical Theory*. Chapman & Hall/CRC.

Mann DG (1993): Patterns of sexual reproduction in diatoms. *Hydrobiologia*, **269-270** (1): 11–20, doi:10.1007/BF00027999.

Mendes P (1993): GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci*, **9** (5): 563–571.

Mendes P (1997): Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci*, **22** (9): 361–363.

Mizuno M (1991): Influence Of Cell-Volume On The Growth And Size-Reduction Of Marine And Estuarine Diatoms. *Journal Of Phycology*, **27** (4): 473–478.

Moreau Y, Aerts S, Moor BD, Strooper BD and Dabrowski M (2003): Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*, **19** (10): 570–577.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P *et al.* (2007): New developments in the InterPro database. *Nucleic Acids Res*, **35** (Database issue): D224–D228, doi:10.1093/nar/gkl841.

Müller (2001): Modellgeschichte ist Kulturgeschichte. URL `http://www.muellerscience.com/ MODELL/Begriffsgeschichte/ ModellgeschichteistKulturgeschichte.htm`, [Online; Stand 23. Oktober 2007].

Müller R (1983): *Zur Geschichte des Modelldenkens und des Modellbegriffs*. aus Modelle - Konstruktion der Wirklichkeit, Herbert Stachowiak (Ed.), Wilhelm Fink Verlag.

Müller T, Philippi N, Dandekar T, Schultz J and Wolf M (2007): Distinguishing species. *RNA*, **13** (9): 1469–1472, doi:10.1261/rna.617107.

Needleman SB and Wunsch CD (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48** (3): 443–453.

Ng A, Jordan M and Weiss Y (2001): On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, **14**.

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997): Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, **10** (1): 1–6.

Nussinov R and Jacobson AB (1980): Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, **77** (11): 6309–6313.

Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus ADME and Fouchier RAM (2006): Global patterns of influenza a virus in wild birds. *Science*, **312** (5772): 384–388, doi:10.1126/science.1122438.

Papin JA, Stelling J, Price ND, Klamt S, Schuster S and Palsson BO (2004): Comparison of network-based pathway analysis methods. *Trends Biotechnol*, **22** (8): 400–405, doi:10.1016/j.tibtech.2004.06.010.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E *et al.* (2007): ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, **35** (Database issue): D747–D750, doi:10.1093/nar/gkl995.

Pauling L, Corey RB and Branson HR (1951): The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, **37** (4): 205–211.

Penrose R (1954): A generalized inverse for matrices. *Proc Camb Phil Soc*, **51**: 406–413.

Pfeiffer T, Snchez-Valdenebro I, Nuo JC, Montero F and Schuster S (1999): METATOOL: for studying metabolic networks. *Bioinformatics*, **15** (3): 251–257.

Piroux N, Saunders K, Page A and Stanley J (2007): Geminivirus pathogenicity protein C4 interacts with Arabidopsis thaliana shaggy-related protein kinase AtSKeta, a component of the brassinosteroid signalling pathway. *Virology*, **362** (2): 428–440, doi:10.1016/j.virol.2006.12.034.

Poolman MG, Venkatesh KV, Pidcock MK and Fell DA (2004): A method for the determination of flux in elementary modes, and its application to Lactobacillus rhamnosus. *Biotechnol Bioeng*, **88** (5): 601–612, doi:10.1002/bit.20273.

R Development Core Team (2006): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Ren Q, Chen K and Paulsen IT (2007): TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer

membrane channels. *Nucleic Acids Res*, **35** (Database issue): D274–D279, doi:10.1093/nar/gkl925.

Ren Q, Kang KH and Paulsen IT (2004): TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res*, **32** (Database issue): D284–D288, doi:10.1093/nar/gkh016.

Rockafellar R (1970): *Convex analysis*. Princeton University Press.

Round FE, Crawford RM and Mann DG (2007): *Diatoms: Biology and Morphology of the Genera*. Cambridge University Press.

Ryals JA, Neuenschwander UH, Willits MG, Molina A, Steiner HY and Hunt MD (1996): Systemic Acquired Resistance. *Plant Cell*, **8** (10): 1809–1819, doi:10.1105/tpc.8.10.1809.

