INTEGRATION UND KOMBINATION BIOINFORMATISCHER METHODEN IN BIOTECHNOLOGIE, SYNTHETISCHER BIOLOGIE UND PHARMAINDUSTRIE

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

vorgelegt von
Dipl. Bioinf. (FH) Beate Krüger
aus Schwalbach am Taunus

Würzburg, Februar 2012

Eingereicht am: .....

Mitglieder der Promotionskommission:

Vorsitzender: .....

1. Gutachter : Professor Dr. Thomas Dandekar

2. Gutachter: Professor Dr. Markus Engstler

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

ERKLÄRUNG

Hiermit erkläre ich ehrenwörtlich, dass ich die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegeben Quellen und Hilfsmittel verwendet habe.
Die Dissertation wurde bisher weder in gleicher noch ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt.

Würzburg, Februar 2012 **Beate Krüger**

DANKSAGUNG

Für meine Doktorarbeit schulde ich sehr vielen Menschen einen herzlichen Dank.

Besonders möchte ich mich bei meinem Doktorvater Herr Professor Thomas Dandekar bedanken. Er hat mich trotz meines FH-Abschlusses und trotz meiner Arbeit bei Boehringer-Ingelheim als Doktoranten bei sich aufgenommen und mir dabei sehr viel Geduld und wertvollen Ratschlägen entgegengebracht hat.

Weiterhin möchte ich mich bei der Firma Boehringer-Ingelheim GmbH und Co KG für ihre flexiblen Arbeitszeiten bedanken, die mir das Arbeiten sehr erleichtert hat.

Ich danke außerdem meinem Freund, der mir stets Mut zugesprochen und mich in meiner Arbeit bestärkt hat.

Und nicht zuletzt danke ich meinen Eltern, die in jeglicher Hinsicht die Grundsteine für meinen Weg gelegt haben.

INHALTSVERZEICHNIS

ZUSAMMENFASSUNG

Zusammenfassung

Die Bioinformatik ist eine interdisziplinäre Wissenschaft, welche Probleme aus allen Lebenswissenschaften mit Hilfe computergestützter Methoden bearbeitet.
Ihr Ziel ist es, die Verarbeitung und Interpretation großer Datenmengen zu ermöglichen. Zudem unterstützt sie den Designprozess von Experimenten in der synthetischen Biologie. Die synthetische Biologie beschäftigt sich mit der Generierung neuer Komponenten und deren Eigenschaften, welche durch die Behandlung und Manipulation lebender Organismen oder Teilen daraus entstehen. Ein besonders interessantes Themengebiet hierbei sind Zweikomponenten-Systeme (Two-Component System, TCS). TCS sind wichtige Signalkaskaden in Bakterien, welche in der Lage sind Informationen aus der Umgebung in eine Zelle zu übertragen und darauf zu reagieren.
Die vorliegende Dissertation beschäftigt sich mit der Beurteilung, Nutzung und Weiterentwicklung von bioinformatischen Methoden zur Untersuchung von Proteininteraktionen und biologischen Systemen. Der wissenschaftliche Beitrag der vorliegenden Arbeit kann in drei Aspekte unterteilt werden:

- Untersuchung und Beurteilung von bioinformatischen Methoden und Weiterführung der Ergebnisse aus der vorhergehenden Diplomarbeit zum Thema Protein-Protein-Interaktionsvorhersagen.
- Analyse genereller evolutionärer Modifikationsmöglichkeiten von TCS sowie deren Design und spezifische Unterschiede.
- Abstraktion bzw. Transfer der gewonnenen Erkenntnisse auf technische und biologische Zusammenhänge. Mit dem Ziel das Design neuer Experimente in der synthetischen Biologie zu vereinfachen und die Vergleichbarkeit von technischen und biologischen Prozessen sowie zwischen Organismen zu ermöglichen.

Das Ergebnis der durchgeführten Studie zeigte, dass Zweikomponenten-Systeme in ihrem Aufbau sehr konserviert sind. Nichtsdestotrotz konnten viele spezifische Eigenschaften und drei generelle Modifikationsmöglichkeiten entdeckt werden. Die Untersuchungen ermöglichten die Identifikation neuer Promotorstellen, erlaubten aber

auch die Beschreibung der Beschaffenheit unterschiedlicher Signalbindestellen. Zudem konnten bisher fehlende Komponenten aus TCS entdeckt werden, ebenso wie neue divergierte TCS-Domänen im Organismus *Mycoplasma*.
Eine Kombination aus technischen Ansätzen und synthetischer Biologie vereinfachte die gezielte Manipulation von TCS oder anderen modularen Systemen. Die Etablierung der vorgestellten zweistufigen Modul-Klassifikation ermöglichte eine effizientere Analyse modular aufgebauter Prozesse und erlaubte somit das molekulare Design synthetischer, biologischer Anwendungen. Zur einfachen Nutzung dieses Ansatzes wurde eine frei zugängliche Software GoSynthetic entwickelt. Konkrete Beispiele demonstrierten die praktische Anwendbarkeit dieser Analysesoftware. Die vorgestellte Klassifikation der synthetisch-biologischen und technischen Einheiten soll die Planung zukünftiger Designexperimente vereinfachen und neue Wege für sinnverwandte Bereiche aufzeigen.

Es ist nicht die Hauptaufgabe der Bioinformatik, Experimente zu ersetzen, sondern resultierende große Datenmengen sinnvoll und effizient auszuwerten. Daraus sollen neue Ideen für weitere Analysen und alternative Anwendungen gewonnen werden, um fehlerhafte oder falsche Ansätze frühzeitig zu erkennen. Die Bioinformatik bietet moderne, technische Verfahren, um vertraute, aber oft mühsame experimentelle Wege durch neue, vielversprechende Ansätze zur Datenstrukturierung und Auswertung großer Datenmengen zu ergänzen. Neue Sichtweisen werden durch die Erleichterung des Testprozederes gefördert. Die resultierende Zeitersparnis führt zudem zu einer Kostenreduktion.

Abstract

The field of Bioinformatics is an interdisciplinary science focusing on the application of computer science to solve problems in different areas of life sciences.

Its scope is to handle and interpret an immense quantity of data and to support computer-aided design approaches of synthetic biological experiments. Synthetic biology deals with the generation of new components and biological characteristics created by manipulation of living organisms or parts of them. Of particular interest are two-component systems (TCS). TCS describe simple and important signalling cascades in bacteria which transfer information from the environment into the cell as a reaction to changes in the environment.

The present thesis is focused on the assessment, applicability and enhancement of bioinformatical methods in order to facilitate analysis of protein interactions and biological systems. The scientific efforts within the thesis can be divided into three aspects:

- Analysis and assessment of bioinformatical methods and enhancement of results from the preceding diploma thesis dealing with protein-protein interaction predictions.
- Analysis of general evolutionary modification possibilities within TCS as well as specific differences and design for the identification of a common approach.
- Abstraction and transfer of the results to technical and biological contexts in order to simplify synthetic biological design experiments. Establishment of comparable vocabulary for both, technical and biological processes as well as different organisms.

The outcome of this thesis revealed that TCS structure is very conserved but that it nevertheless contained some very specific characteristics. New promotor sites were discovered whilst additionally allowing the analysis of the signal binding sites. Missing elements from known TCS could be discovered and a completely new diverged TCS domain in the organism *Mycoplasma* could be identified as well as three general modification possibilities for TCS.

The combination between technological approaches and synthetic biology simplifies the systematic manipulation of TCS or other modular systems. The established two-staged module classification simplifies the analysis of modular processes and thereby the

molecular design of synthetical-biological questions. Concrete examples showed the functionality and usefulness of the classification.

A freely accessible software GoSynthetic provided easy access and application of the developed toolbox.

Not only new concrete scientific findings were provided by the given thesis but also a general approach to identify and analyse TCS and even to create similar analytic procedures. The established classification of biological and technical modules will ease the design of future experiments and reveals new opportunities applicable to similar scientific areas.

It is not the task of Bioinformatics to replace experiments but to analyse the resulting huge amounts of data meaningfully and efficiently. Hence, new ideas for further analysis and alternative cases need to be generated which may finally help to identify erroneous approaches earlier. Bioinformatics offers modern technical methods to amend familiar and sometimes exhausting experimental procedures with promising new approaches for data structuring and analysis of immense quantities of data. New perceptions are encouraged and speedier progress is possible without increasing the experimental coasts.

1. KAPITEL
Einleitung

1.1. Motivation

Geschichte

Die Bioinformatik ist nur auf den ersten Blick eine junge Disziplin. Bekanntheit erlangte sie im Jahr 2001 mit der Veröffentlichung des menschlichen Genoms, ein Ereignis, das in einschlägigen wissenschaftlichen Magazinen wie *Nature* und *Science* als Meilenstein bezeichnet wurde [1].

Aufgabengebiet und Herausforderungen

In Wirklichkeit begann der Feldzug der Bioinformatik schon viel früher mit der Untersuchung der ersten verfügbaren Gen- und Proteinsequenzen. Der Begriff an sich steht für eine interdisziplinäre Wissenschaft, die Probleme aus allen Lebenswissenschaften mit Hilfe computergestützter Methoden adressiert.

Dabei sind sowohl das Forschungsgebiet als auch die genutzten Techniken sehr breit gefächert. Der Schwerpunkt liegt auf der Entwicklung von Algorithmen und Software zur Simulation biochemischer Prozesse und zur Analyse molekularbiologischer Daten.

Mit den Jahren wurden die Sequenzierungstechniken immer weiter verbessert, wodurch eine exponentiell wachsende Datenflut an Protein- und Nukleotidsequenzen entstand. Es ist daher nicht verwunderlich, dass auch die Anzahl der bioinformatischen Algorithmen und Datenbanken stetig wuchs und bis heute immer weiter ansteigt.

Die weitaus größeren Herausforderungen liegen aber noch vor der Wissenschaft: Aus den Sequenzen müssen Gene und Proteine identifiziert und hinsichtlich ihre Funktion charakterisiert werden. Letztendlich ist das Ziel dieser bioinformatischen Ansätze, das komplexe Zusammenspiel der Proteine zu verstehen, welches die Basis zur Entwicklung neuer Medikamente darstellt. Mit Hilfe der Bioinformatik können entscheidende Beiträge zur Lösung dieser Herausforderungen bereitgestellt werden, denn ohne

ausreichende Interpretationsmöglichkeiten sind die Unmengen an generierten Daten nicht sinnvoll verwertbar.
Realisierbar wurden viele der rechen- und speicherintensiven Methoden erst durch die Steigerung der Rechnerleistung und durch die effiziente Nutzung von Datenbanken.

Nutzen in der Pharmaindustrie

Bioinformatisches Wissen hat sich auch die Pharmaindustrie zu Nutze gemacht. Bioinformatiker begleiten den Prozess vom Genom bis zum fertigen Medikament. Die Entdeckung neuer Medikamente ist zwar nach wie vor ein empirischer Prozess, wird aber nicht mehr, wie bei der Entdeckung des Penicillins durch Alexander Flemming im Jahre 1928, dem Zufall überlassen.
Stattdessen wird gezielt nach genetischen Ursachen für Krankheiten geforscht, mittels Hochdurchsatzscreenings nach möglichen Wirkstoffen gesucht, Nebenwirkungen soweit wie möglich reduziert und dadurch die potentiellen Arzneimittelkandidaten optimiert.
Außerdem wird die Individualisierung von Therapien immer bedeutsamer. Dabei wird entweder das Genom des Patienten oder des Erregers sequenziert, um Krankheiten zielgerichteter und sichererer therapieren zu können.

Ziel dieser Doktorarbeit

Die vorliegende Doktorarbeit beschäftigt sich mit der Beurteilung, Nutzung und Weiterentwicklung bioinformatischer Methoden zur Untersuchung von biologischen Prozessen. Der wissenschaftliche Beitrag der vorliegenden Arbeit kann in drei Aspekte unterteilt werden:

- Untersuchung und Beurteilung von bioinformatischen Methoden und Weiterführung der Ergebnisse aus meiner Diplomarbeit zu diesem Themengebiet.
- Analyse des generellen Designs und Modifikationsmöglichkeiten von Zweikomponenten-Systemen (TCS), mit dem Ziel der Manipulation und Nutzung in der synthetischen Biologie.
- Abstraktion der gewonnenen Erkenntnisse auf verschiedenartige technische und biologische Einheiten, um das Design neuer Experimente in der synthetischen Biologie zu vereinfachen und die Vergleichbarkeit von Technik

(Ingenieurwissenschaften) und Biologie zu erhöhen. Entwicklung einer dazugehörigen Software.

Zur Lösung der Aspekte werden folgende drei Ebenen des bioinformatischen Aufgabengebiets genutzt: Die Sequenzanalyse, die Strukturanalyse und die integrative Bioinformatik, welche sich mit der Datenaufbereitung und Speicherung biologischer Daten beschäftigt.

1.2. Biologische Grundlagen

1.2.1. Genkontext

Das traditionelle Prinzip der Analyse von Proteininteraktionen beruht darauf, die Interaktion direkt zu identifizieren. Mit Hilfe der Bioinformatik hat sich eine neue Technik entwickelt, die sogenannten Genkontext-Methoden. Dabei werden die Zusammenhänge von Proteinen auf Basis der Lage ihrer Gene im Genom hergeleitet. Im Folgenden werden die Basistechniken beschrieben.

Gen-Fusion

Das Prinzip der Gen-Fusion geht davon aus, dass ursprünglich separate Gene im Laufe der Evolution miteinander zu einem größeren Gen fusionieren und anschließend gemeinsam weiter existieren. Die Lebenszeit einer solchen Fusion hängt davon ab, ob die fusionierten Gene funktionell eng miteinander verbunden sind. Sinnvoll ist eine Fusion zweier Gene beispielsweise, wenn die kodierten Proteine die Untereinheiten eines größeren Proteins bilden. Der Vorteil besteht darin, dass die Transkription der fusionierten Gene immer gemeinsam stattfindet. Es wäre dagegen ein Nachteil, wenn die kodierten Gene nicht oder nur selten gemeinsam genutzt würden [2]. Die **Abbildung 1.2.1** (rechts) verdeutlicht die Gen-Fusion in unterschiedlichen Organismen. Dargestellt werden Gene (grün, hellgrün, blau, lila), welche in einigen Organismen fusioniert, in anderen Organismen nicht fusioniert vorliegen. Das blaue und das lila Gen sind in zwei Organismen fusioniert, ebenso wie das grüne und das hellgrüne Gen. Auch wenn diese Gene in den anderen Organismen nicht fusioniert sind, so lässt sich doch vermuten, dass die kodierten Proteine auch in den anderen Organismen sehr eng zusammenhängen.

Gen-Nachbarschaft

Funktionell assoziierte Gene sind im Genom häufig in unmittelbarer Nähe zueinander lokalisiert. Diese Nähe bietet die Möglichkeit unterschiedliche Gene gemeinsam zu regulieren. Bei gleicher Orientierung können sie sogar gemeinsam transkribiert werden. Besonders in Prokaryoten kann dieses Phänomen häufig beobachtet werden. Dabei regulieren Transkriptionseinheiten, sogenannte Operons, die gemeinsame Transkription mehrerer Gene. Wenn Gene häufig in einem gemeinsamen Operon vorkommen, ist dies ein deutlicher Hinweis auf eine funktionale Verbindung der kodierten Proteine [3]. Das bei Prokaryoten beobachtete Phänomen kann oft auch auf Eukaryoten übertragen werden [4]. In **Abbildung 1.2.1** (mitte) wird die Gen-Nachbarschaft visualisiert. Die gelben, grünen, hellgrünen und roten Gene besitzen nicht nur die gleiche transkriptionale Orientierung, sondern sind zusätzlich in allen untersuchten Organismen gemeinsam vorhanden oder nicht vorhanden. Das lässt auf eine gemeinsame Funktion schließen.

Phylogenetisches Profil

Die phylogenetische Profildarstellung ist die allgemeinste der Genkontext-Methoden. Bei dieser Methode werden Proteine als miteinander assoziiert gekennzeichnet, wenn deren kodierende Gene in unterschiedlichen Organismen immer gemeinsam auftreten oder immer gemeinsam fehlen. Aus diesem Grund wird die Methode auch *Co-occurrence* (gemeinsames Auftreten) genannt. Das phylogenetische Profil ist nur ein schwacher Hinweis auf die funktionelle Assoziation von Proteinen. Visualisiert ist diese Methode in **Abbildung 1.2.1** (links). Darin wird pro Gen über alle Organismen hinweg die Anwesenheit (grüner Punkt) und Abwesenheit (roter Strich) des Gens dokumentiert.

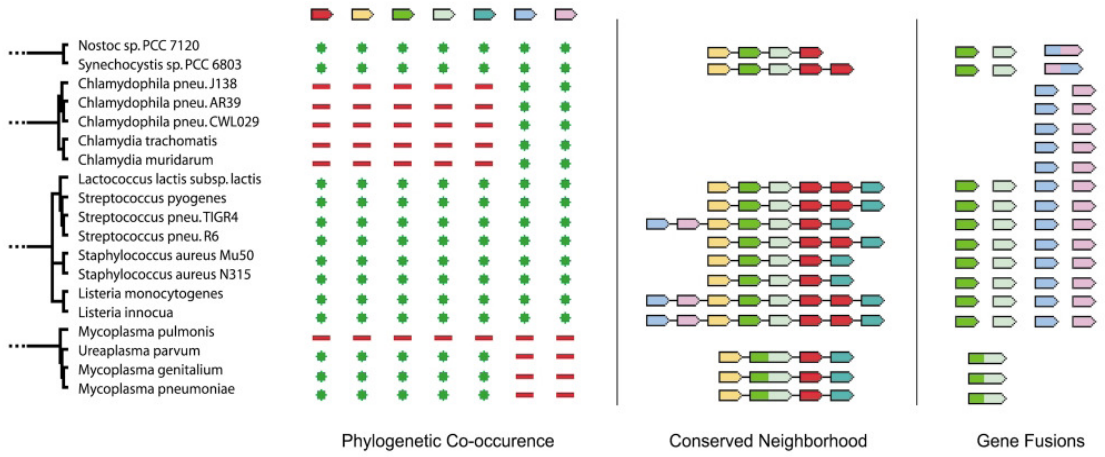

Abbildung1.2.1: Visuelle Darstellung der Genkontext-Methoden [5].
Linker Bereich, das phylogenetische Profil. Grüne Punkte stehen für die Präsenz eines Gens in einem Organismus, rote Linien stehen für seine Abwesenheit.
Mittlerer Bereich, die Gen-Nachbarschaft: Das gelbe, grüne, hellgrüne und rote Gen ist in allen Organismen gemeinsam vorhanden oder nicht vorhanden.
Rechter Bereich, die Gen-Fusion: Das blaue und das lila Gen sind in zwei Organismen fusioniert, ebenso wie das grüne und das hellgrüne Gen.

Präsentation durch Netzwerke

Die Funktion eines einzelnen Proteins kann nur unter Berücksichtigung des Zusammenspiels mit anderen Proteinen ausreichend erforscht und beurteilt werden. Diese Wechselwirkungen werden in sogenannten metabolischen und regulatorischen Netzwerken dargestellt, die unter anderem die Simulation von Stoffwechsel-Wegen im Computer ermöglichen.

1.2.2. Zweikomponenten-Systeme

Zweikomponenten-Systeme (Two-Component System, TCS) stellen sehr einfache Signalkaskaden in Bakterien dar, um Informationen aus der Umgebung in eine Zelle zu übertragen. Lebende Zellen sind darauf angewiesen auf externe Signale (z. B. Druckänderung oder Gradienten von Nahrungsmolekülen) angemessen reagieren zu können.

Diese Systeme bestehen aus einem Transmembranprotein, das als Sensor fungiert, und einem zytoplasmatischen Protein, welches die Regulationen eines zellulären Prozesses steuert [6]. Das Sensorprotein detektiert Änderungen in der Umgebung und bindet Adenosintriphosphat (ATP). Diese Bindung bewirkt eine strukturelle Änderung des

Sensors und ruft den Phosphortransfer auf das passende Regulationsprotein hervor. Das Regulationsprotein selbst verändert ebenfalls die Struktur und löst eine Zellreaktion aus. Die Datenbank KEGG listet bekannte Zweikomponenten-Systeme auf.

Abbildung 1.2.2 verdeutlicht die Signalkaskade in Zweikomponenten-Systemen.

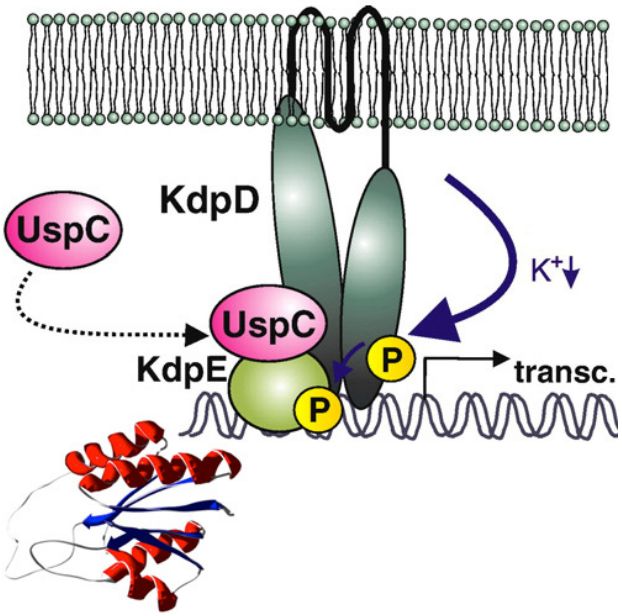

Abbildung 1.2.2: Schematischer Ablauf der einfachsten Signalkaskade in Bakterien, Zweikomponenten-Systeme [7].
Das Sensorprotein KdpD und das Regulationsprotein KdpE kontrollieren die Induktion des kdpFABC Operons. Der Stimulus diese Systems ist Salzstress (Kaliummangel) und stimuliert das System über den periplasmatischen Bereich des Sensorproteins, wodurch dieses phosphoryliert wird (gelb). Das Protein UspC stabilisiert den KdpD/KdpE-Phosphor/DNA-Komplex, die Transkription beginnt.

Sensorprotein

Ein Sensorprotein ist ein Homodimer (zusammengesetzt aus zwei gleichen Einheiten) und besteht aus einer periplasmatischen Sensordomäne und einer Histidinkinase-Domäne (HisKA). Sensordomänen sind in der Lage Änderungen aus der Umgebung (Signale) zu detektieren. Solche Signale können z.B. Lichtmenge, Temperatur, Wasserangebot, Schadstoffe oder oxidativer Stress sein.

HisKA-Domänen wiederum bestehen aus einer Histidin-Phosphotransfer-Domäne (DHp) und einer katalytischen ATP-Bindedomäne (CA), s. **Abbildung 1.2.3**.

Strukturell gesehen wird die Histidindomäne gebildet durch eine vier-alpha-Helix mit einem zentral liegenden Histidin zur Phosphorbindung. Das Dimer wird über hydrophobe Wechselwirkungen zwischen den Histidindomänen ausgebildet.

Die Kinasedomäne entspricht einem alpha/beta-Sandwich. Dieser besteht aus einem hydrophoben Kern und einem sauren Bereich aus Asparaginsäure oder Glutaminsäure, welcher bei der Autophosphorylierung den Phosphor übernimmt.

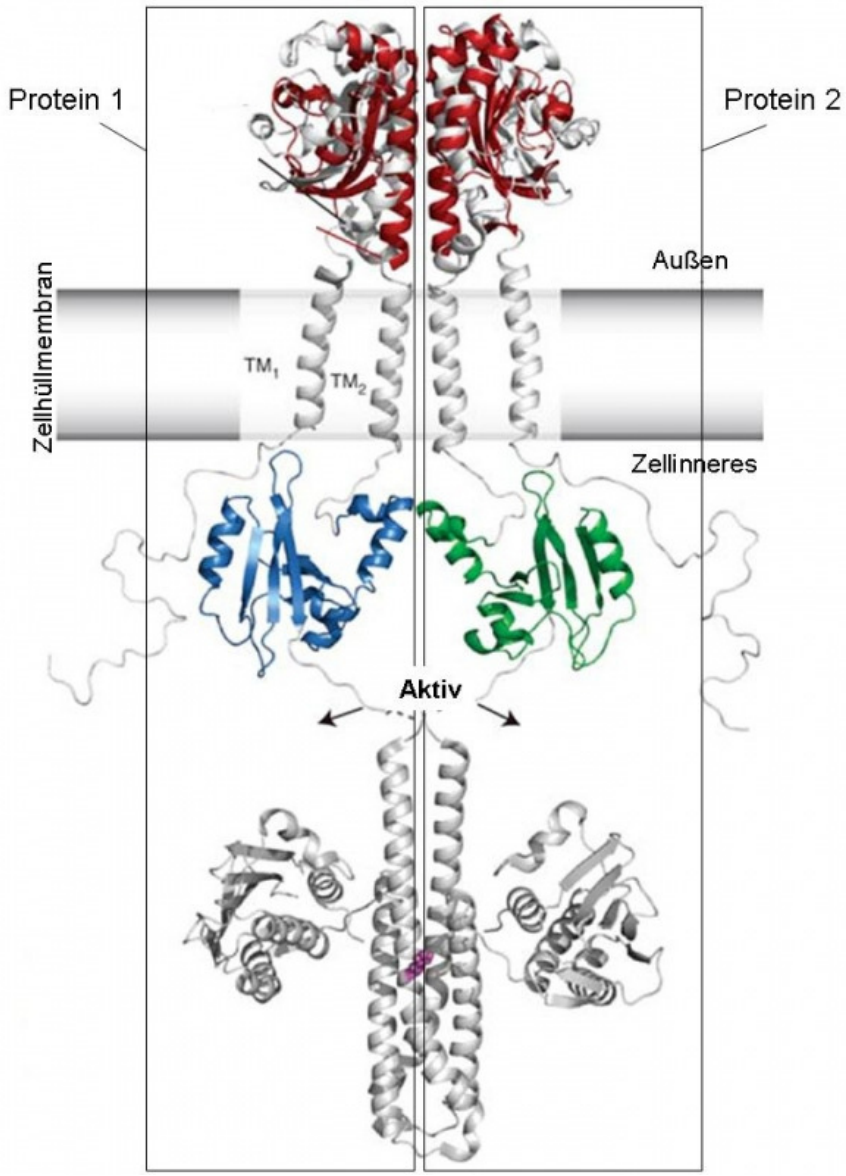

Abbildung 1.2.3: Aufbau Sensorprotein: Kristallstruktur der Histidinkinase DcuS [8].
Kristallstruktur des Homodimer-Sensorproteins DcuS. Es besteht aus der periplasmatsichen Sensordomäne (rot) und der Histidinkinase-Domäne (grau). Die Histidinkinase besteht aus der vier-alpha-Helix Histidindomäne, mit dem konservierten Histidin (magenta) zur Phosporbindung und dem alpha/beta-Sandwich, der Kinasedomäne.

Laut Grebe [9] können HisKA in unterschiedliche Klassen eingeteilt werden. Die Klassifikation ist abhängig von den konservierten Aminosäuremotiven, den N-, G1-, F-, G2- und H-Boxen.

Die H-Box (HExxxP) beinhaltet das phosphorylierte Histidin. Die anderen Aminosäuremotive sind N (NLxxxN), G1 (DxGxG), F (FxPF) und G2 (GxGxGL).

Regulationsprotein

Das Regulationsprotein (RR) ist ein Homodimer und besteht aus einer Empfängerdomäne und einer Effektordomäne, siehe **Abbildung 1.2.4**.

Die Empfängerdomäne, auch Responseregulator-Domäne genannt, nimmt die Phosphatgruppe auf, wodurch das Regulationsprotein seine Struktur verändert. Das dadurch aktivierte Regulationsprotein bewirkt eine zelluläre Antwort, häufig eine Regulation der Genexpression als Transkriptionsfaktor.

Strukturell gesehen wird die Responseregulator-Domäne durch ein alpha/beta 3-Layer-Sandwich gebildet und die Effektordomäne aus einem Helix-Turn-Helix (HTH)-Motiv. Abhängig von der zu regulierenden Funktion können aber auch andere Motive in der Effektordomäne auftreten.

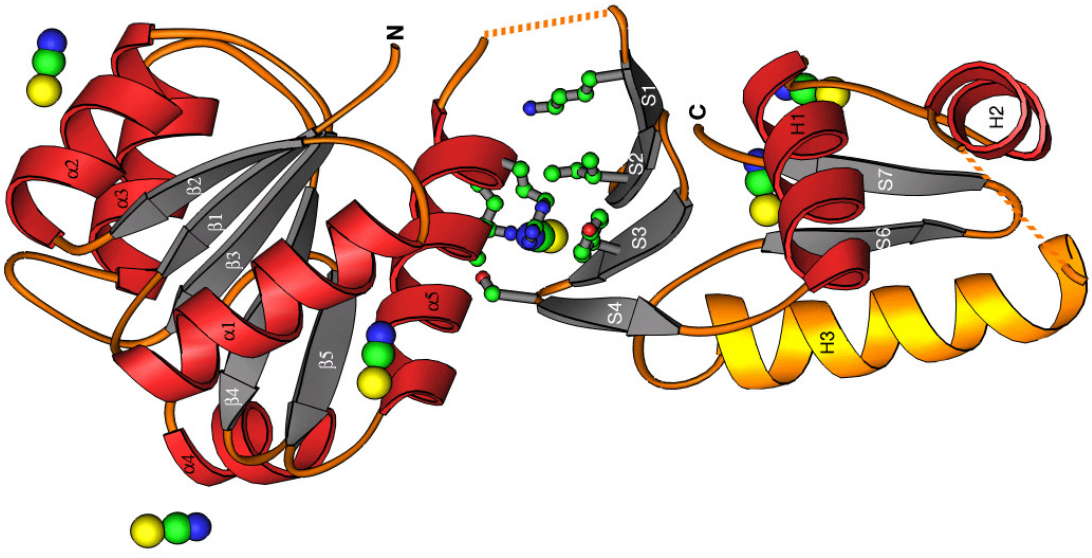

Abbildung 1.2.4: Aufbau Regulationsprotein: Kristallstruktur aus Empfängerdomäne und Effektordomäne.
Regulationsprotein DrrD aus *Thermotoga maritima* besteht aus einem alpha/beta 3-Layer-Sandwich, Responseregulator-Domäne (links) und einer Effektordomäne mit einem Helix-Turn-Helix (HTH) Motiv (rechts), welches an die DNA bindet (gelbe Helix) [PDB 1P2F].

Aktuelle Mutationsanalysen

Mutationen von Zweikomponenten-Systemen wurden bereits Anfang des 21. Jahrhunderts untersucht [10, 11]. Neuere Untersuchungen zeigen, dass Zweikomponenten-Systeme auch für die synthetische Biologie von großem Interesse sind [12, 13]. Ein solches Beispiel stellt das Design eines Zweikomponenten-Systems in *E. coli* dar, welches durch künstliche Veränderung des Sensorproteins auf rotes Licht reagiert [14].

Weitere Untersuchungen haben gezeigt, dass Zweikomponenten-Systeme durch positive Rückkopplungsschleifen gesteuert werden können [15] und dass Sensorproteine bedingt austauschbar sind [16]. Untersuchungen von Skerker im Jahr 2008 zeigen, dass das Osmolaritäts-Sensorprotein EnvZ durch die Sensorproteine RstB, CpxA, AtoS, PhoR und PhoQ ausgetauscht werden kann, ohne dass die Funktionalität des TCS eingeschränkt wird [17].

Geregelt wird die gemeinsame Expression aller regulierten TCS-Gene mit Hilfe eines gemeinsamen Operons, z.B. GlnALG für die Glutaminfamilie, OmpRFC für die Osmosefamilie, PhoPQ für die Phosphatfamilie oder NarGLK für die Nitrat/ Nitritfamilen.

Außerdem wurde beschrieben, dass durch Mehrfachbindungen an Promotorstellen die Spezifität der Aktivierung von Genen erheblich verbessert werden kann [18, 19].

Konnektoren

In Zweikomponenten-Systemen wurden kürzlich neue Komponenten (Proteine oder Domänen), sogenannte Konnektoren, entdeckt. Sie beeinflussen die Funktionalität und Aktivierung von Zweikomponenten-Systemen auf die unterschiedlichsten Weisen. Teilweise sind diese Komponenten organismenspezifisch, teilweise funktionsabhängig [20, 21]. Ein bislang wenig annotierter Konnektor ist das SafA Protein.

Der Konnektor SafA (sensor associating factor A) (früher auch als MG1655 oder b1500 bezeichnet*)* ist verknüpft mit der Signaltransduktion zwischen dem EvgS/EvgA und PhoQ/PhoP Zweikomponenten-System. Seine Expression wird durch das aktivierte EvgA eingeleitet. Das aktivierte EvgS/EvgA System aktiviert wiederum das PhoQ und damit das PhoQ/PhoP Zweikomponenten-System [21], s. **Abbildung 1.2.5** [22].

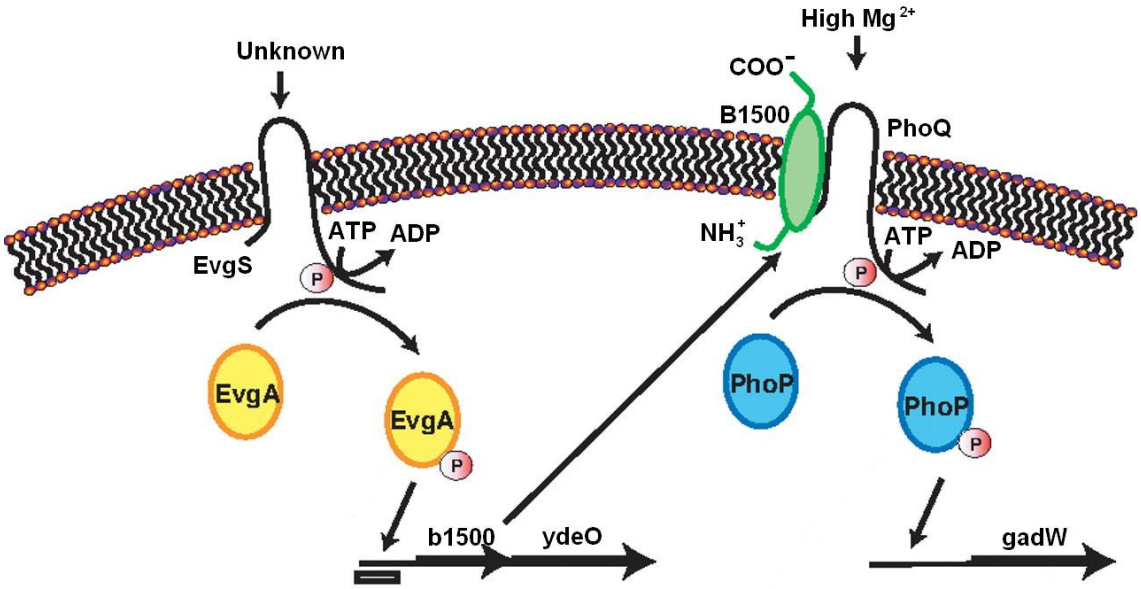

Abbildung 1.2.5: Wirkungsweise SafA. SafA (b1500) verknüpft die Zweikomponenten-Systeme EvgS/EvgA und PhoQ/PhoP miteinander [22].

In diesem Zusammenhang müssen auch die GGDEF-Domäne und die EAL-Domäne erwähnt werden.

Die GGDEF-Domäne ist ähnlich zur Adenylatzyklase und bindet häufig an Signalproteine. Außerdem ist sie vermehrt an regulatorische Domänen gebunden. Weitergehende Funktionen sind derzeit nicht bekannt (Pfam: PF00990).

Die EAL-Domäne kann in diversen bakteriellen Signalproteinen gefunden werden und ist nach ihren konservierten Aminosäuren benannt (Glutaminsäure - E, Alanin - A und Leucin - L) (Pfam: PF00563).

Helicobacter pylori

Helicobacter pylori ist ein säureresistentes, gram-negatives Stäbchenbakterium, welches den menschlichen Magen besiedeln kann. Der spiralig gekrümmte Keim bewegt sich mittels seiner Geißeln fort. Infektionen mit *H. pylori* werden für eine Reihe von Magenerkrankungen verantwortlich gemacht, die mit einer verstärkten Sekretion von Magensäure einhergehen. Darunter fallen beispielsweise die Typ B-Gastritis, etwa 75% der Magengeschwüre und praktisch alle Zwölffingerdarmgeschwüre [23].

Abbildung 1.2.6: Schematische Darstellung des Bakteriums *H. pylori*. Der spiralig gekrümmte Keim bewegt sich mittels seiner Geißeln fort [23].

Mycoplasma pneumoniae

Mycoplasma pneumoniae sind die kleinsten Bakterien, welche außerhalb von Zellen vermehrungsfähig sind und keine Zellwand besitzen (gram-positiv). Sie sind die wichtigster Erreger bei Lungenentzündung, aber auch Tracheobronchitis, Kehlkopfentzündungen, Hirnhautentzündungen, Mittelohrentzündungen und weitere Krankheitsbilder können von *M. pneumoniae* verursacht werden. Zudem wird der Organismus mit Störungen des hämatopoetischen (blutbildenden) Systems, des zentralen Nervensystems, der Leber und der Bauchspeicheldrüse sowie mit kardiovaskulären Syndromen in Verbindung gebracht [24].

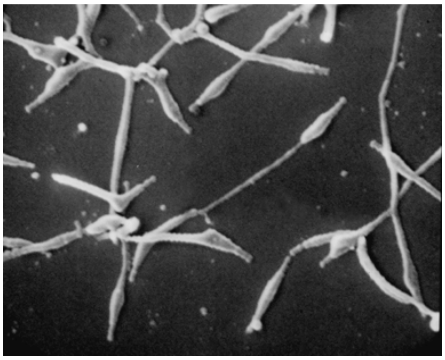

Abbildung 1.2.7: Elektronenmikroskopische Darstellung des pathogenen Organismus *M. pneumoniae* [Bild aus UNC, School of Medicine, Department of Pediatrics].

Vaccinia Virus

Die Herkunft des Vaccinia Virus ist bis heute unsicher. Es könnte aus einer genetischen Rekombination entstanden sein, aber auch aus dem Kuhpockenvirus. Beim Menschen verursacht es eine sehr schwache Infektion, welche zur Resistenz gegen das menschliche Pockenvirus (Variola) führen kann. Für eine Infektion nutzt das relativ große Virus das Abfallwesen der Zelle aus, indem es sich als Zellabfall tarnt. Benachbarte Zellen nehmen die Bruchstücke der vermeintlichen toten Zelle auf, bevor das Immunsystem reagieren kann. Dieselbe Strategie wie Vaccinia Viren nutzen vermutlich auch andere große Viren, wie Herpes- und HI-Viren [25].

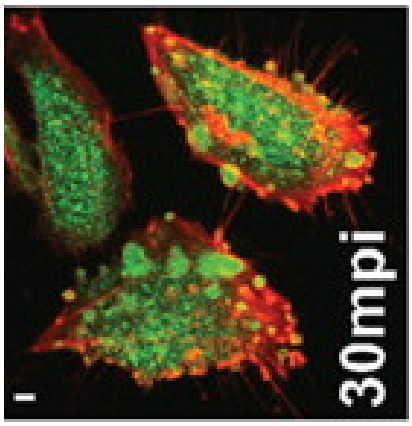

Abbildung 1.2.8: Das Vaccinia Virus wandert entlang der Zellfäden zur Zelloberfläche, wo es die Bildung von Ausstülpungen anregt. Actin fluoresziert rot, das Enzym PAK1 grün [25].

Dies ist lediglich ein Überblick über die in dieser Arbeit genutzten unbekannteren Organismen.

1.2.3. Synthetische Biologie

Die synthetische Biologie ist ein Spezialgebiet der Bioinformatik und wurde bereits 1974 durch den Genetiker Waclaw Szybalski entscheidend geprägt [26]. Sie beschreibt die Generierung neuer Produkte durch die Behandlung und Manipulation lebender Organismen oder Teilen daraus unter Anwendung von Konstruktionsprinzipien der Technik [27]. In der synthetischen Biologie können unterschiedliche Strategien verfolgt werden:

- Integration künstlicher Systeme in Lebewesen um neue Eigenschaften zu erhalten.
- Aufbau chemischer Systeme mit bestimmten Eigenschaften von Lebewesen, analog zu Vorbildern in der Biologie.
- Erzeugung von nur auf die notwendigsten Komponenten reduzierten Organismen für den Nachbau einfacher biologischer Schaltkreise.

Die Aktivitäten in diesem Bereich sind rege. Aufsehenerregende Experimente aus den Jahren 2008 und 2010 beschreiben die erstmalige vollständig synthetische Entwicklung genetischen Erbmaterials des Bakteriums *M. pneumoniae* durch die Forschungsgruppe um Craig Venter [28, 29, 30, 31].

Aufgrund der Aktualität und der Bedeutung des Themas versuchen viele neuere Studien die synthetische Biologie bei ihrer Arbeit zu unterstützen. Datensammlungen mit unterschiedlichen Schwerpunkten, versuchen hier weiterzuhelfen. Im Jahr 2004 fand die erste internationale wissenschaftliche Konferenz zum Thema synthetische Biologie am Massachusetts Institute of Technology (MIT) in den USA statt [32]. Auf deren Basis entstand im Jahr 2009 die „Synthetic Biology Parts List“ [33, 34], welche sich mit der Standardisierung und Beschreibung von wiederverwendbaren Komponenten, sogenannten BioBricks, beschäftigt. Seit 1998 existiert die Gene Ontology (GO) Klassifizierung, eine Standardisierung von Genprodukten und ihren biologischen Prozessen, zellulären Komponenten und molekularen Funktionen [35] Abrufbar ist diese Klassifizierung über den Webdienst AmiGO [36]. Protein-Protein Interaktionen können mit Hilfe der Interaktions- und Genkontext-Datenbank STRING [37] vorhergesagt und in Netzwerken graphisch dargestellt werden.

