Vorhersage von phänotypischen Eigenschaften prokaryotischer Organismen auf Grundlage von Proteindomänendetektion

Bachelorarbeit

vorgelegt von

Stefanie Mühlhausen

aus
Kassel

angefertigt

im Institut für Mikrobiologie und Genetik

an der Biologischen Fakultät
der Georg-August-Universität zu Göttingen

2009

Betreuer: Dr. Peter Meinicke
Zweitbetreuerin: Dr. Maike Tech
Tag der Abgabe der Bachelorarbeit: 19.08.2009

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Liste der verwendeten Abkürzungen und Schreibweisen

SVM	**S**upport **V**ektor **M**aschine
SH2	**S**rc **H**omology **2** (Proteindomäne)
SH3	**S**rc **H**omology **3** (Proteindomäne)
DNA	**D**eoxyribo**n**ucleic **A**cid (Desoxyribonukleinsäure)
Pfam	**P**rotein **Fam**ilies
HMM	**H**idden **M**arkov **M**odell
GOS	**G**lobal **O**cean **S**urvey
RFE	**R**ecursive **F**eature **E**limination (Rekursive Merkmalseliminierung)
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
UFO	**U**ltrafast **f**unctional pr**o**filing
tp	**t**rue **p**ositive (richtig positiv)
fp	**f**alse **p**ositive (falsch positiv)
fn	**f**alse **n**egative (falsch negativ)
DUF	**D**omains of **U**nknown **F**unction (Domänen unbekannter Funktion)
ID	**Id**entifier (eindeutige Identifikationsnummer)
C_4	Monosaccharid mit vier Kohlenstoffatomen
UV	**U**ltraviolettstrahlung
RNA	**R**ibo**n**ucleic **A**cid (Ribonukleinsäure)
$\langle \vec{x} \cdot \vec{y} \rangle$	Skalarprodukt zwischen den Vektoren $\vec{x}$ und $\vec{y}$
$\#\{...\}$	Anzahl der Elemente der Menge $\{...\}$

Kapitel 1

Einleitung

Die Klassifikation von Organismen ist eine zentrale Aufgabe in der Biologie. Charakteristische Merkmale dienen der Einordnung in das bestehende phylogenetische System. Die traditionelle Klassifizierung über den Phänotyp, das äußere Erscheinungsbild eines Organismus, wird durch Kenntnisse über die genetische Ausstattung des Organismus, den Genotyp ergänzt. Der Phänotyp beschreibt die morphologischen, physiologischen, ökologischen und verhaltenstypischen Eigenschaften eines Organismus. Er leitet sich aus der Gesamtheit der Gene des Organismus ab, und wird zusätzlich von Umwelteinflüssen beeinflusst. Somit kann derselbe Genotyp leicht unterschiedliche Phänotypen hervorbringen. [23]

Die phänotypische Klassifikation von Organismen erfordert umfassende Kenntnisse, die durch eingehende Studien erhalten werden können. Dies wirft bei Prokaryoten einige Probleme auf. Schätzungen zufolge sind lediglich 0,1-1% [14] aller Prokaryoten im Labor kultivierbar. Bei dem Großteil der Organismen ist es folglich nicht möglich, den Phänotyp direkt zu beobachten und sie auf Grund dessen zu klassifizieren. Um dennoch eine taxonomische Einordnung vornehmen zu können, wird der Genotyp hinzugezogen. Mittels der DNA-Sequenz erhofft man sich, nah verwandte Organismen zu identifizieren. Das Argument ist, dass eine hohe Sequenzähnlichkeit Hinweis auf geringe phylogenetische Distanz ist. In der Praxis ist es jedoch häufig nicht möglich, prokaryotische Organismen isoliert zu sequenzieren. Spezielle Verfahren wie das Single Cell Sequencing [40] ermöglichen es in Kombination mit bioinformatischen Methoden, aus komplexen mikrobiellen Gemeinschaften Referenzgenome unkultivierter Taxa zu erstellen. Aus diesem Grund spielt die Bioinformatik eine zunehmende Rolle bei der Klassifikation von Organismen, insbesondere bei komplexen Gemeinschaften.

In diesem Zusammenhang ist die Frage interessant, inwieweit der Genotyp Rückschlüsse auf den Phänotyp zulässt. Die vorliegenden Arbeit dient der phänotypischen

Klassifikation von Organismen mit Fokus auf der Detektion von zu Grunde liegenden Proteinfunktionen. Basierend auf bekannten Organismen mit bestimmten phänotypischen Eigenschaften wird eine „Support Vektor Maschine“ (SVM) [3, 5] trainiert. Dabei handelt es sich um ein Verfahren des maschinellen Lernens, das eine Klassifikation unbekannter Daten, in diesem Fall die Einordnung unbekannter Organismen, ermöglicht. Eine SVM findet die Trennebene zwischen zwei Klassen, die den größtmöglichen Abstand („Margin“) zu beiden Klassen hat. Anhand dieser Trennfunktion werden unbekannte Datenpunkte als zu der einen oder anderen Klasse gehörend klassifiziert.

Einzelne Merkmale von Organismen (zum Beispiel die Gramfärbung) werden isoliert von dem restlichen Phänotyp betrachtet und vereinfacht als Zweiklassenproblem modelliert. Das bedeutet, dass jedem Organismus die Ausprägung oder Nicht-Ausprägung des partiellen Phänotypes zugeordnet wird. Dargestellt werden die Organismen als Datenpunkte im mehrdimensionalen Raum. Im Rahmen dieser Arbeit entsprechen die einzelnen Dimensionen spezifischen Proteindomänen. Die Koordinaten der Datenpunkte sind durch die relative Häufigkeit der entsprechenden Proteindomänen gegeben.

Proteindomänen zählen zu den kleinsten Struktureinheiten eines Proteins. Gebildet werden sie aus einer Sequenz von 30-400 Aminosäuren [34]. Abhängig von der räumlichen Struktur, die die Aminosäuresequenz des Proteins einnimmt, bilden sich die komplex gefalteten, globulären Proteindomänen aus. Sie sind unabhängig vom übrigen Protein gefaltet. Zum Teil führen sie eigene Reaktionen aus. Einander ähnliche Proteindomänen können mehrfach in demselben Protein vorkommen (z. B. SH2- und SH3-Domänen in dem menschlichen Onkogen Vav), sowie in evolutionär nahe verwandten Proteinen einer Familie (z. B. in Onkogenen von Mensch und Maus) [30] oder in sehr unterschiedlichen Proteinen mit ähnlicher Funktion. Strukturell und funktionell wichtige Bereiche eines Proteins sind in aller Regel evolutionär stärker konserviert. Generell ist die Struktur eines Proteins besser konserviert als die zugrunde liegende DNA-Sequenz [30].

Ziel der Arbeit soll eine Auswertung der Domänen bezüglich ihrer Wichtigkeit für die Unterscheidung innerhalb eines partiellen Phänotypes (einer Klasse) und zwischen den partiellen Phänotypen sein. Die Menge der für einen partiellen Phänotyp charakteristischen Proteindomänen wiederum kann Rückschlüsse auf charakteristische Stoffwechselwege zulassen. [Im Folgenden wird der partielle Phänotyp verkürzend als Phänotyp bezeichnet, gemeint ist weiterhin der partielle Phänotyp.] Dies wiederum ist eine wichtige Information für die funktionelle Annotation von Proteindomänen. Interessant ist in diesem Zusammenhang die Fragestellung, ob unter den charakteristischen Domänen auch Domänen unbekannter Funktion sind. In diesem Fall könnte deren funktionelle Annotation mit den gewonnenen Informationen ver-

bessert werden. Ebenfalls von Interesse ist die Verteilung der Domänenfamilien unter den charakteristischen Domänen eines Phänotypes, also die Frage, ob bestimmte Familien gehäuft vorkommen.

Eine Datenbank, die sich für die Vorhersage von phänotypischen Eigenschaften auf Grundlage von Proteindomänen besonders eignet, ist Pfam [10]. Pfam ist eine Proteindomänen-Datenbank, in der Datensätze zu Proteinfamilien hinterlegt sind. Der Umfang der Datenbank wird aktuell[1] mit 10340 Proteinfamilien angegeben. Zu den Proteinfamilien werden multiple Sequenzalignments und Profil-Hidden-Markov-Modelle (Profil-HMM) [7] zur Verfügung gestellt. Mit den Profil-HMMs werden Modelle für Proteindomänen realisiert. Die Datenbank ist in zwei Teilbereiche, Pfam-A und Pfam-B, unterteilt. Pfam-A Einträge werden manuell kuratiert. Damit wird eine hohe Qualität der Annotation gewährleistet. Gleichzeitig ist der Umfang der Datenbank durch den großen Arbeitsaufwand limitiert. Zu Pfam-A gibt es eine übergeordnete Hierarchie. Verwandte Familien ähnlicher Struktur, Sequenz oder Profil-HMM sind zu Clans zusammengefasst. Die vollständigere Abbildung der bekannten Proteine ist in Pfam-B hinterlegt. Pfam-B besteht aus automatisch generierten Einträgen. Durch diese Unterteilung liegt die Entscheidung zwischen gesicherten, aber im Umfang eingeschränkten Wissen über Proteindomänen und einer nahezu vollständigen, aber möglicherweise in Teilen fehlerhaften Abbildung der bekannten Proteine beim Nutzer. Grundlage dieser Arbeit sind die unter Pfam-A annotierten Proteindomänen.

Da Proteinfamilien häufig durch strukturelle Bereiche, also Proteindomänen, charakterisiert werden, ermöglicht Pfam die Klassifikation von Proteinen in Proteinfamilien. Neue Sequenzen können auf ihre Ähnlichkeit mit diesen Profil-HMMs untersucht und darauf aufbauend klassifiziert werden.

Wichtige Anwendungsbereiche für die Ergebnisse erschließen sich insbesondere in der Metagenomik. Hierbei handelt es sich um einen Forschungsbereich, der sich mit der Analyse von mikrobiellen Habitaten beschäftigt. Aus einer Umweltprobe (Proben aus Habitaten, wie z.B. hypersaline Mikrobenmatten) wird DNA direkt isoliert und sequenziert, ohne die Organismen vorher im Labor zu kultivieren. Dies führt zu einigen Besonderheiten der Metagenomik. Die Kultivierung von Organismen bietet grundsätzlich die Möglichkeit der taxonomischen Identifizierung, stößt in Habitaten hoher Diversität aber an Grenzen. Diese Grenzen ergeben sich aus oben erwähnten Einschränkungen der Kultivierbarkeit und der Vielzahl an Organismen. Folglich ist die Metagenomik durch die Arbeit mit großen Datensätzen aus DNA-Fragmenten von "anonymen" Organismen gekennzeichnet. Die Datenanalyse erfolgt unter verschiedenen Gesichtspunkten. Während einerseits bei der funktionellen Analyse Aussagen über metabolische Eigenschaften des Metagenomes getroffen werden,

[1] Quelle: http://pfam.sanger.ac.uk/, Abruf am 10. August 2009

steht andererseits bei der phylogenetischen Analyse die Diversität im Vordergrund. In der funktionellen Analyse werden Stoffwechselwege aus der Genannotation abgeleitet. Im Zuge dessen kann die Annotation „hypothetischer Proteine“ verbessert werden [14].

Ein aktuelles Beispiel der sequenzbasierten Analyse von Metagenomen ist der Global Ocean Survey (GOS)-Datensatz [33]. Anhand dieses Metagenomes wurde unter anderem untersucht, ob Korrelationen zwischen ökologischen und genetischen Variablen bestehen. Die zugrundeliegende Fragestellung ist, inwieweit die Stoffwechselwege einer mikrobiellen Gemeinschaft auf deren Anpassung an ihre Umwelt hinweisen. Dabei wird der Begriff der Umwelt als das Zusammenspiel mehrerer, kontinuierlich veränderbarer Variablen verstanden und diese auf Korrelationen mit ausgewählten Stoffwechselwegen hin analysiert. Es wurde gezeigt, dass unweltabhängige Stoffwechselwege (“footprint characteristics”) existieren [11]. Insbesondere Energiestoffwechselwege und autotrophe Prozesse sind mit Umweltvariablen korreliert.

Die funktionelle Charakterisierung wurde auf eine hypersaline Mikrobenmatte aus Guerrero Negro [19] angewandt. Hier wurden verschiedene Schichten auf Phylogenie, funktionelles Potential der Organismen sowie genetische Gradienten zwischen den Schichten untersucht. Es hat sich gezeigt, dass analog zu den bereits untersuchten physikalisch-chemischen Gradienten der verschiedenen Schichten auch genetische Gradienten existieren. So lässt sich entsprechend der in tieferen Schichten abfallenden Sauerstoffgradienten eine Verschiebung der vorherrschenden Genfamilien und Proteindomänen erkennen. Interessant in solchen Zusammenhängen sind uncharakterisierte Proteindomänen, die unter Umständen durch die erzielten Ergebnisse charakterisiert werden können.

In einer weiteren Arbeit wurden Zusammenhänge zwischen metabolischen Eigenschaften und Habitaten aufgedeckt. Dafür werden Organismen mit ähnlichen metabolischen Eigenschaften zu Clustern zusammengefasst. Auf diese Cluster werden Habitate abgebildet. Dabei konnte gezeigt werden, dass sie die Phylogenie der Organismen gut widerspiegeln. Des weiteren ergab die Abbildung der Habitate auf Cluster, dass Korrelationen zwischen den Eigenschaften und Habitaten bestehen. [18]

Kapitel 2

Methoden

Die Vorhersage von phänotypischen Eigenschaften von Prokaryoten erfolgt mit einem Verfahren des maschinellen Lernens, Support Vektor Maschinen. Die Ausprägung eines Phänotypes wird als Summe unabhängiger Eigenschaften aufgefasst. Für jede Eigenschaft wird gesondert überprüft, ob sie auf einen bestimmten, durch sein Proteindomänenprofil gegebenen, Organismus zutrifft oder nicht. Die Organismen werden als Datenpunkte im d-dimensionalen Raum dargestellt. Die Anzahl der Dimensionen entspricht der Anzahl unterschiedlicher Proteindomänenfamilien. Die Koordinate eines Organismus ist durch die relative Häufigkeit von Genen, in denen die entsprechende Domäne vorkommt, gegeben.

Die Organismen werden in zwei Klassen eingeteilt. Haben sie eine bestimmte Eigenschaft, werden sie zu der einen, ansonsten zu der anderen Klasse gezählt. Die Mengen haben die Eigenschaft, sich durch eine lineare Trennfunktion voneinander abgrenzen zu lassen. Die Klassifikation eines neuen Datenpunktes erfolgt anhand dieser Trennfunktion. Abhängig davon, auf welcher Seite der Trennfunktion die Darstellung eines Organismus liegt, wird er zu der einen oder anderen Klasse gezählt.

Es handelt sich um einen hochdimensionalen Datenraum. Etliche Dimensionen tragen nicht zur Diskriminante (der linearen Trennfunktion) bei, z. B. weil die entsprechenden Proteindomänen in keinem Organismus vorkommen. Von daher erscheint es ratsam, die Anzahl der Dimensionen zu reduzieren. Die Dimensionen, die am wenigsten zu der Diskriminaten beitragen, sollen nach und nach eliminiert werden. Diese Technik wird als rekursive Merkmalseliminierung („Recursive Feature Elimination“, RFE) [13] bezeichnet.

DNA-Sequenz und Annotation der Genome wurden der NCBI-Datenbank entnommen. Diese Daten stammen aus einem großangelegten Sequenzierungsprojekt, dem „Prokaryotic Genome Project“ [26]. Bei diesem Projekt handelt es sich um eine Sammlung vollständig bzw. teilweise sequenzierter prokaryotischer Genome. Zu den

Genomen werden einige Informationen bereitgestellt, die von dem Namen des Organismus über den Link zur Genomsequenz bis hin zu phänotypischen Eigenschaften wie der Gramfärbung reichen. Eine vollständige Auflistung aller verfügbaren Eigenschaften ist in Anhang A zu finden. Mittels UFO („Ultrafast Functional Profiling“) [21] wird für jeden Organismus die absolute Häufigkeit der Gene, die mit den Proteindomänen aus Pfam-A korrelieren, bestimmt. Im Folgenden wird dieses Profil als das Domänenprofil eines Organismus bezeichnet.