Saha N, Watson LT, Kafadar K, Onufriev A, Ramakrishnan N, Vasquez-Robinet C and Watkinson J (2004): A general probabilistic model of the PCR process. *Conf Proc IEEE Eng Med Biol Soc*, **4**: 2813–2816, doi: 10.1109/IEMBS.2004.1403803.

Saha N, Watson LT, Kafadar K, Ramakrishnan N, Onufriev A, Mane S and Vasquez-Robinet C (2007): Validation and estimation of parameters for a general probabilistic model of the PCR process. *J Comput Biol*, **14** (1): 97–112, doi:10.1089/cmb.2006.0123.

Sanger F and Tuppy H (1951): The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J*, **49** (4): 481–490.

Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H, Doyle J and Kitano H (2003): Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS*, **7** (4): 355–372, doi: 10.1089/153623103322637670.

Schilling CH, Schuster S, Palsson BO and Heinrich R (1999): Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, **15** (3): 296–303, doi:10.1021/bp990048k.

Schmidt S, Sunyaev S, Bork P and Dandekar T (2003): Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci*, **28** (6): 336–341.

Schneider TD and Stephens RM (1990): Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18** (20): 6097–6100.

Schölkopf B, Smola A and Müller KR (1998): Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, **10** (5): 1299–1319, doi:10.1162/089976698300017467.

Schölkopf B and Smola AJ (2002): *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Schultz J, Maisel S, Gerlach D, Müller T and Wolf M (2005): A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, **11** (4): 361–364, doi:10.1261/rna.7204505.

Schultz J, Milpetz F, Bork P and Ponting CP (1998): SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, **95** (11): 5857–5864.

Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T and Wolf M (2006): The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res*, **34** (Web Server issue): W704–W707, doi:10.1093/nar/gkl129.

Schuster S, Dandekar T and Fell DA (1999): Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, **17** (2): 53–60.

Schuster S, Fell DA and Dandekar T (2000): A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, **18** (3): 326–332, doi:10.1038/73786.

Schuster S and Hilgetag C (1994): On elementary flux modes in biochemical systems at steady state. *Journal of Biological Systems*, **2**: 165–182.

Schuster S, Hilgetag C, Woods JH and Fell DA (2002a): Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol*, **45** (2): 153–181, doi:10.1007/s002850200143.

Schuster S, Pfeiffer T, Moldenhauer F, Koch I and Dandekar T (2002b): Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, **18** (2): 351–361.

Schuster-Böckler B, Schultz J and Rahmann S (2004): HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**: 7, doi:10.1186/1471-2105-5-7.

Schütte R (2001): *Grundsätze ordnungsmäßiger Referenzmodellierung*. Dr. Th. Gabler Verlag.

Schwartz JM and Kanehisa M (2005): A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics*, **21 Suppl 2**: ii204–ii205, doi:10.1093/bioinformatics/bti1132.

Schwartz JM and Kanehisa M (2006): Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*, **7**: 186, doi:10.1186/1471-2105-7-186.

Schwarz R, Liang C, Kaleta C, Kuhnel M, Hoffmann E, Kuznetsov S, Hecker M *et al.* (2007): Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics*, **8** (1): 313, doi: 10.1186/1471-2105-8-313.

Schwarz R, Musch P, von Kamp A, Engels B, Schirmer H, Schuster S and Dandekar T (2005): YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, **6**: 135, doi: 10.1186/1471-2105-6-135.

Schwarz R, Wolf M, Dandekar T and Müller T (2006): A probabilistic model of cell size reduction in *Pseudo-nitzschia delicatissima* (Bacillariophyta). Poster presentation at the 25th Annual Scientific Meeting of the „Deutsche Protozoologische Gesellschaft “, Liebenwalde, Berlin.

Schwarz R, Wolf M and Müller T (2008): A probabilistic model of cell size reduction in *Pseudo-nitzschia delicatissima* (Bacillariophyta). *(in preparation).*

Seibel PN, Müller T, Dandekar T, Schultz J and Wolf M (2006): 4SALE–a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, **7**: 498, doi:10.1186/1471-2105-7-498.