Daneben existieren auch Werkzeuge, welche den Designprozess erleichtern sollen. Diese reichen von einem komplett automatisierten Design [38] über die Erstellung einer eigenen Steuerungs-Programmiersprache [39] bis hin zur Erstellung eines Werkzeugkastens für Basisgeräte in der synthetischen Biologie [40].
Trotz dieser Ansätze ist es bislang nur sehr schwer möglich die benötigten kombinierten Daten über molekulare Prozesse, Funktionen, biologische Bausteine und Interaktionen aus den vielen zur Verfügung stehenden Daten aufzufinden, welches die Grundlage für die Veränderung biologischer Prozesse darstellt.

Dieser kurze Überblick über die Disziplin der synthetischen Biologie verdeutlicht die schnelle Entwicklung und die Bandbreite dieses Themas.

Die Hoffnungen, aber auch die Ängste bezüglich Auswirkungen der synthetischen Biologie sind groß. Dies liegt an der Nähe zum Gebiet des *artificial life* (künstliches Leben), welche aufgrund der Thematik und ihres Anspruchs natürlich hitzige öffentliche Diskussionen auslöst. Die Praxis der synthetischen Biologie wendet sich dagegen von diesen überstiegenen positiven und negativen Utopien ab und widmet sich ganz konkreten, schrittweisen Anwendungen molekularbiologischer Fortschritte für Medizin, Biotechnologie und Pharmazie und die Entwicklung von Software für diese Themengebiete.

2. KAPITEL

Material und Methoden

Dieses Kapitel gliedert sich in die Bereiche Datenbanken, Algorithmen und Techniken, welche während der Dissertation genutzt wurden.

2.1. Datenbanken

Die Datengrundlage dieser Promotion basiert auf öffentlichen, biologischen Datenquellen.
Die genutzten Quellen lassen sich in Sequenz-, Struktur-, Funktions- und Interaktionsdatenbanken unterteilen. Die Wichtigsten davon werden im Folgenden vorgestellt.

2.1.1. Sequenz- und Strukturdatenbanken

Uniprot/Swiss-Prot

Die Uniprot/Swiss-Prot Proteindatenbank (http://www.expasy.ch/sprot/) ist eine 1986 erstellte und seit dem aufwändig gepflegte Proteinsequenzdatenbank. Sie bietet u.a. qualitativ hochwertige Annotationen über Proteinfunktionen, Domänen, posttranslationale Modifikationen und Proteinvariationen [41].

Prosite

Die Prosite Datenbank (http://www.expasy.org/prosite/) enthält eine große Sammlung biologisch bedeutsamer Signaturen, welche in Form von qualitativen und quantitativen Mustern beschrieben werden.

Ein qualitatives Motiv, auch Pattern genannt, ist ein Muster welches durch reguläre Ausdrücke dargestellte konservierte Sequenzen beschreibt.
Ein quantitatives Motiv, auch Profil genannt, beschreibt eine Matrix aus Wahrscheinlichkeiten für Aminosäuren an bestimmten Positionen in einer Sequenz.
Jede Signatur in der Prosite Datenbank beinhaltet biologische Informationen zu Proteinfamilien und Proteindomänen [42].

Pfam

Die Pfam-Datenbank (http://pfam.sanger.ac.uk/) beinhaltet eine Vielzahl multipler Sequenzalinierungen und qualitativen Motiven in Form von Hidden Markov Modellen (HMM). Sie bietet Zugang zu Alinierungen, Organismen- und Sequenzbäumen und funktionellen Informationen zu Domänenfamilien. Derzeit beinhaltet sie über 13672 Proteinfamilien und Domänen [43].

PDB – Protein Data Bank

Die **P**rotein **D**ata **B**ank (PDB; http://www.pdb.org/pdb/) enthält Strukturdaten zu biologischen Makromolekülen. Zusätzlich enthält sie Beschreibungen zur Architektur und Funktionalität der Proteine [44].
PDBSum ist eine web-basierte Oberfläche, welche schematische Diagramme für alle Strukturen der PDB Datenbank enthält. Sie beinhaltet Strukturbilder und Sekundärstrukturinformationen [45].

2.1.2. Funktionsdatenbanken

GO - Gene Ontology

Die **G**ene **O**ntology Datenbank (GO; http://www.geneontology.org/) ist ein bioinformatisches Projekt, um Gen- und Protein-Attribute über unterschiedliche Spezies und Datenbanken hinweg zu standardisieren. Die Datenbank stellt ein kontrolliertes Vokabular zur Beschreibung der Attribute zur Verfügung. Ein Konsortium pflegt diese Angaben [46].

KEGG - Kyoto Encyclopedia of Genes and Genomes

Die **K**yoto **E**ncyclopedia of **G**enes and **G**enomes Datenbank (KEGG; http://www.genome.ad.jp/kegg) bietet sehr vielseitige Informationen. Sie liefert standardisierte und speziesspezifische graphische Darstellungen für zelluläre Prozesse, wie Signal- oder Stoffwechselwege [47].

Prodoric - Procariotic Database of Gene-Regulation

Die Procariotic Database of Gene-Regulation Datenbank (Prodoric; http://prodoric.tu-bs.de/) bietet Informationen über Genregulation, Genexpression und eine Vielzahl an Transkriptionsfaktorbindestellen in Prokaryoten.
Zusätzlich stellt sie einige bioinformatische Werkzeuge für die Vorhersage, Analyse und Visualisierung von Genregulationen zur Verfügung [48].

DBTBS - Database of Transcriptional Regulation in *B. subtilis*

Die **D**ata**b**ase of **T**ranscriptional regulation in ***B****. subtilis* (DBTBS; http://dbtbs.hgc.jp/) enthält eine Sammlung experimentell bestätigter Genregulationsbeziehungen und den dazugehörigen Transkriptionsfaktorbindestellen in *Bacillus subtilis*. Für jeden Transkriptionsfaktor sind die Konsensussequenzen illustriert und über Pfam-Motive klassifiziert [49].

TractorDB - TRAnscription FaCTORs

Die Datenbank **TRA**nscription Fa**CTOR**s (TractorDB; http://www.tractor.lncc.br/) enthält experimentelle Daten aus *Escherichia coli* über transkriptionell regulatorische Systeme. Mit der Zeit wurde sie um andere prokaryotische Spezies und computererzeugter Vorhersagen erweitert [50].

2.1.3. Interaktionsdatenbanken

STRING - Search Tool for Recurring Instances of Neighbouring Genes

Das **S**earch **T**ool for **R**ecurring **I**nstances of **N**eighbouring **G**enes (STRING; http://string.embl.de/) ist eine im Jahr 2000 entwickelte und stetig erweiterte Datenbank der bekannten und vorhergesagten Proteininteraktionen. Die Interaktionen stammen aus Genkontext-Untersuchungen, Hochdurchsatzexperimenten, Co-Expressionsdaten und bekanntem Literatur- und Datenbankwissen (MIPS, BIND, KEGG, BIOCARTA, ArrayProspector, PubMed, DIP, MINT).

Die Datenbank enthält derzeit mehr als 3 Millionen Proteine aus 1100 Organismen. Die Visualisierung der Interaktionen erfolgt in Form von Netzwerken [37].

IntAct

Die Datenbank IntAct (http://www.ebi.ac.uk/intact) bietet eine Vielzahl textueller und graphischer Darstellungen von Proteininteraktionen mit Verbindungen zu GO Annotationen und Domäneninformationen. Sie enthält derzeit über 170.000 binäre und komplexe Interaktionen aus der Literatur, die durch das Swiss-Prot Team überwacht werden [51].

PubMed

Die PubMed Datenbank (http://www.ncbi.nlm.nih.gov/pubmed/) ist eine englischsprachige, textbasierte Meta-Datenbank mit medizinischen Artikeln bezogen auf den gesamten Bereich der Biomedizin. Entwickelt wurde die Datenbank durch das nationale Zentrum für Biotechnologische Informationen (National Center for Biotechnology Information, NCBI). Verzeichnet sind derzeit über 21 Millionen Literaturstellen, sie wächst jährlich um rund 500.000 Dokumente. Jeder Eintrag in PubMed besitzt eine PubMed-ID (PMID) [52].

Weiterhin werden jedem Dokument Schlagworte zugeordnet, sogenannte Medical Subject Headings (MeSH). MeSH ist ein durch die U.S. National Library of Medicine kontrolliertes Vokabular, welches für die Indexierung von Artikeln in MEDLINE/PubMed genutzt wird. MeSH Terminologie beschreibt einen Weg, um eine konsistente Abfrage von Literaturinformationen zu ermöglichen [53].

2.2. Algorithmen

Um Sequenzen und Strukturen zu analysieren werden unterschiedliche Algorithmen benötigt. Dabei handelt es sich um mathematische Algorithmen, welche das Vergleichen von Sequenzen und Strukturen, aber auch das Visualisieren und Vorhersagen ermöglichen.

2.2.1. Sequenzvergleiche

Zum Vergleichen von Sequenzen werden Alinierungen genutzt, welche die Grundlagen aller Sequenzanalysen bilden. Das Ziel ist es, möglichst viele identische oder ähnliche Positionen nebeneinander in den zu vergleichenden Sequenzen zu finden.
Die Ähnlichkeit einer Position wird durch eine sogenannte Substitutionsmatrix ermittelt, welche beschreibt wie ähnlich zwei Aminosäuren oder Nukleotide sind.
Die Qualität einer Alinierung wird mit Hilfe von Vergleichswerten (Scores) und Wahrscheinlichkeiten (E-values) beschrieben. Der Score ist dabei eine Angabe über die Ähnlichkeit der Sequenzen und wird errechnet aus den Werten der Substitutionsmatrize, normiert über die Länge. Der E-value gibt an, wie wahrscheinlich es ist, dass diese Alinierung durch Zufall entstanden ist, d.h. je kleiner der E-value und desto größer der Score, desto ähnlicher ist die Sequenz.
Es können die folgenden Arten von Alinierungen unterschieden werden:

Lokale Sequenzvergleiche

Lokale Sequenzvergleiche (*local alignments*) werden mit dem Smith und Waterman Algorithmus erstellt. Dabei wird versucht, die Anzahl an Übereinstimmungen durch Lücken (Gaps) zu maximieren. So wird der kleinste Abstand zwischen den längsten Teilsequenzen bestimmt. Lokale Alinierungen liefern im Gegensatz zu globalen Alinierungen immer ein Ergebnis [54].

Globale Sequenzvergleiche

Globale Sequenzvergleiche (*global alignments*) werden mit dem Needleman und Wunsch Algorithmus erstellt. Dabei wird versucht, die Anzahl an Übereinstimmung (Matchen) zu maximieren und die Anzahl an Lücken (Gaps) dazwischen zu minimieren. So wird der minimale Abstand zweier Sequenzen bestimmt. Globale Alinierungen sind nur bei einem Vergleich eng verwandter Sequenzen sinnvoll [55].

Multiple Alinierung

Die multiple Alinierung ist die gleichzeitige Analyse mehrerer Sequenzen. Sie liefert dabei genauere Informationen über die Aminosäureverteilung einzelner Positionen. Solche Verteilungen können Aufschluss über konservierte Bereiche in Sequenzen geben.

ClustalW

Die häufigsten multiplen Alinierungen sind globale Alinierungen, die mit heuristischen Methoden errechnet werden.

ClustalW ist ein sehr sensitives Programm zur Berechnung globaler multipler Alinierungen und wurde 1994 von Thomson und Kollegen entwickelt [56].

Blockmaker

Für die Analyse von Proteindomänen oder anderen Motiven werden lokale multiple Alinierungen genutzt.

Blockmaker ist ein Programm zur Berechnung lokaler multipler Alinierungen und wurde 1991 von Henikoff entwickelt. Dabei werden die Sequenzen in kleinere Blöcke zerlegt, welche gemeinsame Motive besitzen und keine Gaps enthalten [57].

Darstellung multipler Alinierungen

Um konservierte Positionen in multiplen Alinierungen hervorzuheben, wird oft eine Konsensussequenz erstellt. Sie besteht aus Zeichen, welche die Stärke der Konservierung mittels z. B. Größe oder Formen angeben.

Für die Darstellung lokaler Alinierungen bietet sich im Besonderen die COBBLER – Darstellung (**CO**nsensus **B**iasing **B**y **L**ocally **E**mbedding **R**esidues) an. Dabei wird eine

Einzelsequenz aus den gefundenen Blöcken genutzt, wobei konservierte Stellen durch die tatsächlich in den Blöcken gefundene Konsensussequenz ersetzt werden [58].
Eine andere Möglichkeit besteht darin, die konservierten Bereiche direkt in der Sequenz durch Größe und farbige Unterlegung hervorzuheben. Diese Formatierung ist beispielsweise möglich mit dem Programm Weblogo (http://weblogo.berkeley.edu) [59].

2.2.2. Sequenzanalyse

BLAST – Basic Local Alignment Search Tool

Das **B**asic **L**ocal **A**lignment **S**earch **T**ool, BLAST ist ein 1990 von Altschul und Kollegen entwickelter heuristischer Suchalgorithmus, um Sequenz-Datenbanken schnell zu durchsuchen. Die Grundlage bildet eine lokale Alinierung, welche allerdings nur ähnliche Positionen als Treffer zählt, die in direkter Nachbarschaft einen zweiten Treffer haben. BLAST filtert und sortiert die Treffer nach ihrem E-value. E-values > 0,001 sind statistisch nicht mehr signifikant [60].

Quantitative Ähnlichkeitssuche

Eine Erweiterung der BLAST-Suche ist der sogenannte PSI-BLAST (**P**osition-**S**pecific **I**terative BLAST). Basierend auf den Treffern einer ersten Suche wird eine multiple Alinierung erstellt, aus deren Profil die nächste BLAST Suche durchgeführt wird. Diese Iteration kann mehrmals durchgeführt werden, wobei die neuen Sequenzen mit in die multiple Alinierung und somit auch in das Profil einbezogen werden [61].

Qualitative Ähnlichkeitssuche

Der sogenannte PHI-BLAST (**P**attern **H**it **I**nitiated BLAST) stellt eine zweite Erweiterung der BLAST-Suche da. Er gibt Auskunft, ob ein bestimmtes qualitatives Motiv in Form eines regulären Ausdrucks in einer Suchsequenz konserviert ist. Dazu wird das Motiv mit der Suchsequenz gegen eine Datenbank verglichen. Als Ergebnis werden nur solche Sequenzen geliefert, welche eine Ähnlichkeit zur Suchsequenz besitzen und welche zusätzlich das Motiv aufweisen [62].

Eine weitere Möglichkeit für qualitative Ähnlichkeitssuchen stellt das an die Prosite-Datenbank angehängte Werkszeug ScanProsite dar. Diese Suche erlaubt zum einen Sequenzen in der Prosite-Datenbank zu scannen und dabei als Ergebnis ein vorhandenes Motiv zu erhalten. Des Weiteren kann auch mit qualitativen Motiven in der Swiss-Prot-Datenbank gesucht werden, um Proteine zu finden, welche zu dem eingegebenen Pattern passen.

2.2.3. Strukturvorhersage

Von vielen Proteinsequenzen ist keine räumliche Struktur bekannt. Proteinfunktionen basieren aber auf der Struktur des Proteins. Um die Funktion oder zumindest Teilfunktionen dieser Proteine vorherzusagen, werden Struktur-Vorhersageprogramme genutzt. Die Grundlage für diese Vorhersagen ist, dass bestimmte Aminosäuren aufgrund ihrer physiko-chemischen Eigenschaften gewisse Sekundärstrukturen bevorzugen, welche wiederum die Bildung gewisser 3D-Strukturen vermuten lassen.

SSPro - Secondary **S**tructure **Pr**edict**i**on

Der Server SCRATCH (http://scratch.proteomics.ics.uci.edu/) beinhaltet eine Sammlung an Programmen zur Vorhersage von Proteinstrukturen. Das hier bevorzugt eingesetzte Programm zur Vorhersage von Sekundärstrukturen ist SSPro.

Für die Vorhersagen in SSPro werden neuronale Netzwerke und PSI-BLAST-Profile genutzt. Geprüft werden die Vorhersagen durch einen großen, nicht-redundanten Testdatensatz, welcher durch drei Tests validiert wird, wodurch 78% der Vorhersagen korrekt sind [63].

PredictProtein

PredictProtein ist ein Internet Service (http://www.predictprotein.org) zur Vorhersage von Proteinstrukturen. Das Ergebnis einer PredictProtein Vorhersage ist eine multiple Sequenzalinierung, ein Prosite Sequenzmotiv und eine vorhergesagte Sekundärstruktur. Regionen niedriger Komplexität, Transmembranhelices, Doppelwendel-Regionen, Disulfidbrücken und viele weitere Strukturmerkmale sind darin ebenfalls annotiert [64].

Porter

Porter ist ein Internet Service (http://distill.ucd.ie/porter/) zur Vorhersage von Sekundärstrukturen. Es basiert auf neuronalen Netzwerken und multiplen Sequenzalinierungen. Seine Zuverlässigkeit wird überprüft durch fünf Validierungen mittels eines großen Testdatensatzes [65].

SSEA – Secondary Structure Element Alignment

SSEA ist ein Webserver (http://protein.cribi.unipd.it/ssea/) zum Vergleich von Sekundärstrukturen. Dabei kann entweder ein Protein gegen eine repräsentative Bibliothek aus bekannten Proteinfaltungen (*database alignment mode*) oder gegen eine einzelne Sekundärstruktur (*one vs. one alignment mode*) verglichen werden [66].

SWISS-MODEL

SWISS-MODEL ist ein vollautomatisierter Server (http://swissmodel.expasy.org/) zur Homologiemodellierung von Proteinstrukturen. Homologiemodellierung bedeutet, dass 3D-Strukturen von Proteinen mit unbekannter Struktur aufgrund von Gemeinsamkeiten in der Sequenz zu Proteinen mit bekannten 3D-Strukturen vorhersagt werden [67].

Strukturvisualisierung - Rasmol

Rasmol ist eine Software zur graphischen Darstellung von Makromolekülen auf Basis der PDB-Datenbank. Sie wurde ursprünglich in den frühen 1990er-Jahren von Roger Sayle entwickelt und wird heute durch eine aktive Nutzergemeinschaft weiterentwickelt [68].

2.3. Software Implementierung

Viele nützliche bioinformatische Programme gibt es bereits, einige wurden hier vorgestellt. Bei sehr spezifischen Fragestellungen werden neben diesen Programmen weitere Analysen benötigt. Um diese erstellen zu können, wurden spezielle Techniken eingesetzt, welche im Folgenden vorgestellt werden.

Perl

Perl ist eine sehr mächtige plattformunabhängige Skriptsprache, welche die Programmiersprache C und Unix-Befehle verbindet. In der vorliegenden Doktorarbeit wurde Perl genutzt, um Dateien aufzubereiten und zu durchsuchen. Den Nutzungsschwerpunkt bilden:

- Motivsuchen in Gensequenzen
- Text-Mining in Textdateien
- Aufbereitung von Dateien zur Erstellung der Webserverseiten
- Erzeugung automatisiert Graphiken in der Seitenbeschreibungssprache Postscript.

Visual Basic for Applications

Visual Basic for Applications (VBA) ist eine Skriptsprache. Sie wurde aus dem BASIC-Dialekt Visual Basic (VB) abgeleitet. Vornehmlich wird sie zur Steuerung von Abläufen innerhalb der Microsoft-Office-Programme verwendet.

Webtechniken

Um die in dieser Doktorarbeit gewonnenen Erkenntnisse zu veröffentlichen, wurde ein Webservice aufgebaut.

Datenbearbeitung und Datenspeicherung für die Visualisierungen wurden mittels Perl und XML (E**x**tensible **M**arkup **L**anguage) realisiert. Weitere Perl-Programme erstellen automatisiert die Seiten des Webservices in der Sprache PHP (**P**HP: **H**ypertext **P**reprocessor). Statistische Graphiken wurden als PostScript-Dateien hergestellt und über eine DG-Bibliothek (dynamic graphic) in Bilder formatiert.

Die asynchrone Datenübertragung mittels AJAX (**A**synchronous **J**avaScript **a**nd **X**ML) regelt den dynamischen Seitenaufbau. Die JavaScript Bibliothek JQuery wird für eine effiziente Datenabfrage genutzt, z.B. für eine dynamische Vorschau und die Baumstrukturdarstellung der Organismen. Interaktionsnetzwerke werden mit Hilfe des Statistikprogramms R und der dazugehörigen IGraph Bibliothek erzeugt. Um eine optimale Darstellung zu erreichen, wird der Fruchterman Reingold Algorithmus [69] verwendet.

3. KAPITEL
Ergebnisse und Interpretationen

Dieses Kapitel beschreibt die detaillierte Implementierung der Datenanalyse, die daraus erzielten Ergebnisse und deren Interpretationen. Es gliedert sich in drei Bereiche: (1) Evaluation bioinformatischer Methoden (2) Untersuchung von Modifikationsmöglichkeiten in TCS und (3) Prozessstrukturanalyse biologischer Systeme mit Erstellung einer zugehöriger Software.

3.1. Evaluierung bioinformatischer Methoden zur Vorhersage des Interaktoms

Experimente zur Untersuchung von Proteinen sind zeit- und kostenintensiv. Aber nur durch genaue Analysen können biologische Grundlagen erforscht, sowie die Wirkungsweise von Medikamenten ergründet werden. Daher ist die Vorhersage von Proteinfunktionen, -strukturen und -interaktionen von großem Interesse für die Grundlagen- und Pharmaforschung. Für diese Analysen existieren bereits viele Programme, jedoch mit unterschiedlicher Qualität und Aussagekraft.
Im Folgenden wird kurz der generelle Ablauf von Analysen mit Vorhersageprogrammen vorgestellt. Der Schwerpunkt liegt hierbei auf der Analyse des Interaktions-Vorhersageprogramms STRING (**S**earch **T**ool for **R**ecurring **I**nstances of **N**eighbouring **G**enes) und im Besonderen auf der Validierung eines sogenannten physikalischen Vergleichswerts.

3.1.1. Evaluation und Eigenschaften verschiedener Vorhersageprogramme

Für die Vorhersage von Proteinfunktionen und Interaktionen wird prinzipiell zunächst nach bekannten Fakten in experimentellen Datenbanken (z.B. Swiss-Prot, IntAct) und Literatur-Datenbanken (z. B. PubMed) gesucht. Für die Datensammlung aus Literaturquellen ist ein mühsames Recherchieren unter Verwendung von Text-Mining-Software notwendig. Wesentlich einfacher verläuft das Suchen in Datenbanken, welche allerdings häufig den Nachteil der Unvollständigkeit mit sich bringen.
Neben der Ansammlung von vorhandenem Wissen aus Datenbanken können eigene Analysen mittels bioinformatischer Methoden aufgestellt werden. Mit Hilfe von Sequenzvergleichen können Ähnlichkeiten zu bereits bekannten Proteinsequenzen identifiziert werden. Programme zur Sequenzalinierung funktionieren sehr gut, allerdings reicht die Untersuchung der Primärstruktur häufig nicht aus. Da die Funktionalitäten meist über die Struktur gesteuert werden, könnten Programme zum Struktur- und Domänenvergleich (PredictProtein, SSPro, SSEA) interessante Hinweise liefern.
Um eine zuverlässige Vorhersagen zu erhalten, ist es allerdings empfehlenswert mehrere unterschiedliche Programme zur Validierung zu nutzen [70].

3.1.2. Vergleich der Datenbank STRING zu anderen Vorhersageprogrammen

Anfang des 21. Jahrhunderts wurde eine neue Technologie zur Vorhersage von Proteininteraktionen entdeckt, die sogenannten Genkontext-Methoden.
Am **E**uropean **M**olecular **B**iology **L**aboratory (EMBL) wurde ein Programm namens STRING entwickelt, welches unter anderem diese neue Technologie nutzt, um Proteininteraktionen vorherzusagen und diese in Form von Netzwerken visualisiert.

Neben STRING gibt es weitere Datenbanken, welche sich mit der Vorhersage von Proteininteraktionspartnern beschäftigen. Um STRING besser einordnen zu können, gibt die folgende Tabelle einen kurzen Überblick über die bekanntesten Produkte mit ähnlicher Funktionalität.

DB-Name	Beschreibung
APID	APID ist ein Webservice, der publiziertes Wissen über Protein-Protein-Interaktionen aus klein- und großmaßstäblichen Experimenten darstellt. Derzeit besitzt die Datenbank fünf Hauptdatenquellen mit über 35.000 Proteinen und 111.000 bewiesenen Proteininteraktionen [71]. http://bioinfow.dep.usal.es/apid/
MetaPPI	MetaPPI ist ein Metaserver für Schnittstellenvorhersagen. Es bietet eine Vorhersagequalität von 70% bei Enzyminhibitoren und Protein-Docking-Stellen [72]. http://scoppi.biotec.tu-dresden.de/metappi/
PIP human protein–protein interaction prediction database	PIP beinhaltet über 37.000 vorhergesagte humane Protein-Protein-Interaktionen. Die Interaktionen wurden über eine Bayesian-Methode aus einer Kombination von Orthologien, Domänenvergleichen, posttranslationalen Modifikationen und subzellulären Lokationen berechnet. PIP ordnet die Vorhersagen nach ihrer Wahrscheinlichkeit [73]. http://bmm.cancerresearchuk.org/~pip/
SPPIDER	SPPIDER benutzt für seine Vorhersagen eine Kombination aus relativen Lösungsmittelzugänglichkeiten mit hoch auflösenden Strukturen. Das SPPIDER ermöglicht die Vorhersage von Aminosäuren auf eine Proteinoberfläche und die Analyse von Protein-Protein-Komplexe auf Basis von 3D Strukturen [74]. http://sppider.cchmc.org
SCOPPI **s**tructural **c**lassification **o**f **p**rotein–**p**rotein **i**nterfaces	SCOPPI klassifiziert und annotiert Domäneninteraktionen aus bekannten Proteinstrukturen [75]. http://www.scoppi.org/
PIC **P**rotein	PIC ist ein Service zur Identifikation verschiedener Interaktionen, wie z.B. Disulfidbrücken, hydrophobe Interaktionen, ionische Bindungen, aromatische Proteininteraktionen.

Interactions **C**alculator	Außerdem ermittelt es die Oberfläche eines Proteins und den Aminosäureabstand zur Oberfläche [76]. http://crick.mbu.iisc.ernet.in/~PIC/
iHop **I**nformation **H**yperlinked **o**ver **P**roteins	iHop umfasst ein Netzwerk über Gen- und Protein- Informationen aus naturwissenschaftlicher Literatur zum Thema Phänotypen, Pathologien und Genfunktionen. Über iHop erhält man Zugriff auf mehrere Millionen PubMed Artikel [77]. http://www.ihop-net.org/UniPub/iHOP/
Cyclonet	Cyclonet ist eine integrierte Datenbank für Zellzyklusregulationen [78]. http://cyclonet.biouml.org/index.html
DIMA	DIMA ist ein Webservice zur Untersuchung von Proteindomänen-Netzwerken [79]. http://mips.helmholtz-muenchen.de/genre/proj/dima2
DOMINE	DOMINE bezeichnet eine Sammlung von belegten und vorhergesagten Proteindomänen-Interaktionen [80]. http://domine.utdallas.edu/cgi-bin/Domine
DroID	DroID ist eine Sammlung an Gen- und Protein-Interaktionen in Drosophila [81]. http://www.droidb.org/
EciD	Die *E. coli* Interaction Database (EcID) stellt Informationen zu Proteininteraktionen des Organismus *Escherichia coli* zusammen. Die Daten basieren auf den Daten von EcoCyc, KEGG und MINT [82]. http://ecid.bioinfo.cnio.es/
hp-DPI	Hp-DPI beschäftigt sich mit Proteininteraktionen im Organismus *H. pylori* [83]. http://dpi.nhri.org.tw/hp/
InterDom	InterDom ist eine Datenbank für vorhergesagte Protein-, Domänen- und Komplex-Interaktionen [84]. http://interdom.i2r.a-star.edu.sg/

Tabelle 3.1.1: Liste einiger bekannter Interaktions-Vorhersageprogramme.
Auflistung der wichtigsten zu STRING ähnlichen, Interaktions-Vorhersageprogrammen mit Beschreibung der Aussageschwerpunkte.

Tabelle 3.1.1 listet die bekanntesten Interaktionsvorhersagedatenbanken auf und beschreibt deren Funktionen und inhaltliche Schwerpunkte.

Daten-bank	Datenquelle	Inter-aktionen	Organis-men	Funktion	Darstellung
APID [71]	BIND, BioGRID, DIP, HPRD, IntAct and MINT	322.579	15	Protein-Protein-Interaktionen	Interaktions-netzwerk
PIP [73]	Yeast-to-hybrid, Kristallographie, Massenspectros-kopie und Affinitäts-aufreinigung	37.000	Human, Ratte, Hefe	Protein-Protein-Interaktionen	Liste der Homologe sortiert nach einem Qualitäts-score
SCOPPI [75]	PDB		analog PDB	Protein-Protein-Interaktionen	Multiple Alinierung
PIC [76]	Vorhersage PDB-Basis			Protein-Protein-Interaktionen	Rasmol-Datei
iHop [77]	Literatur			Protein-Protein-Interaktionen	Text hervor-gehoben
Cyclonet	SBML und CellML		Eukaryoten	Protein-Protein-Interaktionen im Zellzyklus	Textuell
Dima	STRING, Pfam			Interaktionen zwischen Protein-Domänen	Netzwerk
DOMINE	PDB, Pfam	26.219		Protein-Domänen Interaktionen	Textuell
DroID	Hochdurchsatz-daten		*Drosophila*	Protein-Protein-Interaktionen	Textuell
EciD	EcoCyc, KEGG, MINT		*E. coli*	Interaktionen in Protein Komplexen	Netzwerk
hp-DPI		1.462	*H. pylori*	Protein-Protein-Interaktionen	Netzwerk
InterDom	DIP, BIND, PDB			Protein-Domänen Interaktionen	Textuell

Tabelle 3.1.2: Vergleich der wichtigsten Interaktionsvorhersagedatenbanken.
Vergleich der wichtigsten zu STRING ähnlichen Interaktions-Vorhersageprogrammen mit Angabe der Datenquellen, Datenmengen, Organismen und Darstellungen.

Tabelle 3.1.2 zeigt die Unterschiede zwischen STRING und den anderen Interaktionsvorhersagedatenbanken bezüglich Datenquellen, Datenmenge, Organismen, Funktionen und Visualisierungsformen auf. Die größten Unterschiede liegen darin, dass STRING sehr viele unterschiedliche Datenquellen als Basis für seine Interaktionsvorhersagen verwendet und damit auf einer sehr breiten Datenbasis mit vielen Organismen fundiert. Die Datenintegration beruht nicht auf einer Metasuchmaschine, sondern auf einer physikalischen Integration der Datenquellen.

Weitere Vorteile von STRING gegenüber den anderen genannten Programmen sind:

- Die Integration des Genkontext-Prinzips: Damit erweitert STRING seine Vorhersagekraft um ein bislang von anderen Datenbanken nicht genutztes Prinzip.
- Die enorme Flexibilität der Visualisierung (textuell und graphisch) und die große Datenmenge, auf welcher die Berechnung der Interaktionspartner und deren Wahrscheinlichkeit eine wirkliche Interaktion darzustellen, basieren.
- STRING liefert nicht nur Angaben über eine mögliche Interaktion, sondern auch zusätzlich die Zuverlässigkeit der Interaktionsvorhersage.
- Regelmäßige Aktualisierung der Daten.

Damit vereint STRING die wichtigsten Datenquellen und Methoden der oben genannten Programme und erweitert sie um eine weitere mächtige Vorhersage-Methode sowie um eine flexible Darstellungsmöglichkeit.

3.1.3. Erstellung eines neuen Releases der Datenbank STRING (Version 6.3)

Bis zur Version 6.2 lieferte STRING keine Information über die Art der vorhergesagten Interaktion.

Während meines Studiums der Bioinformatik an der FH-Giessen-Friedberg arbeitete ich im Rahmen meiner Diplomarbeit an der Weiterentwicklung von STRING am EMBL in Heidelberg mit. Der Schwerpunkt der Arbeit lag in der Mitgestaltung der STRING Version 6.3. In dieser Zeit wurde ein Algorithmus zur Berechnung eines physikalischen Vergleichswertes (*physical_score*) entwickelt. Dieser beschreibt die Wahrscheinlichkeit für eine physikalische Interaktion (direkte physikalische Interaktion zweier Proteine).

Der physikalische Vergleichswert für eine Interaktion wird zwischen 0 (0%) und 1 (100%) skaliert. Für jede Interaktion der STRING Datenbank wurde der physikalische Vergleichswert berechnet und abgespeichert. Somit kann in der STRING Version 6.3 für jede in STRING vorhandene Interaktion der physikalische Vergleichswert als numerische Wahrscheinlichkeit (zwischen 0 und 1) und graphisch als Farbskala (von rot für physikalisch bis grau für nicht-physikalisch) innerhalb eines STRING Interaktionsnetzwerkes angezeigt werden.

Detaillierte Angaben zu diesem Thema können meiner Diplomarbeit entnommen werden (*Data mining in Protein-Networks: Separating physical from functional interactions*, 2005, FH-Giessen Friedberg).

Die Qualität des physikalischen Vergleichswertes wurde im Rahmen der Doktorarbeit über zwei unterschiedliche Wege evaluiert:

1. Vergleich des Berechnungsalgorithmus mit anderen ähnlichen Algorithmen
2. Bewertung von konkreten Interaktionsbeispielen anhand aktueller Literatur.

3.1.4. Evaluation des physikalischen Vergleichswerts

Zur Bewertung des Algorithmus im Rahmen der Dissertation wurde zunächst der physikalische Vergleichswert auf der gewachsenen Datenbasis von STRING für alle STRING Interaktionen neu berechnet und in der STRING Datenbank abgespeichert [85].

Zur Bewertung der Algorithmusqualität zur Berechnung des physikalischen Vergleichswertes wurde eine ROC-Kurve (**R**eceiver **O**perating **C**haracteristic) erstellt. Die ROC-Kurve ist eine statistische Methode um Ergebnisse zu evaluieren. Sie gibt die Güte eines Systems bezüglich ihres Signal-Rauschverhältnis an [86]. Eine ROC-Kurve nahe der Diagonale deutet auf einen Zufallsprozess hin. Die ideale ROC-Kurve steigt zunächst senkrecht an (die Trefferquote liegt nahe bei 100%, während die Fehlerquote anfangs noch nahe bei 0% bleibt), erst danach steigt die falsch-positive Rate an. Eine ROC-Kurve, die deutlich unterhalb der Diagonalen bleibt, deutet darauf hin, dass die

Werte falsch interpretiert wurden. Je größer die Fläche unterhalb der Kurve ist, desto zuverlässiger ist der Algorithmus.

Zur Untersuchung der Sensitivität des physikalischen Vergleichswerts mittels ROC-Kurve wurde je ein Referenzdatensatz für physikalische Interaktionen (aus MIPS, PDB) und für nicht-physikalische Interaktionen (aus KEGG) erstellt. Alle in STRING gespeicherten physikalischen Vergleichswerte wurden mit den beiden Referenzdatensätzen verglichen. Für jeden physikalischen Vergleichswert der STRING Datenbank wurde die relative Häufigkeitsverteilung in Form von wahr-positiven Vorhersagen auf die falsch-positiven Vorhersagen gegeneinander aufgetragen. Die wahr-positiven Vorhersagen beschreiben die tatsächlich richtig vorhergesagten physikalischen Interaktionen und zeigen damit die Sensitivität an. Die falsch-positiven Vorhersagen stehen für die falsch vorhergesagten physikalischen Interaktionen und beschreiben die Spezifität eines Modells.

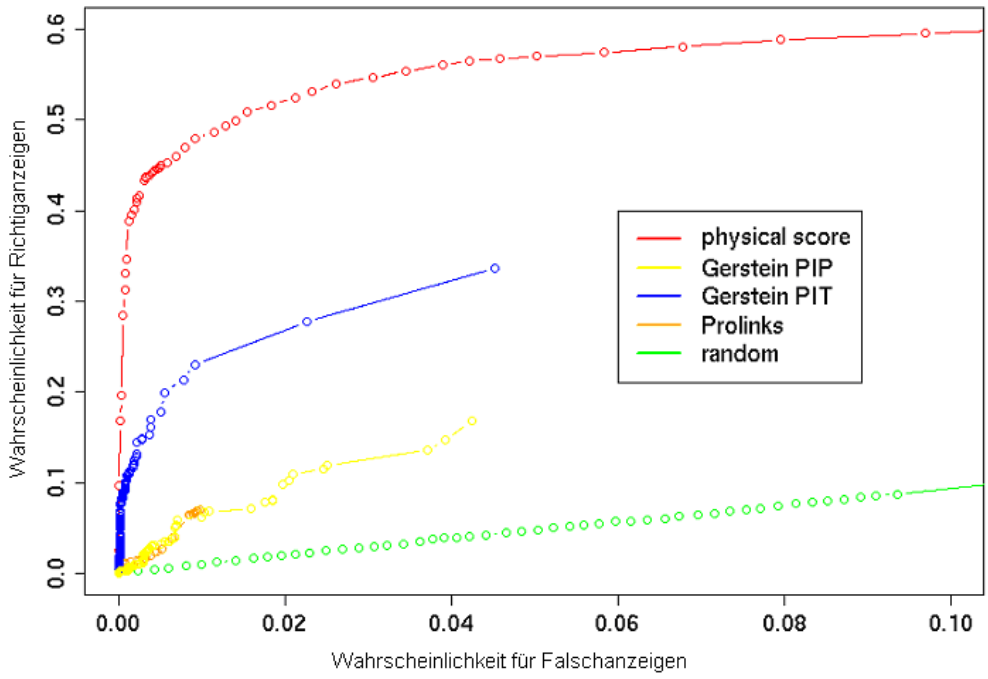

Abbildung 3.1.1: Untersuchung der Datenqualität.
Darstellung eines Ausschnitts einer ROC-Kurve für den physikalischen Vergleichswert der STRING Daten (rot) im Vergleich zu anderen physikalischen Interaktionsvorhersagen von Gerstein (gelb, blau) und Prolinks (orange). Zum besseren Vergleich ist ein Zufallsdatensatz (grün) aufgetragen.

In **Abbildung 3.1.1** ist die ROC-Kurve (rot) für den physikalischen Vergleichswert von STRING abgebildet. Diese Kurve zeigt, dass der physikalische Vergleichswert bereits bei einer sehr geringen falsch-positiven Rate eine sehr hohe wahr-positive Rate besitzt. Zum Vergleich ist eine ROC-Kurve für einen Zufallsdatensatz in grün dargestellt (Diagonale).
Um das Ergebnis für die Qualität des physikalischen Vergleichswertes besser vergleichen zu können, wurden in dem Diagramm drei weitere ROC-Kurven anderer Techniken zur Vorhersage physikalischer Interaktionen dargestellt. Diese ROC-Kurven beinhalten zwei, auf der Bayesian-Methode beruhende, Datensätze von Gerstein; den *PIP*-Datensatz aus den *probabilistic interactomes-predicted* (gelb) und den *PIT*-Datensatz aus der Kombination der *PIP*- und *PIE*- (*probabilistic interactomes-experimentell*) Daten zu einem *PI-total* (blau) [87]. Zusätzlich beinhaltet die Abbildung den Datenbestand der Prolinks Datenbank (orange) [88]. Diese drei Datensätze beschäftigen sich ebenfalls mit der Vorhersage von physikalischen Interaktionen, bedienen sich allerdings anderer Algorithmen und Datensätze.

Der Vergleich mit den Datensätzen von Gerstein (gelb, blau) und den Daten aus der Datenbank Prolinks (orange) zeigt, dass die ROC-Kurve für den physikalischen Vergleichswert (rot) nicht nur besser ist als die des Zufallsdatensatzes (grün), sondern auch besser als die Kurven der Vergleichsdatensätze. Die Güte hinsichtlich Selektivität und Spezifität der berechneten physikalischen Vergleichswerte lässt sich deutlich an der Fläche unter der Kurve belegen.

3.1.5. Evaluation der Ergebnisse durch konkrete Beispiele

Nachdem im vorausgegangenen Abschnitt demonstriert werden konnte, dass die Wahrscheinlichkeit für eine physikalische Interaktion, durch die Berechnung des physikalischen Vergleichswertes mit dem beschriebenen Algorithmus, gut bestimmt werden kann, sollen in einer zweiten Evaluation die vorhergesagten Werte für konkrete Beispiele überprüft werden.

Bei nicht oder nur wenig annotierten Proteinen ist eine Überprüfung des physikalischen Vergleichswertes nur experimentell möglich. Daher soll hier anhand einiger konkreter Beispiele in Prokaryoten und Eukaryoten die Güte des physikalischen Vergleichswertes unter Verwendung von Literaturbeispielen validiert werden. Für eine bessere Überprüfbarkeit durch Homologe, Literatur oder sonstigen Zusammenhänge, umfassen die Interaktionsbeispiele bekannte und unbekannte Interaktionen. Im folgenden Abschnitt werden zunächst die herausgesuchten Interaktionen beschrieben, der von STRING berechnete physikalische Vergleichswert angegeben und die aus der Literatur aufgefundene Evaluation aufgeführt. Darunter wird das Proteininteraktionsnetzwerk aus STRING aufgezeichnet, in dem sich die zu untersuchende Interaktion befindet. Die Linie zwischen den Proteinen (Knoten) zeigt mittels Farbgraduierung den Wert des physikalischen Vergleichswerts und beschreibt die Wahrscheinlichkeit einer Interaktion physikalisch zu sein (rot: zu 100% eine physikalische Interaktion, grau: zu 0% eine physikalische Interaktion).