2.1 Support Vektor Maschinen

Gegeben ist eine Menge X an gelabelten Datenpunkten, wobei für jedes Datum ein Merkmalsvektor $\vec{x}$ und das Label y als Zugehörigkeit zur positiven oder negativen Klasse gegeben ist.

$$\begin{aligned} & X = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\} \\ & \text{mit} \quad \vec{x} = (x_1 x_2 \ldots x_d)^T \\ & \text{und} \quad y_i \in \{-1, 1\} \end{aligned}$$

Angewendet auf die Vorhersage von Phänotypen entspricht X den NCBI-Organismen. Dabei ist $X \in \mathbb{N}^{d \times n}$: Einer Auswahl von insgesamt 1039 komplett sequenzierten Organismen werden Proteindomänen mit Pfam-IDs zwischen PF00001 und PF10797 zugeordnet. Der Merkmalsvektor ist das Domänenprofil, in jeder Dimension ist die Häufigkeit der Pfam-ID eingetragen, die der Dimensionszahl entspricht. Aus den Zusatzinformationen der NCBI-Annotation werden Kategorien generiert. Eine Voraussetzung für die Anwendung von maschinellen Lernen ist eine ausreichende Menge an Trainingsbeispielen pro Klasse. Aus diesem Grund können nicht alle unter NCBI vorgestellten Kategorien zur Generierung von Zweiklassenproblemen verwendet werden. Ein Beispiel hierfür ist die Kategorie „Arrangement“. Die Art des Zellverbandes ist so detailliert dargestellt, dass eine Reduktion auf zwei Klassen kaum durchführbar ist, zumal es zum Teil nur eine Handvoll Organismen pro Klasse gibt. Darüber hinaus können nur die Kategorien verwendet werden, die aller Voraussicht nach durch Domänenprofile modelliert werden. In Anbetracht dieser Tatsachen werden folgende Kategorien generiert:

Zugehörigkeit zur Domäne des Lebens	„Super_Kingdom“
Gramfärbung	„Gram_Stain“
Endosporen	„Endospores“
Beweglichkeit	„Motility“
Sauerstoffbedarf	„Oxygen_Req“

Salztoleranz	„Salinity“
Habitat	„Habitat“
Temperaturbedingungen	„Temp_Range“

Für jede Kategorie (Eigenschaft) wird aus der Menge X eine Teilmenge X' extrahiert. Diese Teilmenge enthält alle Organismen $i = 1, \ldots, n$, die als zu der jeweiligen Kategorie zugehörig annotiert sind. Sie ist weiter unterteilt in zwei disjunkte Untermengen. Die eine Menge enthält alle positiven Beispiele, die andere alle negativen. Ein positives Beispiel ist in diesem Fall ein Organismus, auf den eine bestimmte Eigenschaft, wie zum Beispiel die Bildung von Endosporen oder die Zugehörigkeit zu den Bakterien, zutrifft. Entsprechend ist ein negatives Datenbeispiel ein Organismus, auf den diese Eigenschaft nicht zutrifft. Dies kann, je nach Kategorie, bedeuten, dass ein Organismus eine Eigenschaft nicht hat (z. B. keine Endosporen ausbildet) oder aber eine andere Eigenschaft hat, die von der ersten abgegrenzt werden soll (z. B. kein Bakterium, sondern ein Archaeum ist). Für jede Teilmenge X' wird ein Labelvektor $\vec{y}$ erstellt.

Die beiden Klassen sind linear separabel, das heißt, dass alle Datenpunkte einer Klasse auf derselben Seite einer Hyperebene liegen. In Abbildung 2.1 auf Seite 8 ist dies anschaulich dargestellt.

Gesucht ist eine affin lineare Trennfunktion,

$$
\begin{aligned}
& f : \mathbb{R}^d \to \mathbb{R} \\
& f(\vec{x}) = \langle \vec{w}, \vec{x} \rangle + b \\
& \text{mit} \quad \vec{w}\text{: Normalenvektor} \\
& \text{und} \quad \text{b: Verschiebung gegenüber dem Ursprung}
\end{aligned}
$$

die den Abstand γ der Datenpunkte zur Hyperebene maximiert.

$$
\begin{aligned}
& \text{Maximiere } \gamma \text{ unter der Bedingung} \\
& \forall i : y_i(\langle \vec{w}, \vec{x} \rangle + b) \geq \gamma
\end{aligned}
$$

Dies entspricht mit quadratischer Programmierung der Minimierung von $\|\vec{w}\|^2$.

$$
\begin{aligned}
& \forall i : y_i(\langle \vec{w}, \vec{x} \rangle + b) \geq 1 \\
& \text{mit } \|\vec{w}\| = 1/\gamma
\end{aligned}
$$

Die Maximierung der Mindestabstandes („Margin“) ist also äquivalent zu der Minimierung der quadrierten Norm des Normalenvektors. Da er den Beitrag der einzelnen Dimensionen zur Diskriminanten gewichtet, wird der Normalenvektor auch als Gewichtsvektor bezeichnet.

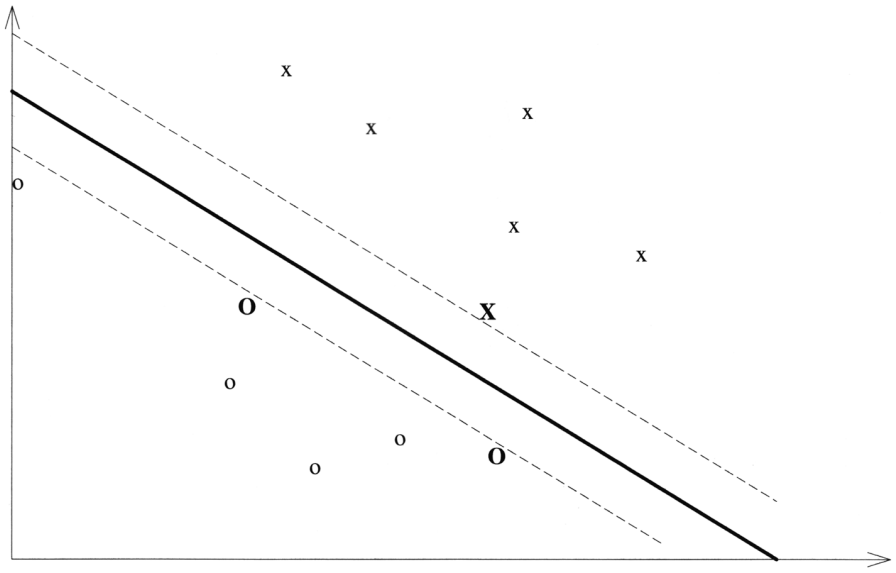

Abbildung 2.1: *Hyperebene mit maximalem Abstand zu den Datenpunkten. Hervorgehoben sind die Support Vektoren der Hyperebene. Quelle: Cristianini, Shawe-Taylor, S. 97*

Die optimale Trennfunktion ist durch einige Datenbeispiele gegeben, die genau den maximalen Mindestabstand zur Hyperebene haben. Diese Datenpunkte werden als Support Vektoren bezeichnet.

Gemäß der Trennfunktion wird ein linearer Klassifikator c realisiert.

$$c(\vec{x}) = \begin{cases} 1, & \text{falls } f(\vec{x}) > 0 \\ -1, & \text{sonst} \end{cases}$$

Die Parameter $\vec{w}$ und b werden entsprechend der oben genannten Bedingung auf bekannten Beispielen optimiert. Dies erfolgt kernbasiert [1, 5], um die Anzahl der zu optimierenden Dimensionen in $\vec{w}$ möglichst gering zu halten.

Dazu wird X' zufällig in eine Trainings- und in eine Testmenge aufgeteilt. Für beide Mengen soll gelten, dass sowohl positive als auch negative Beispiele enthalten sind. Zudem sollte die Kardinalität der Trainingsmenge größer oder gleich der Kardinalität der Testmenge sein. Mit diesen Bedingungen wird gewährleistet, dass eine optimale Diskriminante berechnet werden kann. Die Berechnung der Parameter erfolgt auf der Trainingsmenge. Die Beispiele der Testmenge dienen dazu, die Vorhersagegenauigkeit des Klassifikators zu überprüfen. Dabei treten im Allgemeinen Klassifikationsfehler auf. Zur Abschätzung der Klassifikationsrate werden zwei

Größen der Statistik herangezogen, Sensitivität und Spezifität. Sie sind wie folgt definiert:

$$Sensitivität = \frac{tp}{tp + fn}$$
$$Spezifität = \frac{tp}{tp + fp}$$

wobei mit

tp (true positive): ein korrekt als positiv erkannter Sachverhalt

fp (false positive): ein fälschlicherweise als positiv erkannter Sachverhalt

fn (false negative): ein fälschlicherweise als negativ erkannter Sachverhalt

bezeichnet wird.

Das harmonische Mittel aus Sensitivität und Spezifität ist ein geeignetes Maß für die Güte der Klassifikation bei unterschiedlicher Anzahl an positiven und negativen Beispielen.

Es wird gemäß der Formel

$$H = \frac{2 \cdot Sensitivität \cdot Spezifität}{Sensitivität + Spezifität}$$

gebildet und eignet sich auch zum Vergleich der Klassifikationsraten der verschiedenen Kategorien. In dieser Arbeit wird der Begriff der Performanz stellvertretend für das harmonische Mittel gebraucht.

Die Aufteilung der Organismen in Trainings- und Testmenge erfolgt über den Genus. Um zu erreichen, dass alle Organismen eines Genus entweder zum Training oder zum Testen verwendet werden, wird die Aufteilung über die Genera vorgenommen. Ein vorgegebener Prozentsatz (70 %) aller Organismen soll der Trainingsmenge zugeteilt werden. Zugleich soll die Aufteilung zufällig erfolgen. Dies wird dadurch erreicht, dass über den Genera eine Permutation gebildet wird. Die ersten n Genera werden zum Training benutzt, die übrigen zum Testen. Dabei wird n wie folgt gewählt:

Wähle das kleinste n, so dass gilt: Die Anzahl Organismen in n ist größer oder gleich dem geforderten Anteil Trainingsbeispiele.

Unter den Organismen befinden sich einige, die bislang nicht gültig beschrieben werden konnten. Zu der gültigen Beschreibung einer neuen Bakterienart gehört unter anderem, dass die Art dauerhaft in Reinkultur gehalten werden kann, und in einer Sammlung hinterlegt ist. Ist dies noch nicht der Fall, wird dem Artnamen ein „Candidatus“ vorangestellt. Trotz der (noch) nicht gültigen Beschreibung der Art kann

sie gut charakterisiert sein. Andersherum sind auch die wissenschaftlich gültig beschriebenen Bakterienarten nicht in allen Kategorien gelabelt. Hierdurch motiviert, werden die Candidatus-Organismen ebenfalls berücksichtigt und als Datenbeispiele verwendet.

Bei der hier vorgestellten SVM handelt es sich um eine sogenannte Hard-Margin SVM. Das bedeutet, dass keine Abweichungen („Schlupfvariablen“) vom Mindestabstand zugelassen werden. Dieses Verfahren bietet gegenüber anderen Verfahren wie dem Perzeptron- Lernalgorithmus und Soft-Margin SVMs einige Vorteile. Während der Perzeptron-Lernalgorithmus unter den überabzählbar vielen möglichen Lösungen (Trennfunktionen) eine beliebige findet, findet die SVM die optimale Lösung. Die Anwendbarkeit von SVM ist zunächst auf linear separable Trainingsmengen beschränkt, kann durch die Einführung von Schlupfvariablen aber auf den allgemeineren Fall einer nicht linear separablen Trainingsmenge erweitert werden. Dies macht vor allem dann Sinn, wenn es nur wenige Ausreißer gibt, die bestraft werden sollen. Damit muss zusätzlich ein Hyperparameter, der den Einfluss von Ausreißern bestimmt, optimiert werden. Im Unterschied zu den Modellparametern $\vec{w}$ und b sind Hyperparameter Parameter der zugrundeliegenden A-Priori-Wahrscheinlichkeit.

2.2 Rekursive Merkmalseliminierung

Die rekursive Merkmalseliminierung (RFE) ist ein etabliertes Verfahren zur überwachten Dimensionsreduktion, das in Verbindung mit Klassifikatoren angewandt wird. Bei hochdimensionalen Datenräumen und vergleichsweise wenigen Datenpunkten besteht zum einen die Gefahr des sogenannten „Overfittings“, einer Überanpassung der Trennfunktion an die Daten, und zum anderen treten Probleme bei der Auswertung der Dimensionen auf. In der Bioinformatik steht oft die Frage im Raum, welche Merkmale (das können z. B. Gene, oder auch Proteindomänen sein) typisch für die beiden Klassen sind. Somit sollen aussagekräftige Merkmale detektiert werden. Das Prinzip der RFE besteht darin, rekursiv die Dimensionen (Merkmale) mit dem niedrigsten Gewicht zu verwerfen. Das Verfahren wird solange angewandt, bis auf diese Art eine nach den Gewichten geordnete Liste aller Merkmale erstellt ist. Dazu werden folgende Schritte wiederholt, bis nur noch ein Merkmal übrig ist:

---

repeat
  „trainiere den Klassifikator“
  „berechne $\vec{w}$ “
  „verwirf das Merkmal mit dem geringsten Absolutbetrag“
until „Anzahl übrig gebliebener Merkmale = 0“

Die RFE resultiert in einer geordneten Liste aller Merkmale, wobei das als erstes verworfene Merkmal am Ende der Liste steht.

In der vorliegenden Arbeit wird das Verfahren leicht modifiziert angewandt. Zum einen werden in jedem Iterationsschritt 5% der Merkmale verworfen, anstatt nur eine einziges Merkmal. Der Grund hierfür ist die hohe Anzahl an Ausgangsdimensionen. Da jede Iteration das erneute Trainieren der SVM beinhaltet, würde das Verfahren bei mehr als 10 000 Merkmalen eine extrem lange Laufzeit haben. Werden 5% der verbleibenden Merkmale verworfen und die Rekursion abgebrochen, sobald die Trainingsmenge nicht mehr linear separabel ist, sinkt die Anzahl nötiger Iterationsschritte rapide. Als zweite Abbruchbedingung dient die Anzahl verbleibender Merkmale. Da davon ausgegangen werden kann, dass nicht nur ein einziges Merkmal, sondern vielmehr eine Menge mehrerer Merkmale für die Klassifikation entscheidend ist, wird als Untergrenze das Verbleiben von 10 Merkmalen festgesetzt. Damit ergeben sich nach der Formel

$$10797 \cdot 0,95^x \geq 10$$

mit x: Anzahl nötiger Iterationsschritte

maximal 138 Iterationsschritte. Es hat sich herausgestellt, dass etwa zu diesem Zeitpunkt eine lineare Separabilität der Datenmenge nicht mehr gegeben ist.

Der modifizierte Algorithmus sieht wie folgt aus:

---

for all Kategorien **do**
  „ bilde 20 Zufallspartitionen unter Beachtung des Genus“
  while „ linear separabel AND Anzahl Merkmale $<$ Mindestanzahl“ **do**
    „ trainiere SVM“
    „ sortiere Merkmale“
    „ berechne Performanz“
    „ verwirf die, dem Betrag nach, kleinsten 5% der Merkmale“
    „ speichere übrige Merkmale“
  end while
end for

Die Auswahl der zu verwerfenden Merkmale erfolgt über zwei Statistiken, die miteinander verglichen werden können. Beide Methoden basieren auf der Tatsache, dass der Gewichtsvektor $\vec{w}$ das Gewicht der einzelnen Merkmale angibt. Zum einen erfolgt die Auswahl der Merkmale über das Gewicht der Merkmale in dem Gewichtsvektor, zum anderen über eine t-Teststatistik über den Gewichtsvektor.