Selig C, Wolf M, Müller T, Dandekar T and Schultz J (2007): The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Res*, doi:10.1093/nar/gkm827.

Severini C, Silvestrini F, Mancini P, LaRosa G and Marinucci M (1996): Sequence and secondary structure of the rDNA second internal transcribed spacer in the sibling species Culex pipiens L and Cx-quinquefasciatus Say (Diptera: Culicidae). *Insect Molecular Biology*, **5** (3): 181–186.

Shawe-Taylor J and Cristianini N (2004): *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Smith TF and Waterman MS (1981): Identification of common molecular subsequences. *J Mol Biol*, **147** (1): 195–197.

Smyth GK (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3 (1)**: Article 3.

Snoeijs P, Busse S and Potapova M (2002): The importance of diatom cell size in community analysis. *Journal Of Phycology*, **38** (2): 265–272.

Sonnhammer EL, Eddy SR, Birney E, Bateman A and Durbin R (1998): Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, **26** (1): 320–322.

Stachowiak H (1965): *Denken und Erkennen im kybernetischen Modell*. Springer.

Stachowiak H (1973): *Allgemeine Modelltheorie*. Springer, Wien.

Stelling J, Klamt S, Bettenbrock K, Schuster S and Gilles ED (2002): Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420** (6912): 190–193, doi:10.1038/nature01166.

Stepanova AN, Hoyt JM, Hamilton AA and Alonso JM (2005): A Link between ethylene and auxin uncovered by the characterization of two root-specific ethylene-insensitive mutants in Arabidopsis. *Plant Cell*, **17** (8): 2230–2242, doi:10.1105/tpc.105.033365.

Stoff VA (1969): *Modellierung und Philosophie*. Akademie-Verlag.

Subbarao K and Luke C (2007): H5N1 viruses and vaccines. *PLoS Pathog*, **3** (3): e40, doi:10.1371/journal.ppat.0030040.

Suzuki R and Shimodaira H (2006): Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22** (12): 1540–1542, doi:10.1093/bioinformatics/btl117.

Swofford DL (2003): *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sinauer Associates Inc., Sunderland, Massachusetts.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES *et al.* (1999): Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, **96** (6): 2907–2912.

Thilmony R, Underwood W and He SY (2006): Genome-wide transcriptional analysis of the Arabidopsis thaliana interaction with the plant pathogen Pseudomonas syringae pv. tomato DC3000 and the human pathogen Escherichia coli O157:H7. *Plant J*, **46** (1): 34–53, doi:10.1111/j.1365-313X.2006.02725.x.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997): The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, **25** (24): 4876–82.

Thompson JD, Higgins DG and Gibson TJ (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22** (22): 4673–4680.

Thorburn WM (1918): The Myth of Occam's Razor. *Mind*, **27(107)**: 345–353.

Trawick JD and Schilling CH (2006): Use of constraint-based modeling for the prediction and validation of antimicrobial targets. *Biochem Pharmacol*, **71** (7): 1026–1035, doi:10.1016/j.bcp.2005.10.049.

Urbanczik R (2006): SNA–a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, **7**: 129, doi:10.1186/1471-2105-7-129.

Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J *et al.* (2005): Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol*, **138** (3): 1195–1204, doi:10.1104/pp.105.060459.

Vingron M (2001): Bioinformatics needs to adopt statistical thinking. *Bioinformatics*, **17** (5): 389–390.

Voit EO (2000): *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Pr.

Voit EO (2002): Metabolic modeling: a tool of drug discovery in the postgenomic era. *Drug Discov Today*, **7** (11): 621–628.

Wagner A and Fell DA (2001): The small world inside large metabolic networks. *Proc Biol Sci*, **268** (1478): 1803–1810, doi:10.1098/rspb.2001.1711.

Ward JH (1963): Hierarchical Grouping To Optimize An Objective Function. *Journal Of The American Statistical Association*, **58** (301): 236–&.

Wasserman L (2005): *All of Statistics*. Springer.