Eukaryotische Beispiele

1. Zunächst wird die Interaktion zwischen den beiden RNA-Polymerase II Proteinen RPB2 und RPB4 in Hefe untersucht und in **Abbildung 3.1.2** abgebildet. Der von STRING vorhergesagte physikalische Vergleichswert für diese Interaktion beträgt 1.0, somit ist sie als 100%-ige physikalische Interaktion vorhergesagt. Im STRING Netzwerk ist die Interaktion zwischen RPB2 (rot) und RPB4 (lindgrün) mit einer roten Linie dargestellt, was für eine starke Physikalität der Interaktion steht. Zur Verdeutlichung wurde die Interaktion neben der Darstellung im Netzwerk zusätzlich separat dargestellt. In der Literatur ist annotiert, dass RPB2 und RPB4 Untereinheiten eines größeren Komplexes innerhalb der Polymerase II bilden. Da Untereinheiten tatsächlich direkt physikalisch aneinander binden, konnte die vorhergesagte hohe Physikalität bestätigt werden.

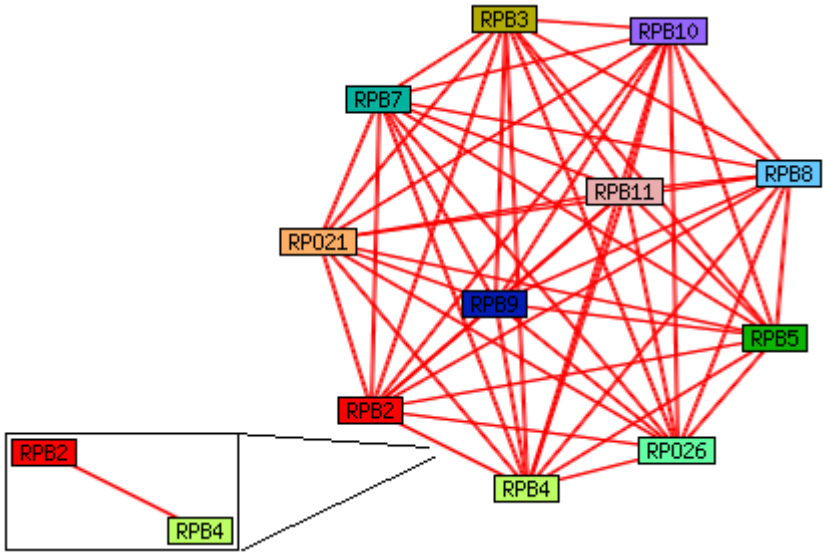

Abbildung 3.1.2: Interaktion zwischen den beiden RNA-Polymerase II Proteinen RPB2 (rot) und RPB4 (grün) in Hefe. Dargestellt ist die Interaktion im Netzwerk mit Ausschnittsvergrößerung der betroffenen Region.

2. a) Auch das zweite Beispiel beschäftigt sich mit Untereinheiten. In diesem Fall handelt es sich um die Untereinheit der Cell Division Control-Proteine CDC12 (rot), CDC3 (hellgrün) und CDC11 (ocker). Diese drei Zellteilungs-Kontroll-Proteine sind Bestandteil eines größeren Komplexes. Der vorhergesagte physikalische Vergleichswert beträgt 1.0 und entspricht damit der biologischen Erklärung.

   b) Zusätzlich bindet CDC11 (ocker) mit einem physikalischen Vergleichswert von 0.9 an das Protein KCC4 (dunkelblau). KCC4 ist vermutlich eine Serin/Threonin-Proteinkinase und ein Homologes zur Proteinkinase Hs11/Nik1 und Gin4. Mit Hilfe von Immunpräzipitation konnte bestätigt werden, dass diese Homologe an KCC4 binden (McMurray et al., 2011). Die hohe Wahrscheinlichkeit der Physikalität wird demnach mittels Literatur bestätigt.

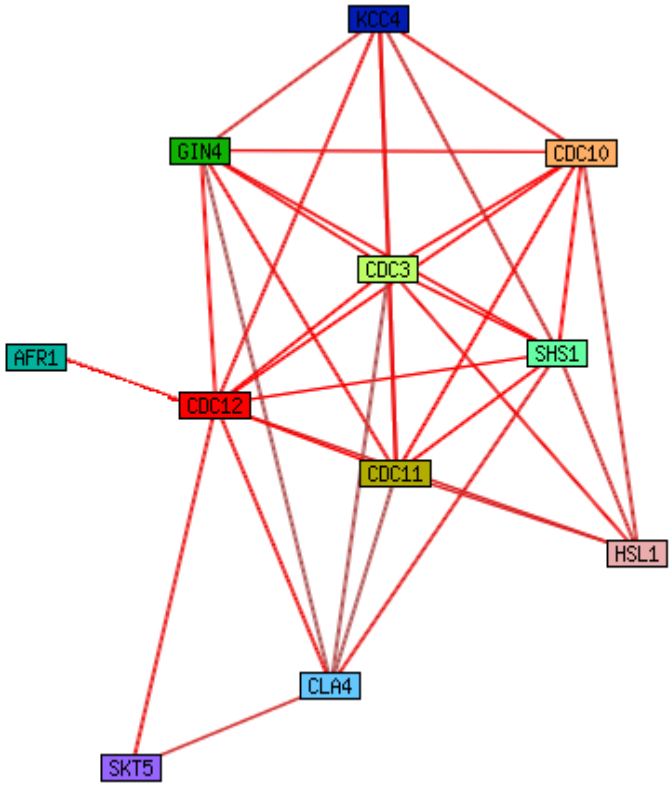

Abbildung 3.1.3: Proteinnetzwerk aus Untereinheiten der Zellteilungskontrolle (Cell Devision Control): CDC12 (rot), CDC3 (hellgrün) und CDC11 (ocker) in Hefe.

3. a) Das prolinreiche Protein Las17 (rot) der Hefe ist ein Homolog des humanen WASP-Proteins (Wiskott-Aldrich Syndrome Protein). Seine Aufgabe ist es, das Aktin ähnliche Protein (Arp2 - hellgrün) zu aktivieren, welches sich in einem Komplex zu Arp3 befindet (Evangelista et al., 2000). Der physikalische Vergleichswert für die Interaktion zwischen Las17 und Arp2 beträgt 0.9. Die Validierung der Interaktion erfolgt in diesem Fall durch die Homologie zu WASP.

   b) Der physikalisch Vergleichswert für die Interaktion zwischen den Proteinen Las17 (rot) und BZZ1 (dunkelblau) beträgt 0.9. BZZ1 ist das WASP/Las17-interagierendem Protein, welches vermutlich an der Regulation der Aktin-Polymerisation beteiligt ist (Soulard et al., 2005). Eine tatsächliche physikalische Interaktion von Las17 und BZZ1 liegt daher nahe.

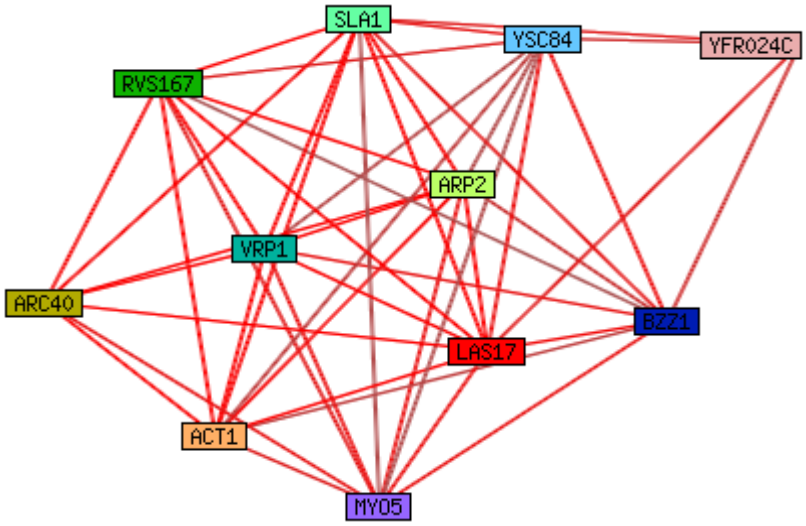

Abbildung 3.1.4: Protein Las17 (rot) im Netzwerk mit Protein Arp2 (hellgrün) und BZZ1 (dunkelblau).

Prokaryotische Beispiele

4. Auch für Prokaryoten funktioniert der physikalische Vergleichswert. RPOA (rot) und RPOB (orange) sind Untereinheiten (A und B) der DNA-gerichteten RNA-Polymerase. Damit sind sie Untereinheiten innerhalb einer quaternären Struktur der *E. coli* RNA-Polymerase. Der berechnete physikalische Vergleichswert beträgt 0.9 und zeigt daher die physikalische Interaktion der beiden Untereinheiten korrekt an.

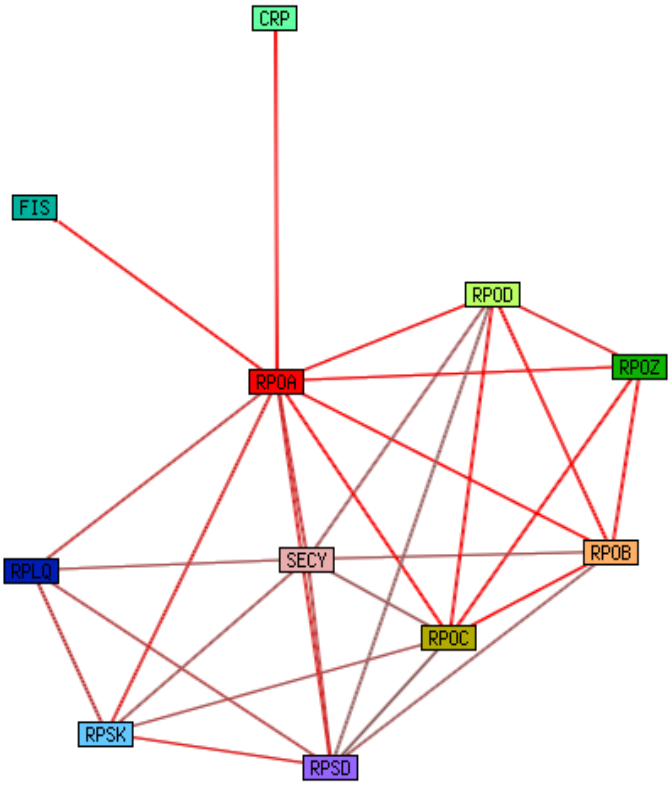

Abbildung 3.1.5: Untereinheiten RPOA (rot) und RPOB (orange) der DNA-gerichteten RNA-Polymerase in *E. coli*.

5. a) Der berechnete physikalische Vergleichswert beträgt 0.9 für die Interaktion zwischen TolB (rot) und TolA (grün). Das periplasmatische Protein TolB ist Teil des Zellhüllkomplexes Tol. TolB interagiert physikalisch mit der TolA-Untereinheit (grün). Aufgrund des Hinweises auf die Untereinheiten, scheint der vorhergesagte Vergleichswert realistisch.

   b) Ein komplexeres Beispiel beschreibt die Interaktion zwischen PAL (orange) und TolB (rot). Der physikalische Vergleichswert beträgt 0.9. Die Evaluation kann in diesem Fall nur über logische Zusammenhänge der Funktionen hergestellt werden. Das PAL Protein ist noch relativ unbekannt und spielt vermutlich eine Rolle in der Aufrechterhaltung der Bakterienhülle, ebenso wie TolB (Walburger et al., 2002).

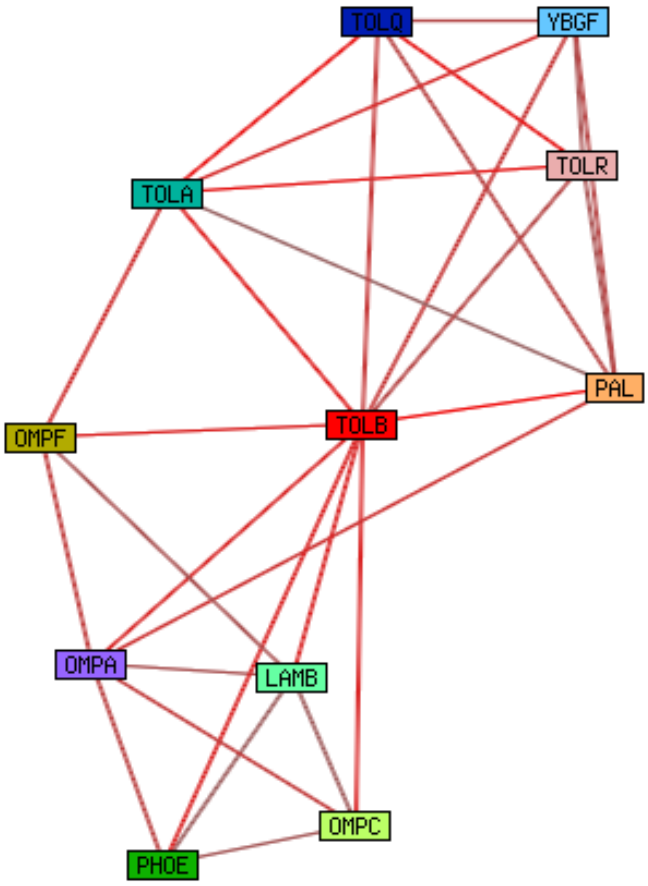

Abbildung 3.1.6: Proteinnetzwerk um das *E. coli* Protein TolB (rot), TolA (grün) und das Protein PAL (orange).

6. a) Der vorhergesagte physikalische Vergleichswert der Interaktion zwischen den Proteinen MelR (rot) und RopA (orange) beträgt 0.7. Das Protein MelR in *E. coli* ist ein Transkriptionsaktivator der RNA-Polymerase (RopA_Ecoli) und bindet somit physikalisch an den alpha-Strang von RopA (Grainger et al., 2004).

   b) Da MelR die Bildung von MelA reguliert, gehört MelA zu den funktionalen (indirekten) Interaktionspartnern von MelR. Physikalisch findet jedoch keine Bindung statt (Grainger et al., 2004). Der physikalische Vergleichswert beträgt daher lediglich 0.3, während der STRING Score für eine Interaktion 0.78 beträgt. Damit ist die Interaktion korrekt als nicht-physikalische Interaktion vorhergesagt.

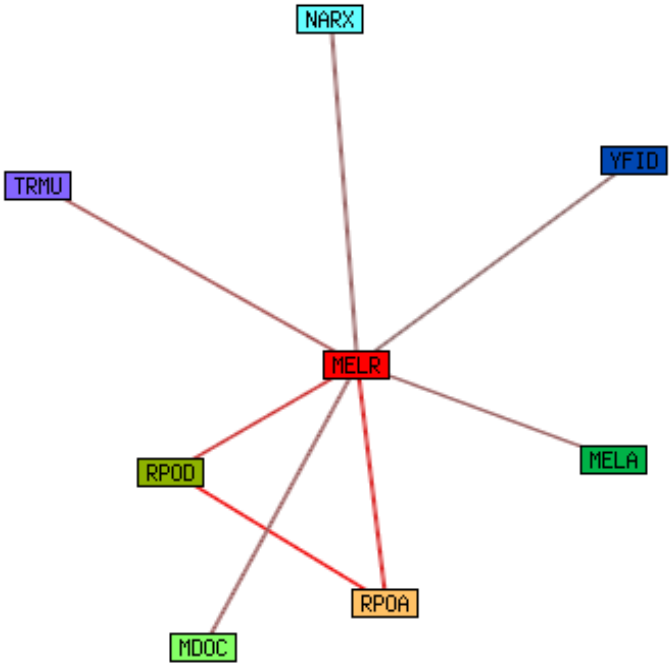

Abbildung 3.1.7: Transkriptionsaktivator MelR (rot) aus *E. coli* im Interaktionsnetzwerk mit MelA (grün) und der RNA-Polymerase (RopA_Ecoli) (orange).

Diese Beispiele zeigen, dass der physikalische Vergleichswert sehr gute Ergebnisse für die relativ einfach zu detektierenden Untereinheiten von größeren Proteinkomplexen liefert (Beispiele 1, 2a, 4).

Sie zeigen weiterhin, dass kaum annotierte physikalische Interaktionen zwischen unterschiedlichen Proteinen ebenfalls aufgedeckt werden können (Beispiele 2b, 3a, 3b, 5a, 5b, 6a). Funktionale Interaktionen werden ebenfalls korrekt durch niedrige physikalische Vergleichswert dargestellt (Beispiel 6b).

Allerdings muss festgestellt werden, dass der physikalische Vergleichswert nur bei gesicherten Interaktionen (mit hohen Wahrscheinlichkeiten, also hohen STRING Scores) zuverlässig funktioniert.

Teile dieses Kapitels wurden bereits veröffentlicht [70, 85].

3.1.6. Automatisierung einer Analysepipeline in der Pharmaindustrie

Auch die Pharmaindustrie nutzt Methoden der Bioinformatik; u.a. für die Automatisierung von Nassexperimenten und zur Reduktion von Experimenten durch *in-silico* Vorhersagen. Die Spannbreite des Einsatzes reicht von der *target identification* (Identifikation von möglichen Therapieansätzen gegen Krankheiten), über die *lead identification* (Detektion und Optimierung chemischer Strukturen, die mögliche Arzneimittelkandidaten darstellen) bis hin zur Vorhersage von Reaktionen chemischer Strukturen mit verschiedenen biologischen Komponenten eines Organismus.

Zur Beantwortung der komplexen Fragestellungen ist die interdisziplinäre Zusammenarbeit zwischen Biologie, Chemie, Bioinformatik und Technik unerlässlich.

Im Rahmen eines Projektes bei der Pharmafirma Boehringer Ingelheim in Biberach war ich führend an der Realisierung eines Automatisierungsprojektes beteiligt. Ziel war es, einen flexibel anwendbaren Arbeitsablauf für die verschiedenen Untersuchungsmethoden zu schaffen, welche die Auswirkungen chemischer Substanzen auf den Körper adressieren. Dieser Ablauf sollte möglichst wenige manuelle Schritte beinhalten, die Prozesse beschleunigen und sie im Prozess fehlerunanfälliger gestalten.

Im Folgenden wird der in diesem Projekt konzipierte modulare Arbeitsablauf genauer beschrieben:

1. Über ein Auftragssystem (Eigenentwicklung der Fa. Boehringer) werden chemische Substanzen für bestimmte Untersuchungsmethoden beauftragt.
2. Mit Hilfe dieses Auftragssystems werden Messpakete (Ansammlung von Substanzen, die in einem Assay in einem Messdurchgang gemessen werden können) zusammengestellt.
3. Die zu untersuchenden Substanzen werden über das Auftragssystem in einem Substanzlager bestellt und nach deren Lieferung der Substanzeingang registriert.
4. Zu jedem Messpaket wird automatisch eine Textdatei mit allen benötigten Parametern für die Messung und Auswertung erzeugt.

5. Diese Textdatei wird in die Messgeräte zu deren Steuerung eingelesen. Die Geräte werden automatisch mit Hilfe der Parameter eingestellt und die Messung wird gestartet.
6. Das Ergebnis der Messungen wird in Textdateien geschrieben.
7. Die Parameterdatei wird gemeinsam mit der Ergebnisdatei in ein Auswertesystem (AssayExplorer® der Firma Accelrys®) geladen.
8. In dem Auswertesystem sind Vorlagen gespeichert wie die unterschiedlichen Analysemethoden auszuwerten sind.
9. Die Parameter, die ausgewerteten Ergebnisse sowie die Rohdaten werden gemeinsam mit Statusinformationen automatisiert in ein Data Warehouse geschrieben.
10. Eine Visualisierungssoftware (Spotfire® der Firma Tibco®) bietet die Möglichkeit, diese Datenbank abzufragen und die Daten zu visualisieren.

Der große Vorteil dieses automatisierten, integrierten Arbeitsablaufs liegt darin, dass durch die Reduktion der manuellen Schritte nicht nur der Datendurchsatz erhöht werden konnte, sondern zugleich wurde der gesamte Ablauf auch wesentlich fehlerunanfälliger. Im Besonderen die Speicherung aller Roh- und Parameterdaten liefert noch viele weitere Möglichkeiten der Datennachbereitung, wie z.B. die Auswertung historischer oder qualitativer Kontrollen.

Das strukturierte Ablegen und Auffinden der Daten legt nicht nur die Grundlage für die Kontrolle der Datenqualität, sondern auch für die Erstellung von *in-silico* Vorhersagen auf Basis chemischer Strukturen, deren chemischen Eigenschaften und den vorhandenen experimentellen Ergebnissen.

Erste Ergebnisse dieses Projektes wurden bereits veröffentlicht [89] und bei der Spotfire Userkonferenz in Japan und Frankreich, sowie bei der SBS Annual Conference in Orlando, Florida vorgestellt.

3.2. Flexibilität und Modifikationen in Zweikomponenten-Systemen (TCS)

Die grundsätzliche Struktur von Zweikomponenten-Systemen ist bereits bekannt [90]. Allerdings besteht die Vermutung, dass Zweikomponenten-Systeme eventuell noch flexibler sind als bisher angenommen und für die synthetische Biologie daher ein ideales System darstellen.

Demzufolge beschäftigt sich dieses Kapitel mit der Flexibilität und den Modifikationsmöglichkeiten von Zweikomponenten-Systemen in Sequenz, Struktur und Funktionalität. Es gliedert sich in vier Unterkapitel:

- Analyse eines konkreten Zweikomponenten-Systems in *Helicobacter pylori.*
- Analyse der generellen Struktur von Zweikomponenten-Systemen, deren Abstraktionsmöglichkeiten und Flexibilität.
- Analyse von Modifikationsmöglichkeiten durch Veränderungen in der Sequenz, anhand einiger Beispielorganismen (*Escherichia coli,* unterschiedlichen *Salmonella*-Stämmen*, Staphylococus aureus, Bacillus subtilis, Salmonella typhimurium* und *Pseudomonas aeruginosa*).
- Analyse der Modifikationsmöglichkeiten durch Domänenvermischung und Domänenentartung am Beispiel der Organismen *Listeria monocytogenes, Legionella pneumophila* und *Mycoplasma pneumoniae*.

3.2.1. Erstellung einer Konsensussequenz für die Bindestelle von ArsR in *Helicobacter pylori*

Der säureresistente Organismus *H. pylori* ist der Auslöser einiger Erkrankungen im sauren Magen-Darm-Trakt. Die pH-abhängige Genregulation und damit die Fähigkeit im sauren Milieu des Magens zu überleben, spiegelt sich im säureinduzierten ArsRS Zweikomponenten-System wieder. Die Tatsache, dass es sich bei dem Regulationsprotein und Transkriptionsregulator ArsR, aber nicht bei dem Sensorprotein ArsS, um einen essentiellen Faktor handelt, deutet darauf hin, dass ArsR auch in seiner nicht phosphorylierten Form eine wichtige Funktion bei der pH-unabhängigen Genregulation außerhalb eines Zweikomponenten-Systems zukommt [91].

Der Transkriptionsregulator ArsR reguliert die Genregulation in unterschiedlichen Zuständen:

- Kontrolle durch ArsR + Phosphor
- Kontrolle durch ArsR + Phosphor + Säure
- Kontrolle durch ArsR + hohe Konzentration von ArsR
- Kontrolle durch ArsR

Ziel dieses Kapitels war es herauszufinden, ob eine allgemeingültige Konsensussequenz für die ArsR Promotorbindestelle in *H. pylori* erstellt werden kann.
Dazu wurden aus bisher unveröffentlichten Expressionsdaten einer anderen Arbeitsgruppe der Universität Würzburg nur diejenigen Gene herausgesucht, welche experimentell gesichert durch ArsR mit Phosphor (ArsR~P) reguliert werden. Diese Gene sind mit ihren Genlängen und Startpunkten in **Tabelle 3.2.1** aufgelistet.

Gen	**Transkriptionsstart relativ zum Startcodon**	**Position der Binderegion relativ zum Transkriptionsstart**	**Länge**
urea	-57	(-21) – (-74)	54 bp
urea		(-105) – (-139)	34 bp
ureI	-65	(-3) – (-50)	48 bp
amiE	-44	(-14) – (-69)	56 bp
amiF	-48	(-13) – (-50)	38 bp
rocF	-29	(-6) – (-67)	62 bp
hp1408	+2	(-22) – (-49)	28 bp
hp0119	-42,-43	(-22) – (-47)	26 bp
hp1186	-55	(-52) – (-98)	46 bp

Tabelle 3.2.1: Experimentelle ArsR~P Bindestellen (abgeleitet aus DNase I-Footprint Analysen). Die Tabelle enthält die Gennamen, Transkriptionsstart, Bindestellen-Start- und -Endpunkt und die Bindelängen.

Zusätzlich wurde die bereits annotierte ArsR~P Bindestelle aus der Datenbank DBTBS herausgesucht. **Tabelle 3.2.2** enthält Sequenz und Position des Promotors.

Lokation	**Genomposition**	**Bindesequenz**
-56:-39	2656932..2656949	AAATAAATTGATTTATTT

Tabelle 3.2.2: Promotorbindestelle von ArsR, bei denen ArsR negativ regulierend wirkt. Annotiert in der Datenbank DBTBS.

Mittels Blockmaker wurde aus den gesicherten ArsR~P Bindestellenbereichen und der in DBTBS annotierten Bindestelle eine Konsensussequenz erzeugt. **Abbildung 3.2.1** zeigt das erzeugte Konsensusmotiv für die ArsR~P-Bindestelle.

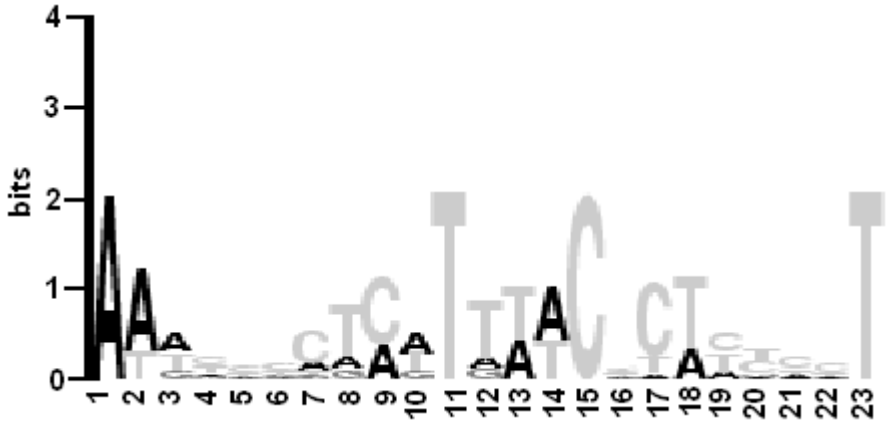

Abbildung 3.2.1:Sequenzlogo für ein mögliches Konsensusmotiv für die ArsR~P-Bindestelle.
Die dazugehörende lokale Alinierung lautet: aaAAATTTCTCATTTACGCTTTAATagttttcttacaa. Die Buchstabengröße zeigt die Konserviertheit der Nukleotide an.

Als Überprüfung der Konsensussequenz wurden weitere mögliche Promotorregionen (im Bereich von -30 bis -60 relativ zum Transkriptionsstart) der durch Expressionsdaten gefundenen Gene herausgesucht. Mittels Visual Basic for Application Skript (*helpy_arsr.vba*) wurden die Sequenzen gemeinsam mit Zufallssequenzen gegen die gesicherte Konsensussequenz verglichen und ein Qualitätsvergleichswert berechnet. Dieser Qualitätsvergleichwert wurde zur besseren Überprüfung zunächst mit einem selbst geschriebenen Algorithmus, anschließend mit einer Alinierung vom EBI überprüft. Für den eigenen Algorithmus wurde aus den Wahrscheinlichkeiten der Konsensussequenz und der *Blossum 62* (Substitutionsmatrix) eine PSSM (*possition specific scoring matrix*) für den Konsensus errechnet. Der Qualitätsvergleichswert wurde berechnet aus der Summe der PSSM-Werte und der zu untersuchenden Sequenz pro Nukleotid. D.h. je höher der Vergleichswert, desto besser passt die Sequenz zum Konsensus.

Abbildung 3.2.2 trägt die berechneten Vergleichswerte gegen Sequenzen auf. Die Sequenzen werden unterschieden in durch Expression gefundene Gene (rot, blau) und selbst erzeugte Zufallssequenzen (grün).

Außerdem wurde zwischen den durch ArsR aktivierten Genen (rot) und reprimierten Genen (blau) unterschieden. Diese Visualisierung zeigt, dass die drei Gruppen durch die Nutzung des Vergleichswerts deutlich voneinander abgetrennt werden können.

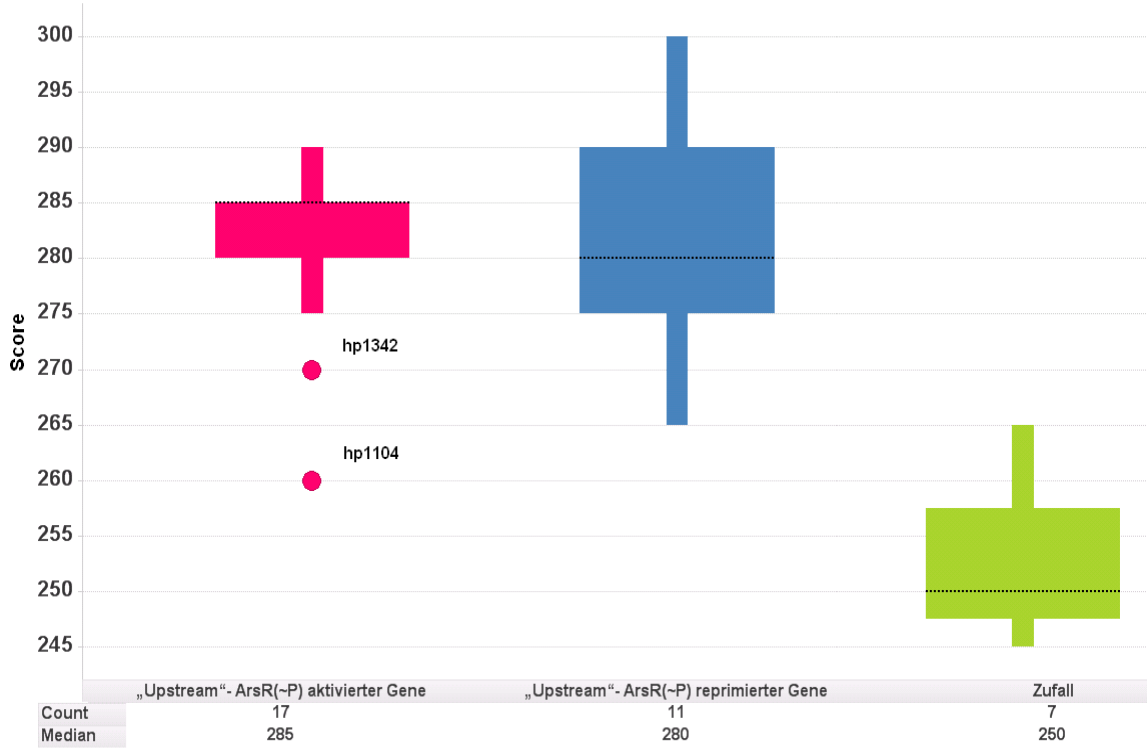

Abbildung 3.2.2: Das Boxplotdiagramm zeigt die Güteverteilung der Konsensussequenz. Die Zufallssequenzen (grün) haben durchgängig einen schlechten Vergleichswert. Die experimentellen Sequenzen (blau, rot) besitzen bis auf hp1104 einen guten Vergleichswert. In rot die Score-Verteilung der ArsR aktivierten Gene, in blau die reprimierten Gene.

Insgesamt ist zu erkennen, dass die Zufallssequenzen einen wesentlich niedrigeren Vergleichswert besitzen als die Expressionsdaten und damit die Konsensussequenz eine gute Spezifität besitzt. Der rote Boxplot visualisiert ArsR aktivierte Gene und der blaue Boxplot beinhaltet die reprimierten Gene. Die Höhe des Vergleichswertes ist für aktivierte und reprimierte Gene vergleichbar. Allerdings ist die Standardabweichung innerhalb des roten Boxplot geringer. Daraus kann geschlossen werden, dass die gefundene Promotorsequenz grundsätzlich für beide Gentypen nutzbar ist, für ArsR aktivierte Gene jedoch etwas besser zu funktionieren scheint. Obwohl hp1104 und hp1342 als ArsR aktivierte Gene markiert sind, besitzen deren Sequenzen einen relativ niedrigen Vergleichswert. Erklärbar ist diese Tatsache dadurch, dass es sich hierbei um Expressionsdaten handelt, bei denen die erhöhte Expression nicht zwangsläufig auf ArsR zurückgeführt werden kann, sondern auch andere Faktoren Einfluss haben könnten. Auf diese Weise können zusätzliche Informationen über die Qualität der Expressionsdaten gesammelt werden.

Eine Unterscheidung auf Kontrolle durch ArsR + Phosphor + Säure konnte aufgrund der geringen Datenmenge nicht gefunden werden.

3.2.2. Genereller Konsensus in TCS und Domänen-Kontext

Um die Konservierung und Flexibilität von Zweikomponenten-Systemen über die Organismen hinweg zu überprüfen, wurde eine große Sequenz-Datensammlung aus öffentlichen Datenbanken angelegt. Daraus wurden Sequenz- und Strukturalinierungen generiert, um aus bekanntem Wissen neue Erkenntnisse über Modifikationsmöglichkeiten und Variabilität ableiten zu können. Die Sequenzen stammen aus den Modellorganismen *E. coli,* unterschiedlichen *Salmonella* Stämmen*, S. aureus, B. subtilis, S. typhimurium, P. aeruginosa, L. monocytogenes* und *L. pneumophila.*

Um einen generellen Überblick über die unterschiedlichen TCS zu erhalten, wurde die Pfam-Datenbank nach bekannten TCS untersucht (insgesamt 11912 Proteinfamilien). Für die Sensorproteinsuche wurden die Domänen HisKA, HATPase_c und Hpt als Anhaltspunkte genutzt. Für das Regulationsprotein wurde nach der Responseregulator-Domäne und nach dem HTA_AraC Motiv gesucht.
Neben den klassischen Sensorproteinen (39854 Proteine) und Regulationsproteinen (48863 Proteine), wurden Hybridtypen gefunden (10147 Hybridkandidaten), also Proteine welche sowohl sensorische als auch regulatorische Funktionen ausüben.
Trotz aller Varianz konnten innerhalb der Proteine insgesamt nur 12 grundsätzlich unterschiedliche Kombinationen in allen Sequenzen gefunden werden.
Sogar in Eukaryoten konnten 109 Proteine mit Teilen aus TCS gefunden werden.
Tabelle 3.2.3 listet die unterschiedlichen Kombinationsmöglichkeiten innerhalb der Sensor- und Regulationsproteine nach ihrer Häufigkeit sortiert auf. Über die Kombinationen zwischen Sensor- und Regulationsproteinen gibt die Tabelle keine Aussage.

Kombinationen im Sensor (Pfam-Familien)	Kombinationen im Regulator (Pfam-Familien)
HisKA + HATPase_c + (n * HAMP + m * PAS + p * Hpt)[1]	Response_reg + Trans_reg_C
HATPase_c	Response_reg * s
HAMP	Response_reg + GerE
His_kinase + HATPase_c	Response_reg + HTH
HisKA + HATPase_c	Response_reg + LytTR
HWE_HK	Response_reg + HisKA
HisKA_2 + HATPase_c	Response_reg + CheB or CheW
HisKA_2	Response_reg + Sigma
HisKA_3	Response_reg + Spo
HisKA	Response_reg + GGDEF
	Response_reg + EAL
	Response_reg + HDOD

Tabelle 3.2.3: Genereller Konsensus. Mögliche Pfam-Familien-Kombinationen im Sensor- und Regulationsprotein.
Alle möglichen Pfam-Familien-Kombinationen für Sensor- und Regulationsproteine sind aufgelistet und sortiert nach der Kombinationshäufigkeit. Kleinbuchstaben zeigen Pfam-Domänen-Replikate an.
[1] m: 0-6, n: 0-10, p: 1-9, s: 1-2

3.2.3. Modifikationsmöglichkeiten in TCS Sequenzen: Bindestellen, Promotoren und Konnektoren

Grundsätzlich scheint es zwei generelle Modifikationsmöglichkeiten für TCS zu geben: (i) Veränderung in der Sequenz oder (ii) Veränderung in der Domänenzusammensetzung. Diese beiden Varianten werden im Folgenden eingehender untersucht.

3.2.3.1 Signalbindestelle im Sensorprotein

Eine Möglichkeit Zweikomponenten-Systeme in der Sequenz zu manipulieren ist die Veränderung des aktivierenden Signals.

Die Signalbindestelle eines Zweikomponenten-Systems befindet sich im periplasmatischen Bereich des Sensorproteins und ist derzeit noch wenig untersucht. Die Datenbank Swiss-Prot enthält jedoch einige wenige Sequenzen. Auf Basis dieser Sequenzen wurde pro Signal über alle untersuchten Organismen hinweg eine Konsensussequenz erstellt. **Tabelle 3.2.4** beinhaltet die Konsensussequenzen aller aufgefundenen Signale über die Organismen hinweg.

Signal	Anzahl Sequenzen	Bindungssequenz
Phosphor	1	**GYLP**
Osmose	4	**NFAILPSLQQFNKVLAYEVRMLMTDKLQLEDGTQLVVPPAFRR EI**yrelg**ISLYTNEAAEEAGLRWAQHYEFLSHQMAQQLGGPTE VRVEVNKSSPVVWLKTWLSPNIWVRVPLTEIHQGDFS**
Stress	6	**LVYKFTAERAGRQSLDDLMNSSLYLMRSELREIPPHDWGKTLK E**mdlnlsfdlrveplskyhlddismhrlrggeiv**ALDDQYTFI QRIPRSHYVLAVGPVPYLYYLHQM**r
Eisen	6	**HESTEQIQLFEQALRDNRNNDRHIMREIRE**
Kupfer	3	**HSVKVHFAEQDINDLKEISATLERVLNHPDETQARRLMTLEDI VSGYSNVLISLADSHGKTVYHSPGAPDIREFARDAIPDKDARG GEVFLLSGPTMMMPGHGHGHMEHSNWRMISLPVGPLVDGKPIY TLYIALSIDFHLHYINDLMNK**
Citrat	4	asfedyltlhvrdmamnqakiiasndsvisavktrdykrlati ankl**QRDTDFDYVVIGDRHSIRLYHPNPEKIGYPMQFTKPGAL EKGESYFITGKGSMGMAMRAKTPIFDDDGKVIGVVSIGYLVSK IDSWRAEFLLP**
Fumarat	4	**SQISDMTRDGLANKALAVARTLADSPEIRQGLQKKPQESGIQA IAEAVRKRNDLLFIVVTDMHSLRYSHPEAQRIGQPFKGDDILK ALNGEENVAINRGFLAQALRVFTPIYDENHISKAQIGVVAIGL ELSRV**tqqindsrw
Nitrat/ Nitrit	8	sslr**DAHAINKAGSLRMQSYRLGYDLPSGEPDKNAHRQMFQQA** lhspvltnlnvwyvpeavk**TRYAHRNANWDGMNNRLQGGDDPW YNENIPNYMNQQDRFTLALDHYQ**erkqffec

Tabelle 3.2.4: Konsensusmotive für Signalbindestelle im Sensor.
Diese Tabelle beinhaltet Konsensussequenzen der Signalbindestellen aus Sensorproteinen. Die erste Spalte beschreibt das Signal, die nächste Spalte beinhaltet die Anzahl an gefundenen Sequenzen. Die Konsensussequenz in der letzen Spalte zeigt die konservierten Stellen. Schwächer konservierte Stellen werden durch Kleinbuchstaben dargestellt.

Die Sequenzvergleiche über unterschiedliche Organismen und TCS-Familien hinweg machen deutlich, dass Signalbindestellen wie erwartet extrem von der Art des Signals abhängen. Zwischen den unterschiedlichen Signalen variierte die Sequenz erheblich.
Innerhalb eines Signals war die Sequenz allerdings unerwartet stark konserviert, selbst zwischen den Organismen und über unterschiedliche Zweikomponenten-Systeme hinweg. In **Tabelle 3.2.5** sind die Sequenzen von unterschiedlichen *E. coli, Salmonella Vibrio, Haemophilus influenzae* und *Shigella* Strängen mittels Alinierung verglichen. Bei den untersuchen Proteinen handelt es sich um die Nitrat/Nitrit Sensorproteine NarX und NarQ. Die Signalboten beider Proteine sind Nitrat/Nitrit, die dazugehörenden Regulatoren sind unterschiedlich.
Dieser Sequenzvergleich zeigt, dass die Sequenzen zwischen den beiden Sensorproteinen und zwischen den Organismen sehr gut konserviert sind.