In der ersten Methode wird für 20 Zufallspartitionen die RFE in Kombination mit einer SVM angewandt. Für jede Iteration der RFE werden die 5% der Merkmale

verworfen, die dem Betrag nach das niedrigste Gewicht im Gewichtsvektor haben. Auf den Zufallspartitionen wird pro Iterationsschritt eine SVM trainiert. Für jeden Parametervektor $\vec{w}$ der SVM wird berechnet, welche Merkmale verworfen werden. Vereinigung und Durchschnitt der Menge der verbleibenden Merkmale wird gebildet und zusammen mit dem über alle Partitionen gemittelten harmonischen Mittel aus Sensivität und Spezifität gespeichert.

Die zweite Methode verfährt umgekehrt. Hier wird pro Iteration der RFE auf 20 gleichbleibenden Zufallspartitionen die SVM trainiert. Zu den 20 resultierenden Gewichtsvektoren wird unter der Annahme der Standardnormalverteilung der Gewichte eine t-Teststatistik aufgestellt. Auf Grund dieser Teststatistik werden die 5% dem Betrag nach geringsten Merkmale verworfen. Mit den entsprechend verbleibenden Merkmalen wird in allen Zufallspartitionen weiter trainiert und selektiert. Pro Iterationsschritt der RFE wird die Menge der verbleibenden Merkmale und das über alle Partitionen berechnete harmonische Mittel aus Sensitivität und Spezifität gespeichert.

Während mit dem Absolutbetrag der Gewichte der einzelnen Merkmale Ausreißer selektiert werden, selektiert die Teststatistik auf Merkmale, deren Gewicht über mehrere Zufallspartitionen kaum schwankt. Die Teststatistik wird gemäß der Formel

$$T_i = \frac{\bar{x}_i - \mu_i}{\sigma_i} \cdot \sqrt{n}$$

berechnet. Ist die Standardabweichung eines Merkmales über alle Partitionen gleich null, so ist die Teststatistik mathematisch nicht definiert. Dieses Merkmal wird entfernt, da sein Gewicht in allen Partitionen gleich ist und somit keine Unterscheidung zwischen positiver und negativer Klasse anhand dieses Merkmales möglich ist.

Das Speichern der selektierten Merkmale pro Iteration bietet gegenüber dem Erstellen einer geordneten Liste mehrere Vorteile. Während in einer geordneten Liste nur das Endergebnis der Merkmalseliminierung enthalten ist, kann man bei der verwendeten Methode das Ergebnis, also die selektierten Merkmale, zu jedem Iterationsschritt untersuchen. Dies ist von Vorteil, da sich das Gewicht der verbleibenden Merkmale in jedem Iterationsschritt verändern kann, und sich somit andere Teilmengen an selektierten Merkmalen ergeben können.

Zudem bietet diese Methode die Möglichkeit, die Teilmenge der selektierten Merkmale bzw. die Menge der eliminierten Merkmale in Abhängigkeit der erzielten Performanz zu betrachten. Wie im nächsten Kapitel (Kapitel 3.4, Seite 23) näher beschrieben wird, haben beide Eliminierungsmethoden einen charakteristischen Kurvenverlauf der Performanz über der Anzahl verbleibender Merkmale. Bei der rekursiven Merkmalseliminierung (RFE) nach der Teststatistik gibt es ein Maximum bzw. ein Plateau maximaler Performanz. Die Performanz der RFE über die zwei-

te Methode nimmt tendenziell mit der Anzahl eliminierter Merkmale ab, schwankt jedoch.

Im Falle der RFE über den Absolutbetrag der t-Teststatistik wird besonderes Augenmerk auf die Merkmale gelegt, die in der Iteration der maximalen Performanz selektiert sind. Bei Eliminierung mittels des sortierten Absolutbetrages der Merkmalsgewichte werden die Merkmale (in dieser Arbeit handelt es sich dabei um Proteindomänen) untersucht, die nach dem größten Rückgang der Performanz selektiert sind. Zusätzlich werden die Merkmale gesondert betrachtet, nach deren Eliminierung die Performanz am stärksten abnimmt.

Kapitel 3

Ergebnisse

3.1 Fragestellung und Ansatz zur Bearbeitung

Das Leitmotiv der vorliegenden Arbeit ist die Verwendung von detektierten Proteindomänen, mit denen phänotypische Eigenschaften von Prokaryoten vorhergesagt werden können. Mittels der NCBI-Datenbank und dem Programm UFO werden gelabelte Datenbeispiele gewonnen. Dabei handelt es sich um Organismen, denen über ihr Domänenprofil ein Merkmalsvektor und über ihre Annotation verschiedene partielle Phänotypen als Label zugeordnet sind. Die Phänotypen werden voneinander getrennt betrachtet. Somit lässt sich die Klassifikationsaufgabe als Zweiklassenproblem auffassen, welches mit geeigneten Verfahren des maschinellen Lernens gelöst werden kann. Die Klassifikation erfolgt mit Support Vektor Maschinen. In Kombination mit einem Verfahren zur Dimensionsreduktion, der rekursiven Merkmalseliminierung („Recursive Feature Elimination“, RFE), werden Dimensionen, die am wenigsten zur Diskriminanten beitragen, verworfen. Das Verfahren selektiert die für die Klassifikation aussagekräftigsten Proteindomänen.

3.2 Zusammenstellung der Kategorien

Aus der NCBI-Datenbank werden Organismen sowie Kategorien ausgewählt. Während die Organismen vollständig sequenziert sein müssen, wird an die Kategorien der Anspruch gestellt, dass es für die einzelnen Klassen genügend Datenbeispiele gibt (in dieser Arbeit wurden mindestens 50 Beispiele in der positiven bzw. negativen Klasse gefordert) und die Performanz, das harmonische Mittel aus Sensitivität und Spezifität, einen festgelegten Schwellenwert (80%) überschreitet. Über das erste Kriterium werden die möglichen Kategorien auf neun reduziert (siehe Tabelle 3.1,

Kategorie	Positiv		Negativ		$\sum +$	$\sum -$	$\sum$ gesamt
Endospores	Yes	100	No	344	100	344	444
Gram_Stain	+	269	-	585	269	585	854
Habitat_1	Aquatic	183	Terrestrial	56	183	56	239
Habitat_2	Specialized	88	Multiple	320	88	320	408
Motility	Yes	467	No	254	467	254	721
Oxygen_Req	Aerobic Facultative	299 379	Anaerobic	174	678	174	852
Salinity	Extreme halophilic Moderate halophilic Mesophilic	8 18 28	Nonhalophilic	175	54	175	229
Super_Kingdom	Archaea	56	Bacteria	983	56	983	1039
Temp_Range	Hyperthermophilic Thermophilic	31 61	Mesophilic	835	92	835	927

Tabelle 3.1: ***Umfang der Kategorien und Klassen.*** *Dargestellt ist die Aufteilung der in den einzelnen Kategorien zu Verfügung stehenden Klassen auf die positive (auch mit „+“) und negative (auch mit „-“ bezeichnete) Klasse. Nicht verwendet wurden die Klassen „Host-Associated“ (332 Vertreter, Kategorie „Habitat“), „Mircoaerophilic“ (30 Vertreter, Kategorie „Oxygen_Req“), „Psychrophilic “ (13 Vertreter) und „Cryophilic“ (1 Vertreter, beide Kategorie „Temp_Range“).*

Seite 16 sowie Anhang A, Seite 49). Aus diesen neun Kategorien sollen Zweiklassenprobleme generiert werden. Die Kategorien sind in der Datenbank zum Teil in mehr als zwei Klassen unterteilt. In diesem Fall ist es notwendig, Klassen zu vereinigen bzw. wegzulassen. Dazu wird nach biologischen Gesichtspunkten vorgegangen, sofern andere Gründe wie z. B. der Umfang der Klasse nicht dagegen sprechen. Für jede Kategorie wird eine positive und eine negative Klasse generiert. Die Aufteilung der ursprünglichen Klassen auf diese beiden ist in der bereits erwähnten Tabelle 3.1 dargestellt.

Eine Reduktion der Klassen ist bei folgenden Kategorien erforderlich: „Habitat_1“ und „Habitat_2“, „Motiliy“, „Salinity“, „Temp_Range“ und „Oxygen_Req“. Die direkte Überführung der Klassen in eine positive und eine negative ist bei den Kategorien „Endospores“, „Gram_Stain“, „Super_Kingdom“ und „Motility“ möglich. Hier besteht bereits die Datenbank-Annotation aus zwei unterschiedlichen Klassen. Die Klassen „Yes“ und „No“, der im NCBI unter „Endospores“ gelabelten Organismen werden ebenso wie „+“ und „−“ der unter „Gram_Stain“ gelabelten Organismen direkt in die positive und die negative Klasse übernommen. „Super_Kingdom“ teilt sich auf in die Klassen „Archaea“ und „Bacteria“. Die Entscheidung, welche der beiden Klassen als positiv bezeichnet werden soll, ist willkürlich, hier wurde „Archaea“ als positiv gewählt. Die Reduktion der zu „Motility“ gehörenden Klas-

sen ist ebenfalls intuitiv ersichtlich. Es gibt drei Annotationen, „Yes“, „yes“ und „Motile“, für dieselbe Klasse. Diese drei Klassen werden zu einer positiven Klasse zusammengefasst. Die verbleibende Klasse „No“ modelliert das Nicht-Vorhandensein der Eigenschaft.

Die Ausprägung der Salztoleranz, die Kategorie „Salinity“, wird unter NCBI zwischen „Extreme halophilic“, „Mesophilic“, „Moderate halophilic“ und „Non-halophilic“ unterschieden. In dieser Kategorie muss die Klassenaufteilung auf Grund der Anzahl an gelabelten Organismen vorgenommen werden. Außer „Non-halophilic“ werden alle Klassen zur positiven zusammengefasst, „Non-halophilic“ verbleibt als negative Klasse.

Bei den verbleibenden drei Kategorien werden nicht alle Klassen in das Zweiklassenproblem überführt, einige Klassen werden ausgeklammert. Grund hierfür ist, dass es in diesen Klassen zu wenig Beispiele gibt oder dass sie keine biologischen Gegensätze modellieren.

Das Problem des zu kleinen Klassenumfanges stellt sich insbesondere bei der Kategorie „Temp_Range“. Der deutlichste Gegensatz der bevorzugten Temperatur ist der zwischen extrem kälteliebenden und extrem wärmeliebenden Organismen. Allerdings gibt es in diesem Datensatz zu wenig Beispiele kälteliebender Organismen. Den Extremen entsprechende Proteindomänen können mit der verwendeten Methodik nicht detektiert werden, der Einsatz von maschinellem Lernen erfordert eine ausreichende Zahl an Datenbeispielen. Somit können die Klassen „Cryophilic“ und „Psychrophilic“ nicht verwendet werden. Die verbleibenden Klassen „Hyperthermophilic“, „Thermophilic“ und „Mesophilic“ werden wie folgt zu zwei Klassen reduziert: Organismen die gemäßigte Temperaturen brauchen, die Klasse „Mesophilic“, werden als negative Klasse bezeichnet. Dem gegenüber werden die wärmeliebenden Organismen, die Klassen „Thermophilic“ sowie „Hyperthermophilic“, als positive Klasse gestellt.

Die unterschiedlichen Sauerstoffbedürfnisse der Organismen werden in der Kategorie „Oxygen_Req“ in vier Klassen zusammengefasst. Diese Klassen sind „Aerobic“, „Anaerobic“, „Facultative“ und „Microaerophilic“. Die kleinste Klasse ist „Microaerophilic“ mit 30 Beispielen, die übrigen drei Klassen enthalten zumindest annähernd sechs mal so viele Beispiele. Verschiedene Reduktionen sind denkbar, um aus den vier Klassen ein Zweiklassenproblem zu generieren. Folgende Aufteilungen werden als biologisch sinnvoll erachtet: aerobe Organismen gegen anaerobe, aerobe und mirkoaerophile gegen anaerobe sowie die Zuordnung der fakultativ anaeroben Organismen zu den aeroben einerseits und den anaeroben andererseits. Am Beispiel dieser Kategorie werden die denkbaren Möglichkeiten anhand der erzielten Performanz (im Sinne von Vorhersagegenauigkeit) evaluiert. Die jeweiligen Performanzen sind in Tabelle 3.2, auf Seite 18 aufgelistet.

Positiv	**Negativ**	**Performanz**
Aerobic, Facultative	Anaerobic	94,38
Aerobic, Microaerophilic	Anaerobic	91,34
Aerobic	Anaerobic, Facultative	68,79
Aerobic	Anaerobic	94

Tabelle 3.2: *Performanz (Angabe in %) verschiedener, biologisch sinnvoller Klassenaufteilungen der Kategorie „Oxygen_Req". Die Einteilung der Organismen in Trainings- und Testmenge wurde unter Beachtung des Genus durchgeführt.*

Die beste Performanz kann durch das Zusammenfassen von „Aerobic" und „Facultative" zu der positiven Klasse gegen „Anaerobic" als negative Klasse erzielt werden.Wird die Klasse „Facultative" hingegen mit der Klasse „Anaerobic" zu der negativen Klasse zusammengefasst und gegen die Klasse „Aerobic" klassifiziert, führt das zu der schlechtesten in dieser Kategorie erzielten Performanz. Die Unterteilung der Klassen in das Zweiklassenproblem „Aerobic" gegen „Anaerobic" führt zu der zweitbesten Performanz. Sie ist geringfügig niedriger als die der Unterteilung, in der die Klasse „Facultative " mit der Klasse „Aerobic" zusammengefasst wird. Dadurch motiviert, wird die Klasse „Facultative" in das Klassifikationsproblem integriert. Die vierte Klasse, „Microaerophilic", wird ausgeklammert, um die Klassifikationsaufgabe nicht zu überfrachten. Die weitere Analyse wird mit der Aufteilung „Aerobic" und „Facultative" als positive Klasse gegen „Anaerobic" als negative Klasse durchgeführt.

Die Kategorie Habitat ist in die fünf Klassen „Aquatic", „Host-associated", „Multiple", „Specialiced" und „Terrestrial" unterteilt. Damit lassen sich zwei Gegensätze darstellen, „Aquatic" gegen „Terrestrial" sowie „Multiple" gegen „Specialized". Dementsprechend wird diese Kategorie in zwei Kategorien aufgeteilt: „Habitat_1" enthält die Klassen „Aquatic" und „Terrestrial", „Habitat_2" die Klassen „Multiple" und „Specialized". Die Bezeichnung der einen Klasse als positiv und der anderen als negativ ist in beiden Fällen willkürlich. In der ersten Kategorie, „Habitat_1", wird „Aquatic" als positive Klasse gekennzeichnet, in der der zweiten „Specialized". Die negativen Kategorien sind „Terrestrial" in „Habitat_1" und „Muliple" in „Habitat_2". Die Klasse „Host-associated" wird in keinem der beiden Zweiklassenprobleme verwendet. Die ursprünglichen fünf Klassen schließen sich gegenseitig nicht aus, es sei denn, die Klasse „Host-associated" wird weggelassen. Organismen, die auf einem Wirt leben, können in verschiedenen Habitaten vorkommen, in terrestrischen genauso wie in aquatischen. Ebenso ist es denkbar, dass diese Lebensgemeinschaft auf ein bestimmtes Habitat angepasst ist und ausschließlich darin leben kann oder in einer Vielzahl von Habitaten existieren kann.

