1 Einleitung

1.1 Bioinformatik

Die Bioinformatik ist ein sich rasant entwickelndes Teilgebiet der Biologie. In ihr vereinen sich sowohl Methoden der Mathematik, als auch der elektronischen Datenverarbeitung, um biologische Fragestellungen zu lösen. Zu den Themengebieten der Bioinformatik gehören die Sequenzanalyse von DNA und Proteinen, die Genomannotation, die Analyse von Genexpressionsdaten, die Beschreibung von metabolischen und Regulationsnetzwerken, vergleichende Genomanalysen und die Vorhersage von Proteinstrukturen. Zur Bearbeitung der spezifischen Fragestellung

Tabelle 1.1: Spezialisierte biologische Datenbanken.

Datenbank	Schwerpunkt	Internetadresse
BRENDA	Enzymdaten	`http://www.brenda-enzymes.info`
EMBL	Nukleotidsequenzen	`http://www.ebi.ac.uk/embl`
KEGG	Metabolische Daten	`http://www.genome.jp/kegg`
PDB	Proteinstrukturen	`http://www.pdb.org`
Prodoric	Transkriptionsfaktoren und regulatorische Daten von Prokaryonten	`http://www.prodoric.de`
Swiss-Prot	Proteinsequenzen	`http://expasy.org/sprot`

stehen eine Vielzahl von Techniken und Konzepten der angewandten Mathematik und Informatik zur Verfügung. So werden beispielsweise „Hidden Markov Modelle (HMM)“ in der Genomannotation eingesetzt um kodierende Bereiche von nicht kodierenden Bereichen der DNA zu unterscheiden [53]. „Supported Vector Machines (SVM)“ werden verwendet um Microarraydaten zu analysieren [26]. Auch Algorithmen der künstlichen Intelligenz wie zum Beispiel „künstliche neuronale Netze (KNN)“ finden in der Bioinformatik Verwendung, um zum Beispiel Promotorsequenzen von bakteriellen Genomen vorherzusagen [40]. Algorithmen sind zum

Teil für konkrete Fragestellungen der Biologie entwickelt worden. So versteht man unter dem Prozess der Alignmenterstellung das Arrangieren von DNA, RNA oder Proteinsequenzen in der Art, dass verwandtschaftliche, strukturelle oder funktionelle Beziehungen zwischen den einzelnen Sequenzen abgeleitet werden können. Man unterscheidet zwischen globalen und lokalen, sowie zwischen paarweisen und multiplen Sequenzalignments. Die wichtigsten Algorithmen, die für dieses komplexe Problem entwickelt wurden sind: Needleman-Wunsch-Algorithmus (globales Alignment) [57], Smith-Waterman-Algorithmus (lokales Alignment) [70], FASTA (heuristisches Verfahren für ein paarweises multiples Alignment) [49] und BLAST (Sammlung von Programmen für heuristische paarweise multiple Alignments) [2]. Neben der Gewinnung neuer Information aus experimentellen Daten, stellt die Speicherung eben dieser experimentellen Daten und der gewonnenen Informationen einen weiteren wichtigen Bereich der Bioinformatik da. Die explosionsartige Entwicklung des Internets hat gerade in dieser Disziplin entscheidenden Einfluss gehabt. Techniken zur Speicherung und Darstellung von Informationen stehen dadurch in einer breiten Fülle zur Verfügung. So werden die Daten in Datenbanken gespeichert und mittels Internet anderen Wissenschaftlern zur Verfügung gestellt. Die meisten biologischen Datenbanken sind dabei auf spezifische Fragestellungen

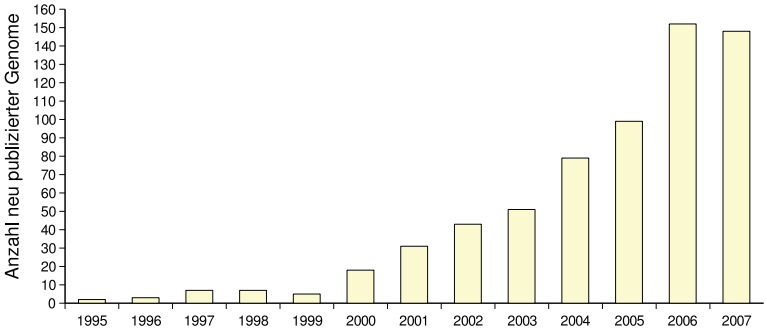

Abbildung 1.1: Dargestellt ist die Anzahl neu publizierter Genome pro Jahr für die letzten 12 Jahre (Quelle: `http://www.genomesonline.org`).