Weiss G and von Haeseler A (1995): Modeling the polymerase chain reaction. *J Comput Biol*, **2** (1): 49–61.

Weiss G and von Haeseler A (1997): A coalescent approach to the polymerase chain reaction. *Nucleic Acids Res*, **25** (15): 3082–3087.

Wikipedia (2007a): -omics — Wikipedia, The Free Encyclopedia. URL `http://en.wikipedia.org/w/index.php?title=-omics&oldid=167455054`, [Online; accessed 29-October-2007].

Wikipedia (2007b): Modell (Begriff) — Wikipedia, Die freie Enzyklopädie. URL http://de.wikipedia.org/w/index.php?title=Modell_%28Begriff%29&oldid=37325187, [Online; Stand 24. Oktober 2007].

Wikipedia (2007c): Palindrom — Wikipedia, Die freie Enzyklopädie. URL http://de.wikipedia.org/w/index.php?title=Palindrom&oldid=38122645, [Online; Stand 12. November 2007].

Wilhelm T, Behre J and Schuster S (2004): Analysis of structural robustness of metabolic networks. *Systems Biology*, **1** (1): 114–120.

Wolf M, Achtziger M, Schultz J, Dandekar T and Müller T (2005): Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, **11** (11): 1616–1623, doi: 10.1261/rna.2144205.

Wu J and with contributions from James MacDonald Jeff Gentry RI (2005): *gcrma: Background Adjustment Using Sequence Information*. R package version 2.2.1.

Wüstneck (1963): Zur philosophischen Verallgemeinerung und Bestimmung des Modellbegriffs. *Deutsche Zeitschrift für Philosophie*, **11**: 1504–1523.

Wüstneck KD (1966): *Methodologische und philosophische Probleme der Modelltheorie und ihrere Anwendung in den Gesellschaftswissenschaften*. Ph.D. thesis, Deutsche Akademie der Wissenschaften, Berlin.

Yeang CH and Vingron M (2006): A joint model of regulatory and metabolic networks. *BMC Bioinformatics*, **7** (1): 332, doi:10.1186/1471-2105-7-332.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA *et al.* (2007): The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol*, **5** (3): e16, doi: 10.1371/journal.pbio.0050016.

Zhang X, Lu X, Shi Q, Xu XQ, Leung HCE, Harris LN, Iglehart JD *et al.* (2006): Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**: 197, doi:10.1186/1471-2105-7-197.

Ziebuhr W, Xiao K, Coulibaly B, Schwarz R and Dandekar T (2004): Pharmacogenomic strategies against resistance development in microbial infections. *Pharmacogenomics*, **5** (4): 361–379, doi:10.1517/14622416.5.4.361.

Zuker M and Stiegler P (1981): Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, **9** (1): 133–148.

Danksagung

Zuvorderst möchte ich meinem Doktorvater, Prof. Thomas Dandekar, danken, dass er mir die Chance gegeben hat, diese Doktorarbeit anzufertigen, mich geduldig in die Materie der metabolischen Netzwerke eingeführt und mir stets seine wertvolle Unterstützung angeboten hat. Auch möchte ich ihm für den großen Freiraum danken, den er mir bei der Bearbeitung dieser Dissertation zugestanden hat. Ohne ihn wäre all das nicht möglich gewesen. Besonders großer Dank gilt auch Dr. Tobias Müller, der letztlich mein Interesse an der Mathematik wieder geweckt hat, und mir zu jeder Zeit ein ebenso verläßlicher Mentor wie erbarmungsloser Badmintonpartner gewesen ist. Ich möchte des Weiteren Dr. Matthias Wolf danken für seine unzähligen Erläuterungen zur Phylogenie und seine ungebrochene Begeisterung für Algen. Prof. Jörg Schultz danke ich für sein stets offenes Ohr für meine biologischen Fragen, für seine unermüdlichen Versuche, mir die Wunder von PERL nahe zu bringen, und für seine tiefgreifenden Einblicke in das Agressionsverhalten domestizierter Wölfe. Dank geht auch an meine Kollegen, die an den hier genannten Projekten mitgewirkt haben, im Speziellen Thorben Friedrich, Philipp Seibel, Steffen Blenk, Chunguang Liang, und Julia Engelmann. Letzterer danke ich insbesondere für ihre Unterstützung und das tolle Bibtex Stylefile, sowie darüber hinaus all jenen, die an den regelmässigen Mittagsseminaren zum Thema „Mensch-Maschine-Interaktionen in 3D Umgebungen und die Auswirkungen schneller Bildabfolgen auf lokalisierte postsynaptische Potentialschwankungen des visuellen Cortex“ teilgenommen haben. Dank gilt auch meinem Vater, Prof. Wilhelm Schwarz, für das gewissenhafte Korrekturlesen dieser Arbeit sowie konstruktiven Kommentaren und Inspiration, insbesondere zu Einleitung und Fazit. Ebenso möchte ich meiner Mutter danken, für ihr Verständnis und ihre Fürsorge. Zum Schluß, aber nicht zuletzt gilt mein ganzer Dank Anna Steidle, die den Glauben an mich nie verloren hat, immer für mich da war und geduldig mit fast schon stoischer Ruhe alle meine Höhen und Tiefen ertragen hat. Dir, liebe Anna, widme ich diese Arbeit.