3.3 Evaluation der Kategorien

Die oben beschriebenen Kategorien werden zum Training einer Support Vektor Maschine verwendet. Als Maß für die Güte der Klassifikation wird das harmonische Mittel aus Sensitivität und Spezifität gebildet. Es wird über 20 Zufallspartitionen gemittelt (vgl. Methoden, Kapitel 2.1 auf Seite 9). Dieser Wert ermöglicht den Vergleich der erzielten Performanzen der Kategorien trotz der unterschiedlichen Anzahl an Datenbeispielen. In dieser Arbeit ist mit dem Begriff „Performanz“ das harmonische Mittel aus Sensitivität und Spezifität gemeint. Ein Wert nahe 100% deutet auf eine hohe Vorhersagbarkeit des Phänotypes über das Domänenprofil hin, die Anzahl falsch klassifizierter Organismen geht gegen null.

Die Performanzen der verschiedenen Kategorien unterscheiden sich deutlich voneinander. Innerhalb einer Kategorie ist die Performanz der über die Genera partitionierten Datenmenge (im Folgenden und in Tabelle 3.3 als „Genus-Korrektur“ bezeichnet) geringer als die der rein zufällig partitionierten Datenmenge. Befinden sich Organismen eines Genus zum Teil in der Trainings- und zum Teil in der Testmenge, wird die Performanz anhand der Klassifikation von Datenpunkten bewertet, die einigen Trainingspunkten sehr ähnlich sind. Die Performanz nimmt in verschiedenen Kategorien unterschiedlich stark ab. Während sie bei „Habitat_2“ deutlich abnimmt von 71,36% bei den zufälligen Partitionen auf 55,81% bei Partitionen nach Genera, bleibt die Performanz bei der Kategorie „Super_Kingdom“ durch die Genus-Korrektur unverändert bei 100%. Die Performanzen der anderen Kategorien verschlechtern sich bei Aufteilungen unter Berücksichtigung des Genus um ungefähr 10%. Ohne Berücksichtigung des Genus erreichen zwei Drittel der Kategorien eine Performanz von über 90%. Ohne Beachtung des Genus bei der Zusammenstellung der Trainings- und Testmenge wird in der Kategorie „Salinity“ mit 63,46% die niedrigste Performanz erzielt, die zweitniedrigste in der Kategorie „Habitat_2“. Die Performanz von „Temp_Range“ liegt ohne Genus-Korrektur über dem Schwellenwert von 80% (vgl. Kapitel 3.2, Seite 15), mit Genus-Korrektur nicht mehr. Alle Kategorien, die ohne Genus-Korrektur eine Performanz von über 90% erreichen, erreichen mit Genus - Korrektur den Schwellenwert von 80%. Um eine Fehleinschätzung der gelernten Diskriminanten zu vermeiden, ist die Genus-Korrektur sinnvoll. Sie liegt den folgenden Darstellungen zugrunde. Die genauen Werte sind in Tabelle 3.3 auf Seite 20 zusammengefasst.

Die Performanzen (bezogen auf die Klassifikation mit Genus- Korrektur) reichen von knapp über 50% bis 100%. Die geringste Performanz wird auf den Kategorien „Habitat_2“ und „Salinity“ mit 55,81% beziehungsweise 58,05% erreicht. Die Werte liegen nur wenig über 50% und sind damit unwesentlich besser als die Performanz einer zufälligen Zuordnung der Testdaten zu der einen oder anderen Klasse. Die Performanz der Kategorie „Temp_Range“ liegt bei 75,73%. Dieser Wert liegt

Kategorie	Ohne Genus-Korrektur	Mit Genus-Korrektur
Endospores	93,64	81,34
Gram_Stain	95	91,44
Habitat_1	91,89	88,04
Habitat_2	71,36	55,81
Motility	92,39	83,85
Oxygen_Req	97,08	94,92
Salinity	63,46	58,05
Super_Kingdom	100	100
Temp_Range	83,2	75,73

Tabelle 3.3: *Performanz der Kategorien (Angaben in %). Die Performanz ist über 20 Zufallspartitionen gemittelt und für eine Aufteilung der Datenbeispiele mit und ohne Beachtung des Genus dargestellt.*

im Grenzbereich der Existenz einer geeigneten, generalisierten Hyperebene. In der vorliegenden Arbeit wird ab einer Performanz von 80% von deren Existenz ausgegangen.

Die übrigen Kategorien haben eine höhere Performanz, sie werden mit rekursiver Merkmalseliminierung auf aussagekräftige Proteindomänen untersucht. Die Performanz (das harmonische Mittel aus Sensitivität und Spezifität) von „Endospores“ liegt mit 81,34% knapp oberhalb des Schwellenwertes. Gleiches gilt für die um wenige Prozentpunkte höhere Performanz von „Motility“, die 83,85% beträgt. Im Gegensatz dazu stehen die auf den verbleibenden Kategorien „Habitat_1“, „Gram_Stain“, „Oxygen_Req“ und „Super_Kingdom“ erzielten Performanzen. Sie liegen deutlich über 80%. Angefangen bei 88,04% der Kategorie „Habitat_1“, nähern sich die Performanzen der einer 100%ig korrekten Klassifikation. „Gram_Stain“ wird mit einer Performanz von 91,44% richtig vorhergesagt, „Oxygen_Req“ erzielt 94,92% und „Super_Kingdom“ 100%. Die Zugehörigkeit zur Domäne des Lebens wird für alle Testbeispiele korrekt vorhergesagt.

Generell lassen sich drei Tendenzen in den auf den Kategorien erzielten Performanzen erkennen. Erstens gibt es Klassifikationsprobleme, die für eine Klassifikation über ihr Domänenprofil nicht geeignet zu sein scheinen. Einige Kategorien, wie z. B. „Salinity“, haben eine äußerst geringe Performanz. Mgliche Gründe werden in Kapitel 4, Seite 35 diskutiert. Die Anzahl der Datenbeispiele pro Klasse wirkt sich auf die Güte der Klassifikation aus: Je weniger Datenbeispiele es pro Klasse gibt,

desto schwieriger wird das Aufstellen einer generalisierten Trennfunktion. Außerdem ist die Qualität der Annotation unter NCBI höchst unterschiedlich. Es gibt keine Kontrollmechanismen, ob die Annotation korrekt ist. Zweitens verschlechtert die Einteilung der Datenbeispiele in Trainings- und Testmenge gemäß des Genus der Organismen die Performanz gegenüber einer rein zufällig vorgenommenen Einteilung deutlich. Grund hierfür ist, dass die korrekte Klassifikation schwieriger wird, sobald alle Organismen eines Phylums (ähnliche Domänenprofile) in einer Menge sind. Drittens ändert sich die Reihenfolge der Performanz der verschiedenen Kategorien. Dies bedeutet, dass die Aufteilung in Trainings- und Testmenge unterschiedliche Auswirkungen auf die Klassifikation hat. Während einige Kategorien äußerst robust gegen die Genus-Korrektur sind (keine Änderung der Performanz bei „Super_Kingdom"), scheint bei anderen keine vernünftige Klassifikation möglich zu sein, sobald alle Organismen eines Genus zum Training bzw. Testen verwendet werden (die Performanz der Kategorie „Habitat_2", spezielles gegen allgemeines Habitat, bricht stark ein und sinkt in einen Bereich, der mit hoher Wahrscheinlichkeit ebenfalls durch „Raten" erzielt werden kann).

3.4 Detektion von diskriminativen Proteindomänen

Eine Auswahl der oben beschriebenen Kategorien wurde auf diskriminative Proteindomänen untersucht. Das Auswahlkriterium ist die erzielte Performanz. Ein großer Anteil an Falschklassifikationen gibt Hinweis darauf, dass die Trennung der Klassen durch die gelernte Diskriminante unzureichend ist. Andersherum deutet ein geringer Anteil falsch klassifizierter Organismen auf das Vorhandensein diskriminativer Proteindomänen hin. Proteindomänen werden im Folgenden als Merkmale bezeichnet, es sei denn, es wird Bezug auf ihre Funktion genommen.

Dadurch motiviert wird ein Schwellenwert von 80% gesetzt. Die Kategorien, auf denen nach der „ Genus-Korrektur" eine höhere Performanz erzielt wird, werden mit rekursiver Merkmalseliminierung (RFE) auf diskriminative Proteindomänen untersucht. Dabei handelt es sich um folgende Kategorien: „Endospores", „Gram_Stain", „Habitat_1", „Motility", „Oxygen_Req" und „Super_Kingdom".

Die RFE wird mit zwei Eliminierungsmethoden verwendet: Zum einen erfolgt die Eliminierung über das Gewicht der Proteindomänen in der Diskriminanten, zum anderen über eine Teststatistik über die Diskriminanten (siehe Seite 11 in Kapitel 2.2, Methoden). Bei dieser Methode handelt es sich um eine neue Variante der RFE. Die Auswahl der Proteindomänen soll anhand beider Eliminierungsmethoden erfolgen. Dazu wird die über 20 Zufallspartitionen gemittelte Performanz betrachtet. Die Eliminierungsmethode mittels Absolutbetrag liefert kein eindeutiges Maximum der Performanz, sondern vielmehr eine stark schwankende Performanz, die ab einem be-

stimmten Eliminierungsschritt deutlich einbricht. Dieser Eliminierungsschritt wird anhand der größten Differenz zweier aufeinander folgender Performanzen ermittelt. Die Merkmale, die in diesem Eliminierungsschritt eliminiert werden, sowie die, die selektiert sind, sollen näher betrachtet werden. Da in jeder Zufallspartition die Eliminierung getrennt von den anderen Partitionen erfolgt, wird die Schnitt- und Vereinigungsmenge der selektierten Merkmale über alle Partitionen pro Iterationsschritt gebildet. Mit Ausnahme der Kategorie „Super_Kingdom“ ist die Schnittmenge in dem interessanten Iterationsschritt eine leere Menge. Die Eliminierung mittels t-Teststatistik macht die Auswertung einfach, da es ein eindeutiges Maximum bzw. ein Plateau maximaler Performanz gibt. In 20 Zufallspartitionen wird auf denselben Merkmalen eine Diskriminante trainiert. Die Merkmale, die zu der maximalen Performanz beitragen, werden untersucht.

Die Menge der oben erwähnten, interessanten Merkmale liegt in der Größenordnung mehrerer hundert Merkmale pro Kategorie. Eine ausführliche Untersuchung würde den Rahmen dieser Arbeit sprengen. Aus diesem Grund wird die Menge der zu untersuchenden Merkmale stark eingeschränkt. Betrachtet werden lediglich diejenigen Merkmale, welche in einer der beiden Eliminierungsmethoden zu den wichtigsten zehn Merkmale gehören und bei der anderen Methode in der Menge der selektierten Merkmale an beliebiger Position enthalten sind. Handelt es sich dabei um mehr als zehn, so werden erneut nur die zehn wichtigsten betrachtet.

Die Wichtigkeit eines Merkmales wird gemäß seines Gewichtes in der Diskriminanten bestimmt: Merkmalen mit größerem Gewicht wird eine größere Bedeutung zugemessen als denen mit geringem Gewicht. Domänen mit negativem Gewicht werden als kontraindikativ für die positive Klasse betrachtet.

Die Eckdaten der oben beschriebenen, ausführlichen Betrachtung sind in Tabelle 3.4, Seite 23 aufgeführt. Dort ist für jede Kategorie angegeben, wieviele Merkmale in dem interessanten Eliminierungsschritt beider Methoden selektiert bzw. verworfen werden. Zusätzlich wird der Anteil Domänen unbekannter Funktion („Domains of Unknown Functions, DUF“ an den selektierten Merkmale der t-Teststatistik angegeben. Eine Auflistung der entsprechenden Pfam-IDs ist in Anhang B (ab Seite 51) hinterlegt. In Anhang C, Seite 59, sind zu jeder Kategorie die DUFs aufgelistet, die in der Merkmalsmenge des interessanten Eliminierungsschrittes der t-Teststatistik enthalten sind. Auf die Merkmale pro Kategorie, die letztendlich untersucht werden, wird in dem jeweiligen Abschnitt eingegangen.

Die Performanzen beider Methoden zeigen einen typischen Verlauf in allen Kategorien. Bei der rekursiven Merkmalseliminierung über den Absolutbetrag nimmt die Performanz beständig ab, mit geringen Schwankungen. Im Gegensatz dazu nimmt die Performanz bei der zweiten Methode über die t-Teststatistik bis wenige Iterationen vor Abbruch stetig zu. Die RFE bricht ab, sobald die Trainingsmenge nicht

Kategorie	RFE über Absolutbetrag		RFE über t-Teststatistik	
	# eliminierte Merkmale	# selektierte Merkmale	# selektierte Merkmale	Anteil DUF
Endospores	18	87	161	12%
Gram_Stain	29	112	124	9%
Habitat_1	34	88	78	8%
Motility	89	202	365	19%
Oxygen_Req	41	156	118	14%
Super_Kingdom	2	64	10	10%

Tabelle 3.4: ***Mit rekursiver Merkmalseliminierung (RFE) selektierte Merkmale:*** *Anzahl der Merkmale in dem Eliminierungsschritt mit der höchsten Aussagekraft. Dies ist bei der Eliminierung über den Absolutbetrag der Schritt, in dem die Performanz am stärksten abnimmt; bei der Eliminierung über die t-Teststatistik der Schritt mit der höchsten Performanz. Die Anzahl Merkmale bei der Eliminierungsmethode über den Absolutbetrag bezieht sich auf die Vereinigungsmenge über alle 20 Zufallspartitionen. Das Symbol # steht für die Anzahl Elemente in einer Menge.*

mehr linear separabel ist oder nur noch 10 Merkmale übrig sind.

Der Kurvenverlauf der RFE nach Absolutbetrag ist in Abbildung 3.1, Seite 24 dargestellt, der der RFE nach t-Teststatistik in Abbildung 3.2, Seite 25.

Im Folgenden wird für jede Kategorie auf die Auswahl der selektierten Merkmale eingegangen, jedoch nicht mehr auf die Performanz. Die Performanz bringt in diesem Kontext keine weiteren Informationen: Über die t-Teststatistik ermittelte Performanzen nehmen in allen Kategorien Maximalwerte nahe 100% an. Die mit der herkömmlichen RFE mit Eliminierung über den Absolutbetrag und der hier vorgestellten, neuen Methode der Eliminierung über die t-Teststatistik erzielten Performanzen sind einander in Tabelle 3.5 auf Seite 26 gegenübergestellt. Aufgelistet sind die besten Performanzen der RFE mit Absolutbetrag und t-Teststatistik und die in den interessantesten Eliminierungsschritten. In der Eliminierungsmethode über den Absolutbetrag ist dies der Schritt mit der größten Abnahme der Performanz. Bei der RFE über die t-Teststatistik ist dieser Schritt zugleich derjenige, in dem die maximale Performanz erzielt wird. Über die RFE mit t-Teststatistik wird in allen Kategorien eine deutlich bessere Performanz erreicht. Vorhersagegenauigkeiten von annähernd 100% bestätigt den Erfolg der neuen Methode, sagt jedoch nichts über die in den einzelnen Kategorien selektierten Merkmale aus. Die Performanzen der RFE mit Absolutbetrag und mit Teststatistik nehmen unterschiedlich stark in den

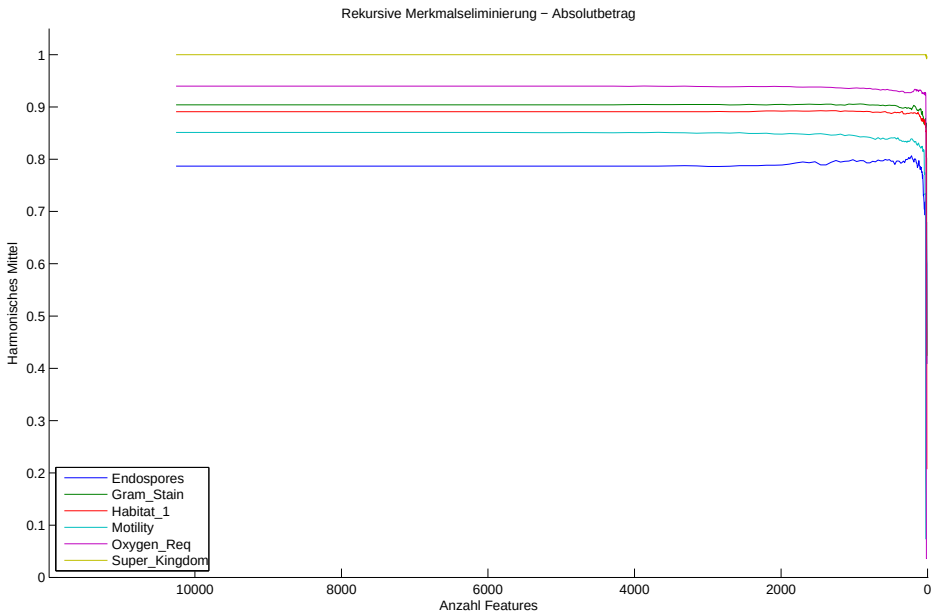

Abbildung 3.1: *Performanz in Relation zu der Anzahl der selektierten Merkmale. Eliminierung mittels RFE, Absolutbetrag*