spezialisiert. In Tabelle 1.1 sind einige wichtige Datenbanken mit dem jeweiligen Schwerpunkt angegeben.

Die Grundlage vieler biologischer Datenbanken bilden Nukleotidsequenzen aus Genomprojekten. Die Anzahl publizierter Genome nimmt stetig zu. In Abbildung 1.1 ist diese Entwicklung graphisch dargestellt. Da die Kosten für die Sequenzierung kompletter Genome immer niedriger werden, wird dieser Trend vermutlich noch an Geschwindigkeit gewinnen. Aufgabe der Bioinformatik ist es hier geeignete Werkzeuge zur Verfügung zu stellen, die eine präzise Annotation der neuen Genome ermöglichen und die Ergebnisse Benutzern zugänglich zu machen.

1.2 Systembiologie

Die Systembiologie ist eng mit der Bioinformatik verbunden, da die Systembiologie, mehr als andere Zweige der Biologie, auf Werkzeuge der Bioinformatik angewiesen ist, um neue Gesetzmäßigkeiten zu entdecken. Ziel der Systembiologie ist es einen Organismus in seiner Gesamtheit zu verstehen und alle untersuchbaren Ebenen in ein Modell zu integrieren, das Vorhersagen über biologische Abläufe erlaubt. Eine herausragende Rolle für die Systembiologie spielen dabei die sogenannten ‚Omics'-Techniken, die eine Analyse der Gesamtheit einer Regulationsebene eines Organismus oder einer Zelle ermöglichen [45]. Für die Transkriptionsebene ermöglichen „Microarrays" einen Blick auf die RNA, die in der betrachteten Zelle zu einem bestimmten Zeitpunkt vorhanden ist. Die Proteomebene wird beispielsweise mit Hilfe von „Zwei-Dimensionalen-Protein-Gelen" oder „Massenspektrometrie" untersucht. Für die Analyse der Metabolite (Metabolomics) in einer Zelle stehen zum Beispiel die „Gaschromtatografie mit Massenspektrometrie (GC/MS)" und „Flüssigchromatographie mit Massenspektrometrie (LC/MS)" zur Verfügung.

Bei der Verarbeitung der dabei entstandenen Datenmengen sind geeignete Algorithmen und Computerprogramme, die beispielsweise Transkriptomdaten gruppieren [7], Unterschiede bei Proteomdaten visualisieren [10] oder Metabolomdaten [43] auswerten, ein unentbehrliches Hilfsmittel geworden.

Ein wichtiges Ziel der Systembiologie ist die Bildung von Modellen für komple-

xe biologische Prozesse. Für die einzelnen Ebenen stehen dafür verschiedene Konzepte zur Verfügung. Das Konzept der „Flux balance analysis“ zum Beispiel kann eingesetzt werden, um den Metabolismus eines Bakteriums quantitativ zu simulieren [42]. Für Transkriptomanalysen stehen verschiedene „Clusteralgorithmen“ zur Verfügung, die Gene nach gleichen Expressionsprofilen einteilen und so Hinweise darauf liefern können, welche Gene einem gemeinsamen Regulon angehören [23].

Die Integration der einzelnen Regulationsebenen zu einem Modell der Zelle oder des Organismus ist die Herausforderung, der sich die Systembiologie in den nächsten Jahren stellen muss [54].