CLUSTAL 2.0.8 Multiple Sequenzalinierung

```
                40...5....50...5....60...5....70...5....80...5....90...5...100
Q8Z4S5_SALTI     SSLRDAEAINIAGSLRMQSYRLGYDLQSGSPQLNAHRQLFQQALHSPVLTNLN-VWYVPEA
Q8XBE5_ECO57     SSLRDAEAINIAGSLRMQSYRLGYDLQSGSPQLNAHRQLFQQALHSPVLTNLN-VWYVPEA
Q8ZN78_SALTY     SSLRDAEAINIAGSLRMQSYRLGYDLQSGSPQLNAHRQLFQQALHSPVLTNLN-VWYVPEA
NARQ_ECOLI       SSLRDAEAINIAGSLRMQSYRLGYDLQSGSPQLNAHRQLFQQALHSPVLTNLN-VWYVPEA
B5R4I7_SALEP     SSLRDAEAINIAGSLRMQSYRLGYDLQSRSPQINAHRQLFQHALNSPVLQNLN-AWYVPQA
Q9KLR7_VIBCH     SSLNDAEAVNVSGSMRMQSYRLAYDIQTQSHDYKAHIFLFENSLYSPSMLALL-DWTVPSD
NARQ_HAEIN       SNKYDAEAINISGSLRMQSYRLLYEMQEQPESVETNLRRYHISLHSSALLEVQNQFFTPNV
NARX_ECOLI       GVQGSAHAINKAGSLRMQSYRL-LAAVPLSEKDKPLIKEMEQTAFSAELTRAA----ERDG
NARX_ECO57       GVQGSAHAINKAGSLRMQSYRL-LAAVPLSEKDKPLIKEMEQTAFSAELTRAA----ERDG
NARX_SHIFL       GVQGSAHAINKAGSLRMQSYRL-LAAVPLSEKDKPLIKEMEQTAFSAELTRAA----ERDG
                 .   .*.*:* :**:*******       . . :.     . :  *. :          .

                          110       120       130       140
Q8Z4S5_SALTI   VKTRYAHLNANWL-EMNNRLSKG-DLPWYQANINNYVNQIDLFVLAL 105
Q8XBE5_ECO57   VKTRYAHLNANWL-EMNNRLSKG-DLPWYQANINNYVNQIDLFVLAL 105
Q8ZN78_SALTY   VKTRYAHLNANWL-EMNNRLSKG-DLPWYQANINNYVNQIDLFVLAL 105
NARQ_ECOLI     VKTRYAHLNANWL-EMNNRLSKG-DLPWYQANINNYVNQIDLFVLAL 105
B5R4I7_SALEP   VKTRYARLHANWL-EMNSRLQDG-DIAWYQTNINNYVDQIDLFVLAL 119
Q9KLR7_VIBCH   IQQDYYQLIERWH-ELKKVLNSD-QKAQYLDQVAPFVSLVDGFVLKL 115
NARQ_HAEIN     LKHSYQNILQRWT-NMEKYARQQ-DVKNYSKQLTDYVADVDYFVFEL 105
NARX_ECOLI     QLAQLQGLQDYWRNELIPALMRAQNRETVSADVSQFVAGLDQLVSGF 140
NARX_ECO57     QLAQLQGLQDYWRNELIPALMRAQNRETVSADVSQFVAGLDQLVSGF 140
NARX_SHIFL     QLAQLQGLQDYWRNELIPALMRAQNRETVSADVSQFVAGLDQLVSGF 103
                     :   *  ::        :      ::  :*  :* :*  :
```

Tabelle 3.2.5: Alinierung der Nitrat/Nitrit Bindestelle der Nitrat/Nitrit Sensorproteine NarX und NarQ.
Dasselbe Signal (Nitrat/Nitrit) in zwei unterschiedlichen Sensorproteinen (NarX, NarQ) in unterschiedlichen *E. coli, Salmonella Vibrio* and *Haemophilus influenzae* Strängen führt zu sehr ähnlichen Sequenzen. Die Nummerierung orientiert sich am Protein Narq_ecoli, Position 40-146. Die Symbole unterhalb der Sequenz zeigen den Grad der Konservierung für diese Aminosäure an, von stark konserviert (*) über mittel (:), schwach (.) bis nicht konserviert ( ).

3.2.3.2 Promotorregionen der durch Zweikomponenten-Systeme regulierten Gene

Im Folgenden soll die Nukleotidsequenz in der Promotorregion des regulierten Gens untersucht werden. Auf diese Weise könnte die vom Zweikomponenten-System regulierte Expression eines Proteins manipuliert werden.

Die DNA-Bindestellen sind von dem zu regulierenden Gen abhängig. Regulationsprotein-bindende Nukleotidstellen sind im Allgemeinen schlecht annotiert. Die Basis dieser Analyse bildet daher eine manuelle Sammlung von Daten aus öffentlichen Datenbanken und eigenen Sequenzanalysen.

Über öffentliche Datenbanken (Prodoric, DBTBS**,** TractorDB, PDB, PDBSum und PubMed) wurden Nukleotidbindestellen für vornehmlich *E. coli K-12*, aber auch andere *E. coli* Stämme und *B. subtilis* analysiert. Weitere Organismen sind bisher kaum untersucht, weshalb die verfügbare Datenmenge relativ klein ist.

Mit Hilfe eines externen Perl-Skripts wurde nach Sequenzähnlichkeiten in den Genomen andere *E. coli* Stämme (*E. coli 536, E. coli CFT073, E. coli K-12, W3110, E. coli O157:H7 EDL933, E. coli K-12 MG1655, E. coli O157:H7 Sakai pO157, E. coli UTI89*) gesucht.

Das Perl-Skript startet eine Motivsuche, wobei die angegebenen Organismen in beide Leserichtungen mit einem Sequenzmotiv nach Treffern untersucht wurden. Zusätzlich kann angegeben werden, wie viele Fehltreffer innerhalb der Sequenz akzeptabel sind. In diesem Fall wurden maximal zwei Fehltreffer innerhalb einer Sequenz zugelassen. Auf diese Art und Weise lieferte das Skript potentielle Regulationsproteinbindestellen im Genom der untersuchten Organismen.

Aus den gefundenen Sequenzen wurden Gen-spezifische Konsensusstellen erstellt. Die Nukleotidsequenzen in *E. coli* sind pro TCS-Familie in **Tabelle 3.2.6** abgebildet.

Regulierte Gene	Konsensus der DNA-Bindesequenz
OmpC	TTTACATTTTGAAACATCT
OmpF	T[GT][GT][TG]TA[CG][AC][TA][AC]TTT[TC]
OmpF/OmpC	TTT[TA]C-TTTT[TG]
NarG1	1 TACCCATTAA 10
NarG2	1 TAACCAT--- 7
NarG3	1 TAATTAT--- 7
NarG4	1 TACTTTA--- 7
NarG5	1 -AGGGGTA-- 7
NarG6	1 TAGGAAT--- 7
NarG7	TTTAACCCGAtcggggtatg
NarK	TAC[TC][CG][CA]T
CitB	agtAATTTAATTaatt
LytT	[TA][AC][CA]GTTN[AG][TG]
LytT	taaggAAATAAAACTGATTTTcacgtca
AlgR	aaatGAATATTTATTCAAat
GlnG/GlnK	tgcaCCACCATGGTGCA

Tabelle 3.2.6: Zielgen-Regulierungsstellen von TCS in *E. coli.*
Die Tabelle zeigt die Konsensi der DNA-Bindesequenzen der Regulationsproteine: Outer membrane pore protein C und F (OmpC/F), Respiratory nitrate Reductase (NarG), Nitrite Extrusion protein (NarK), Citrate utilization protein (CitB), Sensory transduction protein (LytT), Transcriptional regulatory protein (AlgR), Nitrogen regulation protein (GlnG/K).
Die Konsensussequenzen aus den gefundenen potentiellen Bindestellen wurden durch multiple Alinierungen erstellt.

Die Konsensussequenzen der Zielgen Regulierungsstellen für die gram-negativen Organismen *Salmonella typhi* und *Shigella flexneri* sind in **Tabelle 3.2.7** abgebildet.

Familie	Reguliertes Gen	Funktion	Beispiel Organismus	Sequenz
NtrC	GlnH	Transkription Faktor	*Salmonella typhi*	gacat**TTGCACTTAAATAGTGCAC**aaccc
NtrC	GlnA	Transkription Faktor	*Salmonella*	ttcta**TTGCACCAATGTGGTGCTT**aatgt cattg**AAGCACTATTTTGGTGCAA**catag
NtrC	GlnK	Transkription Faktor	*Salmonella*	Ccatt**ATGCACCGTCGTGGTGCGT**ttttc
NtrC	GlnA	Transkription Faktor	*Salmonella*	Ctata**ATGCACTAAAATGGTGCAA**ccttt
NarL	NarK	Transkription Faktor	*Salmonella*	Aatag**CCTACTCATTAAGGGTAAT**aacta
NtrC	GlnG	Transkription Faktor	*Shigella flexneri*	Ctata**ATGCACTAAAATGGTGCAA**cctgt
ArgR	ArgA	Transkription Faktor	*Salmonella*	actaa**TTTCGAATAATAATTCACTA**gtggg
ArgR	ArgC	Transkription Faktor	*Salmonella*	cgtta**ATGAATAAAAATACAT**aatta
Spo	Spo	Transkription Faktor	*B. subtilis*	---------TTTGTCGAATGTAA----------
Spo	Spo	Transkription Faktor	*B. subtilis*	ATTTCATTTTTAGTCGAAAAACAGAGAAAAACA
Spo	Spo	Transkription Faktor	*B. subtilis*	AGAAGATTTTTCGACAAATTCA----------

Tabelle 3.2.7: Zielgen Regulierungsstellen von Zweikomponenten-Systemen in weiteren gram-negativen Bakterien.
Die Tabelle listet Promotorstellen in Zweikomponenten-Systemen von Salmonella, Shigella und *B. subtilis* auf. Großbuchstaben und farbliche Hinterlegungen zeigen stark konservierte Stellen in den unterschiedlichen Stämmen.

An den beiden Tabellen der Promotorsequenzen ist zu erkennen, dass für die durch Zweikomponenten-Systeme regulierten Gene teilweise mehrere Bindestellen vorhanden sind, welche mittels multipler Alinierungen zu Konsensussequenzen zusammengefasst werden können. Ebenfalls auffällig ist, dass die gefundenen Konsensussequenzen sehr kurz sind.

Ein weiteres Perl-Skript (*search_sequence.pl*) nutzte diese erstellten kurzen Konsensussequenzen aus den obigen Tabellen, um im Genom die Region rund um das regulierte Gen nach möglichen Mehrfachvorkommen von Promotorsequenzen abzusuchen. Damit kann das Regulationsprotein mehrfach an die DNA binden und ist in der Lage, die Spezifität der Aktivierung des regulierten Gens erheblich zu verbessern. Das Multimeraufkommen der Regulationsproteine in Zweikomponenten-Systemen vereinfacht die Mehrfachbindung.

Das Skript lieferte die Anzahl der gefundenen Promotorstellen und die Abstände zwischen den Bindestellen. Die Ergebnisse sind in **Tabelle 3.2.8** dargestellt.

Regulationsprotein	Reguliertes Gen	Wieder-holung	Abstand [NS]
Citrate Utilization Protein B (CitB)	Citrate Lyase (CitC)	6	40
Nitrogen Regulationsprotein (NtrC)	Sequenz Glutamin Synthetase (GlnA)	2	63
Nitrogen Regulationsprotein (NtrC)	Nitrogen Regulator Protein (GlnK)	7-12	Variabel
Nitrate/Nitrit Regulationsprotein (NarL)	Respiratorisches Nitrat Reduktase (NarG)	Variabel	Ca. 6
Nitrate/Nitrite Regulationsprotein (NarL)	Nitrite Extrusion Protein (NarK)	Variabel	Variabel
Osmolarity Regulationsprotein (OmpR)	Outer Membran Protein C und F (OmpC/ OmpF)	3	7

Tabelle 3.2.8: Abstände zwischen Promotorstellen in *E. coli.*
Tabelle enthält das regulierende Regulationsprotein, die regulierten Gene die Anzahl an Wiederholungen der Promotorsequenz und die Abstände zwischen den Wiederholungen.

Für die CitC-Expression kann CitB am Promotor oberhalb (upstream) des CitC Gens binden. Die Bindestelle wird sechsmal in Abständen von 40 Basenpaaren (BP) mit kleinen Variationen gefunden.

Für die GlnA-Expression kann NtrC binden. Die Bindestelle wird zweimal im Abstand von 63 BP gefunden.

Für die GlnK-Expression kann NtrC zwischen sieben- und zwölfmal binden.

Für die NarG/NarK-Expression kann NarL binden. Diese Bindestelle wurde mehrfach mit nur kurzen Abständen (ca. sechs BP) gefunden.

Für die OmpC/OmpF-Expression kann OmpR binden. In beiden Fällen existiert die Bindestelle mindestens dreimal mit Abständen von etwa sieben BP. Ähnliches gilt auch für *S. typhimurium*.

3.2.3.3 Autoregulatorische Promotor-Bindestellen in Zweikomponenten-Systeme

Wie von Mitrophanov im Jahre 2008 beschrieben, kann die Bindung des Regulationsproteins nicht nur die Regulierung eines Gens, sondern die Regulation des gesamten Zweikomponenten-Systems in Form einer Rückkopplungsschleife (Feedback-Loop) bewirken [15].
Die dazugehörigen Promotorregionen sind bisher kaum untersucht und annotiert. Interessanterweise befinden sich die Bindestellen nicht nur oberhalb (upstream) des Sensorgens, sondern ebenfalls unterhalb des Gens, welches das Regulationsprotein codiert, wobei beide Bindestellen ähnliche Verhaltensweisen aufzeigen.

In den folgenden gleichartig-aufgebauten sechs Tabellen (**3.2.9 bis 3.2.14**) werden die Promotorsequenzen für die Bindung des Regulationsproteins an das Genom dargestellt. Um die Sequenz leichter zu finden werden Start- und Stoppstelle der Sequenz im Genom angegeben, ebenso wie der dazugehörige Organismus und die Quelle, aus der die Sequenz entnommen wurde. Außerdem wird angegeben, welche Gene durch die Bindung des Regulationsproteins an der Promotorregion aktiviert werden: Das Sensorprotein, das Regulationsprotein und das oder die Gene, welche durch das TCS reguliert werden. Die Tabellen sind aufgeteilt nach TCS-Familien. Jeweils unterhalb der Tabelle befindet sich ein Sequenzlogo basierend auf den in der Tabelle aufgelisteten Sequenzen. Das mit dem Programm Weblogo erstellte Sequenzlogo gibt demnach an, wie konserviert die Promotorbindesequenz für die Aktivierung des Regulationsproteins, des Sensorproteins und für die Aktivierung der TCS-regulierten Gene ist.

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
4055724	4055739	*E. coli K12*	Prodoric	GlnG	glnALG	ACGCCTTTTAGGGGCA
4055747	4055759	*E. coli K12*	Prodoric	GlnG	glnALG	GCACGATGGTGCG
4055768	4055782	*E. coli K12*	Prodoric	GlnG	glnALG	TCACATCGTGGTGCA
4055788	4055802	*E. coli K12*	Prodoric	GlnG	glnALG	GCACTATATTGGTGC
4055820	4055834	*E. coli K12*	Prodoric	GlnG	glnALG	GCACCAACATGGTGC

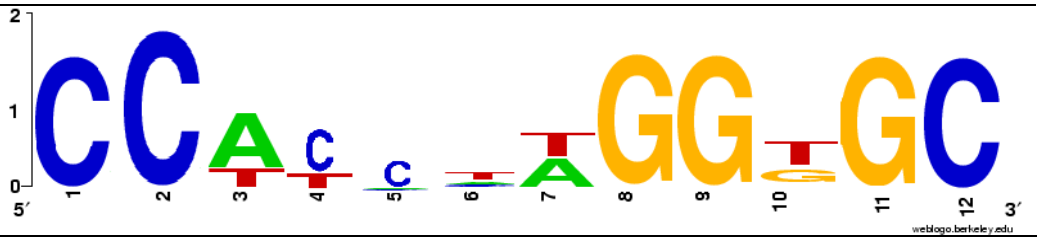

Tabelle 3.2.9: Promotorequenzen für die TCS-Familie GlnL/GlnG.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (GlnL), des Sensorproteins (GlnG) und des regulierten Gens (GlnA).

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
3534321	3534355	*E. coli K12*	Prodoric	OmpR	ompR-envZ	ATTGTTACAAAGCATATTA AACAGCAGCTTAAGTA
3534363	3534400	*E. coli K12*	Prodoric	OmpR	ompR-envZ	TATTCGGCGAAACATTATT GATTCTGTTGATATGATCA
3534427	3534462	*E. coli K12*	Prodoric	OmpR	ompR-envZ	AACAGACAAAGGGAATCAA CGAGATGAAAACGCCCC
986357	986366	*E. coli K12*	Prodoric	OmpR	ompF	TGTAGCACTT
986386	986395	*E. coli K12*	Prodoric	OmpR	ompF	TTTTCTTTTT
986396	986405	*E. coli K12*	Prodoric	OmpR	ompF	GTTACATATT
986406	986415	*E. coli K12*	Prodoric	OmpR	ompF	TTTACTTTTG
2310889	2310898	*E. coli K12*	Prodoric	OmpR	ompC	AGTATCATAT
2310909	2310918	*E. coli K12*	Prodoric	OmpR	ompC	TGAAACATCT
2310930	2310939	*E. coli K12*	Prodoric	OmpR	ompC	TGAAACATCT
2310939	2310948	*E. coli K12*	Prodoric	OmpR	ompC	TTTACATTTT

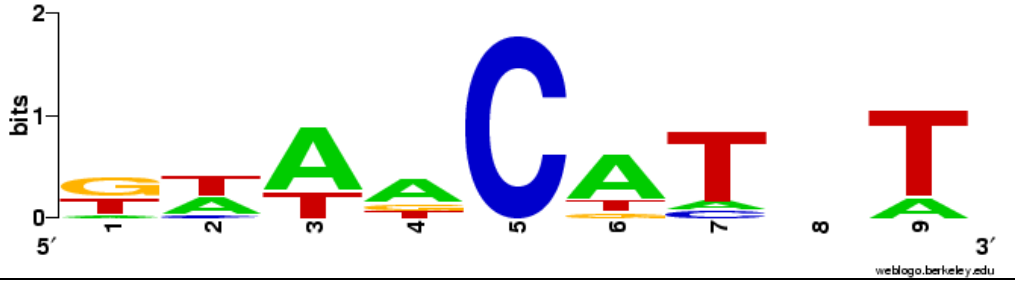

Tabelle 3.2.10: Promotorequenzen für die TCS-Familie EnvZ/OmpR.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (OmpR), des Sensorproteins (EnvZ) und die regulierten Gene (OmpC und OmpF).

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
1189731	1189731	*E. coli K12*	Prodoric	PhoP	phoPQ	TcccctccccgctGGTTTA tttaaTGTTTA
1189749	1189769	*E. coli K12*	Tractor DB	PhoP	phoPQ	TatggGGTTTATTTAATGT TTACCCagcgg
1189730	1189748	*E. coli K12*	Tractor DB	PhoP	phoPQ	GggggTGGTTTATTTAATG TTTAgcggg
2420613	2420679	*E. coli K12*	Prodoric	PhoP	phoPQ	CGCTTTCTAAATttcacaT AACCttcaaaaAGTAAGAA ATGTGAAATGAACGTGCAA TGATATAATT
1319447	1319467	*S. typhi*	Tractor DB	PhoP	phoPQ	TttggGGTTTATTAACTGT TTATCCagaca

Tabelle 3.2.11: Promotorequenzen für die TCS-Familie PhoQ/PhoP.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (PhoP) und des Sensorproteins (PhoQ)

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
1278819	1278825	*E. coli K12*	Prodoric	NarL	NarG/NarL	TACCCAT
1278832	1278838	*E. coli K12*	Prodoric	NarL	NarG/NarL	TACTCCT
1278842	1278848	*E. coli K12*	Prodoric	NarL	NarG/NarL	TACCCAT
1278926	1278932	*E. coli K12*	Prodoric	NarL	NarG/NarL	TAATTAT
1278938	1278944	*E. coli K12*	Prodoric	NarL	NarG/NarL	TAATTAT
1278948	1278954	*E. coli K12*	Prodoric	NarL	NarG/NarL	TAGGAAT
1278956	1278962	*E. coli K12*	Prodoric	NarL	NarG/NarL	TACTTTA
1278970	1278976	*E. coli K12*	Prodoric	NarL	NarG/NarL	TCCCCAT
1276938	1276944	*E. coli K12*	Prodoric	NarL	NarK	TACCCAT
1276958	1276964	*E. coli K12*	Prodoric	NarL	NarK	TACTCCT
1276975	1276981	*E. coli K12*	Prodoric	NarL	NarK	TAACCAC

1276992	1276998	*E. coli K12*	Prodoric	NarL	NarK	TACCGAT
1276998	1277004	*E. coli K12*	Prodoric	NarL	NarK	TACCCTT
1277054	1277060	*E. coli K12*	Prodoric	NarL	NarK	TACTCAC
1277062	1277068	*E. coli K12*	Prodoric	NarL	NarK	TACCCAT
1277072	1277078	*E. coli K12*	Prodoric	NarL	NarK	TATTTAT
1277085	1277091	*E. coli K12*	Prodoric	NarL	NarK	TATCTAT

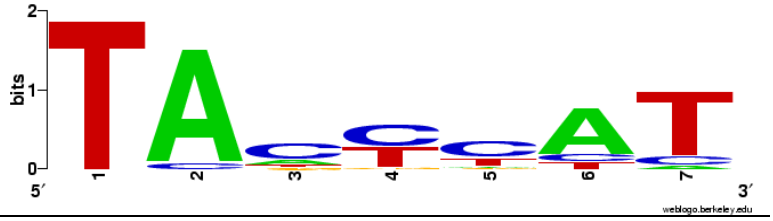

Tabelle 3.2.12: Promotorequenzen für die TCS-Familie NarX/NarL.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (NarL), des Sensorproteins (NarX) und die regulierten Gene (NarG und NarK).

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
4103307	4103322	*E. coli K12*	Prodoric	CpxR	cpxRA	TGTAAAACAACGTAAA
3490009	3490029	*E. coli K12*	Tractor DB	CpxR	cpxRA	CcatcTACGTAAAATTAGG TAAAGGtctga
4039913	4039933	*E. coli K12*	Tractor DB	CpxR	cpxRA	TaaagAGTAAAAGCTTGTA AGCGGCgccac

Tabelle 3.2.13: Promotorequenzen für die TCS-Familie CpxR/CpxA.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (CpxR) und des Sensorproteins (CpxA).

Start	Stopp	Orga-nismus	Daten-Quelle	Binde-protein	Regulierte Gene	Sequenz
1020410	1020529	*B. subtilis*	Prodoric	CitB	citA	gaagccatttgaaatccattt ctattctccctctgattaata tttttaattaattccctttaa aataTTATTatttttttaaata ttataTTTACTa
1021030	1021043	*B. subtilis*	Prodoric	CitB	citA	AGAAAGCGCTTGAA

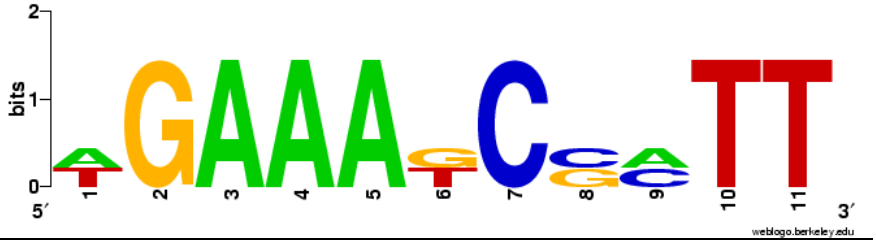

Tabelle 3.2.14: Promotorequenzen für die TCS-Familie CitB/CitA.
Die gefunden Sequenzen sind gültig für die Gene des Regulationsproteins (CitA) und des Sensorproteins (CitB)

Aus den Tabellen ist ersichtlich, dass die Menge an annotierten Sequenzen pro TCS-Familie sehr gering ist. Nur bei wenigen Familien sind die Regulierungsstellen für alle drei Gene vorhanden. Bei den gefundenen Familien bindet allerdings immer das Regulationsprotein an alle Regulierungsstellen. Die Sequenzlogos zeigen, dass die Regulierungsstellen für das Sensorprotein, das Regulationsprotein und die zu regulierenden Gene sehr ähnlich sind. Diese Sequenzen sind teilweise selbst über die Organismen hinweg vergleichbar und damit übertragbar auf weitere TCS-Familien und Organismen.

Auch wenn die aufgelisteten Promotorsequenzen relativ kurz sind, so sind sie doch lang genug, um eine spezifische Bindung des Regulationsproteins zu bewirken.

3.2.3.4 Domänenflexibilität durch Konnektoren

TCS können durch zusätzliche Komponenten (Proteine oder Domänen), sog. Konnektoren beeinflusst werden. Ein kürzlich entdeckter Konnektor (SafA Protein) wurden im Folgenden mit Hilfe von Sequenzvergleichen und Motivsuchen bezüglich seines Vorkommens genauer untersucht, da er Zweikomponenten-Systeme beeinflussen oder miteinander verbinden können und daher sehr interessant ist [20].

SafA Protein

Eine Pfam-Domäne ist für das Protein SafA nicht vorhanden. Mit Hilfe einer iterativen organismusspezifischen Sequenzalinierung wurde daher nach Ähnlichkeiten in anderen Organismen gesucht. Dabei konnte festgestellt werden, dass dieses Protein nicht ausschließlich in *E. coli* Stämmen vorkommt, sondern zusätzlich im *Shigell*a Stamm 2a str. 2457T, *Salmonella typhi, Salmonella typhimurium* und *Mycoplasma pneumoniae* vorkommt. In allen genannten Spezien wurde SafA mit dem TCS Evgs/EvgA-PhoQ/PhoP oder mit unbekannten Proteinen mit vorhergesagter ähnlicher Funktion in Verbindung gebracht. Die gefundenen Sequenzen wurden aliniert und als Sequenzlogo in **Abbildung 3.2.3** zusammengefasst.

Abbildung 3.2.3: Sequenzlogo für den Konnektor SafA.
Das Sequenzlogo wurde mittels Weblogo erstellt, basierend auf den durch PSI-Blast gefunden Sequenzen in *E. coli, S. flexneri, Salmonella und Mycoplasma.*

Die SafA ähnlichen aufgefundenen Proteine waren entweder unbekannte oder hypothetische Proteine. Interaktionsvorhersage-Untersuchungen mit STRING zeigen, dass diese aufgefundenen Proteine entweder direkt mit EvgS oder mit Proteinen ähnlicher Funktion interagieren (siehe **Tabelle 3.2.15**).

Protein	Beschreibung	Organismus	STRING score
NP_310132	Hypothetisches Protein ECs2105	*E. coli 0157*	0,9 zu EvgS
ZP_02799272	konserviertes hypothetisches Protein	*E. coli 0157*	0,9 zu EvgS
YP_540723	Hypothetisches Protein C1714	*E. coli UTI89*	0,9 zu EvgS
NP_837211	Hypothetisches Protein S1655	*S. flexneri*	*0,76* zu EvgS
NP_458304	putative Phosphodiesterase	*S. typhi*	0,65 zu ygiM (evtl. Signal Transduktion Protein)
NP_462516	putative Phosphodiesterase	*S. typhimurium*	0,6 zu Lon

Tabelle 3.2.15 – SafA ähnliche Proteine.
SafA ähnliche Proteine können in mehreren Organismen gefunden werden. Die Tabelle listet diese Proteine gemeinsam mit möglichen Interaktionspartnern inklusive Interaktionswahrscheinlichkeit (STRING Score). Dabei handelt es sich um EvgS oder funktionell ähnliche Proteine.

Das Sequenzlogo und die Interaktionsanalyse machen deutlich, dass die Konnektorsequenzen in den gefundenen Organismen sehr ähnlich und konserviert sind. Der Konnektor scheint TCS-abhängig zu sein und könnte somit als Basis für weitere Organismen eingesetzt werden.

EAL/ GGDEF Domäne

EAL-Domänen treten häufig in Verbindung mit Zweikomponenten-Systemen auf und steht daher in Verdacht sie ebenfalls zu beeinflussen. Daher wurden mit Hilfe von Genkontext-Methoden, Literaturrecherchen und der STRING-Datenbank nach Interaktionspartnern für EAL-Domänen beinhaltende Proteine gesucht. Das Ergebnis der Interaktionspartnersuche zeigt, dass Proteine mit EAL-Domänen vermehrt mit noch unbekannten Proteinen interagieren, aber auch vermehrt mit DNA-bindenden Proteinen, teilweise sogar mit Proteinen, welche eine Responseregulator-Domäne beinhalten.
Ausserdem interagieren sie häufig mit GGDEF-Domänen beinhaltenden Proteinen.
EAL- und GGDEF-Domänen treten häufig in Verbindung mit der Responseregulator-Domäne auf. Demzufolge könnten EAL- und GGDEF-Domänen als potentielle Konnektoren fungieren.
Tabelle 3.2.16 zeigt von STRING vorhergesagte Interaktionspartner zu Proteinen mit EAL-Domäne.

Proteine mit EAL-Domäne	Interaktionspartner
>Q21G90_SACD2 Diguanylate Zyklase/ Phosphodiesterase *Saccharophagus degradans*	Sde_3649 GGDEF Familien Protein
	Sde_2537 hypothetisches Protein
	Sde_3232 hypothetisches Protein
	Sde_3313 mögliche Diguanylate Phosphodiesterase
	Sde_1079 mögliche Diguanylate Phosphodiesterase
	Sde_3648 Formamidopyrimidine-DNA Glycolase
	Sde_0078 GGDEF Domänen Protein
	Sde_3427 mögliche Diguanylate Zyklase (GGDEF)
	Sde_3693 Responseregulator-Domänen Protein (CheY-ähnlich)
	Sde_1063 GGDEF Familien Protein
>A6Q1G4_NITSB Signal Transduktion Responsregulator aus *Nitratiruptor sp.*	dgkA Diacylglycerol Kinase
	NIS_0211 mögliches uncharakterisiertes Protein
	dnaG DNA Primase DnaG
	NIS_0567 mögliches uncharakterisiertes Protein
	NIS_0004 mögliches uncharakterisiertes Protein
	NIS_1647 mögliches uncharakterisiertes Protein

	NIS_1732 mögliches uncharakterisiertes Protein
	NIS_0150 mögliches uncharakterisiertes Protein
	NIS_0136 mögliches uncharakterisiertes Protein
>A1AD34_ECOK1 Mögliches uncharakterisiertes Protein aus *Escherichia coli O1*	yedQ hypothetischen Protein
	yaiC mögliches uncharakterisiertes Protein
	ydeH mögliches uncharakterisiertes Protein ydeH
	yeaP mögliches uncharakterisiertes Protein yeaP
	ycdT vorhergesagte Diguanylate Zyklase
	yfiN vorhergesagte Diguanylate Zyklase
	yneF mögliches uncharakterisiertes Protein yneF
	yeaI mögliches uncharakterisiertes Protein yeaI
	yejA mögliches uncharakterisiertes Protein yejA
	yejB vorhergesagte Oligopeptide Transporter Untereinheit

Tabelle 3.2.16: Proteine mit EAL-Domäne mit ihren vorhergesagten Interaktionspartnern. Die linke Spalte enthält Proteine mit EAL-Domäne, die rechte Spalte beinhaltet von der STRING Datenbank vorhergesagte Interaktionsproteine. Unter diesen Interaktionsproteinen befinden sich vor allem viele nicht-annotierte Proteine, aber auch Proteine mit Responseregulator-Domäne.

Konnektoren können den Ergebnissen zufolge stark divergieren. Die hier vorgestellten Konnektoren könnten sinnvoll in Experimenten der synthetischen Biologie eingesetzt werden. Beispielsweise um Zweikomponenten-Systeme spezifischer zu gestalten oder um sie miteinander zu verbinden. Dazu müsste die organismen- und TCS-spezifischen Eigenschaften der Konnektoren entsprechend angepasst werden.

3.2.4. Modifikation durch Domänenvermischung und Entartung: Erkennung von divergierenden TCS

3.2.4.1 HisKA Substitution

Basierend auf Skerkers Sensorprotein-Austauschexperimenten [17], wurde aus den über 50 gesammelten Sequenzen mittels ClustalW eine Substitutionsmatrix errechnet, welche anzeigt wie gut Sensorproteine gegeneinander austauschbar sind. **Abbildung 3.2.4** zeigt die erstellte numerische Substitutionsmatrix farblich hinterlegt. Eine steigende rote Farbintensität in der Abbildung spiegelt die gesteigerte Austauschbarkeit der Proteine wieder. Es stellte sich heraus, dass der Austausch des Sensorproteins sehr schwierig ist und nur zwischen sehr ähnlichen TCS-Familien und ähnlichen Spezies überhaupt möglich ist. Damit ist auch der Funktionsaustausch mittels Austausch des Sensorproteins nur begrenzt nutzbar. Die Matrix liefert somit eine nützliche Hilfe für die Planung von Austauschexperimenten.

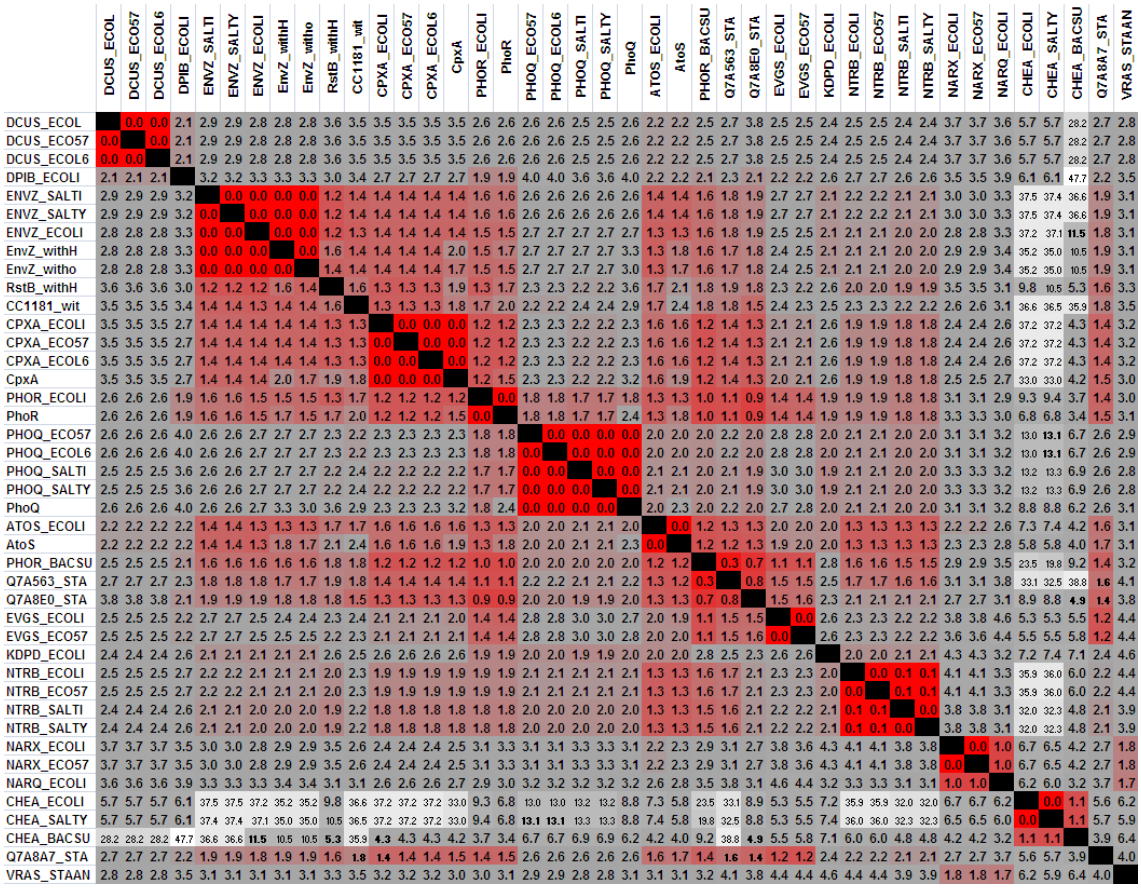

Abbildung 3.2.4: HisKA Substitutionsmatrix. Substitutionsmatrix für die verglichenen Spezies (Swiss-Prot Codes). Die Matrix verdeutlicht die Austauscherfolge zwischen Proteinen mit HisKA-Domäne. Sie wurde aus 50 verschiedenen HisKA-Domänen enthaltenden TCS-Proteinen erstellt. Je farbintensiver die Einfärbung, desto ähnlicher sind die Proteine und desto leichter ist ein Austausch möglich.

3.2.4.2 Erkennung von fehlenden Interaktionspartnern in TCS

Die starke Divergenz zwischen den verschiedenen TCS-Familien macht eine Identifikation von fehlenden TCS-Partnern und erst recht von neuen TCS schwierig.

Um bislang nur teilweise bekannte TCS in *L. pneumophila* und *L. monocytogenes* zu vervollständigen, wurden ausführliche Sequenzvergleiche zu Proteinen bekannter Funktionen aus verwandten Bakteriengenomen durchgeführt. Dazu wurden zunächst die bekannten Teile des TCS aus *L. pneumophila* und *L. monocytogenes* einer bekannten TCS-Familie zugeordnet. Anschließend wurde nach einem sequentiell möglichst eng verwandten Organismus gesucht. Der im Ursprungssystem fehlende Sequenzteil aus dem verwandtem Organismus wurde herausgesucht und anschließend mit diesem Sequenzabschnitt gegen den Ausgangsorganismus verglichen.

Die Ergebnisse sind zusammenfassend in **Tabelle 3.2.17** aufgelistet; neu aufgefundene TCS-Bestandteile sind türkis hinterlegt.

Die Bestätigung der Vorhersage erfolgte allerdings ausschließlich durch gute BLAST Ergebnisse mit sehr niedrigen E-values. Teilweise ist eine generelle Verbindung des gefundenen Proteins zu TCS-Familien bereits bekannt, aber keiner konkreten TCS-Familie oder Interaktionspartnern zugeordnet.

a) *Legionella pneumophila str. Philadelphia 1*

Familie	Identifikation	Sensor	Regulator	Funktion
OmpR	Iterative Sequenzsuche mit Toleranzwert e-30, OmpR Sequenz von *Enterobacter cloacae*	QseC GI:52841522 Bekannt/ annotiert durch PMID 15448271	Möglicher Regulator ähnlich zu QseB GI:52841523	Reguliertes Protein FliC; GI: 52841570; Flagella Regulation
NarL	Iterative Sequenzsuche mit Toleranzwert e-30; NP_288375 *E. coli* O157:H7 str. EDL933	BarA GI: 52842130 Bekannt/ annotiert durch PMID 15448271	Möglicher Regulator ähnlich zu UvrY GI:52842852	Reguliertes Protein CsrA; GI:52841018 Regulator der CO_2 Speicherung
NarL	Iterative Sequenzsuche mit Toleranzwert e-30 in *E. coli* ETEC H10407		Möglicher Regulator ähnlich zu EvgA GI:52840952	Reguliertes Protein EmrY; GI:52841684; Antibiotika Resistenz

b) Listeria monocytogenes, str. EGD-e

Familie	Identifikation	Sensor	Regulator	Funktion
NarL	PSI-Blast Suche mit Toleranzwert e-30 in *E. coli ETEC H10407*	Möglicher Regulator, ähnlich zu EvgS Q4EKW8_LISMO	Möglicher Regulator, ähnlich zu EvgA GI:16804553	Antibiotika-Resistenz EvgS Homolog
OmpR	PSI-Blast Suche in *B. subtilis* mit Toleranzwert 8e-19	Möglicher Sensor ähnlich zu CSSS_BACSU GI:16804620 GI:16803101	Möglicher Regulator, ähnlich zu CSSR_BACSU GI:16804621	Reguliertes Protein HtrA; Serine Protease
OmpR	PSI-Blast Suche in *B. subtilis* mit Toleranzwert e-60;	Möglicher Sensor ähnlich zu ZP_03239257 GI:16803061	PhoP GI:16804539 Bekannt/ annotiert durch PMID 11679669	Virulenz, antimikrobakterielles Peptide Resistenz

Tabelle 3.2.17: Neu annotierte Interaktionspartner in Zweikomponenten-Systemen.
Angegeben sind bekannte und potentielle neue Komponenten (türkis hinterlegt) aus Zweikomponenten-Systemen in (a) *L. pneumophila* und (b) *L. monocytogenes*. Die erste Spalte zeigt die TCS-Familie, die Identifikationsspalte gibt soweit vorhanden das Quellsystem, Methode, Referenzorganismus und Qualität an. Die nächsten drei Spalten beinhalten die Bereiche eines TCS: Sensor, Regulation und die zu regulierende Funktion. Potentiell neue Bestandteile (türkis hinterlegt) beinhalten die Homologe aus einem Referenzorganismus und die Angabe der neuen Komponente. Die angegebenen Identifikationsnummern stammen aus der Genbank Datenbank (Genbank acc. No.: AE017354, bzw. Acc. No.: AE017262; Chien, M. et al, 2004), aus Swiss-Prot ergänzt durch den Listilist-Server (http://genolist.pasteur.fr/ListiList/).

In *L. pneumophila* konnten folgende neue Erkenntnisse gewonnen werden:

- Für die OmpR Familie, welches die Flagellen reguliert, wurde ein mögliches Regulationsprotein, homolog zu QseB neu gefunden.
- Für die Regulation von CsrA wurde ein UvrY ähnliches Regulationsprotein entdeckt.
- Für die Regulation von EmrY, einem antibiotischen Resistenzgen, wurde ein mögliches Regulationsprotein ähnlich zu EvgS aus *E. coli* gefunden.

In *L. monocytogenes* konnten folgende neue Erkenntnisse gewonnen werden:

- Ein Homolog zu den Kalziumsensoren EvgS und EvgA konnte identifiziert werden.
- Außerdem wurde das Zweikomponenten-System CssS/CssR in *L. monocytogenes* gefunden.
- Ein mögliches Sensorprotein zum Regulationsprotein PhoP konnte entdeckt werden.

3.2.4.3 Natürlicher Domänenaustausch

Die Einzelbestandteile von Zweikomponenten-Systemen sind sehr spezifisch für das jeweilige Zweikomponenten-System. Eingehende Analysen mittels ScanProsite, PHI-BLAST, Sequenzanalysen und Literatursuchen konnten allerdings vereinzelt natürliche Beispiele des Domänen- und Funktionsaustausches von TCS-Komponenten aufdecken, bzw. deren Einsatz außerhalb von TCS.