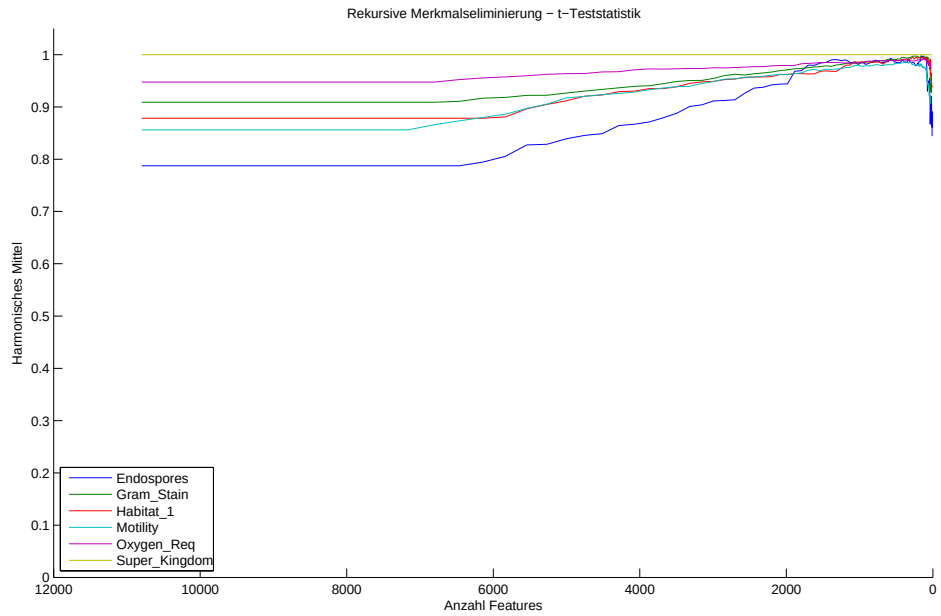

Abbildung 3.2: *Performanz in Relation zu der Anzahl der selektierten Merkmale. Eliminierung mittels RFE, t-Teststatistik*

Kategorie	RFE über Absolutbetrag [1)]	[2)]	RFE über t-Teststatistik [1)]
Endospores	59,94%	80,65%	99,8%
Gram_Stain	68,32%	90,57%	99,78%
Habitat_1	46,05%	89,28%	99,52%
Motility	44,48%	85,13%	98,61%
Oxygen_Req	58,83%	94%	99,23%
Super_Kingdom	100%	100%	100%

Tabelle 3.5: *Performanz der RFE über Absolutbetrag und t-Teststatistik in dem Eliminierungsschritt mit der höchsten Aussagekraft. Die mit [1)] gekennzeichneten Spalten beziehen sich auf die Eliminierungsschritte mit der höchsten Aussagekraft (vgl. Tabelle 3.4, Seite 23). Dies ist im Falle der RFE über die t-Teststatstik zugleich die maximale Performanz. Die maximale Performanz der RFE über den Absolutbetrag ist in der mit [2)] gekennzeichneten Spalte dargestellt.*

verschiedenen Kategorien ab. In den verschiedenen Kategorien werden von vorneherein unterschiedliche Performanzen erreicht. Es wird kein Vergleich der Merkmale zwischen den Kategorien angestrebt: Jede Kategorie repräsentiert ein eigenes Zweiklassenproblem. Damit steht es in der Betrachtungsweise dieser Arbeit für einen unabhängigen, partiellen Phänotyp. Die Auswertung der Merkmale in Bezug auf die Kategorie erfolgt in dem anschließenden Kapitel 4, das auf Seite 35 beginnt.

3.4.1 Endospores

In dieser Kategorie nimmt die Performanz in der Eliminierung über den Absolutbetrag der Gewichte in dem Eliminierungsschritt von 87 Merkmalen in der Schnittmenge über alle Zufallspartitionen zu 69 Merkmalen in der Schnittmenge am stärksten ab. Die maximale Performanz bei der Eliminierung über eine t-Teststatistik wird mit 161 selektierten Merkmalen erzielt. Von diesen 161 Merkmalen sind 12% Domänen unbekannter Funktion (DUF). Die Domänen PF07875 und PF04545 gehören in beiden Merkmalsmengen zu den zehn wichtigsten. Bei diesen beiden Domänen handelt es sich um Domänen, die zu einer Proteinfamilie der Sporenrindenproteine (PF07875) bzw. einer Untereinheit der σ-Faktoren (PF04545) gehören. Weitere sechs Domänen sind in beiden Merkmalsmengen enthalten und zählen in einer der beiden zu den zehn wichtigsten. Dabei handelt es sich um eine Proteindomäne, die die Sporen-DNA vor Abbau schützt (PF00269), eine Domäne eines Quecksilber-

Pfam-ID	Pfam-Familie	Kurzbeschreibung
PF00269	SASP	Sporenprotein
PF04545	Sigma70_r4	RNA-Polymerase Untereinheit
PF07875	Coat F	Sporenprotein
PF00376	MerR	Transkriptionsregulator
PF00082	Peptidase_S8	Proteinspaltung
PF02659	DUF204	hypothetisches Transmembranprotein
PF00296*	Bac_luciferase	Biolumineszenz
PF08534*	Redoxin	u. a. Nukleotidbiosynthese

Tabelle 3.6: *Für die Kategorie „Endospores“ detektierte Proteindomänen, Auflistung absteigend in der Reihenfolge ihres Gewichtes. Mit * gekennzeichnete Domänen haben ein negatives Gewicht, dick gedruckte Domänen gehören in beiden Eliminierungsmethoden zu den 10 wichtigsten.*

abhängigen Transkriptionsregulators (PF00376), eine Peptidasedomäne (PF00082), eine Domäne unbekannter Funktion, die ein hypothetisches Transmembranprotein ist (PF02659) sowie um zwei Domänen mit negativem Gewicht. Diese Domänen sind indikativ für die Nicht-Ausbildung von Endosporen. Es handelt sich um eine Redoxindomäne (PF08534) und eine Domäne, die zu der Familie der Luziferase-ähnlichen Monooxygenasen zählt (PF00296). Die beschriebenen Domänen sind mit Pfam-ID, Pfam-Familie und kurzer Beschreibung der Funktion in Tabelle 3.6 auf Seite 27 zusammengefasst.

3.4.2 Gram_Stain

Die Anzahl selektierter Merkmale in den entscheidenden Eliminierungsschritten beider Methoden ist in dieser Kategorie etwas größer als in der Kategorie „Endospores“. Das bedeutet, dass die maximale Performanz bzw. der größte Verlust in der Performanz in einem früheren Eliminierungsschritt erfolgt ist. In der Kategorie „Gram_Stain“ sinkt die Performanz durch die Eliminierung mittels Absolutbetrag von 29 Merkmalen aus 112 in allen Zufallspartitionen am stärksten. In der Teststatistik wird mit 124 Merkmalen die maximale Performanz erzielt. Von diesen Merkmalen sind 9% Domänen unbekannter Funktion.

Insgesamt zwölf Domänen gehören in einer der beiden Eliminierungsmethoden zu den zehn wichtigsten und sind gleichzeitig in der jeweils anderen Menge. Von diesen

Pfam-ID	Pfam-Familie	Kurzbeschreibung
PF00263*	Secretin	Sekretion Gram-negativer Bakterien
PF02798*	GST_N	Glutathiontransfer
PF03739*	YjgP/YjgQ	unbekannte Funktion
PF08443*	RimK	Translation: Modifikation von Ribosomen
PF00300	PGAM	Phosphatgruppentransfer
PF01520	Amidase_3	Zellwandsynthese Gram-positiver Bakterien
PF01316	Arg_repressor	Repressor der Argininsynthese
PF00529*	HlyD	Sekretion Gram-negativer Bakterien
PF01380*	SIS	Phosphatzuckerbindung
PF08282	Hydrolase_3	Zellwandsynthese Gram-negativer Bakterien

Tabelle 3.7: *Detektierte Proteindomänen der Kategorie „Gram_Stain“. Sie sind in der Reihenfolge ihres Gewichtes aufgelistet. Mit * gekennzeichnete Domänen haben ein negatives Gewicht.*

zwölf Domänen sind die zehn mit dem höchsten Absolutgewicht in Tabelle 3.7 auf Seite 28 aufgeführt. Darunter sind vier Domänen mit positivem Gewicht. Es handelt sich um Domänen von Proteinen, die Phosphatgruppen innerhalb eines Zuckermoleküles transferieren (PF00300), am Zellwandaufbau Gram-positiver Bakterien beteiligt sind (PF01520), des Argininstoffwechselweg-Repressors (PF01316) und an der Bildung der Lipopolysaccharidschicht Gram-negativer Bakterien (PF08282) beteiligt sind.

Die übrigen sechs Proteindomänen haben ein negatives Gewicht und sind somit indikativ für Gram-negative Bakterien. Zwei Domänen werden mit dem Sekretionsstoffwechselweg Gram-negativer Bakterien assoziiert: Secretin (PF00263) und HylD (PF00529). Die anderen Domänen gehören zu Proteinen, die schwefelhaltige, funktionelle Gruppen (speziell Glutathion) auf andere Moleküle übertragen (PF02798) und einem vermutlich membranintegrierten Protein unbekannter Funktion (PF03739). PF08443 ist die ATP-bindende Domäne eines Proteins, dass ribosomale Modifikationen vornimmt, PF01380 eine phosphatzuckerbindende Domäne.

3.4.3 Habitat_1

In dieser Kategorie sind mittels Eliminierung über den Absolutbetrag der Gewichte vor der größten Abnahme der Performanz 88 Merkmale selektiert, in diesem Eliminierungsschritt wurden 34 Merkmale in allen Zufallspartitionen eliminiert. Über die Teststatistik werden 78 Merkmale selektiert. Von diesen 78 sind 8% Domänen unbekannter Funktion.

Sechs Domänen sind in der Schnittmenge der Domänen, die in einer der beiden genannten Mengen zu den zehn wichtigsten zählen. Von diesen, in Tabelle 3.8 (Seite 29) aufgelisteten Domänen gehört eine zu einer Familie chorismatbindender Proteine. Bei den anderen handelt es sich um eine Domäne, die den Transfer der Aminogruppe von Glutamin zu einem anderem Molekül katalysiert (PF00310), eine vermutlich DNA-bindende Domäne der Transposase (PF07282) sowie um eine Domäne, die in einer Vielzahl an Bakterien vorkommt (PF00990). Ihre Funktion ist der Adenylcyclase Katalase homolog. Die übrigen beiden Domänen haben negative Gewichte. Eine Domäne gehört zu einer Familie bakterieller Stressproteine (PF02342), die andere zu einem polysaccharidabbauenden Protein (PF01522).

Pfam-ID	**Pfam-Familie**	**Kurzbeschreibung**
PF00425	Chorismate_bind	Chorismatbindung
PF01522*	Polysacc_deac_1	Polysaccharidabbau
PF00310	GATase_2	Aminogruppentransfer
PF02342*	TerD	Stressprotein
PF07282	Transposase_35	Transposase
PF00990	GGDEF	ATP Spaltung zu cAMP

Tabelle 3.8: *Proteindomänen, die in der Kategorie „Habitat_1“ ermittelt wurden. Die Reihenfolge entspricht der Gewichtung, die Domäne mit dem höchsten Gewicht steht in der ersten Zeile. Mit * gekennzeichnete Domänen haben ein negatives Gewicht.*

3.4.4 Motility

In dieser Kategorie wird, im Vergleich zu den anderen Kategorien, in einer relativ frühen Iteration die maximale Performanz bzw. maximale Abnahme der Performanz erzielt. Aus diesem Grund sind die Mengen der selektierten Merkmale in beiden Eliminierungsmethoden am größten. Über den Absolutbetrag der Gewichte sind

Pfam-ID	Pfam-Familie	Kurzbeschreibung
PF01443*	Viral_helicase1	Basenpaarspaltung, virales Protein
PF00460	Flg_bb_rod	bakterielle Flagellen
PF00565	SNase	Nukleinsäureabbau
PF02503*	PP_kinase	Polyphosphattransferase
PF04964	Flp_Fap	bakterielle Pili
PF01610*	Transposase_12	Transposase
PF01555	N6_N4_Mtase	DNA-Methylierung
PF02879*	PGM_PMM_II	Phosphatzuckertransfer, innerhalb eines Moleküles
PF04545	Sigma70_r4	RNA-Polymerase Untereinheit
PF01839	FG-GAP	Integrinfaltung

Tabelle 3.9: *In der Kategorie „Motility“ selektierte, in der Reihenfolge ihres Gewichtes sortierte Proteindomänen mit kurzer Funktionsbeschreibung. Mit * gekennzeichnete Domänen haben negatives Gewicht.*

in dem Eliminierungsschritt, in dem die Performanz am stärksten abnimmt, 202 Merkmale selektiert, es werden 89 verworfen. Über die t-Teststatistik wird mit 365 Merkmale die maximale Performanz erreicht. Mit 19% ist der Anteil an Domänen unbekannter Funktion in der Merkmalsmenge der Teststatistik am größten.

Die zehn Merkmale mit dem größten Gewicht, die in beiden Mengen vorkommen, sind in Tabelle 3.9, Seite 30 verzeichnet. Das größte Gewicht hat ein Merkmal, das kontraindikativ für Zellbeweglichkeit ist. Es handelt sich um eine virale Helikase-Domäne (PF01443). Unter den zehn analysierten Domänen sind drei weitere mit negativem Gewicht. Dies sind Domänen der Polyphosphatkinase (PF02503), der Phosphogluko- bzw. Phosphomannomutase (PF02879) und eine Transposasedomäne (PF01610). Zwei der Proteindomänen mit positivem Gewicht bilden Teilstrukturen von Flagellen (PF00460) und Fibrillen (PF04964) aus. Beide Strukturen dienen der prokaryotischen Bewegung. Von den übrigen Proteindomänen, die indikativ für Zellbeweglichkeit sind, stehen zwei Domänen in Zusammenhang mit der DNA-Replikation. Diese beiden Domänen sind Untereinheiten der DNA-Methylase (PF01555) und der RNA-Polymerase (PF04545). Des weiteren haben eine Domäne der SNase (PF00565), die Nukleinsäuren umsetzt, und eine Domäne, die an einer speziellen Faltung (β-Propeller) von bestimmten Proteinen, wie Integrin [35], beteiligt ist (PF01839), hohes Gewicht.

3.4.5 Oxygen_Requirement

Die Performanz nimmt in der Kategorie Oxygen_Req am stärksten durch die Eliminierung von 41 aus insgesamt 156 Merkmalen ab. In der anderen Eliminierungsmethode über die Teststatistik wird mit 118 Merkmalen die maximale Performanz erzielt. Bei 14% dieser 118 Merkmale handelt es sich um DUFs.