1.3 Der Sonderforschungsbereich 578: Vom Gen zum Produkt

Der Sonderforschungsbereich 578 ist ein Zusammenschluss verschiedener Institute der Technischen Universität Braunschweig, dem Helmholtz Zentrum für Infektionsforschung und dem Max-Planck-Institut Magdeburg. Der Schwerpunkt des SFB 578 liegt auf der Erforschung der rekombinanten Produktion biotechnologisch oder pharmazeutisch interessanter Proteine mittels gentechnisch veränderter Organismen. Als Wirtssystem für die rekombinante Proteinproduktion dienen dabei *Bacillus megaterium* als Gram positives prokaryotisches System und *Aspergillus niger* als eukaryotisches System.

Innerhalb des SFB werden Forschungsansätze aus verschiedenen wissenschaftlichen Disziplinen gebündelt, um ein möglichst umfassendes Modell des Produktionsprozesses zu erhalten. Der Produktionsprozess soll somit in seiner Gesamtheit verstanden werden und so optimiert werden. Die Forschungsansätze kommen aus den Bereichen der Molekularbiologie, der Mikrobiologie, der technischen Chemie, der Biotechnologie, der pharmazeutischen Technologie, der Bioverfahrenstechnik, der mechanischen, chemischen und thermischen Verfahrenstechnik, der Bioinformatik, der elektrischen Messtechnik und der Mikrotechnik. Um dem interdisziplinären Forschungsansatz gerecht zu werden, unterteilt sich der SFB in vier Projektbereiche, die sich weiter in insgesamt 17 Teilprojekte aufschlüsseln. Die vier Projektbereiche sind Projektbereich A: Molekularbiologie der Produktbildung, Projektbe-

reich B: Systembiotechnologie der Produktbildung, Projektbereich C: Prozesstechnik und Projektbereich D: Anwendungstechnik [32].

1.3.1 Das Teilprojekt B4: Systembiologie der Produkt- und Pelletbildung durch *Aspergillus niger*

Das Teilprojekt B4 mit dem Titel „Systembiologie der Produkt- und Pelletbildung durch *Aspergillus niger*“ beschäftigt sich mit der Herstellung rekombinanter Proteine durch *A. niger*. Dabei wird den Kultivierungsbedingungen, wie zum Beispiel

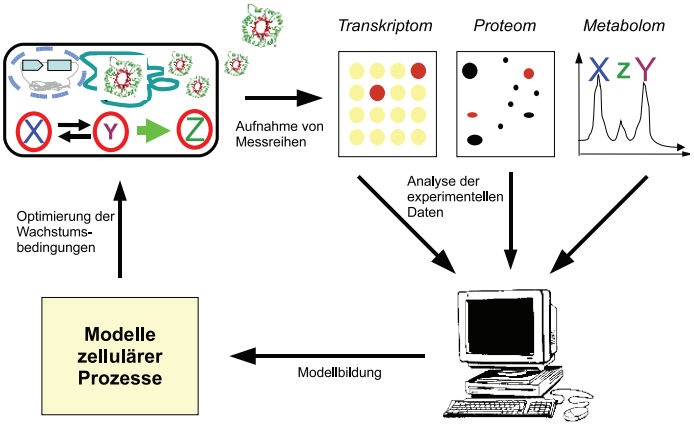

Abbildung 1.2: Dargestellt ist die Kombination aus Tanskriptom-, Proteom- und Metabolomdaten, die mittels bioinformatischer Methoden Vorhersagen über das biologische System erlaubt.

Temperatur, pH-Wert, Sauerstoffpartialdruck und Kohlenstoffquelle besondere Aufmerksamkeit geschenkt, da diese das Wachstum unmittelbar kontrollieren und damit einen erheblichen Einfluss auf die Proteinproduktion ausüben.

Neben der Klärung der Zusammenhänge zwischen Wachstumsbedingungen, Zellmorphologie und Produktivität wird im Teilprojekt B4 ein systembiologischer Ansatz verfolgt, der mit Hilfe der verschiedenen „Omics-Techniken“ die Parameter bestimmt, die für die Produktion entscheidend sind. Zu diesem Zweck werden Un-

tersuchungen des Transkriptoms, Proteoms und Metaboloms mit Techniken der Bioinformatik kombiniert, um die Daten effizient auszuwerten und schließlich Modelle zellulärer Teilprozesse zu generieren.