Erklärungen

Hiermit erkläre ich ehrenwörtlich, dass ich die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Die Dissertation wurde bisher weder in gleicher noch ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt. Außer dem Diplom in Informatik der Fachhochschlue Würzburg habe ich bisher keine weiteren akademischen Grade erworben oder versucht zu erwerben.

Würzburg, 30. Mai 2008

Roland Schwarz

Schriftenverzeichnis

Publikationen

W. Ziebuhr, K. Xiao, B. Coulibaly, R. Schwarz, and T. Dandekar. Pharmacogenomic strategies against resistance development in microbial infections. *Pharmacogenomics*, 5(4):361–379, Jun 2004.

R. Schwarz, P. Musch, A. von Kamp, B. Engels, H. Schirmer, S. Schuster, and T. Dandekar. YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6:135, 2005.

W. Eisenreich, J. Slaghuis, R. Laupitz, J. Bussemer, J. Stritzker, C. Schwarz, R. Schwarz, T. Dandekar, W. Goebel, and A. Bacher. 13C isotopologue perturbation studies of *Listeria monocytogenes* carbon metabolism and its modulation by the virulence regulator prfA. *Proc Natl Acad Sci U S A*, 103(7):2040–2045, Feb 2006.

R. Schwarz, C. Liang, C. Kaleta, M. Kuhnel, E. Hoffmann, S. Kuznetsov, M. Hecker, G. Griffiths, S. Schuster, and T. Dandekar. Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics*, 8(1):313, Aug 2007.

J. C. Engelmann*, R. Schwarz*, S. Blenk, T. Friedrich, P. N. Seibel, T. Dandekar, and T. Müller. Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation. *Bioinformatics and Biology Insights (submitted)*, 2008.
* *both authors contributed equally*

R. Schwarz, M. Wolf, T. Müller. A probabilistic model of cell size reduction in *Pseudo-nitzschia delicatissima* (Bacillariophyta). *(in preparation)*, 2008.

Konferenzbeiträge

R. Schwarz, C. Liang, R. Lampidis, and T. Dandekar. Bioinformatical identification, analysis and prediction of pathogen specific protein ad-

aptations using *L. monocytogenes* as an example. Poster presentation at the 2nd European Conference on Prokaryotic Genomes (ProkaGen), 2005.

R. Schwarz, M. Wolf, T. Dandekar, and T. Müller. A probabilistic model of cell size reduction in *Pseudo-nitzschia delicatissima* (bacillariophyta). Poster presentation at the 25th Annual Scientific Meeting of the „Deutsche Protozoologische Gesellschaft“, Liebenwalde, Berlin, 2006.

J. C. Engelmann, R. Schwarz, S. Blenk, T. Friedrich, P. N. Seibel, T. Dandekar, and T. Müller. Large-scale kernel-based explorative meta-analysis of affymetrix genome arrays. Poster presentation at ISMB / ECCB, 2007.