Pfam-ID	**Pfam-Familie**	**Kurzbeschreibung**
PF02803	Thiolase_C	Fettsäureabbau
PF04095*	NAPRTase	NAD-Synthese
PF03595	C4dic_mal_tran	C_4-Zucker-Transport
PF05656*	DUF805	unbekannte Funktion
PF08666*	SAF domain	Flagellen, Pili
PF00037*	Fer4	Eisen-Schwefel-Cluster-Bindung
PF02837*	Glyco_hydro_2_N	Spaltung glykolytischer Bindungen
PF10604*	Polyketide_cyc2	Polyketidsynthese
PF04055*	Radical_SAM	anaerobe Oxidation
PF00529	HylD	Sekretion Gram-negativer Bakterien

Tabelle 3.10: *Detektierte Proteindomänen der Kategorie „Oxygen_Req". Sie sind in der Reihenfolge ihres Gewichtes aufgelistet. Mit * gekennzeichnete Domänen haben ein negatives Gewicht, die dick gedruckte Domäne gehört in beiden Eliminierungsmethoden zu den 10 wichtigsten.*

Eine Domäne unbekannter Funktion (PF05656) ist unter den zehn Domänen mit dem höchsten Gewicht. In beiden Eliminierungsmethoden zählt die Proteindomäne PF02803 zu den zehn wichtigsten. Es handelt sich um eine Domäne der Thiolase. Dieses Enzym ist am Fettsäureabbau beteiligt. Weitere Proteindomänen, die ein hohes Gewicht haben und indikativ für aerobe beziehungsweise fakultativ anaerobe Organismen sind, sind eine C_4-Dicarboxylat-Transporterdomäne (PF03595) und eine Sektretionsdomäne Gram-negativer Bakterien, HylD (PF00529). Die übrigen sieben Domänen haben ein negatives Gewicht. Die Domäne mit dem zweithöchsten Gewicht limitiert die NAD-Synthese (PF04095). Eine andere Proteindomäne bindet Eisen-Schwefel-Cluster (PF00037). PF02837 ist eine Domäne, die glykosidische Bindungen hydrolisiert. Eine weitere Domäne kommt unter anderem in Flagellen- und Piliusproteinen vor (PF08666). Eine Proteindomäne ist an der Polyketidsynthese (PF10604) beteiligt, eine andere gehört zur START-Superfamilie (PF04055), einer

Proteinfamilie, die unter anderem anaerobe Oxidationen katalysiert.

Die oben genannten Proteindomänen sind in Tabelle 3.10 auf Seite 31 zusammengefasst.

3.4.6 Super_Kingdom

Die rekursive Merkmalseliminierung (RFE) über den Absolutbetrag wird in der Kategorie „Super_Kingdom“ in einem späten Eliminierungsschritt abgebrochen. Diese Methode resultiert in einem Merkmal, das in allen Zufallspartitionen selektiert ist und 64 Merkmalen, die in mindestens einer Partition selekiert sind. Es gibt keine Merkmale, die über beide Methoden zugleich selektiert werden. Die RFE über eine Teststatistik resultiert in einer Menge von zehn selektierten Merkmalen. Darunter ist eine Domäne unbekannter Funktion, der Anteil DUFs liegt damit bei 10%. In dieser Kategorien beträgt die Performanz in jedem Eliminierungsschritt 100%. Die RFE wird abgebrochen, weil die Trainingsmenge nicht mehr linear separabel ist, nicht, weil die maximale Anzahl Iterationen überschritten ist. In dieser Kategorie wird deutlich, dass sich das zweite Abbruchkriterium, eine Mindestmenge selektierter Merkmale, nicht negativ auf die mit der RFE erzielten Ergebnisse auswirkt.

Da über die Teststatistik stabile Performanzen erzielt werden, werden in dieser Kategorie die zehn Merkmale analysiert, die über die Teststatistik ermittelt wurden. Zusätzlich wird das Merkmal betrachtet, dass in der Eliminierung über den Absolutbetrag der Gewichte in allen Zufallspartitionen selektiert ist. Dieses Merkmal, PF04542, ist eine Untereinheit des Sigmafaktors σ^{70}. Die σ-Untereinheit kommt ausschließlich in bakteriellen RNA-Polymerasen vor. Bei diesem speziellen Sigmafaktor handelt es sich um einen generellen Sigmafaktor, der auch als housekeeping-Sigmafaktor bezeichnet wird [12]. Er hat ein negatives Gewicht und ist damit kontraindikativ für die positive Klasse der Archaeen.

Die über die Teststatistik selektierten Merkmale haben positives Gewicht. Alle Merkmale sind an der DNA-Replikation von Archaeen beteiligt. Die Hälfte der Merkmale sind ribosomale Untereinheiten (PF01020, PF00935, PF01015, PF01157, PF01655). Bei einem anderen Merkmal, (PF01194), handelt es sich um eine Untereinheit der RNA-Polymerase von Archaeen. Eine Domäne ist ein Translationsinitiationsfaktor (PF01287). Zwei weitere Domänen sind an der Translation beteiligt. Diese Proteindomänen sind PF02778 und PF09249. Außerdem wird ein Archaeenprotein unbekannter Funktion vorhergesagt (PF04895).

Die beschriebenen Merkmale sind in Tabelle 3.11 auf Seite 33 aufgelistet.

Im folgenden Kapitel (Kapitel 4) werden die hier vorgestellten Ergebnisse mit biologischem Hintergrund diskutiert.

Pfam-ID	**Pfam-Familie**	**Kurzbeschreibung**
PF02778	tRNA_int_endo_N	Translation
PF04895	DUF651	Archaeenprotein unbekannter Funktion
PF01287	elF-5a	Ribosomale Untereinheit
PF01020	Ribosomal_L40e	Ribosomale Untereinheit
PF00935	Ribosomal_L44	Ribosomale Untereinheit
PF01015	Ribosomal_S3Ae	Ribosomale Untereinheit
PF01157	Ribosomal_L21e	Ribosomale Untereinheit
PF01194	RNA_pol_N	RNA-Polymerase Untereinheit, Archaeen
PF01655	Ribosomal_L32e	Ribosomale Untereinheit
PF09249	tRNA_NucTransf2	Translation
PF04542*	Sigma70_r2	RNA-Polymerase Untereinheit, Bakterien

Tabelle 3.11: *Proteindomänen, die in der Kategorie „Super_Kingdom“ über eine t-Teststatistik detektiert wurden. Die mit * gekennzeichnete Domäne hat negatives Gewicht. Diese Domäne ist über den Absolutbetrag der Gewichte selektiert worden. Die Funktion der Domänen ist kurz beschrieben.*

Kapitel 4

Diskussion

In dem vorangegangen Kapitel wurden Proteindomänen vorgestellt, die mit einem kombinierten Ansatz aus Support Vektor Maschinen und rekursiver Merkmalseliminierung als indikativ für bestimmte phänotypische Eigenschaften detektiert wurden. Des weiteren wurde beschrieben, wie aus den phänotypischen Eigenschaften Zweiklassenprobleme konstruiert werden. Noch nicht eingegangen wurde jedoch darauf, was die untersuchten Phänotypen kennzeichnet. In diesem Kaptiel werden zunächst die Phänotypen beschrieben. Davon ausgehend kann fomuliert werden, inwieweit eine bestimmte Proteinausstattung „erwartet“ wird und in welchem Umfang die detektierte Proteinausstattung diese Erwartung erfüllt.

Anhand von Organismen, die diese phänotypischen Eigenschaften haben, werden Proteindomänen detektiert. Diese Domänen lassen Rückschlüsse auf für die Organismen wichtige Proteine zu. Daraus ableitbar sind vorherrschende Stoffwechselwege, sowie, in gewissem Maße, Phänotypen. Einige Phänotypen können besser durch Proteindomänen charakterisiert werden als andere, was sich in der Performanz (dem harmonischen Mittel aus Sensitiviät und Spezifität, vgl. Kapitel 2.1 (Methoden), Seite 9) widerspiegelt. Die Vermutung, dass in diesem Fall die Ausprägung des Phänotyes nicht ausreichend stark über Proteindomänen modelliert werden kann, liegt nahe. Basis für die Analyse ist folgender Zusammenhang: Proteindomänen, die kennzeichnend für die Ausprägung beziehungsweise Nicht-Ausprägung eines Phänotyes sind, haben in entscheidenden Eliminierungsschritten hohes Gewicht. Somit wirkt sich die Beschränkung auf einige wenige Proteindomänen, in der vorliegenden Arbeit die zehn Proteindomänen mit dem höchsten Gewicht, nicht negativ auf die weitere Analyse aus. Im Gegenteil wird erst durch das Einschränken der zu betrachtenden Proteindomänen deren genauere Analyse möglich. Diese Proteindomänen werden durch die Schnittmenge der wichtigsten Proteindomänen zweier Elimierungsmethoden (siehe Methoden 2.2, Seite 11) bestimmt. Die Tatsache, dass die ausgewählten

Proteindomänen in beiden Methoden der Merkmalseliminierung hohes Gewicht haben, unterstreicht die Bedeutung dieser Domänen.

Eine Variante wäre die Fokussierung auf die Eliminierung mittels t-Teststatistik. Diese Eliminierungsmethode ist in der Performanz der ursprünglichen Methode, die ausschließlich auf dem Absolutbetrag der Gewichte basiert, überlegen. Mehr als 99% der Testbeispiele werden korrekt klassifiziert. Ein weiterer Vorteil ist, dass diese Methode robuster gegenüber Schwankungen der Gewichte zwischen den Zufallspartitionen ist. Grund hierfür ist, dass in dieser Methode kleinere Absolutbeträge, die zwischen den Partitionen kaum schwanken, gegenüber hohen, stark schwankenden Gewichten bevorzugt werden.

Der „interessanteste“ Eliminierungsschritt lässt sich eindeutig bestimmen. Unter der Annahme, dass die maximale Performanz mit der Eliminierung aller unwichtigen Proteindomänen korreliert, wird diese Menge betrachtet. „Wichtig“ bedeutet in dem Fall, dass die Proteindomänen die größte diskriminative Eigenschaft haben. Die in der Originalarbeit zur RFE vorgestellte Eliminierung über den Absolutbetrag der Gewichte erzielt auf den vorliegenden Daten keine globalen Maxima der Performanzen. Das deutet zusammen mit der leeren Schnittmenge der über 20 Zufallspartitionen selektierten Merkmale darauf hin, dass das Gewicht der Proteindomänen zwischen den Partitionen stark schwankt. Damit ändert sich die Reihenfolge der Eliminierung in vergleichbaren Iterationsschritten der verschiedenen Partitionen, was die leeren Mengen erklärt.

4.1 Endosporen

Endosporen werden von Gram-positiven Bakterien des Phylums *Firmicutes* gebildet. Sie stellen eine nicht reproduktive Überdauerungsform dar, die durch umweltbedingten Stress induziert wird. Die Resistenz gegen physikalische und chemische Einflüsse wird durch eine mehrschichtige Sporenhülle bedingt. Sie besteht aus einer dicken Peptidoglukanschicht, der Sporenrinde, und der darauf liegenden Sporenhülle. Die Sporenhülle ist eine Multiproteinschicht. Ihre Zusammensetzung ist, wie die der übrigen Schichten, speziesspezifisch und variiert stark zwischen den Spezies [15]. Sie enthält typischerweise das Protein Calciumdipicolnat in hohen Konzentrationen. Dieses Protein macht die Hitzestabilität der Spore aus. Der Sporenkern enthält Cytoplasma und einige Enzyme, er ist jedoch nicht metabolisch aktiv. Die DNA ist an das Sporenprotein SASP gebunden. Dieses Protein bewirkt eine Konformationsänderung der DNA und schützt sie dadurch vor chemischem und enzymatischem Abbau. Dieser Schutzmechanismus ist ein Grund für die hohe Resistenz der Sporen gegenüber UV-Strahlung.

Die Sporulation ist eine Phase hoher Genaktivität. Sie beginnt mit asymmetrischer Zellteilung und einer Replikation des kompletten Genomes. Die Genaktivität wird von vier σ-Faktoren, Untereinheiten der RNA-Polymerase, kontrolliert [36]. Unter anderem werden Mantelproteine, Stoffwechselproteine und Transportproteine [8] von σ-Faktoren kontrolliert.

Viele detektierte Proteindomänen können der Synthese von sporentypischen Strukturen wie dem Mantel oder der Regulation der Synthese zugeordnet werden. Zwei Domänen haben in beiden Methoden der Merkmalseliminierung ein hohes, positives Gewicht. Bei diesen Domänen handelt es sich um die σ_4-Domäne und eine des Sporenhüllenproteins Coat F.

Die σ_4-Domäne ist eine stark konservierte Struktur innerhalb der σ-Faktoren. Eine weitere detektierte Domäne wird mit einem prokaryotischen Transkriptionsregulator assoziiert. Dabei handelt es sich um MerR, eine mit Quecksilber assoziierte Transkriptionsdomäne. Damit sind zwei der acht betrachteten Proteindomänen mit dem höchsten Gewicht mit der sporentypischen Transkriptionsregulation assoziiert. Ebenfalls in deutlichem Zusammenhang mit der Sporulation steht die Coat F-Domäne, die an der Synthese der Sporenhülle beteiligt ist.

Bei den weiteren für die Endosporenbildung indikativen Domänen handelt es sich um eine SASP-Domäne, eine Untereinheit der Peptidase und eine Domäne unbekannter Funktion. SASP ist ein für Endosporen spezifisches Protein, an das die Sporen-DNA gebunden ist. Außerdem wird eine Peptidaseuntereinheit detektiert. Sie gehört zu der Familie der Serinproteasen, die die Spaltung von Peptidbindungen katalysieren und damit Proteine abbauen [16]. Hier ist der Zusammenhang zur Sporulation weniger offensichtlich erkennbar. Möglicherweise ist während der Sporulation und der damit verbundenen Neusynthese sporentypischer Proteine [9] der verstärkte Abbau anderer Proteine von Vorteil für den Organismus. Die Domäne unbekannter Funktion gehört zu einer Familie hypothetischer Transmembranproteine. Gerade bei dieser Domäne können nur Vermutungen über den Zusammenhang zur Sporulation angestellt werden. Es ist zu vermuten, dass während der Sporulation der Transport durch Membranen eine besondere Rolle spielt [9]. Genauso gut kann es sich bei diesem Protein um ein in der Membran verankertes Protein gänzlich anderer Funktion handeln.

Die beiden Domänen negativen Gewichtes sind weitaus schwieriger einzuordnen. Es handelt sich um eine zur Redoxinfamilie gehörende Proteindomäne und um eine Monooxygenasedomäne. Proteine der Redoxinfamilie dienen unter anderem der Nukleotidbiosynthese, aber auch der Regulation des Calvinzyklus (Thioredoxin) [22]. Die Monooxygenase ist der bakteriellen Luziferase ähnlich. Durch die Oxidation langkettiger Alkohole freiwerdende Energie wird in Form von Licht frei. Die Monooxygenase kommt bei Gram-negativen *Firmicutes* des verwendeten Datensatzes

nicht vor. Inwieweit daraus auf eine allgemeine Tendenz geschlossen werden kann, bleibt offen. Darüber hinaus ist bei beiden Proteinen kein Zusammenhang zu der Nicht-Ausbildung von Endosporen erkennbar.

Die detektierten, für die Sporulation indikativen Proteindomänen entsprechen den Erwartungen an Proteindomänen, die für die Sporenbildung spezifisch sind.

4.2 Gramfärbung

Bakterien habe eine starre, formgebende Zellwand. Ein wichtiger Bestandteil dieser Zellwand ist Peptidoglykan. Über Aufbau und Zusammensetzung der Zellwand lassen sich Bakterien unterscheiden: In einer differenzierenden Färbemethode der Gram- Färbung, die auf den dänischen Arzt Hans Christian Gram zurückgeht, nehmen sie je nach Aufbau der Zellwand eine von zwei Farben an. Anhand dessen werden Bakterien in Gram-positive und Gram-negative Bakterien eingeteilt.