In einem iterativen Prozess soll das Modell letztendlich mit den experimentellen Daten angepasst und optimiert werden, um Aussagen über Produktionsengstellen treffen zu können, die dann aufgelöst werden können. In Abbildung 1.2 ist dieser Prozess schematisch dargestellt[1,2].

1.4 Die Schimmelpilzgattung *Aspergillus*

Die Gattung *Aspergillus* gehört zur Klasse der „Echten Schlauchpilze (Ascomycetes)“. Ihnen gemein ist die Ausbildung eines schlauchähnlichen Gebildes (Ascus), in dem die Ascusspore gebildet wird. Die Ascusspore stellt das Endprodukt der sexuellen Fortpflanzung der Ascomyceten dar. Die Ascomyceten zeichnen sich außerdem durch ein septiertes Myzel aus [66]. Die Gattung der Aspergilli umfasst sowohl Arten, die saprotroph leben, als auch opportunistische Krankheitserreger. Alle Krankheiten, an denen *Aspergillus*-Arten beteiligt sind, werden als Aspergillose bezeichnet. Im Allgemeinen versteht man darunter seltene Infektionen durch *Aspergillus*, die sich meist in der Lunge oder den Ohren manifestieren. Meist sind daher auch Symptome am Atmungsapparat festzustellen, jedoch können sich auch Läsionen an Leber, Darm oder anderen Stellen des Körpers bilden [69].

Aspergilli sind in der Natur weit verbreitet und kommen im häuslichen Bereich auf altem Brot, Käse oder Obst vor. Sie wachsen als filamentöse Pilze. Jedes Filament wächst hauptsächlich an der Spitze durch Verlängerung der terminalen Zelle. Die einzelnen Filamente bezeichnet man als Hyphe. Durch multiples Verzweigen der Hyphen und dadurch, dass sich auf diese Weise viele verschiedene Hyphen miteinander verweben, entsteht ein Teppich aus Pilzfäden, den man als Myzel bezeichnet [56]. Verschiedene *Aspergillus*-Arten finden in der Biotechnologie Verwendung, um zum Beispiel Zitronensäure, Sojasoße oder Essig herzustellen. Auch in der genetischen Forschung gehören verschiedene *Aspergillus*-Arten mittlerweile zu den

[1] `http://sfb578.tu-braunschweig.de/seiten/tp/tpb4.html`
[2] `http://www.tu-braunschweig.de/ibvt/forschung/projekte/sfb578-b4`

etablierten Modellorganismen [1].

1.4.1 *Aspergillus niger*

Aspergillus niger hat seinen Namen von der der Farbe seiner Konnidien, der asexuellen Vermehrungsform der Ascomyceten, die im Allgemeinen tiefschwarz sind. Unter Kupfermangel jedoch verfärben sich diese Sporen zuweilen auch schon mal gelb [69]. Der filamentöse Pilz findet in der Biotechnologie viele Anwendungen. So wird *A. niger* beispielsweise eingesetzt, um Zitronensäure für die Lebensmittelindustrie oder Pharmabranche herzustellen. Außerdem wird mit Hilfe von *A. niger* Gluconsäure produziert, die als Träger von Kalzium oder Natrium ebenfalls in der Lebensmittelindustrie Verwendung findet.

Das Genom von *A. niger* wird auf eine Größe zwischen 35,5 und 38,5 Megabasenpaare geschätzt, die sich auf 8 Chromosomen verteilen [6]. Der GC-Gehalt wird mit 52% angegeben [47].

A. niger ist ein Bodenbewohner und hat als saprotroph lebender Organismus Anteil am globalen Kreislauf des Kohlenstoffs. Als Modellorganismus dient er unter anderem zur Untersuchung der eukaryotischen Proteinsekretion, des Einflusses verschiedener Umweltfaktoren auf das Ausscheiden biomasseabbauender Enzyme, der entscheidenden molekularen Mechanismen zur Entwicklung von Fermentationsprozessen und Regulationsstrukturen der Pilzmorphologie [6].