Domänenaustausch im Sensorprotein

Mit Hilfe einer Motiv-Suche mit der HisKA-Domäne gegen die Swiss-Prot Datenbank konnten natürliche Beispiele identifiziert werden, bei denen HisKA-Domänen auch außerhalb von Zweikomponenten-Systemen vorliegen:

- Eine HisKA-Domäne ohne ein Zweikomponenten-System kommt im Branched-Chain alpha-Ketoacid Dehydrogenase Komplex (BCKD) der Maus vor. Bei diesem Komplex führt die Bindung von ATP zu einer Umstrukturierung der Loop-Region der Nukleotidbindestelle. Die strukturellen Änderungen führen des Weiteren zur Bildung einer aromatischen Verbindung im Grenzbereich zwischen der Nukleotidbindedomäne und der vier-alpha-Helix, wo eine Bewegung des oberen Bereiches der zwei Helices bewirkt wird, um die Enzymaktivität zu verändern [92]. Das entspricht einer HisKA Domäne ohne dazugehöriger RR-Domäne, welches zu einer neuen zellulären Antwort in Form einer Änderung der Enzymaktivität führt.
- Eine weitere HisKA-Domäne ist im Phytochrom A des *Populus tremuloides* zu finden. Es existiert in zwei Formen, welche reversibel durch Licht ineinander überführt werden können: Die Pf-Form absorbiert maximal im roten Lichtbereich und das Pfr absorbiert maximal im violetten Bereich. Die Umwandlung von Pf zu Pfr leitet eine Vielzahl an morphogenetischen Antworten ein, während die Rückumwandlung von Pfr to Pf den Prozess stoppt. Pfr ist ein genereller Transkriptionsfaktor und kontrolliert die Expression zahlreicher Zellkerngene, u.a. die kleinen Subeinheiten der Ribulose-Bisphosphate Carboxylase, Chlorophyll A/B Bindungsprotein, die Protochlophyllide Reduktase und rRNA. Außerdem kontrolliert es die Expression seines eigenen Gens in einer negativen Rückkopplungsregulation [93].

Domänenaustausch im Regulationsprotein

Die Responseregulator-Domäne von Zweikomponenten-Systemen konnte ebenfalls in anderen Zusammenhängen gefunden werden. Allerdings konnte der Domänenaustausch im Regulationsprotein deutlich seltener beobachtet werden als im Sensorprotein. Dennoch konnten solche Beispiele analog zum Suchprinzip mit der HisKA-Domäne selbst in Eukaryoten gefunden werden.

- Ein Beispiel für eine Responseregulator-Domäne ausserhalb eines TCS-Kontextes bildet die Serin/Threonin Proteinkinase Ppk18 in *Schizosaccaromyces pombe,* der sogenannten „Teilungshefe". Ppk18 spielt eine ausschlaggebende Rolle in der Zellproliferation und dem Zellwachstum in Abhängigkeit des Ernährungsstatus. Die Responseregulator-Domäne befindet sich im C-Terminus des zytoplasmatischen Proteins (gut konservierte PROSITE Signatur PS50110) und ist das Ziel von Rapamycin (TOR). TOR wiederum aktiviert Ppk18 durch Phosphorylierung, enthält aber keine typische HisKA-Domäne [94].

  Die Sequenzzusammensetzung von TOR zeigt jedoch eine Ähnlichkeit zur Pfam-Domäne HATPase_c.

In **Tabelle 3.2.18** werden die aufgefundenen TCS-Komponenten ohne direkten TCS-Zusammenhang aufgelistet.

Domäne	Protein	Kontext	Funktion
HisKA	Pyruvat Dehydrogenase Kinase	Glukose-Meta-bolismus, in *S. cerevisiae*	Inhibiert den mitochondrischen Pyruvat Dehydrogenase Komplex durch Phosphorylierung der E1 Alpha-Untereinheit und bildet damit einen Teil der Regulation des Glukose-Metabolismus.
HisKA	Adenylat Zyklase	Sporen-bildung, in *E. coli*	Durch die Produktion von cAMP wird die cAMP- abhängige Proteinkinase (PKAs) aktiviert, welche wiederum die Sporenproduktion anstößt.
HisKA	BCKD-Kinase	Katabolischer Stoffwechsel-weg von Valin, Leucin und Isoleucin, in der Maus	BCKD-Kinase katalysiert die Phosphorylierung und Inaktivierung des Branched-chain alpha-ketoacid Dehydrogenase-Komplexes. Diese wiederum ist das Schlüsselenzym des katabolischen Stoffwechselwegs von Valin, Leucin und Isoleucin, s.o.

HisKA	Phytochrom A	Regulatorischer Photorezeptor in *Populus tremuloides*	Phytochrom A ist ein regulatorischer Photorezeptor, welcher in zwei Formen existiert, die mittels Licht reversibel ineinander überführt werden können: Es kontrolliert die Expression eine Vielzahl von nuklearen Genen, s.o.
Response Reg	Gleitendes Motilität Protein Z	Chemosensorisches System, in *Myxococcus*	Das Protein Z wird als Antwort auf die vom Frz chemosensorischen System aufgenommen Umweltsignale benötigt, welches einem Bakterium die Fortbewegung ermöglicht.
Response Reg	Adenylatzyklase	Sporenbildung, in *Caulobacter*	Durch die Produktion von cAMP wird die cAMP- abhängige Proteinkinase (PKA) aktiviert, welche wiederum die Sporenbildung anstößt.
Response Reg	Serin/Threonin-Proteinkinase ppk18	*Schizosaccaromyces pombe*	Serin/Threonin-Proteinkinase ppk18, s.o.

Tabelle 3.2.18: Natürliche Beispiele des Domänenaustausches. Die Tabelle zeigt natürliche Bespiele, bei denen HisKA-Domänen und Responseregulator-Domänen ohne den jeweils anderen Teil des Zweikomponenten-Systems vorliegen. Die unterschiedlichen Funktionen und Zusammenhänge sind ebenfalls beschrieben.

Aus dieser Beobachtung lässt sich schließen, dass Teile und Funktionen von Zweikomponenten-Systemen sehr vielseitig einsetzbar sind und somit auch auf andere Wege für die synthetische Biologie nutzbar gemacht werden können. Selbst bei Eukaryoten konnten Teile aus TCS aufgefunden werden, was möglicherweise auf degenerierte Zweikomponenten-Systeme zurückgeführt werden könnte.

Die Beispiele betonen die Vielfalt der Einsatzmöglichkeiten von TCS oder einzelnen Teilen daraus.

Funktionelle Zusammenhänge der aufgefundenen Komponenten zu Zweikomponenten-Systemen konnten nicht komplett ausgeschlossen werden, sind allerdings aufgrund der beschriebenen Funktionen unwahrscheinlich.

3.2.4.4 Divergierendes TCS: Eine mögliche neue TCS-Familie in *Mycoplasma pneumoniae*

Die Kombination aus vorhergesagten Sequenzmotiven und Strukturanalysen bietet eine Möglichkeit selbst stärker degenerierte TCS aufzufinden. Im Folgenden wird eine neue mögliche TCS-Familie für den bislang als TCS-frei bezeichneten Organismus *M. pneumoniae* vorgestellt und der dazugehörige Identifikationsweg beschrieben.

Mögliches Sensorprotein MPN013

Degenerierte TCS-Systeme können nicht mit einfachen Sequenzsuchen aufgefunden werden, da direkte Sequenzähnlichkeiten (z.B. mit BLAST) alleine (ohne weiteres Vorwissen) keine signifikanten Treffer liefern.

Als TCS-Identifikationsstartpunkt wurde im Organismus *M. pneumoniae* eine PSI-BLAST Suche durchgeführt, welche erst nach mehreren Iterationen neben einigen UPF (**U**ncharacterized **P**rotein **F**amilies) auch einen Zusammenhang des Proteins MPN013 zu einer HisKA-Domäne aufdeckte. MPN013 gehört zu der nicht annotierten Proteinfamilie DUF16, welches nur in *Mycoplasma* bekannt ist (Swiss-Prot: Y014_MYCPN).

Um diesen ersten Hinweis auf MPN013 als potentielles Sensorprotein zu festigen, wurden Analysen der Primär-, Sekundär- und Tertiärstruktur durchgeführt und auf TCS-Tauglichkeit untersucht.

Zur Überprüfung der Primärstruktur wurde mittels einer PSI-BLAST Analyse das *M. pneumoniae* Protein MPN013 als mögliches Sensorprotein identifiziert. Die ähnlichste gesicherte HisKA-Sequenz zu MPN013 ist die Sequenz des Sensorproteins NarX aus *Psychrobacter arcticus* (PSI-BLAST E-value $6x10^{-13}$ nach fünf Iterationen).

Anschließend wurden die Sekundär- und Tertiärstruktur von MPN013 analysiert, um zu überprüfen, ob sie der typischen Struktur eines Sensorproteins entsprechen. Eine Homologiemodellierung mittels SWISS-MODEL ergab das Template PDB:2ba2 (Kristallstruktur von MPN010, ein weiteres Mitglied der DUF16 Familie) für MPN013 als verwandte Sekundär- und Tertiärstruktur. Dieses Template beschreibt eine 4-alpha-Helix für den N-terminalen Bereich von MNP013, so wie es für gewöhnliche HisKA-Domänen in Sensorproteinen üblich ist. C-terminal besitzt MPN013 eine zusätzliche

zweite Domäne, ebenfalls typisch für HisKA-Domänen. Detailliert kann die Struktur folgendermaßen beschrieben werden:
Die Sekundärstruktur von MPN013 beginnt wie alle Sensorproteine mit einer unspezifischen Domäne (1-120), welche eine Signalbindestelle repräsentieren könnte. Daran knüpft direkt eine alpha-Helix Struktur an (130-165). Unterstützt wird dieses Ergebnis nicht nur durch das Homologiemodell, sondern ebenfalls durch Sekundärstruktur- und Tertiärstrukturvorhersagen mittels PredictProtein [64] und Porter [65]. Diese Analyse ist i. B. für den C-terminalen Bereich interessant, da dieser bislang keinem bekannten Homologiemodell zugeordnet werden konnte. Der C-terminale Sequenzabschnitt besteht laut Vorhersagen aus einer Mischung aus Helices, Faltblättern und Loops. Auch wenn die Ergebnisse der Sekundärstruktur-Vorhersageprogramme nicht komplett übereinstimmten, so wurde die Ähnlichkeit zu einem alpha/beta-Sandwich dennoch von einer Alinierung der Sekundärstruktur mit der Software SSEA [66] bestätigt (ähnlich zu Ornithin Transcarbamylase aus *E. coli* [PDB: 1akm]).

Um die typischen Charakteristika von TCS weiter zu verdeutlichen, zeigt **Abbildung 3.2.5** eine multiple Alinierung der Primärstruktur von MPN013 mit bekannten Sensorproteinen (NarX von *P. arcticus,* NarX von *E. coli*). Um auch die Sekundärstruktur zu vergleichen, wurde das Sekundärstrukturtemplate für HisKA 2c2a (violett) (HisKA853 von *Thermotoga maritima*) ebenfalls an die Primärstrukturen aliniert, gemeinsam mit der Sekundärstruktur des Homologiemodels für MPN013 (PDB:2ba2) (grün). Rote Punkte oberhalb der alinierten Sequenzen zeigen besonders wichtige und konservierte Aminosäuren in der HisKA-Domäne. Die schwarzen Konservierungssymbole unterhalb der Sequenz zeigen die Konserviertheit der Aminosäuren in der Alinierung aller gelisteten Sequenzen. Wichtige konservierte Aminosäuren in den bekannten Sensorpoteinen und in MPN013 wurden durch gelbe Markierungen hervorgehoben: Der erste gelbe Bereich der Alinierung markiert die stark konservierte Histidin-Umgebung, welche den Phosphor bindet und zum Regulationsprotein transferiert. Diese Histidin-Umgebung wurde im Bereich der 4-alpha-Helix lokalisiert. Bereits der Vergleich zwischen *E. coli, P. arcticus* und MPN013 zeigte, dass dieser Bereich variabel in seiner Position und Umgebung ist. Der Sekundärstrukturvergleich verdeutlicht, dass das Histidin am Ende der alpha-Helix lokalisiert sein muss. Der dritte und vierte Bereich enthält die ATP-Bindestelle und ist

stärker konserviert. Diese beiden Bereiche können mit dem ATP-Bindestellen PFAM-Motiv, bestehend aus den Aminosäurepattern Glu/Asn-X-Ile/Leu-X-Asn/Ala-X und Asp/Glu-X-Gly/Ser-X-Gly/Glu-Ile, beschrieben werden. Im Besonderen in den gelben Bereichen passen auch die Sekundärstrukturen zwischen MPN013 und den HisKA-Domänen gut zusammen. Für den C-terminalen Bereich von MNP013 konnte keine gesicherte Sekundärstruktur vorhergesagt werden.

Vergleiche mit der HisKA Subklassifizierung von Grebe [9] zeigen eine neue Histidinumgebung für das MPN013. Diese ist klar zu unterscheiden von den bisher beschriebenen Subklassen. Die stärkste Ähnlichkeit besteht zu einer Mischung aus den Klassen HK3b und HK11. Die Autophosphorylierungsstelle beinhaltet die konservierten Aminosäuren Histidin und Arginin, ähnlich wie in der Subklasse HK11. Innerhalb der ATP-Bindestelle ähnelt die neue HisKA-Subklasse aufgrund des konservierten Glycin der HK3b Subklasse.

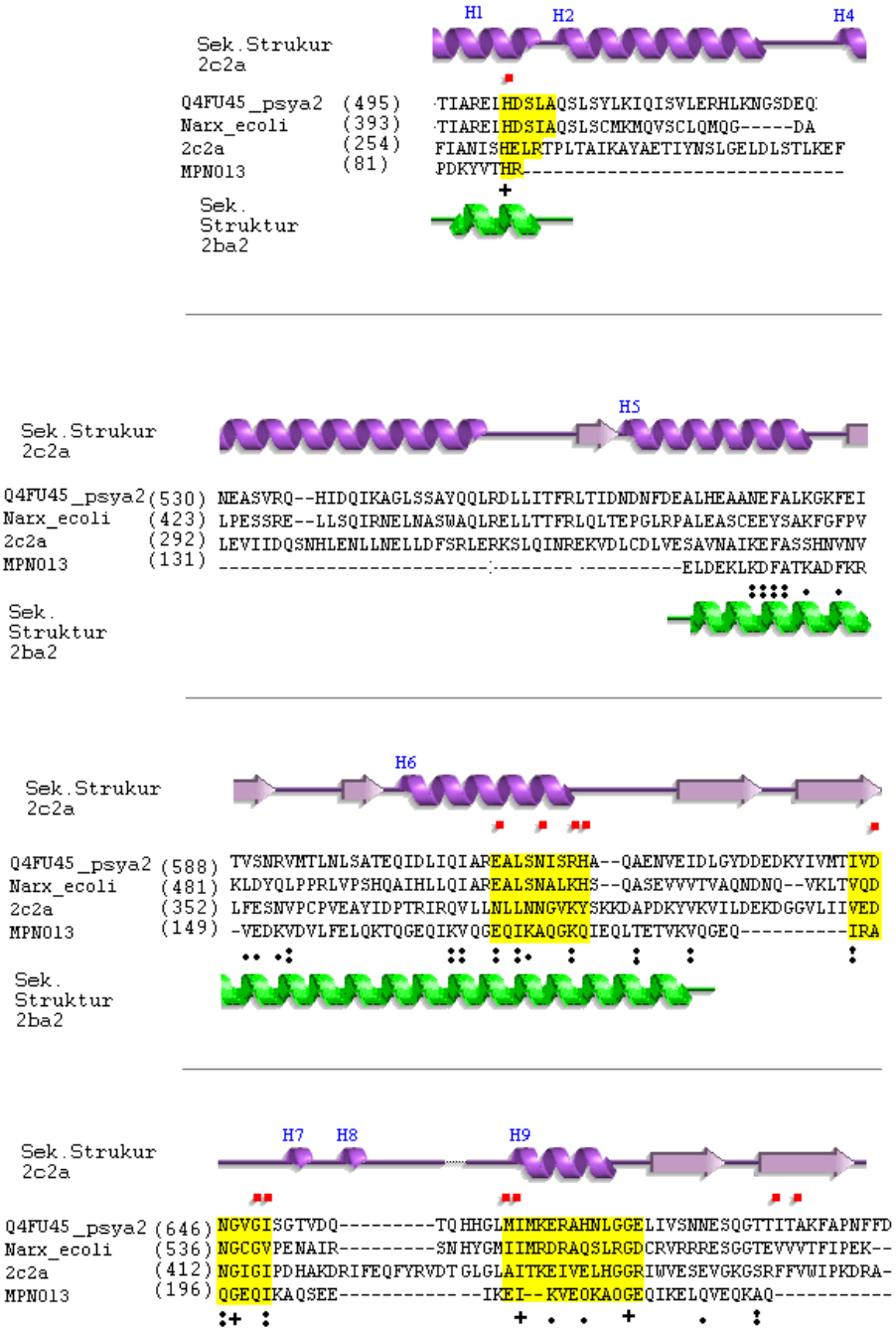

Abbildung 3.2.5: Vergleich des potentiellen Sensorproteins MPN013 mit bekannten Sensorproteinen.
Alinierung von MPN013 mit bekannten Sensorproteinen (NarX aus *E. coli*, Histidinkinasen-Homologiemodell 2c2a und sequenziell ähnliches Sensorprotein Q3Fu45 aus *P. arcticus*) in seiner Primär- und Sekundärstruktur (Aminosäureposition ist angegeben). Symbole **+ : .** (stark nach schwach) zeigen die Konserviertheit der Aminosäuren innerhalb der Alinierung. Oberhalb der Alinierung werden wichtige Aminosäuren für ein Sensorprotein angezeigt (rote Punkte). Gelbe Bereiche markieren besondere Ähnlichkeiten zwischen den Sequenzen. Die Sekundärstruktur für HisKA-Domänen (violett) ist oberhalb der Alinierung angebracht. Spiralen visualisieren alpha-Helices, während Pfeile für beta-Faltblätter stehen. Die alpha-Helices sind von H1 bis H9 durchnummeriert. Das Homologiemodell b2a2 für MPN013 befindet sich unterhalb der Alinierung (grün).

Zusammengefasst kann gesagt werden, dass Primär-, Sekundär- und Tertiärstruktur hohe Ähnlichkeiten zu Sensorproteinen im Allgemeinen aufweisen, auch wenn die Gesamtstruktur von MPN013 nicht komplett mit einem typischen Sensorprotein übereinstimmt. Der Sekundärstrukturvergleich zeigt, dass die Struktur eines Sensorproteins flexibler zu sein scheint als bisher angenommen. Die Zwischenregionen variieren stärker als bisher angenommen. MPN013 folgt den Bildungsregeln für bekannte TCS und unterstützt damit die Vorhersage.

Mögliches Regulationsprotein MPN014

Zusätzliche Hinweise für das degenerierte TCS in *Mycoplasma* konnten über die Suche nach dem zugehörigen Regulationsprotein gefunden werden.

Diese Suche wurde initiiert durch einen organismenspezifischen iterativen PSI-BLAST mit NarL von *P. arcticus*. NarL ist das korrespondierende Regulationsprotein zur HisKA aus NarX des Organismus *P. arcticus,* welches, wie in der Sensorproteinanalyse erwähnt, eine hohe Ähnlichkeit zu MPN013 aufweist. Eine organismenspezifische PSI-BLAST-Suche in *Mycoplasma* mit NarL ermittelte auf Primärstruktur-Ebene eine hohe Ähnlichkeit zu dem Protein MPN014 aus *Mycoplasma*. Dieses Ergebnis wurde weiterhin bestätigt durch Gen-Nachbarschaft Betrachtungen [85, 95], welche ebenfalls für TCS gültig sind [96]. Dabei liegen das Gen des Sensorproteins und das Gen des Regulationsproteins im Genom direkt hintereinander, was für MPN013 (ORF: D12_orf257; Genposition 14992 – 15765, positiv orientiert) und für MPN014 (ORF: DnaE (PR00162757) Genposition 15867 – 16505, positiv orientiert) zutrifft.

Um diese Hypothese auch auf Sekundär- und Tertiärstruktur-Ebene zu bestätigen, wurde ein Homologiemodell für MPN014 errechnet. Eine Sekundärstrukturalinierung zeigte eine Homologie zu NarL von *P. arcticus* und dem generellen Strukturtemplate PDB:1p2f (TM_0126 von *T. maritima*) für RR in TCS.

MPN014 ist bereits als mögliche Topoisomerase/DNA-Primase annotiert, wodurch der ebenfalls für Regulationsproteine typische DNA-Bindecharakter von MPN014 bestätigt wird.

Für einen detaillierten Vergleich wurde in **Abbildung 3.2.6** MPN014 in einer multiplen Alinierung mit NarL von *E. coli* und *P. arcticus* verglichen und die dazugehörigen Sekundärstrukturen (PDB:1p2f) darüber dargestellt (violett). Der Sequenzvergleich zeigte große Ähnlichkeit zu NarL in *P. arcticus*, NarL in *E. coli* und

MPN014 in *M. pneumoniae* (konservierte Aminosäuren in der Alinierung sind mittels schwarzer Symbole unterhalb der Alinierung hervorgehoben).
Das Phosphor-bindende alpha/beta 3-Layer-Sandwich (H1 bis H5) wird ebenso abgebildet wie die DNA-bindende alpha-Helix (H6 bis H11). Die Alinierung identifizierte konservierte Aminosäuren (bunt unterlegt). Der zweite Teil von MPN014 zeigte kein HTH-Motiv, aber die hohe Ähnlichkeit zum Toprim-Motif und seine Verbindung zum Proteincluster CLSK542094 (DNA-Primase related protein) unterstützt die These der DNA-Bindung, so wie es für viele als Transkriptionsfaktor wirkende Regulationsproteine in TCS üblich ist.

Allerdings wurde durch diesen Vergleich ebenfalls ersichtlich, dass MPN014 in einigen Punkten stark von einem typischen Regulationsprotein abweicht:

- Die Sequenz beinhaltet nur schwache hydrophobe Aminosäuren in der Region des beta-Faltblatts 1.
- Das konservierte saure Aminosäurenpaar für die Metallionenbindung während der Phosphorylierung ist bei MPN014 vorhanden, allerdings nur in Form von Glutaminsäure mit einem folgenden Glutamin anstelle einer Asparaginsäure (grün).
- Hydrophobe Aminosäuren für das beta-Faltblatt 3 und die folgende konservierte Asparaginsäure der Phosphorylierungsstelle (gelb) sind ebenso vorhanden wie hydrophobe Aminosäuren für das beta-Faltblatt 4. Das darauf folgende konservierte Serin oder Threonin, welches die Phosphorgruppe bindet und die Konformationsänderung bewirkt, wurde ersetzt durch ein Asparagin (zweite Zeile, mitte).

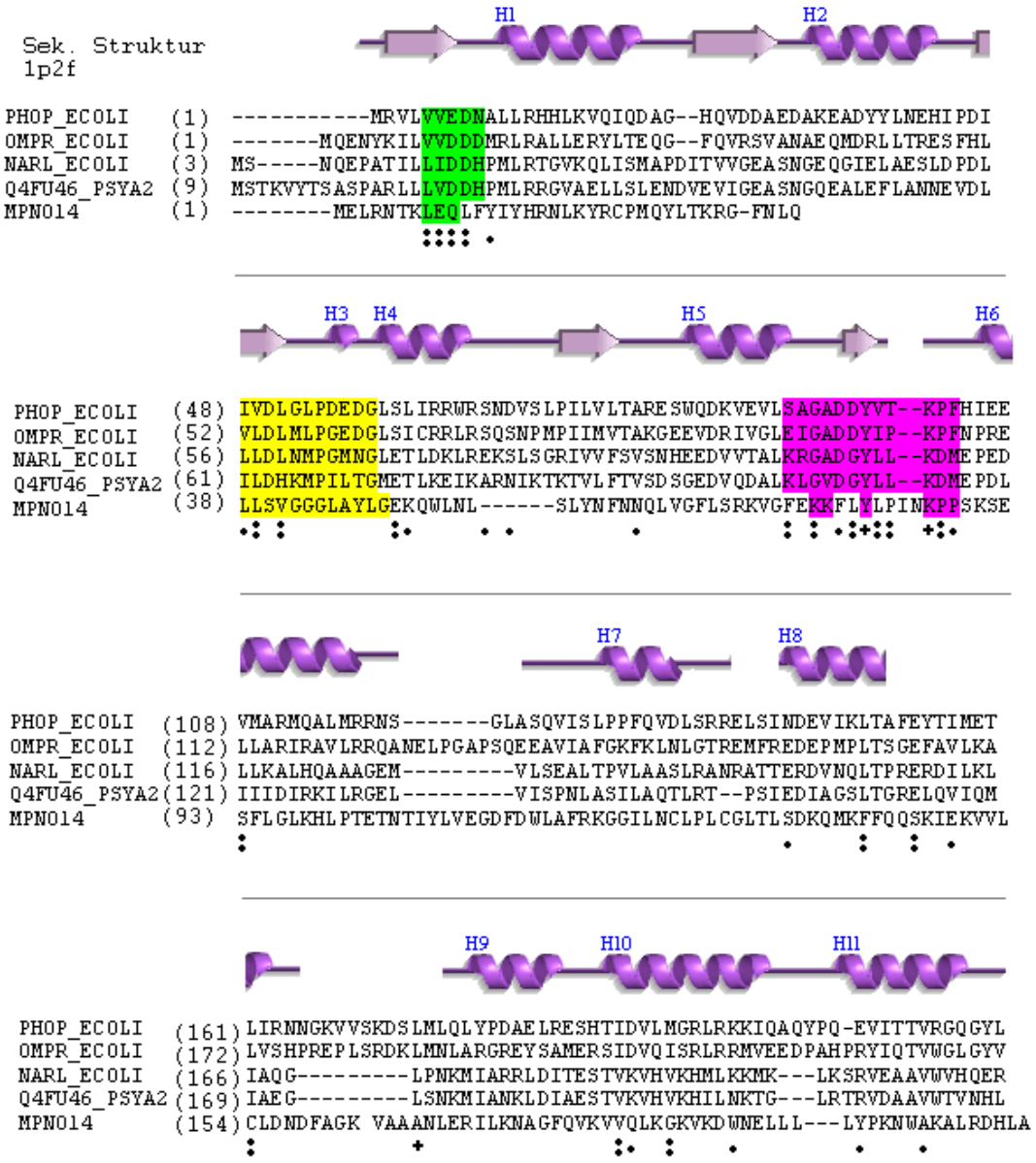

Abbildung 3.2.6: Vergleich des potentiellen Regulationsproteins MPN014 mit bekannten Regulationsproteinen.
Alinierung von MPN014 mit bekannten Regulationsproteinen (NarL, PhoP, OmpR aus *E. coli* und einem sequenziell ähnlichen Regulationsprotein Q4FU46 (NarL) aus *P. arcticus*) in seiner Primär- und Sekundärstruktur aliniert (Aminosäureposition ist angegeben). Die Symbole **+ : .** (stark nach schwach) zeigen die Konserviertheit der Aminosäuren innerhalb der Alinierung. Bunt hinterlegte Bereiche (grün, gelb, magenta) markieren besondere Ähnlichkeiten zwischen den Sequenzen. Die Sekundärstruktur für Responseregulator-Domänen ist oberhalb der Alinierung angebracht (violett). Spiralen visualisieren alpha-Helices, während Pfeile für beta-Faltblätter stehen. Die alpha-Helices sind von H1 bis H11 durchnummeriert. Das Phosphor bindende alpha/beta 3-Layer-Sandwich (H1 bis H5) wird ebenso dargestellt wie die DNA-bindende alpha-Helix (H6 bis H11).

Basierend auf der Summe der Ähnlichkeiten zu bekannten TCS können MNP013 und MPN014 trotz ihrer Abweichungen als degenerierte TCS angesehen werden.
Aufgrund der Familienzugehörigkeit können zusätzlich die ganze DUF16 Familie als mögliche Sensorproteine (MPN139, MPN138, MPN137, MPN130, MPN127, MPN104, MPN038, MPN013, MPN010, MPN655, MPN524, MPN504, MPN501, MPN410, MPN368, MPN344, MPN287, MPN283, MPN204) und die DNA-Primase Familie (MPN014, MPN353) als Regulationsproteine für *M. pneumoniae* betrachtet werden.

Untersuchungen im verwandten Organismus *Mycoplasma genitalium* konnten nur ein homologes Gegenstück zum MPN014 Protein in Form der DnaGp Familie (DNA Primase Familie) identifizieren. Eine analoge Familie zum Sensorprotein konnte dagegen nicht identifiziert werden.

Eine Veröffentlichung zu diesem Thema ist derzeit bei *Bioinformatics and Biology Insights* eingereicht.

3.3. Prozessstruktur biologischer Systeme

Nassexperimente zur Untersuchung von Proteinfunktionen oder ganzen biologischen Systemen sind sehr teuer und komplex. Daher ist es sinnvoll, vorhandenes Wissen als Basis für die Vorhersagen neuer Informationen zu nutzen, um Nassexperimente gezielter einsetzen zu können.
Eine Möglichkeit, vorhandenes Wissen über Proteinsequenzen und deren Funktionen wiederverwendbar und neu interpretierbar zu gestalten, wurde im vorherigen Kapitel anhand von TCS aufgezeigt.
Die beschriebene Vorgehensweise deckt sehr gut die Analyse konkreter Sequenzen in biologischen Systemen ab. Die synthetische Biologie versucht jedoch häufig, gesamte Funktionseinheiten in komplexen biologischen Systemen auszutauschen oder zu manipulieren. Um hierbei die synthetische Biologie besser unterstützen zu können, muss zunächst das komplexe System in seine funktionalen und organisatorischen Bestandteile zerlegt werden. TCS beispielsweise sind sehr einfache biologische Systeme, da sie nur aus zwei Einheiten (Detektionseinheit, Antworteinheit) zusammengesetzt sind, welche direkt miteinander in Verbindung stehen und reguliert werden. Bei komplexeren Systemen ist diese manuelle Zerlegung in Einheiten jedoch sehr mühsam und erfordert aufwändige Suchen.

Dieses Ergebniskapitel beschäftigt sich mit der Aufstellung einer abstrakten zweistufigen, strukturierten Modul-Klassifikation (Strukturierung), bei welcher technische/künstliche und biologische Denkweisen interdisziplinär gegenübergestellt und geclustert werden. Das Ziel ist die Beschleunigung und Vereinfachung der Prozessanalyse zu Beginn der Entwicklung von synthetischen Experimenten, ohne manuell die Unmengen an zur Verfügung stehenden Datenbanken manuell zu durchsuchen. Es soll durch die Zerlegung komplexer Prozesse in funktionale und organisatorische Einheiten die Planung von synthetisch-biologischen Experimenten vereinfachen. Zur Veranschaulichung und besseren Einsetzbarkeit der Klassifikation wurde eine Software entwickelt, welche für Technik (Ingenieurwesen) und Biologie gleichermaßen nutzbar ist und neue Blickwinkel eröffnen soll. Dazu werden folgende Schlüsselfragen analysiert:

- Wie können technische Prozesse biologischen Prozessen oder Prozesse in unterschiedlichen Organismen gegenübergestellt werden?
- Gibt es bestimmte Bereiche, bei denen die Technik der Biologie überlegen ist oder anders herum?
- Ist es möglich, den Aufwand und den Erfolg von Designexperimenten abzuschätzen?

Zur Beantwortung dieser Fragen gliedert sich die Arbeit in die folgenden Bereiche:

a) Erstellung und Validierung einer hierarchischen, zweistufigen Klassifikation auf biologischer und technischer Basis.
b) Untersuchung der statistischen Verteilung dieser Klassifikation in unterschiedlichen Organismen und in unterschiedlichen Bereichen der Forschung.
c) Vorstellung einer neu entwickelten Software (GoSynthetic) zur Beschreibung und zum Vergleich von biologischen und technischen Prozessen und deren Visualisierung.
d) Abwägung der Nutzbarkeit dieser Klassifikation und der Software mittels praktischer Beispiele.

Dieses Kapitel befasst sich mit der Behandlung eines typischen Problems der synthetischen Biologie, nämlich der Veränderung und dem Neudesign von Prozessen und Organismen. Die Standardisierung und Gruppierung der zweistufigen Klassifikation bilden eine Grundlage für Zeitersparnis bei der Prozessidentifikation und vereinfacht die Sammlung von bekanntem Wissen.

3.3.1. Design und systematische Kategorisierung biologischer Prozesse

Um die gestellten Fragen systematisch zu beantworten, wurde zunächst eine abstrakte, zweistufige, hierarchische Klassifikation zur Beschreibung von Prozessen etabliert. Die Klassifikation beinhaltet als Ausgangspunkt die Funktionen der Definition von Leben: Demnach muss ein Lebewesen die Fähigkeit besitzen, sich zu verändern und fortzupflanzen, es benötigt einen Stoffwechsel und eine abgegrenzte Struktur [97].

Die Verfeinerung und Vervollständigung der Klassifikation basierte auf einem iterativen Text-Mining Prozess, welcher als Grundlage die Gene Ontologie (GO) nutzt.

Die Nutzung von GO-Begriffen bietet zwei große Vorteile. Zum einen beinhaltet GO eine sehr gute, vorstrukturierte Datengrundlage, da sich ein ganzes Konsortium mit dessen Etablierung beschäftigt, zum anderen ermöglicht es ein vereinfachtes Text-Mining, da GO bereits aus Schlagworten besteht und somit die normalen Text-Mining Probleme (wie z.B. unterschiedliche Bezeichner beschreiben die gleichen Funktionen oder unterschiedliche Schreibweisen) ausgeschlossen werden konnten.

Alle GO-Begriffe wurden über Text-Mining den Basiseinheiten der Klassifikation zugeordnet. Aus GO-Begriffen, welche im ersten Evaluationsprozess keiner Einheit der Klassifikation zugewiesen werden konnten, wurden iterativ neue oder veränderte Einheiten innerhalb der Klassifikation erstellt. Im Anschluss an die Neugliederung der Einteilungen wurde der gesamte Evaluationsprozess wiederholt. Die iterative Evaluation an sich ist auf dem Text-Mining Prinzip begründet, welches mittels Shell- und Perl-Skripten (*buzzword.sh* und *buzzword_search.pl*) automatisiert wurde. Am Ende erfolgte eine manuelle, stichprobenartige Überprüfung der Klassifikation.

Als Ergebnis des iterativen Prozesses konnte für die Biologie die folgende Haupteinteilung innerhalb der Klassifikation aufgestellt werden: „*Signaltransduktion*“, „*Regulation*“, „*Pfad*“, „*Umwandlung*“, „*Kontainer*“, „*Systemstatus*“, „*Komplex*“. Analog dazu wurde eine technische Einteilung aufgestellt und den biologischen Einheiten zugewiesen. Die technischen Einheiten lauten: „*Detektion*“, „*Kontrollsystem*“, „*Prozessstruktur*“, „*Umbau*“, „*Behälter*“, „*Prozessphase*“ und

„*Baueinheit*“. Die detaillierte Klassifikation mit Beispielen und einer Gegenüberstellung Biologie zu Technik ist in nachfolgender **Tabelle 3.3.1** enthalten.

Die Einheiten der ersten Hierarchiestufe können weiterhin in Untereinheiten unterteilt werden, welche die Bedingungen der ersten Hierarchiestufe genauer beschreiben. Beispielsweise kann die Einheit „*Pfad*“ in die Untereinheiten „*Linear*“ und „*Kreis*“ unterteilt werden.

BIOLOGISCHE EINTEILUNG ***(EN) – (DE)***	***TECHNISCHE EINTEILUNG*** ***(EN) – (DE)***
Sensing - Signaltransduktion Biologische Systeme besitzen Sensoren zur Aufnahme von Umwelteinflüssen und zur Reaktion darauf. Mögliche Einflussparameter: Temperatur, Ernährungszustand, Osmose. *Biologische Beispiele:* Zweikomponenten-Systeme *Untereinheiten: (en) – (de)* **Transmitter - Überträger** **Receiver - Empfänger** **Communication - Kommunikation**	**Detection - Detektion** Detektoren reagieren auf Änderungen in der Umwelt. Beispielsweise Temperatur-änderungen, Luftdruck oder Luftfeuchte. *Technisches Beispiel:* Zentralheizung mit Thermostat, Klimaanlage *Untereinheiten: (en) – (de)* **Sensor - Fühler** **Adjuster - Einsteller** **Reporter – Berichter**

Regulation - Regulation Regulation biologischer Systeme zur Aufrechterhaltung, Abdämpfung, Verstärkung und Abschwächung von speziellen Effekten. *Biologische Beispiele:* Blutgerinnung ist positiv, verstärkend reguliert *Untereinheiten: (en) – (de)* **Activation - Aktivierung** **Inhibition - Hemmung**	**Control System - Kontrollsystem** Nutzung von Kontrollsystemen zur Aufrechterhaltung, Abdämpfung und Verstärkung von speziellen Eigenschaften. *Technisches Beispiel:* Regelkreise, negative Rückkopplung in Warmwassermischer, Verstärker im Radio. *Untereinheiten: (en) – (de)* **Stimulation - Stimulation** **Reduction - Reduktion**
Pathway - Pfad Organisation biologischer Prozesse in Pfade, welche ablaufende Reaktionen zusammen-fassen. Reaktionen können z.B. linear oder zyklisch sein. *Biologische Beispiele:* Glykolyse ist ein linearer Pfad mit einer Verzweigung. Der Citratzyklus ist zyklisch *Untereinheiten: (en) – (de)* **Linear - Linear** **Branching - Verzweigung** **Cycle - Kreis** **Cascading - Kaskadieren** **Network - Netzwerk**	**Process Structure - Prozessstruktur** Prozessstrukturen beschreiben geometrische Wege von Prozessen oder Wege von transportierten Produkten. Wege können z.B. linear oder zyklisch sein. *Technisches Beispiel:* Zyklischer Prozessverlauf im Kühlschrank (Kältekreislauf), lineare Lichtausbreitung *Untereinheiten: (en) – (de)* **Conveyer - Förderband** **Splitting - Aufsplittung** **Circulation - Zirkulation** **Cascading - Kaskadieren** **Meshwork - Geflecht**

Transformation - Umwandlung Produkte werden nicht verbraucht, sondern umgewandelt, meist in eine andere Energieform. Die Transformation beschäftigt sich mit der Energieproduktion, dem Verbrauch und der Umwandlung von Stoffen in andere Formen. *Biologische Beispiele:* Nahrung wird in Energie in Form von ATP umgewandelt. *Untereinheiten: (en) – (de)* **Anabolism - Anabolismus** **Catabolism - Katabolismus** **Metabolism - Metabolismus** **Degradation - Zersetzung** **Synthesis - Synthese**	**Conversion - Umbau** In der Technik müssen Produkte recycled werden. Energie wird „produziert“, „verbraucht“ und dabei in eine andere Energieform umgewandelt. *Technisches Beispiel:* Müllrecycling und Energieumwandlung in Photovoltaikanlagen oder Windkraftwerken. *Untereinheiten: (en) – (de)* **Energy Consumption - Energieverbrauch** **Energy Production - Energieproduktion** **Remodeling - Umstrukturierung** **Disassembling - Zerlegung** **Production - Produktion**
Container - Kontainer Kontainer werden benötigt, um Produkte zu transportieren oder Informationen zu speichern. *Biologische Beispiele:* Vakuolen zum Substanztransport. Viren, DNA, Sporen zur Informationsspeicherung und –weitergabe. *Untereinheiten: (en) – (de)* **Information Storage - Erbinformation** **Transport Container - Träger** **Transport Route - Transportweg**	**Box - Behälter** Behälter werden zur Produktspeicherung oder zum Transport benötigt. Produkte können abstrakte Informationen oder konkrete Warengüter sein. *Technisches Beispiel:* DVD, LKW, Lagerhalle. *Untereinheiten: (en) – (de)* **Data Medium - Informationsspeicher** **Carrier - Beförderer** **Highway - Transportweg**

System state - Systemstatus Biologische Entitäten besitzen unterschiedliche Stati innerhalb ihres Lebenszyklus, z.B. wachsen, vermehren, metabolisieren, sterben. *Biologische Beispiele:* Eine Zelle ist aktiv, inaktiv oder kann sterben. *Untereinheiten: (en) – (de)* **Active - Aktiv** **Dying - Tod** **Inactive - Inaktiv** **Out Of Control – Ausser Kontrolle**	**Process phase - Prozessphase** Mit einer Prozessphase wird der Zustand oder Abschnitt eines technischen Systems beschrieben. *Technisches Beispiel:* Ein Prozess kann warten, aktiv, beendet, an, aus oder auch unkontrollierbar sein. *Untereinheiten: (en) – (de)* **On - An** **Off - Aus** **Standby - Warten** **Broken - Defekt**
Complex - Komplex Biologische Systeme bestehen aus Komplexen. Die komplette Funktionalität wird durch das Zusammenspiel aller Untereinheiten eines Komplexes erreicht. *Biologische Beispiele:* Haemoglobin im Komplex mit Haem, Rubisco, ... *Untereinheiten: (en) – (de)* **Adjuvant - Helfer** **Barrier - Barriere** **Subunit - Untereinheit** **Whole Complex - Komplex**	**Assembly - Baueinheit** Technische Geräte bestehen aus mehreren Baueinheiten, welche bereits funktionstüchtig sind, deren Gesamtfunktion aber erst im Zusammenspiel erreicht wird. *Technisches Beispiel:* Ein Auto besteht aus Motor, Fahrgestell, Reifen,... *Untereinheiten: (en) – (de)* **Cooperator - Helfer** **Obstacle - Widerstand** **Component - Komponente** **Entity - Einheit**

Tabelle 3.3.1: Beschreibung der Klassifikation mit ihren Einheiten und deren Untereinheiten in Biologie und Technik. Die Tabelle beschreibt die zweistufige, hierarchische Klassifikation mit allen Einheiten und Untereinheiten. Zur Verdeutlichung werden die Einheiten beschrieben und Beispiele aufgelistet. Biologische und technische Einheiten werden gegenübergestellt.

Das Prinzip der Klassifikation kann beispielhaft am Prozess der Glykolyse verdeutlicht werden: Die Glykolyse wird von der Klassifikation durch die Einheiten „*Regulation*“ und „*Pfad*“ beschrieben. Weiterhin wird der Prozess als hauptsächlich „*Linear*“ beschrieben, beinhaltet aber auch die Einheit „*Verzweigung*“ aufgrund des Verzweigungspunkts (Fructose-1,6-bisphosphat) innerhalb der Glykolyse. Des Weiteren ist die Einheit der „*Umwandlung*“ in Form des Energiehaushaltes stark vertreten, da sich die Glykolyse mit der Energieproduktion aus Zucker beschäftigt.