Die Zellwand Gram-negativer Bakterien besteht aus zwei Schichten: dem Peptidoglykan und einer äußeren Membran. Peptidoglykan ist eine aus Zuckern aufgebaute, einfache Zellwandschicht. Die Zuckermoleküle sind über vier Aminosäuren miteinander verknüpft. Es handelt sich bei den Zuckermolekülen unter anderem um N-Acetylglucosamin und N-Acetylmuraminsäure, andere Zellwandbestandteile variieren von Spezies zu Spezies [4]. Die Peptidoglykanschicht ist in den periplasmatischen Raum zwischen Cytoplasmamembran und äußerer Membran eingebettet. Das Periplasma ist über Transportproteine für kleine wasserlösliche Proteine, sogenannte Porine, mit der äußeren Membran verbunden. Die äußere Membran besteht aus Phospholipiden und darauf aufgelagerten Lipopolysacchariden. Die Lipopolysaccharide bestehen aus einem lipophilen, nach innen gerichteten Teil und einem hydrophoben Teil. Der lipophile Teil hat antigenischen Charakter und toxische Eigenschaften, die zu seiner Bezeichnung als Endotoxin führen.

Synthese und Stoffwechselwege zum Aufbau der Gram-positiven Zellwand unterscheiden sich von denen der Gram-negativen Zellwand. Die Peptidoglykanschicht Gram-positiver Bakterien ist um ein vielfaches dicker als die Gram-negativer Bakterien. Sie sind aus ähnlichen Zuckern aufgebaut, und auch die Verknüpfung der Zucker erfolgt über eine Aminosäurekette. Deren Zusammensetzung ist jedoch bei Gram-positiven Bakterien eine andere als bei Gram-negativen. Darüber hinaus sind in der Peptidoglykanschicht anionische Polymere eingelagert. Diese, sowie zellwandassoziierte Proteine ragen zum Teil aus der Peptidoglykanschicht heraus. [4]

Auf die Zellwand vieler Prokaryoten aufgelagert sind Oberflächenstrukturen. Dabei handelt es sich um Proteinstrukuren unbekannter Funktion, die jedoch bekanntermaßen zu der antigenischen Variabilität einiger Bakterien beitragen [25].

Einige der detektierten Proteindomänen können direkt Zellwandstrukturen Gram-positiver bzw. Gram-negativer Bakterien zugeordnet werden. Darunter sind zwei Proteindomänen, die an der Sekretion von Proteinen beteiligt sind: Secretin und HlyD. Die beiden entsprechenden Proteine sind spezifisch für Gram-negative Bakterien. Sie werden folgerichtig negativ, also kontraindikativ für Gram-positive Bakterien, gewichtet. Zwei weitere, detektierte Proteindomänen sind an der Zellwandsynthese beteiligt. Hierzu gehört Hydrolase_3. Diese Proteindomäne ist an der Synthese der Lipopolysaccharidschicht Gram-negativer Bakterien beteiligt. Dennoch ist diese Proteindomäne positiv gewichtet. Ebenfalls positives Gewicht hat die Domäne Amidase_3. Die Amidase ist an der Synthese der N-Acetylmuraminsäure beteiligt. Obwohl diese Struktur sowohl in Gram-positiven als auch in Gram-negativen Bakterien vorkommt, ist sie vor allem für Gram-positive Bakterien spezifisch [25]. Insofern ist es nachvollziehbar, dass sie in dem Klassifikationsproblem für die positive Klasse, Gram-positive Bakterien, indikativ ist. Grund hierfür kann die dickere Peptidoglykanschicht Gram-positiver Bakterien sein.

Bei zwei anderen Domänen handelt es sich um Proteindomänen, die am Umbau von Zuckermolekülen beteiligt sind. Eine der Domänen, PGAM, ist indikativ für Gram-positive Bakterien. PGAM transferiert Phosphatgruppen innerhalb eines Zuckermoleküls. Mit negativem Gewicht wird eine Phosphatzucker-bindende Domäne (SIS) vorhergesagt. Man kann davon ausgehen, dass diese Domänen am Aufbau der Zellwand, insbesondere der Peptidoglykanschicht beteiligt sind.

Die übrigen Proteindomänen sind schwerer Gram-positiven bzw. Gram-negativen Bakterien zuzuordnen. Drei der vier Domänen werden mit sehr hohem Gewicht für die negative Klasse vorhergesagt. Darunter ist beispielsweise eine Domäne, die ein ribosomales Protein modifiziert (RimK). Diese Domänen haben gemeinsam, dass kein Zusammenhang zu dem Zellwandaufbau festgestellt werden konnte.

Betrachtet man die analysierten Merkmale als Ganzes, so fällt folgendes auf: Es wird zwar anhand von einigen, mit der Zellwand in Verbindung stehenden Domänen klassifiziert, jedoch nicht aussschließlich. Gemessen an der Zahl der Proteindomänen, die eindeutig an der Zellwandsynthese Gram-positiver oder Gram-negativer Bakterien beteiligt sind, ist es verwunderlich, dass unter den Domänen mit dem höchsten Gewicht einige sind, die nicht spezifisch für Gram-positive oder negative Bakterien zu sein scheinen.

4.3 Habitat

Eine weitere, in der vorliegenden Arbeit untersuchte Eigenschaft von Prokaryoten ist das Vorkommen in aquatischen und terrestrischen Habitaten. Terrestrische Lebens-

räume können als lebensfeindlicher für Prokaryoten betrachtet werden: Verglichen mit aquatischen Habitaten unterliegen terrestrische stärkeren Veränderungen. Lebewesen, die in diesem Habitat vorkommen, sind einer Reihe von umweltbedingten Stressoren ausgesetzt, wie beispielsweise der Gefahr plötzlicher Austrocknung [31]. Eine der beiden Domänen, die indikativ für terrestrische Habitate sind, (TerD), gehört zu einer Familie bakterieller Stresshormone.

In dieser Kategorie werden Proteindomänen stark gewichtet, die Prokaryoten auf sehr allgemeine Art charakterisieren. Unter den Domänen mit positivem Gewicht, die indikativ für das Besetzen von aquatischen Habitaten sind, ist die GGDEF-Domäne. Sie ist der Adenylcyclase Katalase-Domäne homolog und kommt in einer Vielzahl an Bakterien vor [29]. Diese Domäne kann jedoch nicht ausschließlich aquatischen Habitaten zugeordnet werden. Gleiches gilt für die anderen Domänen mit positivem Gewicht, wie z. B. der GATase_2. Diese Domäne transferiert die Aminogruppe von Glutamin zu anderen Molekülen. Die detektierten Domänen scheinen indikativ für Prokaryoten im Allgemeinen zu sein. Zum Teil sind die Proteine jedoch nicht spezifisch für Prokaryoten. Ein Beispiel für ein solches Protein ist die Transposase. Transposasen schneiden die für sich kodierenden DNA-Bereiche mit umliegenden Bereichen aus dem DNA-Strang aus und integrieren sie an anderer Stelle. Es wird vermutet, dass diese Proteine über größere Veränderungen im Genom zu evolutionären Ereignissen wie Artaufspaltungen geführt haben [24, 28].

Die in dieser Kategorie detektierten Proteindomänen lassen sich nicht vollständig den zugrunde liegenden Klassen zuordnen. Vor allem die positiv gewichteten Domänen sind weniger indikativ für aquatische Habitate sondern vielmehr für das Vorkommen von Prokaryoten an sich (in einem beliebigen Habitat).

4.4 Beweglichkeit

Eine weit verbreitete Eigenschaft unter Prokaryoten ist die Beweglichkeit. Wie die Fortbewegung erreicht wird, unterscheidet sich stark zwischen Bakterien und Archaeen, aber auch innerhalb der Domänen des Lebens. Die am weitesten verbreitete Bewegungsform sind Flagellen. Weniger verbreitete Formen der Bewegung sind Fibrillen (Pili) und Gleiten mit Hilfe eines Schleimfilmes.

Flagellen sind äußerst komplexe Strukturen, deren Synthese eine Vielzahl von Proteinen erfordert. Bakterielle Flagellen bestehen aus Basalkörper, Haken und Filament. Das aus dem Protein Flagellin aufgebaute Filament führt propellerartige Drehungen aus und erzeugt damit eine Bewegung. Es ist über Basalkörper und Haken mit der Zellwand verbunden. Aufgebaut sind Flagellen aus mehr als 20 verschiedenen Proteinen, die Zusammensetzung ist von Spezies zu Spezies unterschiedlich

[2]. Ungefähr 30 weitere Proteine sind für die Regulation und den Zusammenbau der aufbauenden Proteine erforderlich, z. B. für die Sekretion der Proteine zum richtigen Zeitpunkt.

Die Flagellen der Archaeen sind anders aufgebaut als bakterielle. Wichtige Unterschiede liegen in der Zusammensetzung des Filamentes und der Verankerung in der Membran. Archaerielles Flagellin ist einem bestimmten bakteriellen Fibrillentyp (Typ IV Pili) ähnlich. Eine dem bakteriellen Basalkörper ähnliche Struktur gibt es in Flagellen von Archaeen nicht.

Die zu den Archaeen gehörenden Spirochaeten bewegen sich typischerweise mit Axialfibrillen. Fibrillen sind aus anderen Proteinen zusammengesetzt als Flagellen. Ein Filament erzeugt ebenfalls eine drehende Bewegung, ist jedoch anders zusammengesetzt als das bakterielle Filament. Auch die Verankerung in der Zellwand erfolgt über die gleiche Struktur, die jedoch wiederum aus anderen Proteinen zusammengesetzt ist. Beide Fortbewegungsformen werden über Protonengradienten angetrieben.

Von den detektierten Merkmale sind zwei Proteindomänen strukturgebend für prokaryotische Fortbewegungsformen. Die eine Proteindomäne, Flg_bb_rod, ist Teil des Basalkörpers von bakteriellen Flagellen, die andere, Flp_Fap, eine Komponente der Pili. Die übrigen gehören zu den Proteinen, die an der Regulation und dem Aufbau der Fortbewegungsorganellen beteiligt sind. Darunter sind die σ_4-Domäne, eine Domäne einer Untereinheit der bakteriellen RNA-Polymerase. Diese Domäne wurde ebenfalls für die phänotypische Eigenschaft „Endosporen“ detektiert. Der σ-Faktor hat ein großes Regulon, zu dem unter anderem viele Gene der Bildung von Endosporen zählen. Somit wird ein Regulator der Sporulation in der Kategorie Bewegung hoch gewichtet. Paredes hat am Beispiel von *Bacillus subtilis* gezeigt, dass beide Prozesse eng miteinander verbunden sind: sie sind entgegengesetzt reguliert [27].

Ebenfalls in Zusammenhang mit der Bewegung von Zellen steht Integrin. Integrin gehört zu einer Familie von Zelladhäsionsproteinen. Sie sind an der Weiterleitung von intrazellulären Signalen zur Steuerung von Zellbewegungen beteiligt. Eine Proteindomäne, die an der Faltung von Integrin beteiligt ist, ist unter den detektierten Proteindomänen.

Für zwei negativ gewichtete Domänen konnte kein Zusammenhang zur Kontraindikativität für prokaryotische Beweglichkeit hergestellt werden. Bei diesen Domänen handelt es sich um Domänen der Polyphosphatkinase und der Phosphoglucomutase.

Die am höchsten gewichtete Domäne ist ein Domäne der viralen Helikase. Es ist verwunderlich, dass diese Domäne am höchsten gewichtet wird. Virale Domänen sollten nicht in so vielen prokaryotischen Organismen vorhergesagt werden, dass anhand dieser Domänen klassifiziert wird.

In dieser Kategorie wird in einem sehr frühen Eliminierungsschritt, also mit vielen Merkmalen, die höchste Performanz erzielt. Eine Erklärung hierfür liefert die Tatsache, dass es sich bei Fortbewegungsorganellen um sehr komplexe Strukturen handelt, an deren Synthese viele Proteine beteiligt sind. Aus diesem Grund ist die hohe Anzahl an Merkmalen nachvollziehbar.

4.5 Sauerstoffbedarf

Die Klasseneinteilung der Kategorie Oxygen_Req (Sauerstoffbedarf) wurde nicht ausschließlich anhand biologischer Gesichtspunkte entschieden. Es gibt einige Kombinationen aus den Klassen, die biologisch sinnvoll sind. Das Weglassen der beiden Klassen „Microaerophilic“ und „Facultative“ sorgt ohne Zweifel für ein Zweiklassenproblem, dass biologisch nicht anfechtbar ist. Jedoch steht zu vermuten, dass die auf der Unterteilung „Aerobic“ – „Anaerobic“ erzielten Ergebnisse keine neuen Erkenntnisse bringen. Es gibt eine Reihe gut untersuchter Merkmale, anhand derer sich Aerobier von Anaerobiern unterscheiden. So haben aerobe Organismen im Gegensatz zu anaeroben beispielsweise Enzymklassen, die reaktive Sauerstoffspezies abbauen können. Katalase ist ein solches Enzym, das Wasserstoffperoxid zu Sauerstoff und Wasser umsetzt. Dieses Enzym wird sogar in einem mikrobiologischen Differenzierungstest, dem Katalasetest, zwischen Aerobiern und Anaerobiern verwendet. Viel interessanter hingegen ist die Frage, anhand welcher Proteindomänen klassifiziert wird, sobald Organismen mit anderem Sauerstoffbedarf hinzugenommen werden.

Die Organismen der beiden anderen Klassen stehen mit ihrer Proteinausstattung bezüglich der Sauerstofftoleranz zwischen den aeroben und anaeroben. Fakultativ anerobe Organismen nutzen Sauerstoff als Elektronenakzeptor, falls er vorhanden ist. Anderenfalls stellen sie ihren Metabolismus auf Fermentation bzw. anaerobe Atmung um. Mikroaerophile verstoffwechseln Sauerstoff, können jedoch nur bei Konzentrationen wachsen, die deutlich geringer sind als der Sauerstoffpartialdruck der Luft.

Die detektierten Proteindomänen haben zum größten Teil negatives Gewicht. Somit sind sie kontraindikativ für aerobe bzw. fakultativ anaerobe Organismen. Eine Proteindomäne kann eindeutig anaeroben Stoffwechselwegen zugeordnet werden. Dabei handelt es sich um Radical_SAM. Enzyme dieser Proteinfamilie modifizieren andere Enzyme (Sulfatasen) posttranslational, was zu deren Aktivierung führt. Ein Großteil der radikalen SAM Enzyme hat sauerstoffempfindliche Cluster wie Eisen-Schwefel-Cluster. Die katalytische Eigenschaft des Enzymes ist nur unter strikt anaeroben Bedingungen gegeben. Ebenfalls unter den detektierten Proteindomänen ist Fer4. In dieser Familie sind Proteine zusammengefasst, die Eisen-Schwefel-Cluster

bindende Domänen haben.

Für die übrigen Domänen gibt es keinen Hinweis darauf, dass sie typisch für Anaerobier sind. Dazu zählen folgende Domänen: NAPRTase ist eine Proteindomäne, die an dem ersten Schritt der NAD-Synthese über die Wiederverwendung von Abbauprodukten (der sogenannte „Salvage Pathway“) beteiligt ist [6]. Dabei handelt es sich um einen für alle Lebewesen wichtigen Stoffwechselweg. Es ist nicht ersichtlich, warum diese Domäne in dieser Kategorie detektiert wurde. Insbesondere der Zusammenhang zu anaeroben Organismen kann nicht hergestellt werden. Ebenfalls in allen Domänen des Lebens kommen die Proteindomänen Polyketide_cyc2 und die SAF-Domäne vor. Die erstgenannte Domäne ist an der Synthese von Polyketiden, einer Gruppe Sekundärmetabolite, beteiligt. Sie gehört zu der START-Proteinfamilie. Diese weit verbreitete Familie ist essentiell für enzymatische Aktivität. Sie bindet Liganden und ermöglicht damit Interaktionen zwischen Proteinen und kleinen Molekülen [17]. Die SAF-Domäne katalysiert eine Reihe unterschiedlicher Reaktionen. Sie ist Teil verschiedenster Proteine, darunter sind Frostschutzproteine genauso wie Flagellenproteine.