1.4.2 Weitere *Aspergillus*-Arten

Die Gattung *Aspergillus* umfasst mehr als 185 Arten [27]. Da die einzelnen Arten zum Teil eine große Bedeutung in der Biotechnologie oder als Krankheitserreger haben, wurden inzwischen die Genome von mehreren Arten aufgeklärt. Einige dieser *Aspergillus*-Arten werden im Folgenden kurz vorgestellt:

Aspergillus nidulans

Aspergillus nidulans hat große Bedeutung als Modellorganismus in der Genetik erlangt. Der Grund dafür beruht unter anderem darauf, dass er im Gegensatz zu vielen anderen *Aspergillus*-Arten einen gut untersuchten sexuellen Fortpflanzungszyklus

besitzt. Das Genom von *A. nidulans* ist ca. 30 Megabasenpaare groß und besitzt ca. 10000 proteinkodierende Sequenzen auf 8 Chromosomen. Der GC-Gehalt beträgt 50% [27].

Aspergillus fumigatus

Aspergillus fumigatus kann Allergien auslösen, als opportunistischer Krankheitserreger wirken oder auch primärer Krankheitsauslöser sein. Der Pilz ist sehr weit verbreitet und man findet ihn beispielsweise häufig in Häusern und Wohnungen aber auch in Komposthaufen. Das Genom von *A. fumigatus* ist 29,4 Megabasenpaare groß und auf 8 Chromosomen verteilt. Der GC-Gehalt beträgt 49,9% [58].

Aspergillus oryzae

Aspergillus oryzae hat besonders in Japan große wirtschaftliche Bedeutung, da er zum Beispiel eingesetzt wird, um Reis zu Wein zu fermentieren (Sake) oder Sojasoße herzustellen. Durch seine Fähigkeit große Mengen Enzym, wie zum Beispiel Amylasen und Proteasen, zu sekretieren, spielt er auch bei der heterologen Proteinproduktion eine Rolle [46]. Das Genom von *A. oryzae* ist 37,2 Megabasenpaare groß und ebenfalls auf 8 Chromosomen verteilt. Der GC-Gehalt beträgt 48,2% [51].

1.5 Heterologe Proteinexpression

Unter heterologer Proteinexpression versteht man die Übertragung eines Fremdgens in einen Wirtsorganismus und die Expression des Gens, mit der Absicht, dass das entsprechende Protein von dem Wirtsorganismus synthetisiert wird. Diese Technik ist in der Biotechnologie weit verbreitet und bildet die Grundlage für die Herstellung vieler Stoffe für die Pharma- und Lebensmittelindustrie. So beruhen beispielsweise biotechnologische Verfahren zur Herstellung von Insulin, vieler Antibiotika, Ascorbinsäure, Chymosin (Enzym des Labferments), etc. auf dieser Technik.

Um einen Wirtsorganismus zur Produktion eines für ihn fremden Proteins zu bewegen, sind mehrere Schritte erforderlich. Zunächst muss das Zielgen aus dem Organismus isoliert werden. Hat man das Zielgen isoliert wird der Wirtsorganismus

mittels eines geeigneten Vektors mit dem Zielgen transformiert, so dass die Gensequenz nun zum Ablesen im Wirt vorliegt. Anschließend sollte der Wirt das gewünschte Protein synthetisieren und gegebenenfalls ins Medium sekretieren. Handelt es sich um einen Wirt, der das Produkt nicht ins Medium sekretiert, schließen sich noch ein Zellaufschluss und ein Reinigungsschritt an (siehe Abbildung 1.3).