Mit Hilfe dieser aufgestellten Klassifikation können im weiteren Verlauf alle biologischen Prozesse einheitlich beschrieben und mit der technischen Denkweise oder ganzen elektrotechnischen Prozessen verglichen werden.
Die GO-Begriffe an sich wären zu detailliert und zu stark auf die Biologie bezogen, als dass sie für die Übertragung auf andere Denkweisen nutzbar wären. Erst die erstellte abstrakte Klassifikation ermöglicht die große Vielfalt und Komplexität der Biologie einfach, aber vor allem vergleichbar zu anderen Wissenschaften darzustellen. Die Assoziation zu den technischen Begriffen ebnet und erleichtert somit den Weg zu anderen Denkweisen. Auf diese Art und Weise könnten recht einfach weitere Wissenschaftsbereiche und deren Vokabular zu dieser Klassifikation hinzugefügt werden.
Auch eine Aktualisierung oder Veränderung der Klassifikation ist mit Hilfe des Text-Minings relativ einfach möglich, da nur die Zuordnung entsprechend angepasst werden muss.

3.3.1.1 Strukturvergleich und Validierung mittels Gene Ontology, MIT BioBricks und COG

Um die Vollständigkeit der erzeugten Einheiten in der Klassifikation zu überprüfen, wurde eine auf Text-Mining bezogene Validierung gegen Gene Ontology, MIT BioBricks und *Clusters of Orthologous Groups of proteins* (COG) durchgeführt.
Bei einer Validierung gegen GO deckte die aufgestellte Klassifikation 95% aller Prozesse und Funktionen der GO-Klassifizierung ab.
Zusätzlich wurde die Abdeckung der Einheiten gegen die MIT BioBricks [32, 34] überprüft. Dabei stellte sich heraus, dass alle BioBricks durch die Klassifikation

abgedeckt werden können, die BioBricks dagegen kein Äquivalent für die Einheiten „*Prozessphase*“ und „*Prozessstruktur*“ besitzen.

Dieses Phänomen ist durch die gegensätzlichen Ausgangspunkte der BioBricks (tatsächliche biologische Einheiten) und der Klassifikation (Prozesse in Biologie und Technik) zu erklären. Die vollständige Zuordnung zwischen der Klassifikation und den BioBricks ist **Tabelle 3.3.2** zu entnehmen.

TECHNISCHE EINHEITEN	*BIOBRICKS*
Baueinheit	Protein domain
Komponente	Protein coding sequences/ Ribosome Binding Sites
Komponente	Conjugation/ Coliroid
Einheit	Translational units
Widerstand	Terminators
Behälter	Plasmids
Beförderer	Plasmid backbones
Informationsspeicher	DNA/Protein coding sequences
Transportweg	Motility and chemotaxis
Detektion	Motility and chemotaxis
Einsteller	Receivers and senders/ Measurement devices
Berichter	Reporters/ Measurement devices
Fühler	Measurement devices/
	Cell-cell signalling and quorum sensing
Kontrollsystem	
Reduktion	
Stimulation	Promoters
Umbau	DNA recombination
Zerlegung	
Energieverbrauch	
Energieproduktion	
Produktion	Protein generators/ Protein coding sequences/ Ribosome Binding Sites
Umstrukturierung	Biosynthesis

Prozessphase	
Defekt	
Aus	
An	
Warten	
Prozesstruktur	
Kaskade	
Zirkulation	
Förderband	
Geflecht	
Aufsplittung	

Tabelle 3.3.2: Auflistung technischer Prozesse und deren Abdeckung mittels BioBricks des MIT. Die erste Spalte enthält die technischen Einheiten der Klassifikation. Die zweite Spalte beinhaltet die zugeordneten BioBricks. Leere Bereiche konnten nicht zugeordnet werden.

Neben GO und MIT wurde die Klassifikation gegen die COG-Klassen [98] validiert, welche automatisiert den Einheiten zugeordnet wurden. Dabei konnten auf Anhieb 80% der COG-Klassen zugeordnet werden. Die fehlenden 20% wurden gebildet von Komplexen ohne weitere Funktionsangabe und sind daher nicht automatisch abdeckbar.

3.3.2. Untersuchung der Unterschiede in Organismen und verschiedenen Forschungsfeldern

Im folgenden Kapitel wird die aufgestellte und auf Vollständigkeit überprüfte Klassifikation auf unterschiedliche Organismen und verschiedene Forschungsfelder angewendet und seine Verteilung darin überprüft.

3.3.2.1 Verteilung der Klassifikation in unterschiedlichen Organismen

In diesem Abschnitt wird die Frage beantwortet ob unterschiedliche Organismen, unterschiedliche Funktionsschwerpunkte setzen. Für einen Vergleich der Verteilung der Einheiten aus der Klassifikation in unterschiedlichen Organismen wurden die Prozesse von 17 Beispielorganismen unter Anwendung der Klassifikation untersucht (*Arabidopsis thaliana, Bos taurus, Caenrohabditis elgans, Candida albicans, Danio rerio, Drosophila melangoaster, Escherichia coli, Gallus gallus, Homo sapiens,*

Listeria monocytogenes, Mus musculus, Pseudomonas aeruginosa, Rattus norvegicus, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Staphylococcus aureus, Vaccinia Virus).

Die Verteilung der Einheiten wird in **Abbildung 3.3.1** dargestellt (erstellt mit dem Skript *organism.sh*). Die Y-Achse zeigt die prozentuale Verteilung der Einheiten normiert pro Organismus. Die X-Achse zeigt die biologischen und technischen Einheiten.

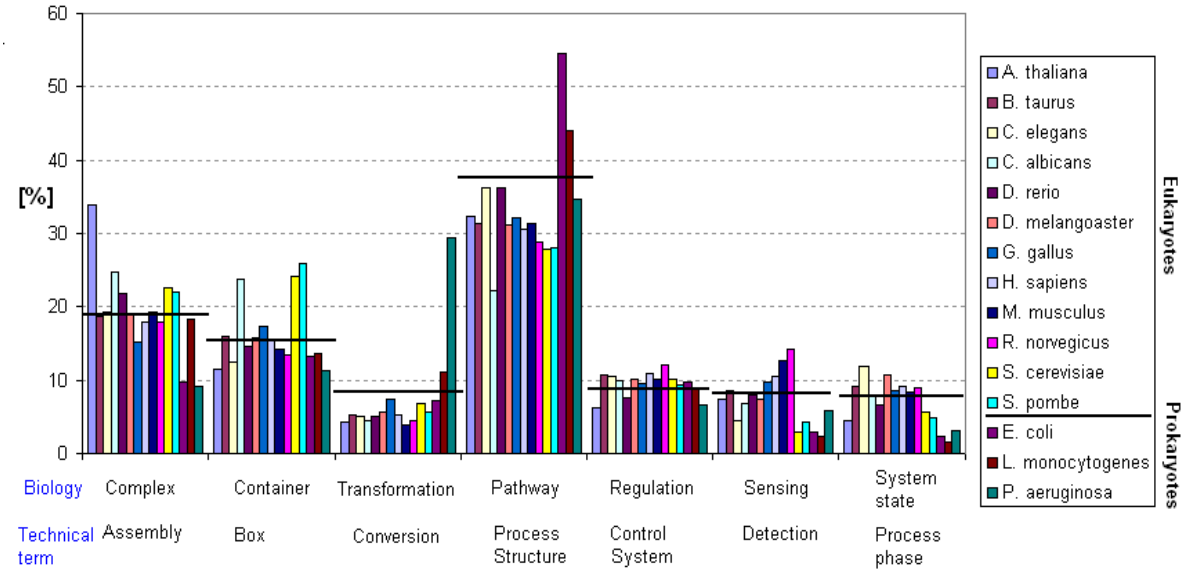

Abbildung 3.3.1: Einheitenverteilung im Organismenvergleich.
Die Y-Achse zeigt die prozentuale Verteilung der Einheiten normiert pro Organismus. Die X-Achse zeigt die biologischen und technischen Einheiten. Waagrechte, schwarze Balken markieren die Mittelwerte pro Einheit.

Insgesamt kann für alle Organismen gezeigt werden, dass die Einheit „*Pfad*“/„*Prozessstruktur*“ die am stärksten vertretene Einheit in allen Organismen ist. Sie beschreibt den geometrischen Fluss zwischen einzelnen biologischen/technischen Einheiten oder auch den Informationsfluss (z.B. metabolischer Zyklus im Citratzyklus). Die zweithäufigste Einheit ist die „*Kontainer*“/„*Behälter*“ Einheit, welche Transportvehikel beschreibt. Die „*Komplex*“/„*Baueinheit*“ Einheit bildet ebenfalls einen dominanten Bereich. Es umfasst Komplexe und Fraktionen in biologischen und technischen Einheiten.

Die Abbildung zeigt außerdem signifikante Unterschiede zwischen den Organismen, i. B. zwischen Eukaryoten und Prokaryoten. Statistische Berechnungen (Studentische Verteilung, einseitiger T-TEST mit zwei Stichproben bei gleicher Varianz) bestätigen den optischen Eindruck. Die Gleichheit bei zwei normalverteilten Gruppen (Eukaryoten

gegen Prokaryoten) wurde überprüft. Der p-Wert gibt hierbei an ob die gefundenen Unterschiede statistisch signifikant sind. Dies ist der Fall bei einem p-Wert kleiner/gleich 0.05.
Prokaryoten (*E. coli, L. monocytogenes, P. aeroginosa*) legen ihren Schwerpunkt auf die Einheit "*Komplex*" (p-Wert: 0.04), "*Umwandlung*" (p-Wert: 0.04), "*Pfad*" (p-Wert: 0.02) und "*Systemstatus*" (p-value: 0.01), während Eukaryoten (*G. gallus, H. sapiens, M. musculus, R. norvegicus, B. taurus*) einen besonderen Fokus auf die Einheiten "*Signaltransduktion*" (p-Wert: 0.02) und "*Regulation*" (p-Wert: 0.05) legen.
Während die Biochemie von Bakterien komplexer ist (vielfältige Stoffwechselleistung: Atmung, Gärung, Chemotrophie; hohe physiologische Flexibilität und Stabilität unter extremen Bedingungen: Temperatur, Druck, pH) und daher vor allem "*Pfad*", "*Komplex*" und "*Umwandlung*" benötigen, sind Eukaryoten insgesamt komplexer im Aufbau (Zellkern, Zellkompartimente) und besitzen eine höhere Proteinvielfalt (alternatives Splicen), weshalb besonders „*Signaltransduktion*" und „*Regulation*" benötigt werden.

Zusätzlich kann festgestellt werden, dass die Einheit „*Komplex*" in Pflanzen stärker vertreten ist und „*Kontainer*" verstärkt in Pilzen, „*Umwandlung*" und „*Pfad*" dominieren in Bakterien und „*Signaltransduktion*" in höheren Organismen. Die „*Regulation*" ist bei allen Arten stark vertreten.

Die beschriebene Einheitenverteilung basiert auf der GO-Klassifizierung und kann analog zu der Einheitenverteilung auf Basis von Proteinen beobachtet werden.

3.3.2.2 Strukturvergleich zwischen Biologie, Synthetischer Biologie und Technik

Des Weitern wurde überprüft, ob in unterschiedlichen Bereichen der Forschung (generelle Biologie, synthetische Biologie, Elektronik) eine unterschiedliche Verteilung der Einheiten beobachtet werden kann. Zur Erstellung der im Folgenden genannten Zahlen wurde in diesen Bereichen der Literatur separat nach den Einheiten aus der aufgestellten Klassifikation gesucht, gezählt und prozentual ins Verhältnis zueinander gesetzt.
Der obere linke Bereich in **Abbildung 3.3.2** zeigt die Einheitenverteilung der allgemeinen Biologie, basierend auf MeSH Begriffen.

Der obere rechte Bereich in **Abbildung 3.3.2** zeigt die Einheitenverteilung bezogen auf die synthetische Biologie betreffende Literatur. Die Auswahl der Literatur erfolgte über MeSH-Begriffe.

Der untere Bereich in **Abbildung 3.3.2** zeigt die Einheitenverteilung der technischen Literatur. Die Literatur ist dabei begrenzt auf Veröffentlichungen in den technischen IEEE Magazinen (Journal of the Institute of Electronical and Electronics Engineers).

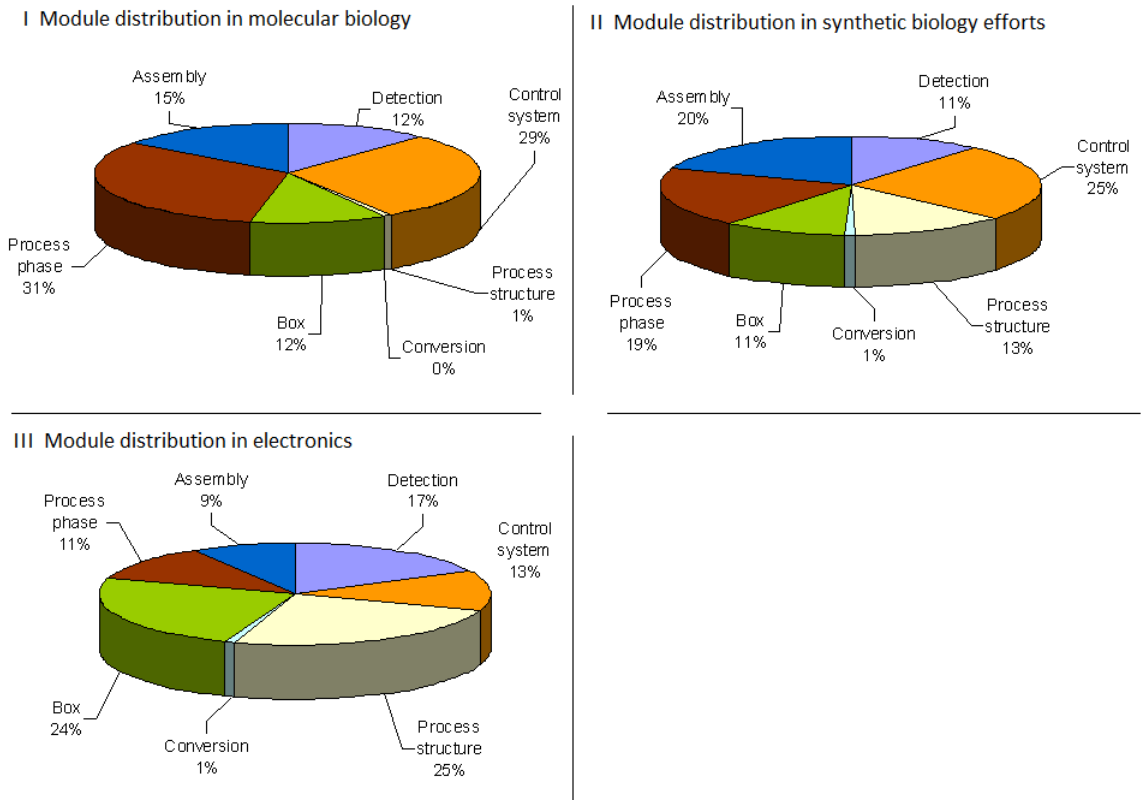

Abbildung 3.3.2: Einheitenverteilung in unterschiedlichen Bereichen der Literatur. Verglichen werden die generelle Biologie (oben links), synthetische Biologie (oben rechts) und die Elektronik (unten links).

Basierend auf den beschriebenen Beobachtungen kann festgestellt werden, dass in der generellen Biologie „*Prozessstruktur*“ unterrepräsentiert und „*Prozessphase*“ überrepräsentiert sind, während in der synthetischen Biologie genau das Gegenteil beobachtet werden kann. In der Elektrotechnik liegen die Themenschwerpunkte in der „*Detektion*“, „*Prozessstruktur*“ und „*Behälter*“ wohingegen „*Kontrollsystem*“ eher schwach vertreten war.

Demzufolge scheinen „*Prozessstruktur*“ in der synthetischen Biologie und der Technik überrepräsentiert zu sein, wohingegen der wichtige Bereich „*Prozessphase*“ im Gegensatz zur allgemeinen Biologie stark unterrepräsentiert ist. Die „*Prozessstruktur*“

schließt die Analyse und Modifikation des Prozessflusses ein. Die „*Prozessphase*“ dagegen beschreibt den Status eines Prozesses oder eines Systems. Demzufolge deckt die Literatur der Elektronik über diese beiden Einheiten 36% aller Themen ab, wohingegen die Literatur der Molekularbiologie darüber 32% abdeckt.

Die „*Kontrollsystem*“ Einheit scheint in der synthetischen Biologie (25%) stärker vertreten zu sein als in der Technik (13%).

Die beobachteten Differenzen reflektieren die unterschiedlichen Aufgabenschwerpunkte der verschiedenen Ansätze. Während in der generellen Literatur der Molekularbiologie über die Beschreibung und Manipulation von allen Einheiten diskutiert wird, so wird in der Technik und in der synthetischen Biologie der Forschungs- und damit Veröffentlichungsschwerpunkt auf die Manipulation und die *de novo* Generation von Einheiten gesetzt.

Daraus kann weiterhin geschlossen werden, dass die technische Manipulation der Einheit „*Prozessstruktur*“ weniger komplex ist als die Manipulation der Einheit „*Prozessphase*“ und daher auch seltener gefunden werden kann. Während in der Elektronik die meisten Bereiche eher konstant und träge reagieren, muss in der synthetischen Biologie verstärkt auf die „*Kontrollsystem*“ Einheit zurückgegriffen werden, um i. B. veränderte Systeme zu beherrschen. Dieses bereits von Boyle und Silver beobachtete Phänomen [99] konnte hiermit nicht nur quantitativ bestätigt werden, sondern es konnten zusätzlich die einbezogenen Untereinheiten überprüft werden.

3.3.3. Eine Software zum Prozessdesign in der synthetischen Biologie - GoSynthetic

Nach der Beschreibung des Grundprinzips und der Vorstellung des breiten Einsatzbereiches der Klassifikation, wird im Folgenden die Software GoSynthetic vorgestellt. Sie stellt die beschriebene Klassifikation zur Verfügung und integriert sie in bekanntes Wissen. Die Software liefert eine schnelle und einfache Sofortabschätzung über zu analysierende Prozesse. Zudem kann sie die generelle Datensammlung erheblich erleichtern und ermöglicht Vergleiche zu anderen Organismen oder Technologien.

Alle beschriebenen Daten sind in der Software GoSynthetic (http://gosynthetic.bioapps.biozentrum.uni-wuerzburg.de/index.php) abrufbar.

3.3.3.1 Funktions- und Oberflächenbeschreibung

Das Bildschirmfoto (**Abbildung 3.3.3**) vermittelt einen graphischen Eindruck der Software GoSynthetic.

Prinzipiell können darin alle interessanten Prozesse als Suchbegriffe verwendet werden, um sie in ihrem biologischen Kontext in den 17 genannten Modellorganismen zu untersuchen.

Die Software kann leicht über den hierarchischen Modulbaum der Klassifikation durchsucht werden. Dieser enthält eine Zuordnung zwischen biologischen und technischen Begriffen, die dazugehörige GO-Klassifizierung, COG-Familien, sowie alle zugehörigen Proteine und Interaktionsinformationen. Querverweise zu den öffentlichen Proteindatenbanken Craig Venter Institute, The Arabidopsis Information Ressource (TAIR), WormBase, Candida Genome Database (CGD), Flybase, EcoCyc & EcoliHub, Mouse Genome Database (MGD), PseudoCAP, Uniprot, Rat Genome Database (RGD), Saccharomyces Genome Database (SGD), Sanger GeneDB, Zebrafish Model Organism Database (ZFIN) sind ebenfalls vorhanden.

Für einen einfachen Zugriff existieren drei unterschiedliche Suchszenarien, die sogenannte Stichwortsuche, die Prozesssuche und die Sequenzsuche. Die Suchergebnisse werden durch Texthervorhebung gekennzeichnet, eine graphische Statistik sowie ein Interaktionsnetzwerk sind ebenfalls erhältlich:

a) *Stichwortsuche:* Die Stichwortsuche ist zunächst ähnlich der AmiGO-Suche (Oberflächensuche von GO). Das Ergebnis ist eine Liste aller Einheiten, welche einen GO-Eintrag mit dem Suchbegriff enthalten. Der Suchbegriff ist farblich gekennzeichnet. Dieses Verfahren ist einerseits sehr einfach, andererseits trotzdem sehr nützlich für die Suche nach GO-Annotationen und Funktionen.
b) *Prozesssuche:* Die Prozesssuche befasst sich mit der Suche nach biologischen und technischen Prozessen. Im Hintergrund wird zunächst eine reguläre Ausdruckssuche mit einem Suchbegriff gestartet. Alle in dem Suchbegriff enthaltenden GO-Begriffe mit den dazugehörigen Proteinen werden zu der Suche assoziiert. Diese neue Suche schließt die in der vorherigen Suche aufgefundenen GO-Begriffe und Proteine mit ein. Damit können sowohl über die Prozesssuche als auch mit dem Suchbegriff assoziierte Prozesse, Funktionen und Proteine aufgefunden werden. Die Ergebnisdarstellung erfolgt textuell als Liste, aber auch in einer graphischen Statistik in Form von Kreisdiagrammen.

Dabei wird ein Kreisdiagramm für die erste Klassifikationsstufe der Einheiten erstellt und jeweils ein eigenes Kreisdiagramm pro Haupteinheit zur Darstellung seiner Untereinheiten-Verteilung.

c) *Sequenzsuche:* Um nicht auf die Modellorganismen beschränkt zu sein, existiert eine weitere Suchoption - die Sequenzsuche. Dabei kann der Benutzer eine beliebige Sequenz eingeben und erhält über eine automatisch im Hintergrund ausgeführte BLAST-Suche die Modulanalyse des ähnlichsten Proteins aus den implementierten Modellorganismen. Die Analyse des ähnlichsten Proteins erfolgt über eine Prozesssuche. Das Ergebnis wird daher ebenfalls in Form von Kreisdiagrammen der Module dargestellt. Diese Sequenzsuche ist sehr variabel einsetzbar. Beispielsweise können damit sehr leicht unterschiedliche Proteine in ihrer Funktion untersucht und vergleichen werden, aber auch fusionierte Proteine können damit untersucht werden.

Eine weitere Möglichkeit von GoSynthetic wird durch die Verknüpfung zwischen der Protein-Protein Interaktionsdatenbank IntAct, der Gene Ontology und der zweistufigen Klassifikation im Interaktionsmodus (Oberflächenbereich *VRelation*) geschaffen. Diese Anbindung ermöglicht die Identifikation von Interaktionen zwischen Proteinen und Verbindungen zwischen biologischen Prozessen in Form von Prozessinteraktionsnetzwerken. Als Suchbegriffe dienen Prozesse, Funktionen oder Teile daraus. Im Hintergrund werden alle Proteine zu dem Suchbegriff und die zugehörigen Interaktionen herausgesucht. Die Ergebnisdarstellung erfolgt als graphisches Proteininteraktionsnetzwerk. Es kann zwischen den technischen und den biologischen Einheiten leicht gewechselt werden. Dies erlaubt die einfache Konstruktion von eigenen, neuen Netzwerken auf Basis der gewünschten technischen oder biologischen Einheiten. Innerhalb des Netzwerkes verdeutlichen rote Kreise Aktivierungsprozesse, blaue Kreise Inaktivierungsprozesse und grüne Kreise weisen auf eine Relation zum Metabolismus hin. Konkrete Anwendungsbeispiele können dem Kapitel 3.3.4 entnommen werden.

Das Suchergebnis des Prozessinteraktionsnetzwerkes kann nach Cyptoscape [100] exportiert werden, um dort die dynamische Visualisierung und Manipulation der molekularen Interaktionsnetzwerke fortzusetzen. Cytoscape ist eine Quellcode-offene Plattform zur Modifikation von Netzwerken und um weitere Informationen

(Genexpressionsmatrizen oder andere Daten) zu jedem einzelnen Element des Netzwerks dazu zufügen. In diesem beschriebenen Fall könnten außerdem die Interaktionen zwischen den Proteinen modifiziert und weitere Layouts hinzugefügt werden. Selbst neue Proteine könnten in das vorhandene Netzwerk eingebaut werden.

GoSynthetic kann damit nicht nur analysieren, sondern kann weiterhin als eine Art Ideengeber für die Prozessentwicklung und das Design neuer Experimente im Bereich der synthetischen Biologie genutzt werden.
Ein in GoSynthetic verfügbares Tutorium hilft bei den ersten Schritten mit der Software.

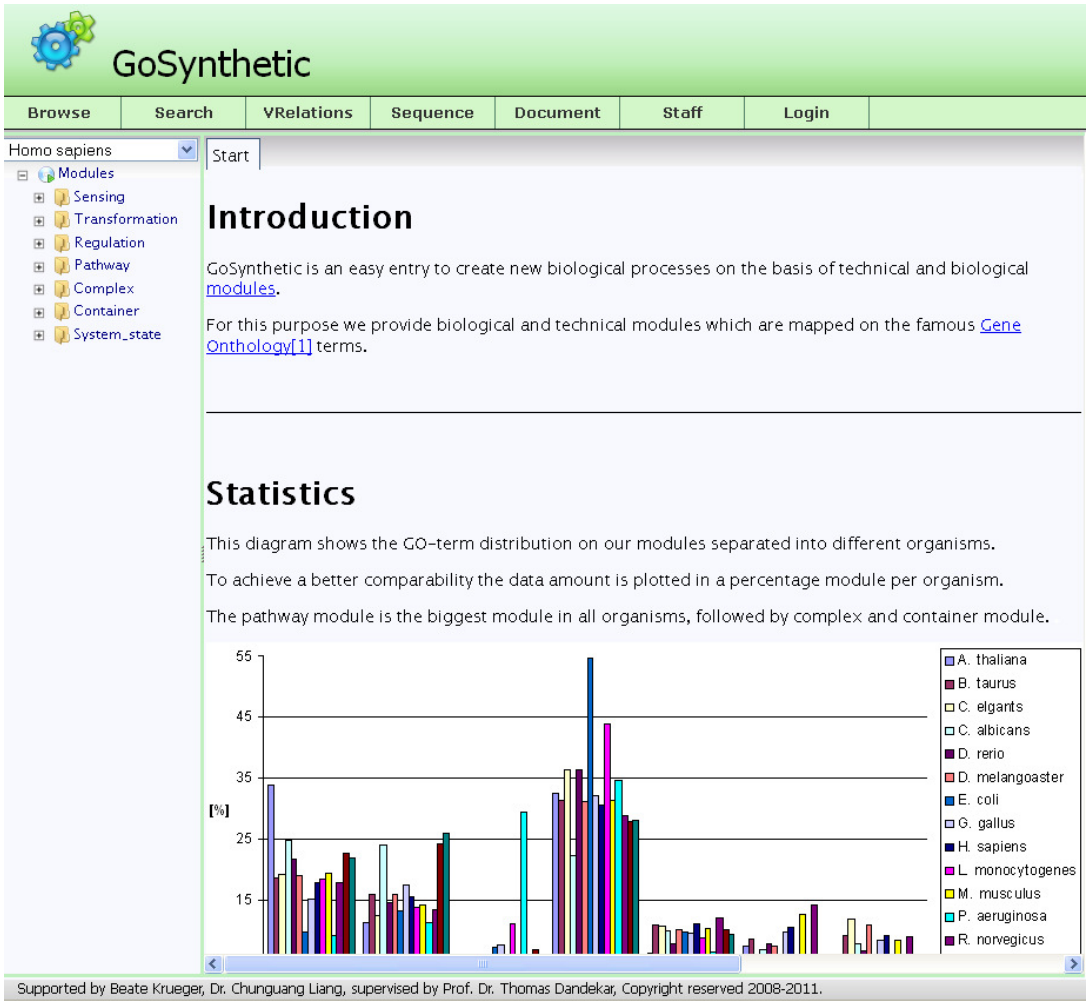

Abbildung 3.3.3: Startseite von GoSynthetic. Im linken Bereich ist die Klassifikation in Form eines Modul-Baumes dargestellt. Die Organismen können über die Auswahlbox ausgewählt werden. Der Bereich *Browse* dient der strukturierten Suche nach Einheiten, im Bereich *Search*, sind die beiden unterschiedlichen Suchformen zu finden. Der Bereich *VRelation* beinhaltet die Prozessinteraktionssuche und der Bereich *Sequence* die Sequenzsuche. Das Tutorium ist im Bereich *Document* abgelegt.

3.3.3.2 Datensammlung

Die Hauptdatenquelle der Software basiert auf den Daten des Gene Ontology Servers (Stand September 2011). Weitere Datenquellen wurden schrittweise dazu geladen und miteinander verbunden. Zu diesen Datenquellen zählen die Interaktionsdatenbank IntAct, die funktionellen Annotationen von KOG (Eukaryotic Orthologous Groups)/ COG (Clusters of Orthologous Groups of proteins) [98], NCBI Mesh terms und die BioBrick Liste des MIT.

Zahlenmäßig beinhaltet die Software derzeit 17 Organismen, 28.544 GO-Annotationen, 1.102 COG/KOG-Annotationen, 23.536 Interaktionen und 1.073.589 Protein-Funktionsrelationen.

Alle Datenquellen wurden als Textdateien von den dazugehörigen Datenservern übernommen, mit Hilfe von Perlskripten aufbereitet und in eine eigene Datenbank geladen (Skripte *erstelle_organism_files.pl, add_description.pl*). Die Verknüpfung der Daten konnte über allgemeine Proteinbezeichnungen und GO-Annotationen erreicht werden. Die Zuordnung zu der zweistufigen, biologischen und technischen Klassifikation erfolgte ebenfalls mit Hilfe eines iterativen Text-Mining Skriptes (*buzzword_search.pl*) in mehreren Stufen, wurde aber anschließend manuell stichprobenhaft überprüft.

3.3.3.3 Datenspeicherung und Verarbeitung

Die gesammelten und aufbereiteten Daten wurden in einer relationalen Datenbank gespeichert, welche in ihrem Aufbau der zweiten Normalform folgt. Die zweite Normalform beschreibt eine allgemeine Form der Datenablage in Datenbanken, welche folgende Bedingungen erfüllen muss [101]:

1. alle Werte in einer Tabelle beziehen sich auf den Schlüssel
2. der Primärschlüssel besteht aus nur einem Attribut

Diese Normalisierung bietet die Möglichkeit, beliebige neue Datenquellen einfach hinzuzufügen.

Das Datenbankmodell wird in der Form eines Entity-Relationship Diagramms (Beschreibung der Datentabellen in der Datenbank zueinander) in **Abbildung 3.3.4** dargestellt.

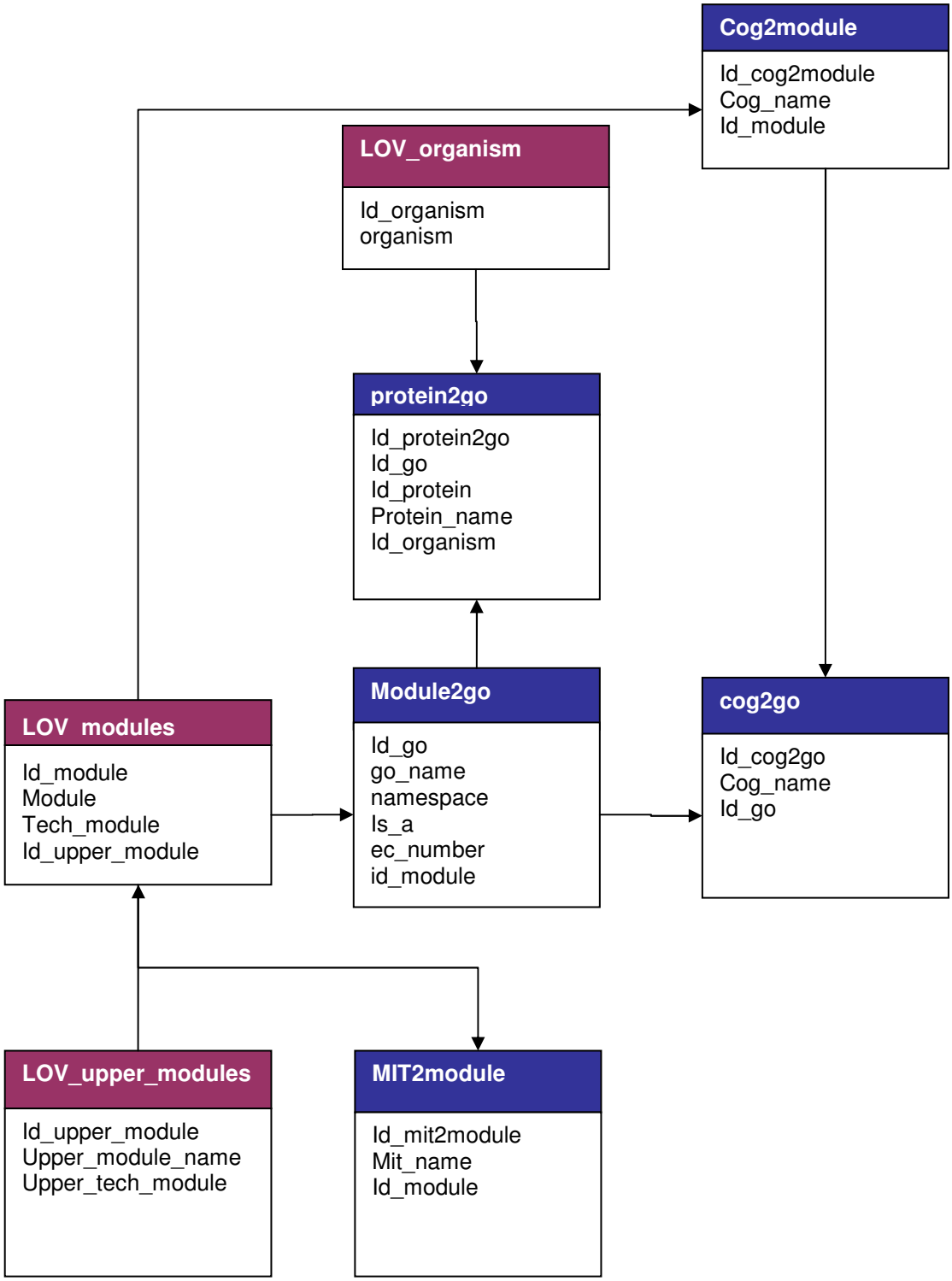

Abbildung 3.3.4: Entitiy-Relationship Diagramm der GoSynthetic Datenbank.
Rot markiert sind die Tabellen welche Masterdaten enthalten (Organismen und Moduleinheiten). In blau sind die Tabellen mit den eigentlichen Daten markiert. Die Pfeile stellen 1 : n Beziehungen zwischen den Tabellen dar.

Um aus der Software-Oberfläche einen schnellen Datenzugriff zu gewährleisten, wurden die Daten für die Anzeige auf der Oberfläche pivotiert, aufbereitet und mit Perl-Skripten (*relation_search.pl, intact_search.pl, protein_search.pl*) auf die spezifischen Abfragedaten eingeschränkt.
Aus diesem Grund können allerdings derzeit spezielle Abfragen ausschließlich direkt in der Datenbankumgebung vorgenommen werden.

3.3.3.4 Datenvisualisierung und Performance

Die Datenaufbereitung und Datenspeicherung in GoSynthetic wurde mittels Perl-Skripten und XML-Zwischenspeicherung realisiert. Für die Datenaufbereitung und Darstellung wurde auf weitläufig genutzte Techniken wie PostScript zurückgegriffen (*create_piechart.pl, intact_plot.pl*). Die Software selbst wurde mittels moderner Webtechniken (AJAX) realisiert. Details können dem Unterkapitel Software-Implementierung entnommen werden.

Aufgrund der beschriebenen Datenaufbereitung ist die Performance des Webservers sehr gut, die Antwortzeiten liegen trotz der großen Datenmengen im Sekundenbereich. Damit ist die Performance dieser Software vergleichbar mit ähnlichen Internetanwendungen wie beispielsweise STRING, Entrez und AmiGO.

3.3.4. Anwendungsbeispiele aus der synthetischen Biologie

Das folgende Kapitel beschreibt die Funktionsweise von GoSynthetic, deren Einsatzmöglichkeiten und Anwendbarkeit auf Basis von Anwendungsbeispielen aus der synthetischen Biologie.

Das Kapitel untersucht folgende fünf Fallbeispiele:

1. Analyse der Glykolyse: Generierung eines schnellen Überblicks über einen biologischen Prozess und dessen Vergleich zwischen Organismen.
2. Analyse eines technischen Kreislaufs: Zeigt den Weg von einem technischen Prozess zu einer biologischen Implementierung.
3. Analyse des Lotuseffekts: Zeigt den Weg von einem biologischen Prozess zu einem möglichen technischen Nachbau.
4. Analyse eines Netzwerkknotenpunktes: Zeigt die Erfolgswahrscheinlichkeit für die Modifikation eines bestehenden biologischen Prozesses.
5. Analyse eines modifizierten Organismus: Zeigt das einfache Auffinden von Verbesserungsmöglichkeiten des Designs im einem modifizierten Organismus.

Die einfacheren Beispiele sind komplett über die Oberfläche von GoSynthetic zu reproduzieren, komplexere Beispiele sind als Datenbasis in der Datenbank vorhanden, aber noch nicht vollständig über die Oberfläche abgedeckt.

Einheitenbezeichungen aus der Klassifikation werden aufgrund der Übersichtlichkeit im Folgenden durch Hochkommata gekennzeichnet.

3.3.4.1 Analyse der Glykolyse

Ziel und Hintergrund

Um mit der Handhabung und der Ausdrucksweise, aber auch mit der Prozessbeschreibung der Klassifikation vertraut zu werden, soll in diesem ersten Beispiel die Prozessanalyse als Ergebnis eines sehr einfachen Anwendungsfalls, nämlich der Glykolyse, vorgestellt werden. Wie bereits beschrieben können mit GoSynthetic Analysen gestartet werden, welche nach Proteinen, Prozessen, GO-Klassifizierung und/oder COG-Begriffen suchen oder nach der vorgestellten Klassifikation. Als Ergebnis werden, je nachdem welche Suche ausgewählt wurde, die Ergebnisse tabellarisch angezeigt und gegebenenfalls um weitere Informationen ergänzt.

Softwaresuche und gelieferte Ergebnisse

In diesem Glykolyse-Beispiel wurde eine Prozesssuche durchgeführt. Auf der Oberfläche muss dazu lediglich der Begriff „glycolysis“ eingegeben werden.

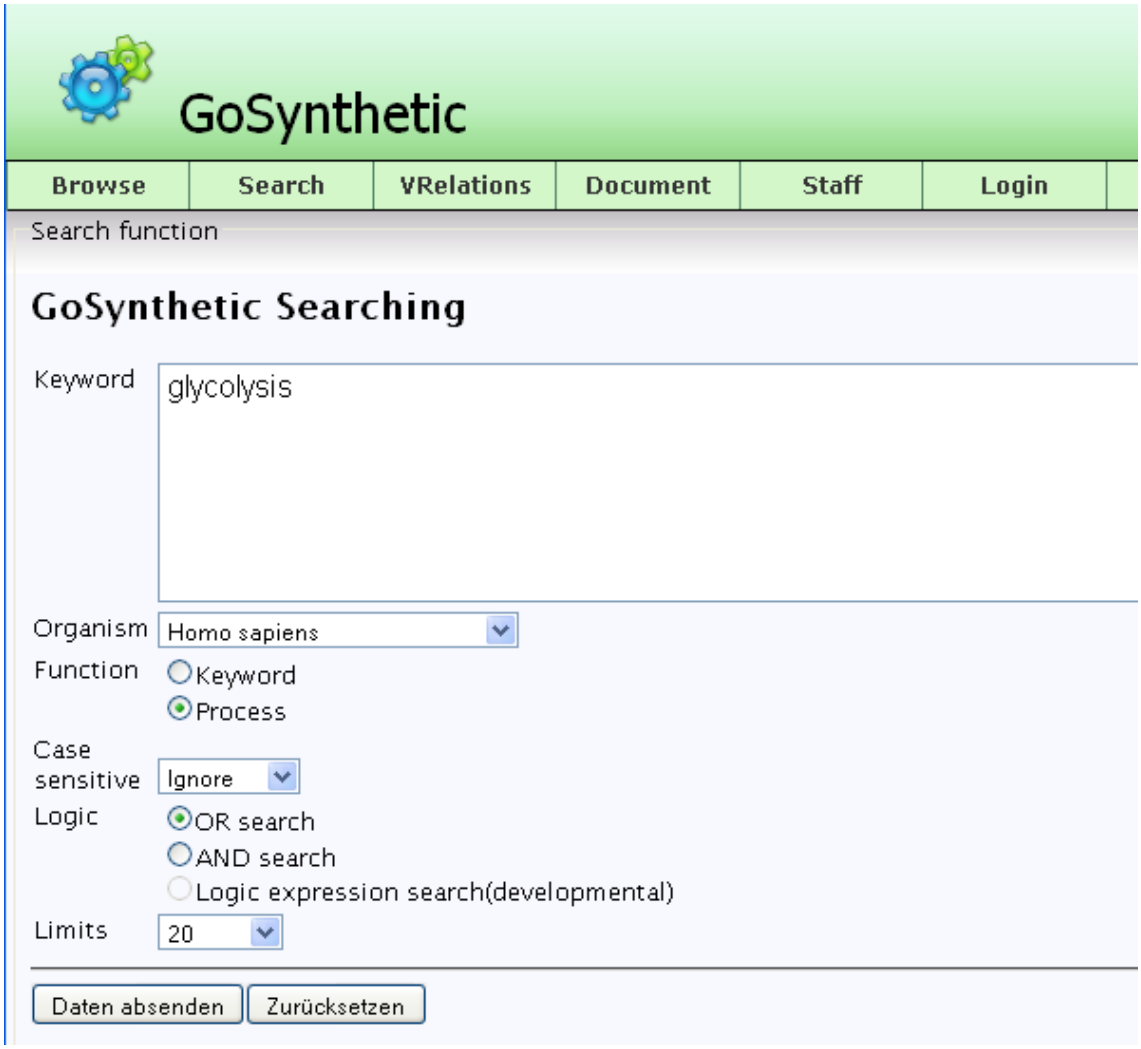

Abbildung 3.3.5: Suchoberfläche von GoSynthetic. Um den Prozess der Glykolyse zu untersuchen, muss der Bereich *Search* ausgewählt, der Begriff „glycolysis“ in den Suchbereich eingetippt, sowie der Organismus und die Suchart der Prozesssuche ausgewählt werden.