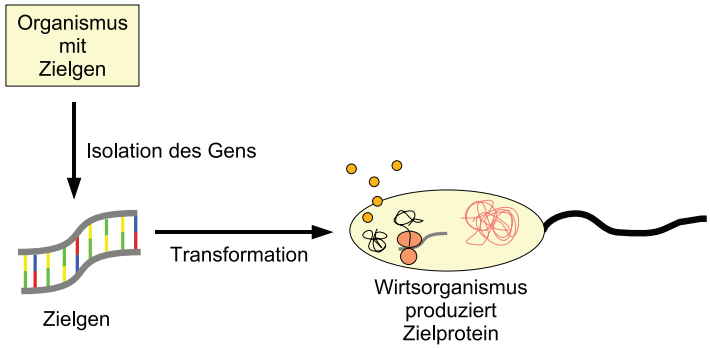

Abbildung 1.3: Schematische Darstellung der heterologen Proteinproduktion. Nachdem das Zielgen aus dem Organismus isoliert wurde, kann der Wirtsorganismus damit transformiert werden. Einige Wirtsorganismen sekretieren das gewünschte Zielprodukt direkt ins Medium. Andernfalls schließt sich ein Zellaufschluss und ein Reinigungsschritt an.

Bei dieser stark vereinfachten Beschreibung heterologer Proteinproduktion können eine Reihe von Problemen auftreten. Hat man erstmal das Gen isoliert und konnte die Sequenz stromabwärts eines geeigneten Promotors kloniert werden, wird der Wirtsorganismus mit dem Zielgen transformiert. Wenn dieser Schritt erfolgreich verläuft, ist das leider noch keine Garantie dafür, dass das Zielprodukt synthetisiert wird. Ein wichtiger Punkt der berücksichtigt werden muss ist, dass das Gen außerhalb seines natürlichen genomischen Kontextes steht. Auf die in Abbildung 1.3 dargestellte Vorgehensweise werden beispielsweise regulatorische Elemente, die sich stromaufwärts oder stromabwärts des betrachteten Gens befinden, gänzlich außer Acht gelassen. Weiterhin muss man berücksichtigen, dass der Wirtsorganismus und der Organismus, aus dem das Gen stammt, unter Umständen unterschiedliche regulatorische Elemente bevorzugen. Letztlich kann sich auch der verwendete geneti-

sche Code in einigen Punkten unterscheiden [31].

Auf eine besondere Art der Regulation innerhalb der Sequenz und die damit verbundenen Schwierigkeiten der heterologen Proteinexpression, wird im Folgenden eingegangen: die „codon usage“.

1.5.1 Kodonnutzung und heterologe Proteinproduktion

Der genetische Code ist degeneriert, das bedeutet, dass bis zu sechs Kodons für die gleiche Aminosäure kodieren können. Abhängig von dem jeweiligen Organismus, werden die Kodons unterschiedlich häufig für die betrachtete Aminosäure eingesetzt [36, 37]. Beispielsweise verwenden GC-reiche Organismen mehr Kodons die ‚G‘ und ‚C‘ enthalten als Kodons die ‚A‘ und ‚T‘ enthalten. Aus dieser Bevorzugung bestimmter Kodons ergibt sich, dass es optimale und suboptimale Kodons für jeden Organismus gibt. Die „codon usage“ verschiedener Gene eines Organismus hat großen Einfluss auf die Expressivität des Gens [30]. Die Bevorzugung bestimmter Kodons ist nicht universell gültig, sondern für jeden Organismus einzigartig. Der Grund für dieses Ungleichgewicht der einzelnen Kodons beruht auf der unterschiedlichen Anzahl der entsprechenden tRNAs in den verschiedenen Organismen.

Die „codon usage“ spielt somit eine entscheidende Rolle bei der heterologen Proteinexpression. Kodons im Zielgen, die in dem Wirtsorganismus selten eingesetzt werden, können zu einer geringen Translationsrate der mRNA, einer verringerten mRNA-Stabilität und manchmal so gar zu einem verfrühten Abbruch der Translation führen [72, 29]. In *Escherichia coli* wurde festgestellt, dass falsche Aminosäuren eingebaut werden können, wenn seltene Kodons für die Aminosäure Arginin verwendet werden [15].

Prinzipiell gibt es zwei mögliche Lösungen für dieses Problem. Die erste Möglichkeit ist es dem Wirtsorganismus die entsprechenden tRNAs zuzuführen, die andernfalls nur in unzureichenden Mengen vorliegen [14]. Die andere Möglichkeit, auf die hier näher eingegangen wird, besteht darin die Gensequenz „de novo“ zu synthetisieren und damit den Gegebenheiten des Wirtsorganismus anzupassen.