Als Ergebnis werden die in **Abbildung 3.3.6** aufgeführten Kreisdiagramme präsentiert. Im ersten Diagramm gewinnt man einen Überblick über die erste Hierarchieebene. Die kleineren Kreisdiagramme zeigen die Aufteilung in die Unterhierarchie. Gleich das Hauptdiagramm zeigt, dass die Glykolyse einen stark “*regulierten*“ „*Pfad*“ darstellt.

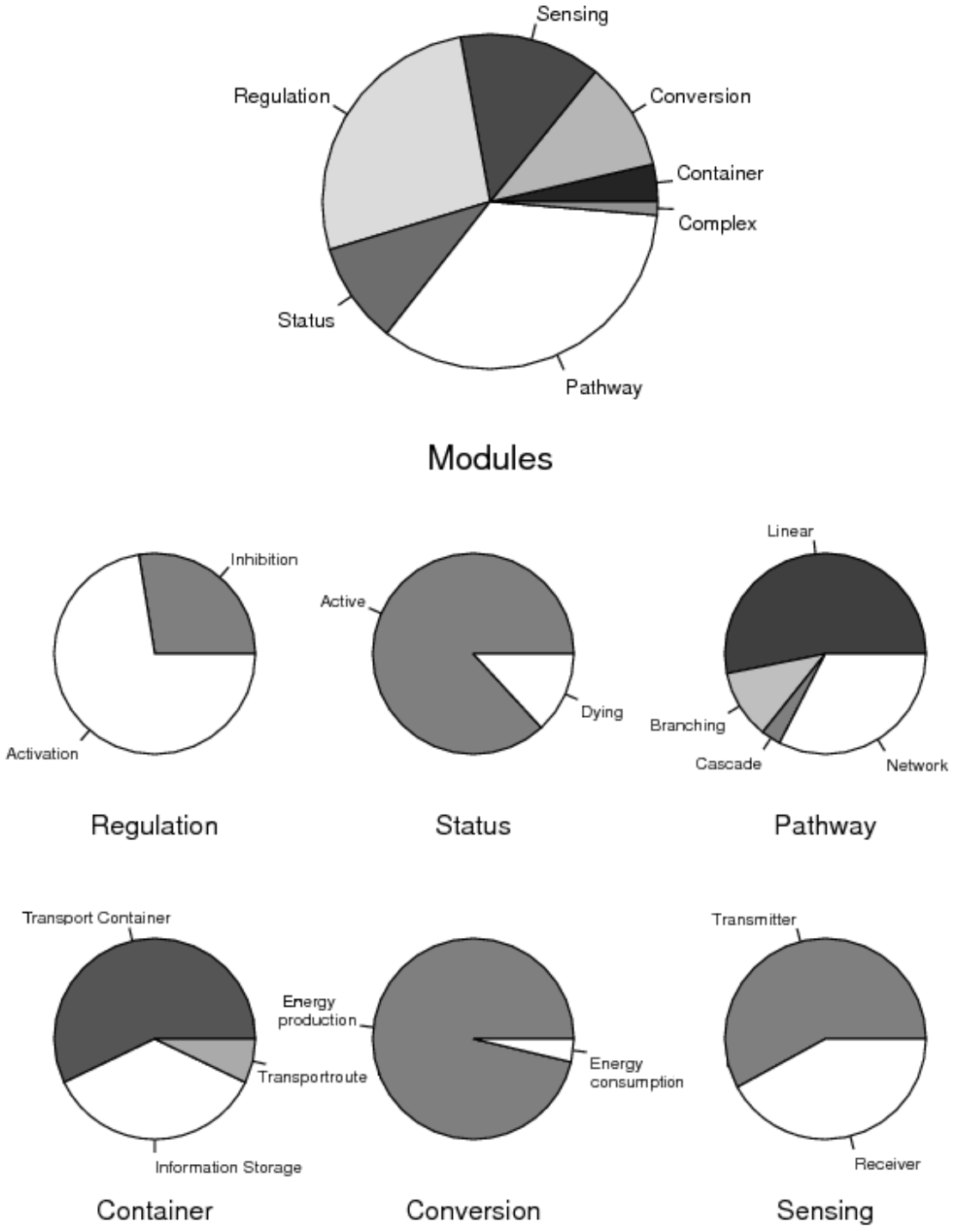

Abbildung 3.3.6: Einheitenverteilung in der Glykolyse. Das oberste Kreisdiagramm visualisiert die Verteilung der Einheiten der ersten Hierarchiestufe im Menschen. Die weiteren Kreisdiagramme beschreiben die Verteilung innerhalb dieser Einheiten, aufgeteilt in die Untereinheiten (Submodule).

Zusätzlich kann eine Interaktionssuche (*VRelation*) mit dem Suchbegriff „Glycolysis“ durchgeführt werden. **Abbildung 3.3.7** zeigt alle mit Interaktionen aufgefundenen menschlichen Proteine, welche in Verbindung zur Glykolyse stehen, deren Protein-Protein-Interaktionen und Verbindung zu anderen Prozessen. Aktivierende und inaktivierende Proteine sind farblich hervorgehoben.

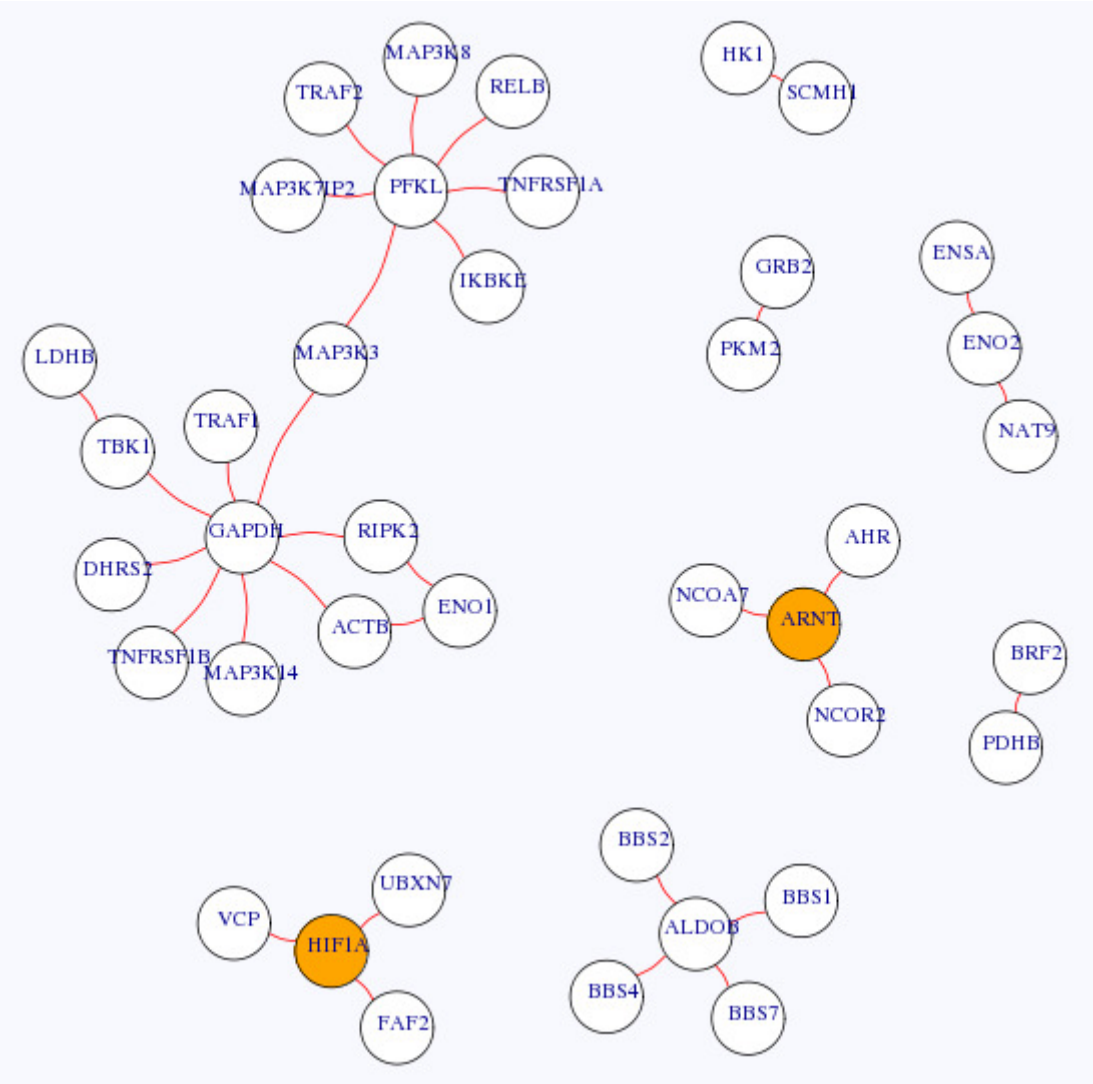

Abbildung 3.3.7: Interaktionen der Glykolyse (VRelation): Eingegebener Suchbegriff „glycolysis“, ausgewählter Organismus „Homo sapiens“. Die Netzwerk Visualisierung zeigt die interagierenden Proteine innerhalb des Suchprozesses und mit anderen Prozessen. Aktivierende Proteine sind in orange hervorgehoben.

Interpretation

Die Untereinheiten erklären den Prozess im Detail: Die Glykolyse stellt einen stark „*aktivierten*“ Prozess dar. Die Prozessstruktur ist vornehmlich „*linear“*, besitzt aber ebenfalls einen „*Verzweigungs*“ Bereich. Die Hauptaufgabe der Glykolyse besteht in der „*Umwandlung*“, vor allem in Untereinheiten der “*Energieproduktion*”. Die prozentuale Aufteilung der Prozesse ist ebenfalls Ergebnis der Analyse und kann den Kreisdiagrammen entnommen werden.

Dieses Beispiel zeigt, dass GoSynthetic in der Lage ist, bekannte und unbekannte Prozesse grob in ihrer Funktionsweise zu kategorisieren und somit schnell einen Überblick über den Prozess zu verschaffen.

Das Protein Glyceraldehyde-3-phosphate Dehydrogenase (GAPDH) erscheint über die Interaktionssuche als zentrales Protein der Glykolyse. Neben seiner linearen metabolischen Funktion innerhalb der Glykolyse interagiert es mit weiteren Proteinen, wie z.B. der Mitogen-aktivierenden Proteinkinase 14 (MAPK14) und dem Tumor Nekrose Faktorrezeptor (TNFR54B).

Komplexere Zusammenhänge zur Glykolyse beinhalten die Proteine Aryl Hydrocarbon Rezeptor nuklearer Translokator (ARNT) und Hypoxia-induzierte Faktor 1-alpha (HIF1A), welche in diesem Zusammenhang beide einen aktivierenden Charakter besitzen. Die Interaktion der Aldolase B (ALDOB) mit den Bardet-Biedl Syndrom Proteinen 1,2,4,7 (BBS1, BBS2, BBS4, BBS7) und somit die Verbindung zwischen der Glykolyse und der autosomal-rezessiv vererbbaren Genmutation des Bardet-Biedl Syndroms kann einfach über die Interaktionssuche der GoSynthetic Software erfolgen und liefert wertvolles Wissen für mögliche biotechnologische Experimente.

Über die gleichen Sucheinstellungen, aber Auswahl eines anderen Organismus, können die Unterschiede zwischen Organismen untersucht werden. In diesem Fall wurde ein Vergleich der Glykolyse in *Escherichia coli, Pseudomonas aeruginosa, Saccharomyces cerevisiae* und *Rattus norvegicus* vorgenommen, s. **Abbildung 3.3.8**. Unterschiede in der Prozessstruktur wurden ersichtlich: Während *P. aeruginosa* besonders auf die Einheit „*Transformation“* fokussiert, legen Hefe und Ratte den Schwerpunkt auf Einheit „*Regulation*“.

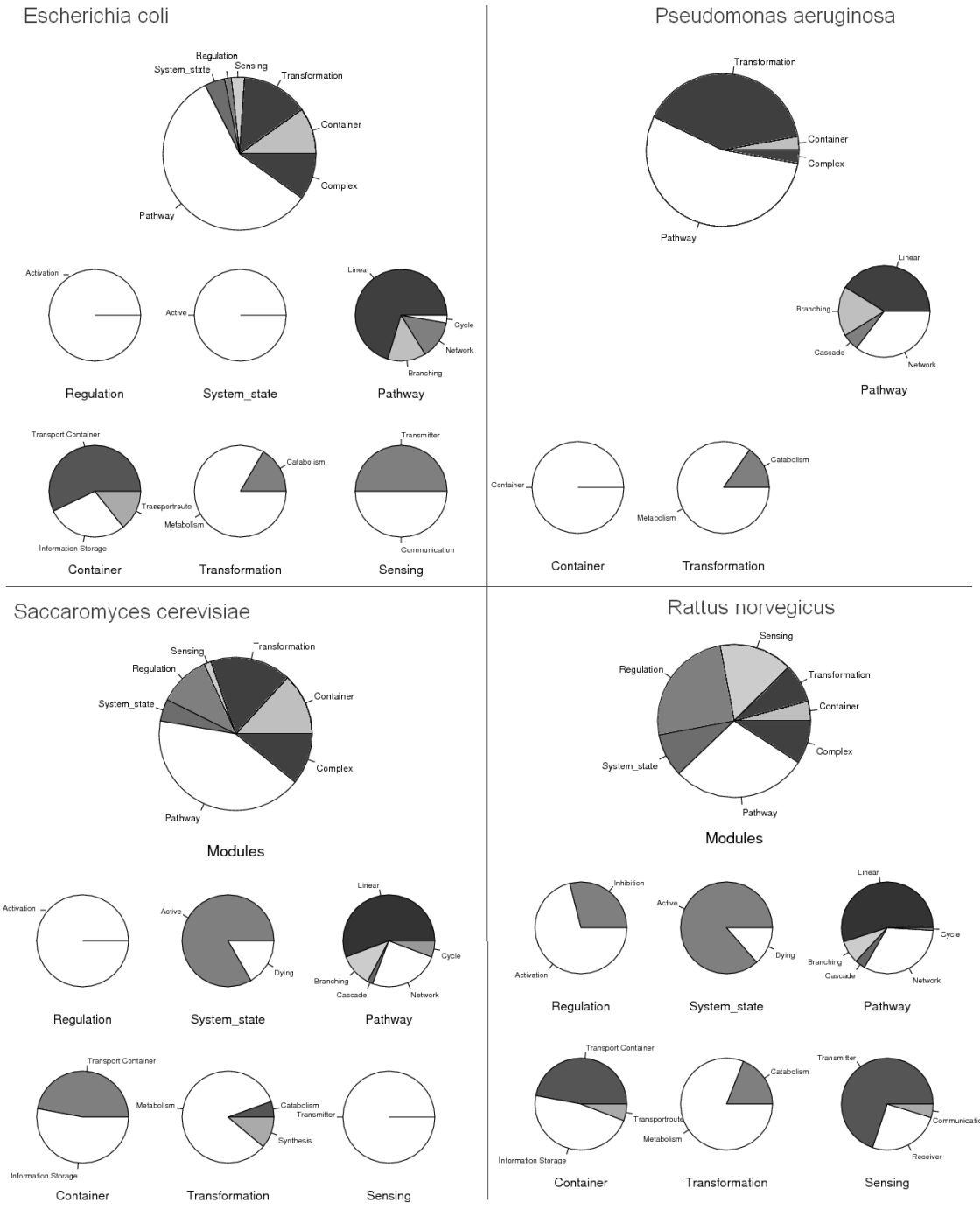

Abbildung 4.3.8: Vergleich der Glykolyse in unterschiedlichen Organismen. Kreisdiagramme für den Glykolyseprozess in den Organismen *Escherichia coli, Pseudomonas aeruginosa, Saccharomyces cerevisiae* und *Rattus norvegicus.*

3.3.4.2 Biologische Implementierung eines technischen Kreislaufs

Ziel und Hintergrund

Mit Hilfe von GoSynthetic können nicht nur biologische Prozesse analysiert werden, sondern es können auch technische Prozesse biologisch nachgebaut werden. In diesem Beispiel wird mit Hilfe der Software die Umsetzung der technischen Temperaturkontrolle z.B. mittels Thermostat in die Biologie vorgestellt.

Softwaresuche und gelieferte Ergebnisse

Um die biologische Implementierung der Temperaturkontrolle zu planen, müssen Proteine herausgesucht werden, welche in der Lage sind, die Temperatur zu detektieren und zu regulieren. Die erste Funktion kann durch die Einheit „*Signaltransduktion/Überträger*“ beschrieben werden und als Temperatursensor eingesetzt werden. Daher muss mit der Software zunächst im Hierarchiebaum ein Organismus herausgesucht werden und dann der Bereich „*Signaltransduktion/ Überträger*“ selektiert werden. Hier wird nach dem Begriff „*Temperatur*“ gesucht. Alternativ kann eine Stichwortsuche mit den beiden Begriffen durchgeführt werden.

Interessanterweise ist die Ergebnisliste der Proteine aus der Suche begrenzt auf den Modelorganismus Maus relativ klein: nur das Homeobox Protein ähnliche *1-Prrxl1* und der Nervenwachstum-Faktor Rezeptor *Ngfr* wurden als Proteine mit den benötigten Eigenschaften identifiziert. Beide Proteine reagieren auf Temperaturänderungen und wirken als regulative Elemente während des Transkriptionsprozesses.

Um auch die zweite Funktion der Temperaturkontrolle zu erfüllen, muss als nächstes ein Protein gefunden werden, welches die Regulation der „*Temperatur“* übernehmen kann. Um dies zu finden, wird in der Stichwortsuche nach „*Temperatur*“ und „*Regulation*“ gesucht. Die Proteinergebnisliste ist wiederum relativ klein: *Adrb1* (beta-1 adrenergene Rezeptor), *Adrb2* (beta-2 adrenergene Rezeptor) und *Adrb3* (beta-3 adrenergene Rezeptor).

Auf der Suche nach Temperaturaktivatoren (Stichwortsuche mit den Einheiten „*Temperatur*“ und „*Aktivierung*“) findet man im letzen Schritt der Suche das uncharakterisierte Protein *Apln* und das Protein *Slc27a1*, ein Mitglied der löslichen Transporter-Familie (Fettsäuretransporter).

Um die ausgewählten Proteine auf ihre Passgenauigkeit zu untersuchen, können diese separat analysiert werden. Dazu kann direkt im Hierarchiebaum der Softwareoberfläche oder mit der Prozess-Suche analysiert werden.
Das Protein *Ngrf,* welches für die Temperaturdetektion zuständig ist, wird durch die folgenden Einheiten beschrieben: „*Fühler*“ (tatsächliche Temperaturdetektion), „*Informationsspeicherung*“ und „*Beförderer*“ (die Detektion der Temperaturänderung muss weitergeleitet werden zur Reaktionseinheit). Zur Regulation der Körpertemperatur mittels *Adrb1* werden die Einheiten „*Überträger*“ (für die Aktivierung der Temperaturregelung), „*Kreis*“ (Generierung einer positiven Rückkopplung), „*Regulation*“ (Antwort auf die neue Information des Detektors), „*Reduktion*“ und „*Netzwerk*“ (Hitzeverteilung) benötigt. Für die Erreichung des Temperaturanstiegs mittels *Slc27a1* werden die Einheiten „*Energieproduktion*“ (die tatsächliche Energieproduktion) und ein Temperatur „*Beförderer*“ (Temperaturweiterleitung) benötigt.

Interpretation

Mit dem beschriebenen Vorgehen über die Software ist eine einfache und schnelle Sammlung aller benötigten Komponenten für einen technischen Prozess einfach möglich. Sie liefert die Grundlagen für darauf folgenden, spezifischen Analysen. Dieses technische Beispiel zeigt, dass GoSynthetic nicht nur zum Studium und besseren Verständnis eines biologischen Systems genutzt werden kann, sondern ebenfalls als Startpunkt zum Design eines Experiments der synthetischen Biologie, welches einen technischen Prozess nachahmt oder einen anderen Kontrollzyklus darstellt.

3.3.4.3 Beispiel zur biologischen Implementierung einer selbstreinigenden Eigenschaft (Lotuseffekt)

Ziel und Hintergrund

In diesem dritten Beispiel soll der Weg von einem biologischen Prozess (Lotus-Effekt) zu einem technischen Prozess bzw. dessen Verbesserungsmöglichkeiten beschrieben werden, also genau der entgegengesetzte Weg zum vorherigen Beispiel.

Dazu muss zunächst die biologische Realisierung des Effekts untersucht werden. Biologisch betrachtet basiert der Lotuseffekt auf einer speziellen wachsbedeckten Oberfläche. Dieses Wachs ist noppenartig auf der Oberfläche angebracht und vergrößert sie somit. Aufgrund dieser Oberflächenvergrößerung nimmt die Adhäsion von Wasser und Schmutz zur Oberfläche ab [102].

Softwaresuche und die gelieferten Ergebnisse

Da die Software keine fertigen Effekte speichert, kann in diesem Fall nicht nach dem Prozess „Lotuseffekt“ gesucht werden. Stattdessen kann der Lotuseffekt nur genauer untersucht werden, indem basierend auf dem oben beschriebenen Wissen die Wachsproduktion selbst genauer analysiert wird. Erst mit diesem Vorwissen kann mit der Software ähnlich wie im Thermostatbeispiel nach den Einheiten „*Netzwerk*“ in Kombination mit dem Begriff „*Wax*“ gesucht werden oder eine Prozesssuche mit dem Begriff „W*ax*“ durchgeführt werden.

Als Ergebnis liefert die Software die Bestandteile der Wachsproduktion, die betroffenen Funktionen und die beteiligten Proteine (in den Kreisdiagrammen): die Wachsproduktion wird durch die Einheit „*Anabolismus*“, verknüpft mit dem Acetyl-CoA und dem „*Überträger*“ Einheit beschrieben und durch Licht beeinflusst. Über die Interaktionssuche können außerdem zugehörige Proteine angezeigt werden. Die in *A. thalinana* identifizierten Zentralproteine sind z.B. die Fettsäure-Hydroxylase *Cer1* und das *Wax2* Protein.

Interpretation

Aus technischer Sicht betrachtet wird demnach ein Sensor („*Fühler*“) benötigt, welcher auf Licht reagiert und unter einer größeren Menge an Energieverbrauch („*Energieverbrauch*“) die Wachsproduktion vorantreibt. Das hergestellte Wachs muss an die Oberfläche transportiert werden („*Beförderer*“).

In der Technik wird die Vergrößerung der Oberfläche derzeit nicht mittels Wachs, sondern über die Nanotechnologie realisiert.

3.3.4.4 Modifizierung eines Netzwerks mit einem zentralen Knotenpunkt, WASP

Ziel und Hintergrund

Dieses vierte Beispiel legt den Schwerpunkt auf die abstrakte, funktionelle Analyse eines zentralen Knotenpunktes und erlaubt die Identifizierung der wichtigsten Proteine. Zusätzlich zeigt dieses Beispiel die Abgrenzung zu den Programmen AmiGO und STRING auf.

Das Wiskott-Aldrich Syndrom ist ein Immundefekt, welcher zu einer erhöhten Ekzembildung und Blutungsneigung führt. Ausgelöst wird dieses Syndrom durch einen genetischen Defekt des Wiskott-Aldrich Syndrom Proteins (Swiss-Prot WASP_HUMAN). WASP wirkt als Modifikator und Interaktionspartner des Zytoskeletts. Gleichzeitig vereinfacht es die Infektion mit dem Vaccinia Virus, indem es als Adapter zwischen dem Virus und dem neuralen Wiskott-Aldrich Syndrom Protein WASL fungiert.

Softwaresuche und gelieferte Ergebnisse

Als Ergebnis der Prozesssuche nach dem Begriff WASP liefert die Software die bekannten Kreisdiagramme mit den Haupteinheiten „*Regulation* und „*Aktivierung*“. WASP agiert auch als „*Empfänger*“ während der Signaltransduktion zum Vaccinia Virus, da es das Signal der Zelloberfläche auf das Aktin des Zytoskeletts überträgt.

Die „*Prozessstruktur*“ ist hauptsächlich „*Linear*“ und ein „*Netzwerk*“, da WASP die Polymerisation des Aktins und die Neuverteilung induziert.

Ergebnisvergleich mit AmiGo und STRING

Im Vergleich zu GoSynthetic liefert die direkte AmiGO-Suche keine Information über die Funktion von WASP, sie gibt lediglich Hinweise zur Domänenzusammensetzung. Eine STRING Suche hingegen liefert mögliche Interaktionspartner zu WASP, aber auch keine Einordnung über die Funktion von WASP oder seinen Prozesszusammenhang.

GoSynthetic hingegen liefert nicht nur die beschriebenen Prozesse, sondern kann über die Interaktionssuche die funktional und physikalisch zu WASP zugehörigen Proteine identifizieren: Abelson-interactor-2 Protein, Aktin-depolymerisierungs Faktor Protein, Ena/Vasodilator-stimuliertes Phosphoprotein-ähnliches Protein (EVL), Appetit-regulierendes Hormon, sekretionsfördernder Wachstumshormon Rezeptor Type 1, Neurales Wiskott-Aldrich Syndrome Protein, WAS/WASL Interaktionsprotein Familie Mitglied 1 (WIPF1). EVL und WIPF1 interagieren mit der Thymidine-Protein Kinase HCK.

Interpretation

Aus dieser Sammlung kann geschlossen werden, dass GoSynthetic neben den Interaktionspartnern zusätzlich Proteine identifiziert, welche funktional mit den Einheiten und Prozessen des Suchproteins verknüpft sind. Damit sind bereits solche Einheiten und Markerproteine identifiziert, welche bei einer Veränderung des WASP Prozesses besonders beachtet und in ihrer Funktion überprüft werden müssen.

Die Untersuchung eines zentralen Knotenpunktes findet viele weiter interessante Anwendungen, darunter z.B. die Untersuchung des Entzündungsprozesses. Neben der modularen Analyse des Prozesses über die Kreisdiagramme, kann zusätzlich die Interaktionssuche dienen (**Abbildung 3.3.9**). Darin wird ersichtlich, dass der Entzündungsprozess im Menschen nicht nur sehr komplex ist und viele interagierende Proteine enthält, sondern sehr stark reguliert ist. Die Tumor Nekrose Faktor Proteine 1 (blau) und 2 (orange) (TNFRSF1/2) dienen als zentrale Knotenpunkte des Entzündungsprozesses und regeln deren Aktivierung und Inhibition. Zwischen diesen beiden Gegenspielen befindet sich die Rezeptor-interagierende Serin/Threonine-Proteinkinase 2 (RIPK2), welche wiederum mit der CASP8-induzierten Apoptose und der Aktivierung des Transkriptionsfaktors NF-kappa-B verbunden ist.

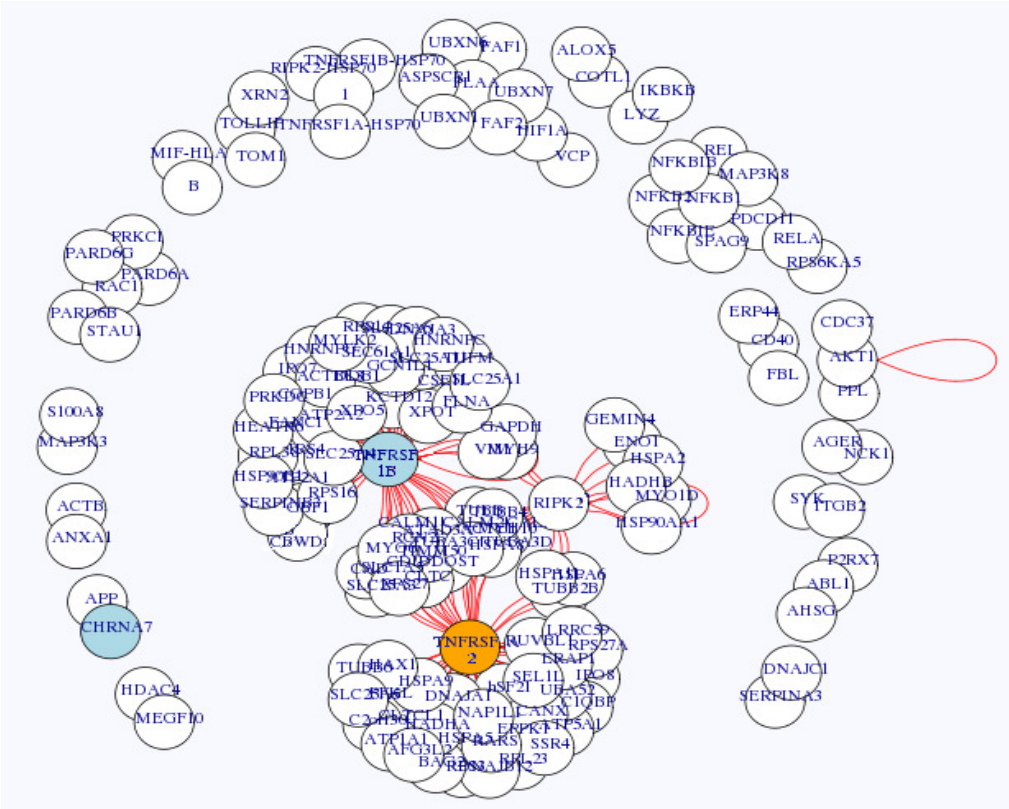

Abbildung 3.3.9: Interaktionsansicht des Entzündungsprozesses.
Analyse des Inflammationsprozesses im Menschen. Suchbegriff: „inflammation“, Organismus „human“. Aktivierende Proteine sind orange, inhibitorische Proteine in blau gekennzeichnet.

3.3.4.5 Design eines onkolytischen Vaccinia Virus

Ziel und Hintergrund

Dieses Beispiel zeigt, wie GoSynthetic genmanipulative Veränderungen in einem beliebigen Organismus (hier ein onkolytisches Vaccinia Virus) unterstützen kann.

Virale Infektionen können einen großen Einfluss auf den betroffenen Organismus und sein Immunsystem haben. Die Grundidee der Gentherapie mittels onkolytischen Viren besteht darin, dass bevorzugt Krebszellen von der viralen Infektion betroffen sind und nicht die gesunden Zellen. Aufgrund dieser Infektion und der Vermehrung der Viren sterben die Krebszellen ab. Nach Abtötung der Krebszellen ist das Immunsystem des betroffenen Organismus wieder in der Lage, das Virus selbst zu bekämpfen. Erste Versuche mit onkolytischen Viren in den 90er Jahren waren jedoch entmutigend. Denn obwohl die Krebszellen bekämpft wurden, erlag der Wirtsorganismus anschließend der Virusinfektion [103]. 2009 gelang es, ein genmanipuliertes Vaccinia Virus Lister GLV 1h68 zu entwickeln, welche die Gefahr für den Wirt-Organismus reduzierte [104]. Im Gegensatz zum Wildtyp Vaccinia Virus Western Reserve (WR) produziert das GLV 1h68 die Proteine (i) beta-Galactosidase ß-gal (ersetzt Thymidinkinase J2R), (ii) RUC-GFP (ersetzt das sekretorische Protein F14.5L) und (iii) Glucuronidase (ersetzt Hämagglutinin A56R).

Softwaresuche und gelieferte Ergebnisse

Mit Hilfe der Klassifikation wurden die sich ergebenden Effekte auf die Umgebung und Unterschiede im Proteom des originalen Vaccinia Virus WR im Vergleich zum neuen Virus GLV 1h68 analysiert. Die GoSynthetic-Oberfläche genügt in diesem Fall nicht, da das modifizierte Vaccinia Virus nicht in der Software integriert ist. Dennoch kann über die Proteinliste der beiden Viren die zweistufige, hierarchische Klassifikation angewendet werden, indem die Proteinfunktionen der Klassifikation mittels Text-Mining zugeordnet werden. Dies ermöglicht die Untersuchung des Einheitenvorkommens innerhalb der beiden Viren und den prozentualen Vergleich zwischen ihnen, zeigt aber auch wie einfach neue Organismen der Datenbank zugefügt werden können.

Abbildung 3.3.10 fasst den Einheitenvergleich des originalen und modifizierten Virus zusammen. Die Abbildung stellt die prozentualen Einheitenunterschiede zwischen Vaccinia Virus WR und GLV 1h68 dar. Das Vaccinia Virus (WR) ist optimiert für eine Wirtsinfektion. Es legt seinen Schwerpunkt auf die Einheiten

„*Informationsspeicherung*“ und „*Umwandlung*“. Im Gegensatz dazu erscheint das neue GLV 1h68 harmloser und stärker reguliert, da es weniger Informationen in seine Umgebung abgibt. Modular betrachtet legt GLV 1h68 seinen Schwerpunkt auf die Einheiten „*Regulation*“, „*Empfänger*“ und „*Hemmung*“. Der Tendenz folgend könnte daraus geschlossen werden, dass eine weitere Reduktion der Viralität durch Reduktion der Einheiten „*Transportcontainer*“, „*Informationsspeicherung*“, „*Umwandlung*“ und „*Hemmung*“ erreicht werden könnte.

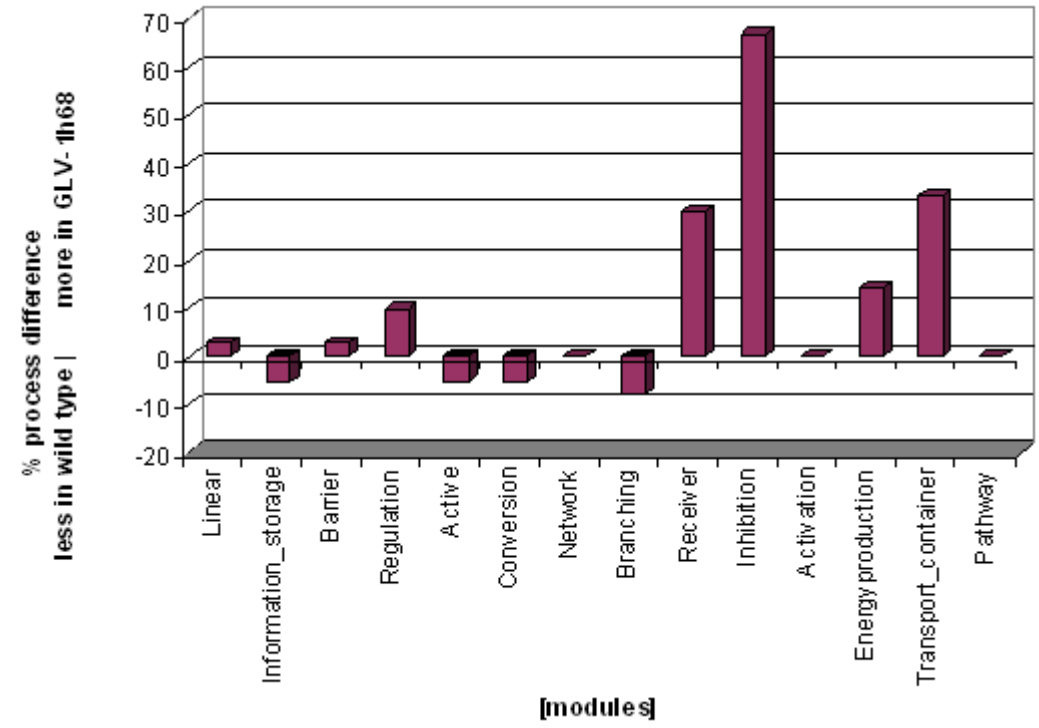

Abbildung 3.3.10 Unterschiede in der Einheitenverteilung in den Proteinen aus Wildtyp Vaccinia Virus Western Reserve (WR) und Vaccinia Virus GLV 1h68. Auf der x-Achse sind die Einheiten aufgetragen, wohingegen auf der y-Achse die prozentualen Unterschiede zwischen den beiden Viren aufgezeigt wird. Der Wildtyp wurde als 100% angenommen.

Interpretation

Um die Gründe der reduzierten Viralität des GLV 1h68 zu verstehen, wurden im Weiteren die einzelnen veränderten Gene mit Hilfe der Klassifikation im Detail untersucht:

Durch den Ersatz des Proteins Thymidinkinase J2R durch das beta-Galactosidase GusA Protein werden die Einheiten „*Informationsspeicherung*“, „*Umwandlung*“ und „*Netzwerk*“ durch „*Verzweigung*“ und „*Linear*“ ersetzt.

Durch den Austausch des sekretorischen Proteins F14.5L mit dem GFP Protein wird die Einheit „*Informationsspeicherung*“ ersetzt durch „*Energieproduktion*“, „*Linear*“ und „*Verzweigung*“.
Der Austausch des Hämagglutinin A56R durch die zellzerstörende Glucuronidase führt zur Ersetzung der „*Träger*“-Einheit durch die Einheiten „*Energieverbrauch*“, „*Linear*“ und „*Verzweigung*“.

Experimentelle Validierung
Zur Validierung des Vaccinia Virus-Designs existieren experimentelle Daten aus einer anderen Forschungseinheit der Universität Würzburg. Die schrittweise Erzeugung des Tripel-Mutanten GLV 1h68 wurde ebenso beschrieben wie die Untersuchung aller Mutanten auf seine Wirksamkeit und die Überlebensrate des Wirtes [104].

Eine Veröffentlichung zu diesem Thema befindet sich derzeit bei *PLOS One* in Revision.
Erste Ergebnisse wurden bereits bei der Konferenz “Streamlined and Synthetic genomes” im November 2009 in Valencia, Spanien vorgestellt.

4 KAPITEL
Diskussion

Die vorgestellte Dissertation umfasst ein weites Gebiet der Bioinformatik und reicht von Sequenz- und Strukturanalysen von Genen und Proteinen, deren Interaktionen bis hin zu generellen Interpretationen. All diese Themen sind zentrale Gebiete der heutigen Bioinformatik und kompetitiv arbeiten hieran viele Gruppen im informatischen Wettstreit um die besten Ergebnisse. Um die vorliegende Arbeit besser beurteilen zu können, sollen im Folgenden die eigenen Ergebnisse im Vergleich interpretiert und hinsichtlich ihrer Nützlichkeit, Limitation und Neuheit durchleuchtet werden.

4.1 Vorhersagen mit bioinformatischen Methoden

Die Bioinformatik hat in den letzten Jahren viele Programme zur Vorhersage von Proteininteraktionen hervorgebracht (z.B. STRING, iHop). Ein entscheidender und häufig vernachlässigter Punkt ist die kritische Evaluation der Methoden gegeneinander und die Verknüpfung der Methoden und Programme.

Haupterkenntnisse

Die beschriebene Analyse von bioinformatischen Methoden und deren Evaluation als Weiterführung meiner Diplomarbeit zeigen, dass die Weiterentwicklung, Konkretisierung und Kombination vorhandener Programme zum Erleichterten und vor allem schnelleren Auffinden wichtiger Informationen dienen kann.

Die vorgestellte Methodik, im Besonderen mit dem Schwerpunkt der Genkontext-Methoden und dem physikalischen Vergleichswert, liefert eine wertvolle Grundlage für die Untersuchung neuer Sequenzen und erweitert die Analyse vorhandener Sequenzen. Am Beispiel der Einführung des physikalischen Vergleichswertes kann das Gebiet der

Proteininteraktionsvorhersage erweitert und zuverlässig immer weiter verfeinert und konkretisiert werden.

Limitationen und Möglichkeiten

Auch die Pharmaindustrie nutzt die Vorteile der Bioinformatik und die damit in Zusammenhang stehenden Programme, Automatisierungs- und Vorhersage-möglichkeiten immer intensiver (z.B. EMBOS, iHop). Sie werden benötigt, um gezielter nach möglichen biologischen Angriffspunkten für Medikamente zu suchen, deren Wirkungsweise zu verstehen und mögliche Nebenwirkungen frühzeitig zu erkennen und ggf. reduzieren zu können. Die Validierung und Selektion dieser vielen bioinformatischen Programme ist sehr mühsam. Besonders hilfreich sind diese Programme, wenn sie für ihre Vorhersagen zusätzlich die Qualität und Verlässlichkeit der Vorhersage beschreiben. Der in diesem Zusammenhang validierte physikalische Vergleichswert leistet einen Beitrag, um diese Art der Untersuchungen zu erleichtern. Aber der erforderliche Aufwand lohnt sich, denn letztendlich kann mit der Bioinformatik die Planung von Experimenten verbessert, die Aussagekraft der Ergebnisse erhöht und Fehler reduziert werden, wodurch wiederum die Datenqualität verbessert wird.

Im Besonderen der Einsatz von eigenen und öffentlichen Datenbanken (z.B. Nucleic Acids Research, NAR [105]) als Speichermedium für Roh- und Parameterdaten liefert noch viele weitere und oft bisher nicht erschlossene Möglichkeiten. Nicht nur die Qualität der Untersuchungen kann damit automatisiert analysiert werden, sondern neue Vorhersagemethoden auf Basis tatsächlicher Untersuchungen können erstellt und verbessert werden.

Quintessenz

Zusammenfassend kann daher gesagt werden, dass die Methoden der Bioinformatik Einzug in die gesamte biologische und chemische Forschung gefunden haben und aus dem täglichen Leben in vielen wissenschaftlichen Bereichen nicht mehr wegzudenken sind. Aber erst die hier vorgestellte Kombination aus validierten Methoden und übersichtlicher Aufbereitung führt zu einer Arbeitserleichterung oder Verbesserung der Qualität. Der vorgestellte physikalische Vergleichswert kann hierbei eine große Hilfe sein.

4.2 Modifikationsuntersuchung in Zweikomponenten-Systemen

Grundlagen über Funktion und Aufbau von biotechnologisch interessanten Zweikomponenten-Systemen sind bereits bekannt und beschrieben [106, 107]. Dieses Wissen kann in aufbereiteter Form in öffentlichen Datenbanken wie z.B. MiST2 [108] abgerufen werden. Bei genauerer Betrachtung wird allerdings schnell klar, dass individuelle TCS weitere detaillierte Untersuchungen der Sequenz und Struktur benötigen, um Protein- und DNA-Designexperimente gezielter zu unterstützen. Die zu klärenden Herausforderungen dabei sind, fehlende Details in Zweikomponenten-Systemen aufzudecken und deren Flexibilität zu untersuchen. Aus diesem neuen und dem bekannten Wissen sollen allgemeingültige Modifikationsmöglichkeiten für Zweikomponenten-Systeme abgeleitet werden.

Haupterkenntnisse

Basierend auf breit angelegten TCS Untersuchungen in unterschiedlichen Organismen, konnten drei grundlegende Modifikationsszenarien in TCS für die Veränderung von zellulären Antworten oder Funktionen entdeckt werden: (i) Modifikation von Sequenzen innerhalb einer TCS-Domäne, zur Veränderung von Promotoren oder Signalen (ii) Austausch kompletter TCS-Domänen, um Funktionen aus unterschiedlichen TCS zu mischen (iii) Zufügen neuer Komponenten, welche TCS Funktion beeinflussen.

Konkret konnte in der vorgestellten Analyse festgestellt werden, dass Zweikomponenten-Systeme aus wenigen verschiedenen Grunddomänen zusammengesetzt sind. Die Erkennungssequenzen der Signalbindestellen sind hauptsächlich vom Stimulus und weniger vom System oder dem Organismus abhängig. Des Weiteren konnten neue Promotorsequenzen identifiziert werden. Neue TCS–Partner in *Listeria* und *Legionella* wurden ebenso aufgefunden wie eine potentiell neue, divergierte TCS-Familie im Organismus *Mycoplasma*, welcher bisher als TCS-frei galt [109]. Drei untersuchte Konnektoren zeigen weitere Modifikationsmöglichkeiten z.B. über die Verbindung von unterschiedlichen TCSs miteinander.

Aus diesen Erkenntnissen kann geschlossen werden, dass Standard-TCS einerseits stark konserviert sind (wenige Pfam-Familien), sich im Detail allerdings stark

unterscheiden (Konnektoren, Promotorstellen, Signale). Die starke Abhängigkeit der Signalbindestellen vom Signal ist nicht unerwartet, die extreme Unabhängigkeit vom Organismus allerdings schon, da sich auch die Zweikomponenten-System Familien zwischen Organismen sequentiell stark unterscheiden. Die Kombination aus MPN013/MPN014 in *Mycoplasma* als potentielles divergiertes TCS zeigt eine hohe Sequenz- und Strukturähnlichkeit zu anderen TCS auf. Allerdings fehlen MPN013 sowie MPN014 einige essentielle Bereiche, um als typisches TCS zu gelten. Die aufgefundenen Ähnlichkeiten legen dennoch die Vermutung nahe, dass es sich um ein entartetes TCS handeln könnte.

Limitationen und Möglichkeiten

Um die aufgedeckten Ergebnisse weiter zu bekräftigen, könnte die Anzahl untersuchter Organismen erhört und der Einsatz von weiteren bioinformatischen Programmen in Betracht gezogen werden. Allerdings ist zu erwarten, dass die theoretischen Ergebnisse nicht stark variieren, da die Datenbasis bereits jetzt sehr breit ist.

Die konkreten Sequenzen und Proteine der Studie können direkt in der Biotechnologie benutzt werden. Die etwas abstrakteren Erkenntnisse der Modifikationsmöglichkeiten dienen zum allgemeinen besseren Verständnis von TCS.

Modifikationen im Bereich des Sequenz- oder Domänenaustausch sind schwierig vorherzusagen und Modifikationen können ungeahnte Reaktionen hervorrufen. Die beschriebenen Sequenzen in den Signal- und Promotorbereichen scheinen dennoch gute Ergebnisse zu versprechen. Eine weitere vielversprechende Modifikationsmöglichkeit scheint in den SafA ähnlichen Proteinen gefunden zu sein, da sie kein TCS direkt modifizieren, sondern als Brücke zwischen unterschiedlichen TCS dienen.

Allerdings ist bei dieser Studie zu beachten, dass es sich um eine rein bioinformatische Analyse handelt. Auch wenn niedrige E-values und gute Vergleichswerte in den Sequenzanalysen von einer guten Datenqualität zeugen, so wurden die Ergebnisse nicht experimentell bestätigt.

Quintessenz

Das Gesamtergebnis dieser breit gefächerten TCS-Analyse bietet nicht nur die beschriebenen neuen Erkenntnisse und Modifikationsmöglichkeiten, sondern demonstriert zusätzlich eine generelle Vorgehensweise zur Analyse, Modifikation und Verifizierung neuer TCS Ergebnisse auf.

4.3 Strukturierung biologischer Systeme

Die synthetische Biologie beschäftigt sich mit der Herausforderung, komplexe Systeme besser zu verstehen und die Grundlagen für deren gezielte Manipulationen zu untersuchen. Es existiert eine Vielzahl an experimentellen und technischen Ansätzen, um die Planung des Designprozesses in der synthetischen Biologie zu vereinfachen [28-36]. Der Schwerpunkt der hier vorgestellten Analyse zur Unterstützung des Designprozesses beschäftigt sich mit den folgenden Fragen:

- Unterliegen biologische und technische Prozesse ähnlichen Strukturierungsprinzipien?
- Wie hängen Biologie und Elektrotechnik oder unterschiedliche Organismen miteinander zusammen und wie können diese Erkenntnisse genutzt werden?

Haupterkenntnisse

Zur Lösung dieser Fragestellung wurde eine zweistufige Klassifikation/Strukturierung über Prozesse und Verhaltensweisen erstellt und gegen bekannte Konzepte wie Gene Ontology, COG und MIT-BioBricks validiert [34, 35]. Dabei konnte herausgefunden werden, dass die erstellte Klassifikation die bekannten Konzepte abdeckt und diese sogar erweitert.

Als Fortführung der bisherigen Ansätze [36, 38, 108, 110, 111] wurde die Themenüberlappung in der bisherigen biologischen Forschung, der synthetischen Biologie und der Elektrotechnik überprüft. Es wurden dabei Unterschiede aufgezeigt, aber auch Gemeinsamkeiten, bei denen besonders aus der Elektrotechnik gewachsene Strukturen in der Biologie genutzt werden können.

Als Ergebnis konnte festgestellt werden, dass die Klassifikations-Einheiten "*Prozessstruktur*" und „*Regulation*" in der synthetischen Biologie und der

Elektrotechnik stärker vertreten sind. Demzufolge ist es am vielversprechendsten und derzeit am häufigsten genutzt, relativ einfache Organismen mit den Regulationsmöglichkeiten der Technik zu manipulieren.
Die Ergebnisse der vorliegenden Analyse sind benutzerfreundlich in der Software GoSynthetic mit einer zugrunde liegenden Datenbank aufbereitet. Diese Software führt die Erfahrungen der Technik und der Biologie für neue Experimente der synthetischen Biologie zusammen, bereitet sie statistisch auf und bietet viele praktische Suchfunktionen.

Die Alleinstellungsmerkmale von GoSynthetic im Vergleich zu den bereits existierenden Programmen sind:

- Die neue, zweistufige Klassifikation besitzt einen hohen Abstraktionsgrad, kann aber mittels der vorhandenen Verknüpfungen zu Proteinen, Funktionalitäten, BioBricks und Interaktionen sehr detailliert werden.
- Komplexe biologische Prozesse können einfach in deren Eigenschaften und Einzelteile zerlegt, abstrahiert und auch für technische Prozesse übernommen werden.
- Der Vergleich zu anderen Organismen (selbst modifizierten Organismen) oder Techniken ist einfach möglich, ebenso das Auffinden von Verbindung zu anderen Prozessen, welche mit den zu untersuchenden Prozess zusammenhängen.
- Eine Erweiterung der aufgestellten Klassifikation um zusätzliche Datenquellen, Organismen oder das Vokabular anderer Wissenschaftsbereiche ist einfach möglich, ebenso wie die Abänderung der Klassifikation selbst.
- Die Software vereinfacht das interaktive Design von künstlichen und komplexen Prozessen.

Das Ziel ist, Wissenschaftler im Bereich der synthetischen Biologie und anderernForschungsbereichen zu neuen Sichtweisen und Experimenten anzuregen und sie dabei zu unterstützen. Die Anwendbarkeit dieser Software wurde anhand einiger Beispiele (WASP, Vaccinia Virus, Lotus-Effekt, Thermostat) ausführlich erläutert.
Aber auch für andere Forschungsfelder wie die Pharmaindustrie könnte GoSynthetic nützlich sein. Denn immer wieder fällt in einer relativ späten Phase der Forschung auf,

dass das Indikationsgebiet für eine Wirksubstanz nicht optimal ist. In solchen Fällen werden neue Indikationsgebiete für die Wirksubstanz gesucht. Die Interaktionssuche von GoSynthetic könnte einfach und schnell neue, ungewöhnliche Zusammenhänge auffinden. Ein zunehmend wichtiger Bereich in der Pharmaindustrie ist die Biopharmazie. Sie nutzt Prinzipien der synthetischen Biologie und somit könnte GoSynthetic auch im Designprozess der Biopharmazie eingesetzt werden.

Limitationen und Möglichkeiten

Eine ähnliche Analyse der genannten Beispiele ist grundsätzlich auch ohne die beschriebene Software möglich, würde aber komplexe und zeitaufwändige Such- und Analysevorgänge über viele Protein-, Funktions- und Interaktions-Datenbanken benötigen. Die vereinfachte Suche bildet einen Rahmen und wird erst über die Kombination der vorhandenen Datenbanken und das Zufügen der Abstraktion der zweistufigen Klassifikation ermöglicht.

Bei dieser Studie muss darauf hingewiesen werden, dass die genutzten Daten auf der GO-Klassifizierung basieren. D.h. bestehende Verzerrungen innerhalb der GO-Klassifizierung bestehen auch in den hiesigen Analysen. Jedoch wurde dieser Einflussparameter durch die zusätzliche Verwendung von COG und MIT-Daten minimiert und könnte durch eine weitere Vergrößerung der Datenbasis nochmals reduziert werden.

Die Anzahl der Organismen ist derzeit auf 17 Modelorganismen begrenzt. Auch sie könnte erweitert werden, würde aber vermutlich keine größeren Veränderungen im Ergebnis aufweisen, da es sich um eine abstrakte Klassifikation handelt bei der Unterschiede zwischen einzelnen Stämmen oder sehr ähnlichen Organismen nicht auffallen, bzw. durch die Sequenzsuche von GoSynthetic abgefangen werden kann.

Eine Schwierigkeit bei der Nutzung von GoSynthetic zeigt sich bei der Auswertung des Lotus-Effektes. Um diese Analyse realisieren zu können, musste bereits biologisches Wissen über die Ursachen des Effekts bekannt sein. Eine weitere Limitation wird im Beispiel des onkolytisch wirkenden Virus aufgezeigt. Auch wenn die genetischen Veränderungen theoretisch recht gut beschrieben und vorhergesagt werden können, so ist die Klassifikation nicht in der Lage die Wirkung auf einen gesamten Organismus

abzuschätzen. Im Besonderen der Einfluss des Immunsystems müsste hierbei beachtet werden, was die Software allerdings nicht leisten kann.
Während die Limitation des Ansatzes durch fehlende Informationen wie im Beispiel des Lotus-Effekts durch Hinzunahme bekannter Struktur-Wirkungs-Daten reduziert werden könnte, so wird der Einfluss weiterer Einwirkungsfaktoren (wie die des Immunsystems) in diesem Ansatz nicht realisiert werden können.

Quintessenz

Die Anwendungsfälle der Klassifikation sind sehr vielfältig. Schon die hier vorgestellten Beispiele machen klar, dass es sich bei der Datenbank um eine Expertensoftware handelt, wohingegen die Oberfläche für jeden sehr einfach zu bedienen ist.
Im Besonderen für sinnverwandte Bereiche der synthetischen Biologie ist der Einsatz der Software sehr vorteilhaft, da synthetische Biologie, Biopharmazie und Elektrotechnik in einigen Bereichen eng miteinander verwandt sind oder zumindest ähnlichen Prinzipien folgen. Die vorgestellte Klassifikation bildet die Grundlage für das Erzielen gemeinsamer Strategien und den Austausch von Erfahrung und Vorgehensweisen in unterschiedlichen Technologien und geht dabei weit über die Funktionen der Gene Ontology hinaus.

4.4 Generelle Bedeutung und Ausblick

Die Grundvoraussetzung der Integration und Kombination bioinformatischer Methoden in Biotechnologie, synthetischer Biologie und Pharmaindustrie ist es, biologische Systeme zu verstehen, um sie später gezielt nutzen und gegebenenfalls verändern oder beeinflussen zu können.
Das Gesamtergebnis der vorliegenden Doktorarbeit umfasst die Techniken der abstrakten Analyse bis hin zu der Entwicklung einer neuen Software.

Begonnen wurde mit der Analyse, Weiterentwicklung und Kombination bekannter bioinformatischer Methoden. Im Anschluss daran wurden mit Hilfe dieser Methoden Zweikomponenten-Systeme auf deren Flexibilität untersucht.
Die Software spannt den Bogen zwischen Theorie und Praxis und erweitert die bestehenden bioinformatischen Programme um einen technischen Blickwinkel.

Die Beispiele zeigen die Vielseitigkeit und breite Einsetzbarkeit der Bioinformatik, nicht nur für Biologen, sondern auch für Bioingenieure und selbst für andere Wissenschaftsbereiche, welche aus den gewachsenen Strukturen der Biologie Erfahrungen in ihre eigenen Bereiche einbeziehen können (Beispiel Vaccinia Virus).

Allerdings muss auch gesagt werden, dass es sich hierbei um Ergebnisse der theoretischen Bioinformatik handelt. Experimentelle Untersuchungen sind weiterhin unerlässlich und werden auch in Zukunft immer nötig sein. Die Vorhersagen und entwickelten Programme sind maximal so gut wie die experimentellen Daten, auf denen sie basieren. Auch die Vorhersagekraft der bioinformatischen Programme beschränkt sich auf den Bereich, der von den experimentellen Daten erfasst wurde. Je weiter der experimentell abgedeckte Bereich verlassen wird, desto geringer wird die Aussagekraft der bioinformatischen Methoden und Vorhersagen.

Es ist nicht die Hauptaufgabe der Bioinformatik, Experimente zu ersetzen, sondern Experimente zu priorisieren. Weiterhin unterstützt sie die Auswertung vorhandener großer Datenmengen, um daraus neue Ideen für weitere Analysen und Anwendungen zu gewinnen und damit Irrwege früher zu erkennen und gezielter bestimmte Richtungen einzuschlagen. Die Bioinformatik bietet die Möglichkeit die vertrauten, aber oft

mühsamen Wege durch die neuen und vielversprechenden Möglichkeiten der Datenstrukturierung und der Auswertungen mit Hilfe der modernen Informatik zu ergänzen.

Die potentiellen Einsatzmöglichkeiten der Bioinformatik gehen weit über die derzeitigen Anwendungen hinaus. Denn erst der Einsatz der Informatik ermöglicht es in den Biowissenschaften Modelle zu entwickeln und mit ihnen zu rechnen noch vor den Experimenten im Reagenzglas.

ANHANG

A Literaturnachweis

1. Venter, J. C. et al. (2001) **The Sequence of the Human Genome**. *Science* 291, 304–1351
2. Huynen M, Snel B, Lathe W 3rd, Bork P. (2000) **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res.* 10(8):1204-10.
3. Huynen M, Snel B, Lathe W, Bork P. (2000) **Exploitation of gene context.** *Curr Opin Struct Biol.* 10(3):366-70.
4. Galperin MY, Koonin EV. (2000) **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol.* 18(6):609-13.
5. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. (2003) **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res.*; 31(1):258-61.
6. Yamada S, Shiro Y**.** (2008) **Structural basis of the signal transduction in the two-component system.** *Adv Exp Med Biol*; 631:22-39.
7. Heermann R., Weber A., Mayer B., Ott M., Hauser E., Gabriel G., Pirch T., Jung K. (2008) **The Universal Stress Protein UspC Scaffolds the KdpD/KdpE Signaling Cascade of Escherichia coli under Salt Stress.** *Journal of Molecular Biology*. 386; 134-148
8. Etzkorn M., Kneuper H., Dünnwald P., Vijayan V., Krämer J., Griesinger C., Becker S., Unden G, Baldus M. (2008) **Plasticity of the PAS domain and a potential role for signal transduction in the histidine kinase DcuS**. *Nat. Struct. Biol.* 15: 1031-1035.
9. Grebe TW, Stock JB. (1999) **The histidine protein kinase superfamily.** *Adv Microb Physiol*; 41:139-227.
10. Salis H, Kaznessis YN. (2006) **Computer-aided design of modular protein devices: Boolean AND gene activation.** *Phys Biol*; 295-310.
11. Robinson VL, Buckler DR, Stock AM. (2000) **A tale of two components: a novel kinase and a regulatory switch.** *Nat Struc Biol*; 7:626-33.
12. Drubin DA, Way JC, Silver PA. (2007) **Designing biological systems**. *Genes Dev*; 21: 242-254.
13. Pleiss J. (2006) **The promise of synthetic biology**. *Appl Microbiol Biotech*; 73: 735-739
14. Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, Voigt CA. (2005) **Synthetic biology: engineering Escherichia coli to see light.** *Nature*; 438:441-442.
15. Mitrophanov AY, Groisman EA. (2008) **Positive feedback in cellular control systems**. *Bioessays*; 30:542-55.
16. Ninfa AJ. (2007) **Using two-component systems and other bacterial regulatory factors for the fabrication of synthetic genetic devices**. *Methods Enzymol.*; 422:488-512.

17. Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT. (2008) **Rewiring the specificity of two-component signal transduction systems.** *Cell*. 13;133(6):1043-54

18. Huang L, Tsui P, Freundlich M. (1992) **Positive and negative control of ompB transcription in *Escherichia coli* by cyclic AMP and the cyclic AMP receptor protein**. *J Bacteriol*.; 174:664-70.

19. Jubelin G, Vianney A, Beloin C, Ghigo JM, Lazzaroni JC, Lejeune P, Dorel C. (2005) **CpxR/OmpR interplay regulates curli gene expression in response to osmolarity in *Escherichia coli*.** *J Bacteriol*.; 187:2038-2049.

20. Kato A., Mitrophanov AY, GroismanEA. (2007) **A connector of two-component regulatory systems promotes signal amplification and persistence of expression**. *Proc Natl Acad Sci U S A*.; 104(29):12063-8.

21. Eguchi Y, Ishii E, Hata K, Utsumi R. (2011) **Regulation of acid resistance by connectors of two-component signal transduction systems in Escherichia coli.** *J Bacteriol.*;193(5):1222-8.

22. Eguchi Y, Itou J, Yamane M, Demizu R, Yamato F, Okada A, Mori H, Kato A, Utsumi R (2007). **B1500, a small membrane protein, connects the two-component systems EvgS/EvgA and PhoQ/PhoP in Escherichia coli.** *Proc Natl Acad Sci U S A* 104(47); 18712-7.

23. Herbarth O., Bauer M., Fritz G. J., Herbarth P., Rolle-Kampczyk U. et al. (2007) **Helicobacter pylori colonisation and eczema**. *Journal of Epidemiology and Community Health*;61:638-640

24. Pachkov M, Dandekar T, Korbel J, Bork P, Schuster S**.** (2007) **Use of pathway analysis and genome context methods for functional genomics of Mycoplasma pneumoniae nucleotide metabolism.** *Gene.* 15;396(2):215-25.

25. J. Mercer, A. Helenius (2008) **Vaccinia Virus Uses Macropinocytosis and Apoptotic Mimicry to Enter Host Cell**. *Science*, *320*, 531-535

26. Szybalski W. (1974) **In Vivo and in Vitro Initiation of Transcription**, *Control of Gene Expression*. New York: pp 23-24 and pp. 404-417.Plenum Press,

27. Serrano L. (2007) **Synthetic biology: promises and challenges.** *Mol Syst Biol.*; 3:158.

28. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA 3rd, Smith HO, Venter JC. (2010) **Creation of a bacterial cell controlled by a chemically synthesized genome.** *Science*; 329(5987):52-6.

29. Lartigue C, Glass JI, Alperovich N, Pieper R, Parmar PP, Hutchison CA 3rd, Smith HO, Venter JC. (2007) **Genome transplantation in bacteria: changing one species to another.** *Science*; 317(5838):632-8.

30. Camacho DM, Collins JJ. (2009) **Systems biology strikes gold.** *Cell*. 3; 137(1):24-6.

31. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, Cosma MP. (2009). **A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches**. *Cell*. 3; 137(1):172-81.

32. Lu TK, Khalil AS, Collins JJ. **Next-generation synthetic gene networks.** *Nat. Biotechnol.* 2009 Dec;27(12):1139-50.

33. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, Endy D. **Measuring the activity of BioBrick promoters using an in vivo reference standard.** *J Biol Eng.* 2009 Mar 20;3:4.

34. Shetty RP, Endy D, Knight TF Jr. (2008) **Engineering BioBrick vectors from BioBrick parts.** *J Biol Eng.* 14: 2-5.

35. The Gene Ontology Consortium. (2011) **The Gene Ontology: enhancements for 2011**. *Nucleic Acids Res*. **18**.

36. Balsa-Canto E, Banga JR. (2011) **AMIGO, a toolbox for advanced model identification in systems biology using global optimization**. *Bioinformatics*. 15;27(16):2311-3.

37. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. (2011) **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**; *Nucleic Acids Res*.; 39:D561-8.

38. Beal J, Lu T, Weiss R. (2011) **Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks**. *PloS One*.; 6(8):e22490.

39. Marchisio MA, Stelling J. (2011) **Automatic design of digital synthetic gene circuits.** *PLoS Comput Biol*.;7(2):e1001083.

40. Lou C, Liu X, Ni M, Huang Y, Huang Q, Huang L, Jiang L, Lu D, Wang M, Liu C, Chen D, Chen C, Chen X, Yang L, Ma H, Chen J, Ouyang Q. (2010) **Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch.** *Mol Syst Biol.*; 6:350.

41. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. Yeh, L.L. (2006) **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res*. 34, D187-D191.

42. Sigrist C. J. A, Cerutti L., Castro E.,Langendijk-Genevaux P. S., Bulliard V., Bairoch A., Hulo A. **PROSITE, a protein domain database for functional characterization and annotation** (2009) *Nucleic Acids Res*., **38**: D161-D166.

43. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. (2010) **The Pfam protein families database.** *Nucleic Acids Research,* 38:D211-D222

44. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman

HM, Bourne PE. (2011) **The RCSB Protein Data Bank: redesigned web site and web services.** *Nucleic Acids Research*, 39:D392-401

45. Laskowski R A, Chistyakov V V, Thornton J M (2005) **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.** *Nucleic Acids Res*., 33, D266-D268.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) **The Gene Ontology Consortium. Gene ontology: tool for the unification of biology.** *Nat. Genet*; 25(1):25-9.
47. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. (2010) **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res*. 38, D355-60
48. Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Münch R. (2009) **PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes**. *Nucleic Acids Res.* 37:D61-5
49. Sierro N., Makita Y., de Hoon M.J.L. and Nakai K. (2008) **DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information.** *Nucleic Acids Res.*, 36
50. Pérez AG, Angarica VE, Vasconcelos AT, Collado-Vides J. (2007) **Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes.** *Nucleic Acids Res.*;35:D132-6
51. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, (2010) **The IntAct molecular interaction database in 2010.** *Nucleic Acids Res*.; 38:D525-31.
52. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L. (2004) **Database resources of the National Center for Biotechnology Information: update.** *Nucleic Acids Res.*;32: D35-40.
53. Nelson, Stuart J., Zeng, Kelly; Kilbourne, John (2009) **Building a Standards-Based and Collaborative E-Prescribing Tool - MyRxPad**. *Proceedings of the 2009 IEEE*
54. Smith, T. Waterman M. S. (1981). **Comparison of biosequences**. *Advances in Applied Mathematics* 2:482-489
55. Needleman, S. B., Wunsch C. D. (1970). **A general method applicable to the search for similarities in the aminoacid sequence of two proteins**. *J Mol Biol*, 48:443-453.
56. Thompson, J. D., Higgins, D. G., Gibson, T. J. (1994). **Clustal W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acid Res*, 22:4673-4680.
57. Henikoff, S. Henikoff, J. (1991). **Automated assembly of protein blocks for database searching.** *Nucleic Acids Res*, 19:6565-6572

58. Henikoff, S. Henikoff, J.G. (1997). **Embedding strategies for effective use of information from multiple sequence alignments.** *Protein Science*. 6:698-705

59. Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004). **WebLogo: A sequence logo generator,** *Genome Research*, 14:1188-1190,

60. Altschul, S. F., Gish, W., Myers, W. M. E. W., Lipman, D. J. (1990). **Basic local alignment search tool.** *J Mol Biol*, 215:403-410

61. Altschul, S. F., Koonin, E. V. (1998) **Iterated profile searches with PSI-BLAST – a lool for discovery in protein databases.** 23:44-447

62. Zhang, Z., Schaeffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V., Altschul, S. F. (1998). **Protein sequence similarity searches using patterns as seeds.** *Nucleic Acids Res*, 26:3986-3990.

63. J. Cheng, A. Randall, M. Sweredoski, P. Baldi, (2005) **SCRATCH: a Protein Structure and Structural Feature Prediction Server**, *Nucleic Acids Research*, vol. 33:72-76.

64. B Rost, G Yachdav and J Liu (2004) **PredictProtein: The PredictProtein Server**. *Nucleic Acids Research* 32:321-326.

65. G.Pollastri, A. J. M. Martin, C. Mooney, A. Vullo. (2007) **Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information.** *BMC Bioinformatics*, 8:201.

66. Paolo Fontana, Eckart Bindewald, Stefano Toppo, Riccardo Velasco, Giorgio Valle and Silvio C.E. Tosatto**.** (2005) **The SSEA Server for Protein Secondary Structure Alignment.** *Bioinformatics*, 21(3): 393-395.

67. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. (2009). **The SWISS-MODEL Repository and associated resources.** *Nucleic Acids Res.*; 37:D387-92.

68. Bernstein HJ. (2000) Recent changes to **RasMol, recombining the variants.** Trends *Biochem Sci*., 25(9):453-5.

69. Fruchterman TMJ, Reingold EM. (1991) *Graph Drawing by Force-Directed Placement. Software, Practice and Experience*; 21:1129–1164.

70. Krüger B, Dandekar T. (2009) **Bioinformatical approaches to detect and analyze protein interactions.** *Methods Mol Biol*.; 564:401-31.

71. Hernandez-Toro J., Prieto C. and De Las Rivas J. (2007). **APID2NET: unified interactome graphic analyzer.** *Bioinformatics*

72. Bingding Huang and Michael Schroeder (2008), **Using protein binding site to improve protein-protein docking Gene**; 422(1-2):14-21.

73. McDowall, MD, Scott, MS and Barton, GJ. (2009) **PIPs: Human protein-protein interactions prediction database.** *Nucleic Acids Research* 37:D651-D656

74. A. Porollo, J. Meller. **Prediction-based Fingerprints of Protein-Protein Interactions Proteins:** *Structure, Function and Bioinformatics* (2007) 66: 630-45.

75. Winter C, Henschel A, Kim WK, Schroeder M. (2006) **SCOPPI: a structural classification of protein-protein interfaces.** *Nucleic Acids Res*.; 34:D310-4.

76. K. G. Tina, R. Bhadra and N. Srinivasan, (2007) **PIC: Protein Interactions Calculator**, *Nucleic Acids Research*, 35:W473–W476.

77. Hoffmann, R., Valencia, A; (2004) **A Gene Network for Navigating the Literature**. Nature *Genetics* 36, 664

78. Kolpakov F., Poroikov V., Sharipov R., Kondrakhin Y., Zakharov A., Lagunin A., Milanesi L. and Kel A. (2007) **Cyclonet - an integrated database on cell cycle regulation and carcinogenesis.** *Nucleic Acids Res*. 35

79. Pagel P, Oesterheld M, Stümpflen V, Frishman D. (2006) **The DIMA web resource – exploring the protein domain network**. *Bioinformatics*; 22(8): 997-998

80. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, and Jothi R. **DOMINE: A comprehensive collection of known and predicted domain-domain interactions.** *Nucleic Acids Research*, Vol 39 (Database Issue), D730-735, 2011.

81. Pacifico S, Liu G, Guest S, Parrish JR, Fotouhi F, Finley RL Jr. (2006) **A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for Drosophila**. *BMC Bioinformatics* 7:195.

82. Andres Leon E, Ezkurdia I, García B, Valencia A, Juan D. (2009) **EcID. A database for the inference of functional interactions in E. coli**. *Nucleic Acids Res*.;37(Database issue):D629-35.

83. Lin C.Y., Chen C.L., Cho C. S., Wang L. M. Chang C. M. Chen P. Y. Lo C. Z, Hsiung C. A. (2005) **hp-DPI: Helicobacter pylori Database of Protein Interactomes—embracing experimental and inferred interactions.** *Bioinformatics* 21 (7): 1288-1290.

84. Ng S. K., Zhang Z., Tan S. H., Lin K. (2003) **InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes.** *Nucl. Acids Res.* 31 (1): 251-254.

85. von Mering C, J Jensen L, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P. (2006) **STRING 7--recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res*.

86. Greiner M., Pfeiffer D., Smith R. D. (2000) **Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests.** *Preventive Veterinary Medicine*. Volume 45

87. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. (2003) **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science*; 302(5644):449-53.

88. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. (2004) **Prolinks: a database of protein functional linkages derived from coevolution.** *Genome Biol*.; 5(5):R35.

89. Luippold A.H., Arnhold T., Joerg W, Krueger B., Suessmuth R.D. **Application of a Rapid and Integrated Analysis System (RIAS) as a High-Throughput Processing Tool for in vitro ADME Samples by LC-MS/MS.** *J. Biomol. Screening. Manuscript accepted (2010).*

90. Laub MT, Goulian M. (2007) **Specificity in two-component signal transduction pathways.** *Annu Rev Genet*.; 41:121-45. Review.

91. Müller S, Götz M, Beier D. (2009) **Histidine residue 94 is involved in pH sensing by histidine kinase ArsS of Helicobacte pylori.** *PLoS One*; 4(9):e6930.

92. Huang YS, Chuang DT. (2000) **Regulation of branched-chain alpha-keto acid dehydrogenase kinase gene expression by glucocorticoids in hepatoma cells and rat liver.** *Methods Enzymol.*; 324:498-511.

93. Müller R, Fernández AP, Hiltbrunner A, Schäfer E, Kretsch T. (2009) **The histidine kinase-related domain of Arabidopsis phytochrome a controls the spectral sensitivity and the subcellular distribution of the photoreceptor.** *Plant Physiol.*;150(3):1297-309

94. Nakashima A, Sato T, Tamanoi F. (2010) **Fission yeast TORC1 regulates phosphorylation of ribosomal S6 proteins in response to nutrients and its activity is inhibited by rapamycin.** *J Cell Sci.*; 123(Pt 5):777-86.

95. Gaudermann P, Vogl I, Zientz E, Silva FJ, Moya A, Gross R, Dandekar T. (2006) **Analysis of and function predictions for previously conserved hypothetical or putative proteins in *Blochmannia floridanus*.** *BMC Microbiol*; 6:1.

96. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 23, 324–328

97. Bedau MA. (2010) **An aristotelian account of minimal chemical life.** *Astrobiology***.** 10(10):1011-20.

98. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. (2003) **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics*; 4:41.

99. Boyle PM, Silver PA. (2009) **Harnessing nature's toolbox: regulatory elements for synthetic biology.** *J R Soc Interface*.6 Suppl 4:S535-46. Review

100. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M**.** (2007) **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc.*; 2(10):2366-82.

101. **SQL**. Einführung in SQL und relationale Datenbanken. SPC TEIA Lehrbuch Verlag (2002)

102. Koch K, Barthlott W. Philos –(2009) **Superhydrophobic and superhydrophilic plant surfaces: an inspiration for biomimetic materials.** *Transact A Math Phys Eng Sci.* 28; 367(1893):1487-509. Review.

103. Kaneda Y. **A non-replicating oncolytic vector as a novel therapeutic tool against cancer.** *BMB Rep.* 2010 Dec;43(12):773-80. Review.

104. Zhang Q, Liang C, Yu YA, Chen N, Dandekar T, Szalay AA. (2009) **The highly attenuated oncolytic recombinant vaccinia virus GLV-1h68: comparative genomic features and the contribution of F14.5L inactivation**. *Mol Genet Genomics*. 282(4):417-35.

105. Galperin MY, Cochrane GR**.** (2011) **The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection**. *Nucleic Acids Res.*; 39(Database issue):D1-6.

106. Masher, T., Helmann, J., and Unden, G. (2006) **Stimulus perception in bacterial signal-transducing histidine kinases.** *Microbiol Mol Biol Rev*.;70(4):910-38.

107. Whitworth, DE and Cock,JA (2009) **Evolution of prokaryotic two-component systems: insights from comparative genomics.** *Amino Acids* 37:459–466.

108. Ulrich LE, Zhulin IB. (2010) **The MiST2 database: a comprehensive genomics resource on microbial signal transduction.** *Nucleic Acids Res*; 38:D401-7.

109. Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L. (2009) **Transcriptome complexity in a genome-reduced bacterium.** *Science*. 27;326(5957):1268-71

110. Haseloff J, Ajioka J. (2009) **Synthetic biology: history, challenges and prospects**. *J R Soc Interface*, Suppl 4:S389-91.

111. Lammers CR, Flórez LA, Schmeisky AG, Roppel SF, Mäder U, Hamoen L, Stülke J. (2010) **Connecting parts with processes: SubtiWiki and SubtiPathways integrate gene and pathway annotation for Bacillus subtilis**. *Microbiology*; 156(Pt 3):849-59.

B Genutzte Software

SOFTWARE	VERSION	NUTZUNG
Perl	5.8.8	Datensammlung, -aufbereitung und Text-mining
PostScript	1.5	Visualisierung und Präsentation
HTML	4.01	Integration der Ergebnisse in die Synthetic biology Software-Oberfläche
Spotfire®	2.2	Datenvisualisierung
Visual Basic for Application Microsoft Excel	2002 SP3	Datenaufbereitung und Programmierung
Microsoft Access	2002 SP3	Datenspeicherung
NCBI BLAST, PSI-BLAST und PHI-BLAST	2.2.24	Sequenzähnlichkeitssuchen
EBI ClustalW		Multiple Alinierungen
Blockmaker		Lokale Sequenzähnlichkeitssuchen
Weblogo	2.8.2	Erstellung Sequenzlogos
SSPro	4.0	Vorhersage Sekundärstruktur
PredictProtein		Vorhersage Proteinstrukturdetails
SSEA	08 /2004	Alinierung Sekundärstrukturen
SWISS-MODEL		Homologiemodellierung
Rasmol	2.6	3D-Darstellung Proteinstrukturen

C Abkürzungen

ABKÜRZUNG	BEDEUTUNG
Asp	Asparagin
ATP	**A**denosin**tri**p**h**osphat
BLAST	Basic Local Alignment Search Tool
BP	Basenpaare
COBBLER	**CO**nsensus **B**iasing **B**y **L**ocally **E**mbedding **R**esidues
DBTBS	Database of transcriptional regulation in *B. subtilis*
EnvZ, OmpR, NarL, NarX, NtrC, PhoB	Proteine aus bekannten TCS: Osmolaritäts-Sensorprotein EnvZ, Transkriptional regulatorisches Protein OmpR, Nitrate/Nitrit Response

	Regulator Protein NarL, Nitrate/Nitrit Sensorprotein NarX, Nitrogen Assimilation regulatorisches Protein NtrC, Phosphat regulatorisches transkriptionales Regulationsprotein PhoB
EBI	European Bioinformatics Institute
GO	Gene Ontology
His	Histidin
HisKA	**H**istidin**k**in**a**se
HTH	**H**elix-**T**urn-**H**elix
KEGG	Kyoto Encyclopedia of Genes and Genomes
MeSH	Medical Subject Headings
MIT	Massachusetts Institute of Technology
PAS	Ubiquitär vorkommende Strukturdomäne in Verbindung mit Signalproteinen, benannt nach den drei vorkommenden Proteinen: **P**eriod circadian protein, **A**h receptor nuclear translocator protein, **S**ingle-minded protein
PDB	Protein Data Bank
PHI-BLAST	Pattern Hit Initiated BLAST
Prodoric	**PRO**cariot**IC D**atabase Of Gene-**R**egulation
PSI-BLAST	Position-Specific Iterative BLAST
ResponseReg	Responseregulator-Domäne des Regulationsproteins in TCS
ROC	**R**eceiver **O**perating **C**haracteristic
RR	Regulationsprotein
SSEA	Secondary Structure Element Alignment
SSPro	**S**econdary **S**tructure **Pr**edict**io**n
STRING	Search Tool for Recurring Instances of Neighbouring Genes
TCS	Two-Component System (Zweikomponenten-System)
TransReg	Transkriptions- Regulator Domäne

D Lebenslauf

Persönliche Daten

Name:	Beate Krüger
Geburtstag:	07.03.1982 in Frankfurt/ Main
Anschrift:	Feldstraße 11, 65824 Schwalbach Tel.: 06196/81989 E-mail: beate.krueger82@web.de

Schulbildung/Studium

1988 – 1992	Besuch der Grundschule in Schwalbach
1992 – 2001	Besuch des Gymnasiums St. Angela (priv. staatl. anerkannt) in Königstein Abschluss Abitur (Note: 1,4)
2001-2005	Besuch Fachhochschule Giessen-Friedberg Studium der Bioinformatik Abschluss Diplom (Note: 1,4)
Seit 2006	Julius-Maximilians Universität in Würzburg Promotion

Praktische Erfahrungen

2008	Praktikum bei Sanofi-Aventis (1 Monat)
2005	Praxissemester bei Merck KGaA (6 Monate)
2005	Diplomarbeit am European Molecular Biology Laboratory (EMBL) (6 Monate)
seit 2006	Festanstellung bei Boehringer Ingelheim Pharma GmbH und Co KG in Biberach a.d. Riss als Systemanalytiker, Schwerpunkt Business Consultant

Sonstige Qualifikationen

Sprachkenntnisse:	Englisch, fließend Französisch, Grundkenntnisse
Computerkenntnisse:	Datenbanken: Oracle, MS-SQL, My-SQL Skriptsprachen: Perl, Unix, R (Grundlagen) Programmiersprachen: C#, Java, Delphi, Pascal
Sonstiges:	HTML, XSLT, Xpath, XML ,UML

Hobbies

Tanzen, Lesen

Schwalbach, Februar 2012

E Publikationen

- von Mering C, J Jensen L, Kuhn M, Chaffron S, Doerks T, **Kruger B**, Snel B, Bork P. (2006) STRING 7—Recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*

- **Krüger B**, Dandekar T.; (2009) Bioinformatical approaches to detect and analyze protein interactions. *Methods Mol Biol*. 564:401-31;
  Impact Faktor 13

- Luippold AH, Arnhold T, Jörg W, **Krüger B**, Süssmuth RD. (2011) Application of a rapid and integrated analysis system (RIAS) as a high-throughput processing tool for in vitro ADME samples by liquid chromatography/tandem mass spectrometry. *J Biomol Screen*;16(3):370-7.

- Liang C, **Krüger B**, Schaack D, Audretsch C, Nilla S, Boyanova D, Bernhardt J, Moya A, Lalk M, Schuster S, Völker U, Dandekar T. A systems perspective on bacterial adaptation exemplified on S. aureus fermentation metabolism.
  Eingereicht bei *Biological Chemistry*

- **Krueger B**, Liang C, Dandekar T. GoSynthetic software on natural and engineered molecular processes.
  In Revision bei *PLOS One*

- **Krueger B**, Friedrich T, Förster F, Bernhard J, Gross R, and Dandekar T. Different evolutionary modifications as a guide to rewire two-component systems
  Eingereicht bei *Bioinformatics and Biology Insights*

Konferenzen:

- **B. Krüger**, C. Liang, M. Naseem, T. Dandekar**;** Workshop: "*Streamlined and Synthetic genomes***"**, Nov 2009 Valencia, Spain; "Streamlining genomes: Several different roads and different results"

- Arnhold T., **Krüger B**.; "*7th Japan Spotfire Users Forum*" Nov 2009 Tokio, Japan. "*Data handling, visualization, and communication in the pharmaceutical research environment with Tibco Spotfire and TS Web Player*"

- Arnhold T., **Krüger B**.; "*Tibco Spotfire European Users Conference*" Mar 2011 Paris, Frankreich. "*Data handling, visualization, and communication in the pharmaceutical research environment*"

- Arnhold T., **Krüger B**., Luippold A, Jörg, W., Klinder, K; "*SBS Annual Conference*", Mar 2011 Orlando, Florida "*The optimization of instrumentation, workflow and data analysis for in vitro ADME assays: A business case*"

- C. Liang, **B. Krüger**, D. Schaack, C. Audretsch, S. Nilla, D. Boyanova, J. Bernhardt, A. Moya, M. Lalk, S. Schuster, U. Völker, T. Dandekar. *A systems perspective on bacterial adaptation exemplified on S. aureus fermentation metabolism.* Biological Chemistry: Highlight Issue to the GBM Meeting "Molecular Life Sciences 2011"