Die Proteindomäne mit dem höchsten Gewicht ist indikativ für aerobe Organismen. Ihre Bedeutung wird dadurch bestätigt, dass sie in beiden Eliminierungsmethoden zu den wichtigsten Merkmalen gehört. Diese Domäne wird in der Pfam-Datenbank als C4dic_mal_tran geführt. Sie gehört zu einem Transporter von C_4-Zuckern wie Dicarboxylat oder Malat. Dieser Transporter ist Teil des mitochondiralen Transportsystems [32]. In den Mitochondrien findet unter aeroben Bedingungen der energetisch wichtige, letzte Schritt des Glukoseabbaues (die Endoxidation) statt. Bei der Endoxidation handelt es sich um einen wichtigen Stoffwechselweg aerober Organismen. Insofern ist es nicht verwunderlich, dass eine zu diesem Stoffwechselweg gehörende Proteindomäne sehr hoch gewichtet wird.

Zum Teil ist keine Verbindung zwischen den hoch gewichteten Proteindomänen und dem Sauerstoffbedarf erkennbar. In dieser Kategorie wurde davon ausgegangen, dass anhand von Proteindomänen klassifiziert wird, die sich eindeutig der einen oder anderen Klasse zuordnen lassen. Zudem wurde von einigen Proteindomänen vermutet, dass sie ein hohes Gewicht bekommen würden. Ein Beispiel, in dem das nicht der Fall war, sind katalasespezifische Domänen. Insofern sind die detektierten Domänen zum Teil unerwartet.

4.6 Domäne des Lebens

In der Evolution haben sich drei phylogenetische Abstammungslinien entwickelt, Bacteria, Archaea und Eukarya. Sie werden als Domänen des Lebens bezeichnet.

Zwei der drei Abstammungslinien, Bakterien und Archaeen sind prokaryotisch. Dennoch sind Archaeen näher den Eukaryoten als den Bakterien verwandt [39]. Eine Vielzahl phänotypischer Eigenschaften korreliert mit der Domänenzugehörigkeit. Prominente Unterschiede bestehen in dem Zellwand- und Membranaufbau sowie der RNA-Polymerase. Kennzeichnend für die Zellwand von Bakterien ist Peptidoglykan. Dieses Molekül findet sich weder in der Zellwand von Archaeen noch in der von Eukaryoten. Die Zellwand von Archaeen ist aus Pseudopeptidoglykan, Polysacchariden und Proteinen aufgebaut, die der Eukaryoten aus Polysacchariden, Cellulose oder Chitin. Membranlipide der Archaeen sind über Etherlipide miteinander verknüpft, Lipide von Eukaryoten und Bakterien über Esterbindungen [39]. Typische Unterschiede gibt es auch in der Enzymausstattung, hier sei die RNA-Polymerase genannt. Bakterien haben eine einzige RNA-Polymerase mit vier Untereinheiten. Dem gegenüber stehen mehrere RNA-Polymerasen bei Archaeen und Eukaryoten. Sie sind sowohl in Eukaryoten als auch in Archaeen aus einer größeren Anzahl Untereinheiten aufgebaut [20].

Diese Unterschiede spiegeln sich in den detektierten Proteindomänen wider. Sie sind so grundlegender Natur, dass in dieser Kategorie die Klassenzugehörigkeit aller Organismen richtig vorhergesagt wurde. Das einzige Merkmal, das in der Merkmalseliminierung über den Absolutbetrag der Gewichte in allen Zufallspartitionen selektiert ist, ist eine Untereinheit des bakteriellen RNA-Polymerasekomplexes. Die σ^{70}-Untereinheit gehört zu den housekeeping-Genen der Bacteria. Als housekeeping-Gene werden die Gene bezeichnet, die konstitutiv exprimiert werden [12]. Sie sind in Prozesse involviert, die ständig in der Zelle ablaufen.

Fünf der zehn über die t-Teststatistik detektierten Proteindomänen sind ribosomale Proteindomänen. Auch sie gehören zu den housekeeping-Genen einer Zelle. Diese ribosomale Untereinheiten haben in Archaeen und Eukaryoten eine ähnliche Struktur [20]. Außerdem werden einige an der Translation beteiligten Proteindomänen detektiert, darunter ein Translationsintiationsfaktor und Proteine, die an der Synthese der t-RNA beteiligt sind. An den detektierten Domänen zeigt sich die gemeinsame Abstammung der Archaea und Eukarya. Das archaerielle Transkriptionssystem ist dem eukaryotischen homolog [20], unterscheidet sich aber in Struktur und Funktion von dem bakteriellen.

In dieser Katgorie gibt es eine so große Anzahl an Unterschieden zwischen den beiden Klassen, dass offensichtlich Proteindomänen detektiert werden, die zahlenmäßig den anderen überlegen sind. Dabei handelt es sich um housekeeping-Gene, die auf Grund ihrer Bedeutung für grundlegende Prozesse der Zellen mehrfach im Genom vorkommen.

Kapitel 5

Ausblick

Der vorgestellte Ansatz zur Vorhersage von phänotypischen Eigenschaften auf Grundlage von Proteindomänendetektion zielt auf die funktionelle Annotation von Organismen ab. Eine Übertragung auf Metagenome ist leicht möglich. Im Unterschied zur Analyse eines einzigen Genomes liegt in der Metagenomik das Augenmerk nicht auf der Klassifikation, sondern vielmehr auf den Proportionen der vorkommenden phänotypischen Eigenschaften. Für das Beispiel des Sauerstoffbedarfs bedeutet dies, das nicht die Klassifikation der einzelnen Organismen als Aerobier bzw. fakultative Anaerobier oder als strikte Anaerobier im Vordergrund steht. Wichtiger für die Charakterisierung des Metagenomes ist der Anteil aerober Organismen. Durch das Verhältnis der einzelnen Klassen eines partiellen Phänotypes ist eine Zuordnung zwischen dem Metagenom und seinen kennzeichnenden Eigenschaften möglich.

Eine biologische Herangehensweise an die detektierten Proteindomänen ist ebenfalls möglich. So sind unter den analysierten Proteindomänen die Domänen unbekannter Funktion biologisch äußerst interessant. Sie haben ein hohes Gewicht in den Diskriminanten und bleiben über viele Elimierungsschritte erhalten. Das lässt darauf schließen, dass sie eine Bedeutung für Organismen des jeweiligen Phänotypes haben. Somit lassen sich weitere Aussagen über ihre Funktion treffen. An dieser Stelle ist es sinnvoll, die über bioinformatischen Methoden aufgestellte Theorie experimentell zu überprüfen. Mit weiteren Hinweisen auf die Funktion der Proteindomänen kann deren Annotation verbessert werden.

Untersuchungen über Korrelationen zwischen ökologischen und genetischen Variablen sind Gegenstand aktueller Forschung [11, 19]. Es ist zu erwarten, dass die Ergebnisse vielversprechende Antworten bezüglich offener Fragestellungen der Anpassung von Organismen an ihre Umwelt liefern können.

Die Anwendung von Korrelationsanalysen zur Detektion von Zusammenhängen

zwischen genetischen Profilen und metabolischen Netzwerken ist eine neuere Entwicklung. Es hat sich als sinnvoll erwiesen, genetische Profile über die Häufigkeit von Genprodukten, d. h. Proteinen, zu modellieren.

Als ein Anwendungsziel solcher Arbeiten kann die Umweltökologie genannt werden. Letztlich soll es ermöglicht werden, Vorhersagen über die Anpassung von Organismen an ihre Umwelt zu treffen. Diese Vorhersagen ermöglichen neue Blickwinkel. So können durch Hinweise auf Anpassungen an die Umwelt Aussagen über Veränderungen, die der Organismus an seiner Umwelt vornimmt, getroffen werden. Ökologische Effekte von mikrobiellen Gemeinschaften zu verstehen, wird in Zeiten globaler Klima- und Umweltveränderungen immer wichtiger. Insofern kann die Bioinformatik einen wichtigen Beitrag zu aktuellen, fachübergreifenden Themen leisten.

Die längerfristige Perspektive der funktionellen Charakterisierung von Genomen und Metagenomen ist, neben der Umweltökologie, die Medizin [30]. Die funktionelle Charakterisierung von Genomen und Metagenomen kann in der modernen Diagnostik bedeutsam werden. Weitere Forschung wird zeigen, ob die funktionelle Charakterisierung zu einem anerkannten Diagnoseverfahren der Medizin werden kann. So könnte es beispielsweise möglich sein, über die Analyse von mikrobiellen Metagenomen der Schleimhäute oder des Darmes eine erste Aussage über krankhafte Veränderungen zu treffen [38, 37].

Kapitel 6

Zusammenfassung

In der vorliegenden Arbeit wurden bestimmte phänotypische Eigenschaften auf indikative Proteindomänen untersucht. Dazu wurden aus den partiellen Phänotypen Zweiklassenprobleme modelliert. In den so entstandenen Kategorien wurden mittels eines kombinierten Ansatzes aus Support Vektor Maschine und rekursiver Merkmalselimierung Proteindomänen detektiert, welche typisch für den entsprechenden Phänotyp sind. Von diesen Proteindomänen wurde eine Auswahl getroffen, die mit Bezug auf den jeweiligen partiellen Phänotyp ausgewertet wurden. Die Auswahl wurde in zwei Eliminierungsmethoden, über den Absolutbetrag und der t-Teststatistik der Gewichte, getroffen. Von den Proteindomänen, die in beiden Methoden detektiert wurden, wurden die zehn mit dem höchsten Gewicht untersucht.

In allen Kategorien konnten mit dieser Methode Proteindomänen detektiert werden, die den Phänotypen zugeordnet werden können. Zwischen den Kategorien gibt es Unterschiede in der Zuverlässigkeit der Vorhersage und der Aussagekraft der detektierten Domänen. So sind die in der Kategorie Super_Kingdom detektierten Proteindomänen eindeutig der einen oder anderen Klasse zuzuordnen. In der Kategorie Habitat_1 hingegen konnte keine eindeutige Verbindung zwischen den detektierten Domänen und dem Vorkommen in dem einen oder anderen Habitat hergestellt werden. Obwohl die Proteindomänen offensichtlich nicht mit der Zugehörigkeit zu der einen oder anderen Klassen korrelieren, wird eine relativ hohe Performanz erzielt.

An dieser Stelle wird deutlich, dass die festgelegte Mindestrate der Klassifikation willkürlich gewählt ist. Es stellt sich die Frage, ob in anderen Kategorien, die auf Grund des Schwellenwertes nicht weiter analysiert wurden, aussagekräftigere Domänen detektiert werden können. Insbesondere die Kategorie Temp_Range ist hierfür ein Kandidat. Es ist zu vermuten, dass sich bevorzugte Temperaturbedingungen über die Anpassungen des Organismus an diese Bedingungen in dem Domänenprofil widerspiegeln. Basierend auf diesen Erfahrung sollte das Auswahlkriterum in

zukünftigen Analysen angepasst werden.

Auf alle Kategorien betrachtet wurden viele Regulationsproteine detektiert. Grund hierfür ist deren Bedeutung für die Organismen. Regulationsproteine sind hoch exprimiert und dementsprechend stark in den Domänenprofilen vertreten.

Im Großen und Ganzen wurden mit der verwendeten Methodik Proteindomänen detektiert, die dem Vorwissen über typische Proteindomänen für die jeweiligen Phänotypen entsprechen. Somit kann davon ausgegangen werden, dass über die Auswahl der Domänen mit dem höchsten Gewicht aussagekräftige isoliert werden können. Die Erwartung, also die Detektion von aussagekräftigen Proteindomänen, wurde erfüllt.

Im Rahmen dieser Arbeit konnte nicht geklärt werden, inwieweit Fehler in der Vorhersage der Domänenprofile Auswirkungen auf die Bestimmung diskriminativer Proteindomänen haben. Es besteht die Möglichkeit, dass durch UFO vorhergesagte Proteindomänen eine leicht verfälschte Darstellung der tatsächlichen Domänenprofile sind.

Anhang A

Prokaryotic Genome Project

Die vollständige Auflistung der Eigenschaften, die im Rahmen des Prokaryotic Genome Project annotiert werden. Die in dieser Arbeit verwendeten Kategorien sind dick gedruckt. Falls abweichende Abkürzungen verwendet wurden, sind sie in Klammern angegeben.

Organism

Kingdom (Super_Kindom)

Genome Size

GC content

Gram stain (Gram_Stain)

Shape

Arrangement

Endospores

Motility

Salinity

Oxygen Requirements (Oxygen_Req)

Habitat (Habitat_1, Habitat_2)

Temperature Range (Temp_range)

Pathogenic in

Disease

Anhang B

Rekursive Merkmalseliminierung

Vollständige Liste der Pfam-IDs, die mittels rekursiver Merkmalseliminierung ermittelt werden. Zu der Eliminierungsmethode über den Absolutbetrag sind die Pfam-IDs aufgeführt, die in dem Selektionsschritt vor der stärksten Abnahme der Performanz selektiert wurden sowie die Pfam-IDs, die in diesem Schritt elimiert wurden. Für jede Kategorie sind Vereinigungsmenge über alle Zufallsdistributionen und, wenn die Menge nicht leer ist, Schnittmenge aufgeführt. Außerdem sind für die zweite Eliminierungsmethode über die t-Teststatistik die Pfam-IDs aufgeführt, die in dem Selektionsschritt, in dem die maximale Performanz (harmonisches Mittel aus Sensitivität und Spezifität) erzielt wurde, selektiert waren.

Endospores

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF01314 PF02852 PF08279 PF00704 PF03610 PF00296 PF00501 PF00942 PF03553 PF00395 PF01391 PF08534 PF08239 PF04122 PF00563 PF00389 PF00860 PF02588

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF00269 PF04672 PF03323 PF04545 PF04149 PF03144 PF02230 PF07875 PF00376 PF00082

Selektierte Merkmale (t-Teststatistik)

PF00662 PF06798 PF09925 PF06026 PF02599 PF02463 PF08812 PF06826

PF01467 PF02978 PF09547 PF10414 PF01351 PF07486 PF00146 PF09546
PF00318 PF00478 PF00115 PF02583 PF02534 PF08461 PF01725 PF00037
PF02562 PF04127 PF00237 PF00238 PF00281 PF02606 PF08668 PF02806
PF02436 PF01796 PF01075 PF00861 PF06888 PF00884 PF04456 PF00269
PF04012 PF02107 PF06898 PF01869 PF02108 PF05681 PF06835 PF00067
PF07873 PF02589 PF02687 PF00263 PF03946 PF05137 PF00347 PF04413
PF02355 PF05127 PF02649 PF04613 PF08298 PF03060 PF00133 PF08218
PF03802 PF01148 PF10369 PF04453 PF06686 PF00069 PF00270 PF06750
PF03419 PF03739 PF00561 PF02261 PF07441 PF02223 PF00256 PF09555
PF04542 PF02119 PF03862 PF01509 PF09855 PF02588 PF04998 PF01061
PF03364 PF04560 PF07451 PF08378 PF07241 PF01258 PF09551 PF00252
PF03418 PF04026 PF06993 PF00082 PF04565 PF00572 PF00673 PF00410
PF09578 PF01842 PF01746 PF01127 PF00768 PF00773 PF01975 PF00189
PF00453 PF02470 PF08264 PF10135 PF02656 PF02142 PF01227 PF04365
PF00472 PF02661 PF00155 PF07561 PF00211 PF04715 PF03551 PF05532
PF05949 PF02657 PF03958 PF02518 PF00989 PF00298 PF00152 PF10035
PF01381 PF02954 PF09334 PF02424 PF10764 PF03422 PF00376 PF07715
PF00106 PF05000 PF00623 PF02405 PF00501 PF04983 PF07875 PF00165
PF02659 PF02879 PF00408 PF00296 PF00440 PF08534 PF00005 PF02880
PF04545

Gram_Stain

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF07486 PF09587 PF02836 PF01276 PF07859 PF00781 PF02018 PF03781
PF00437 PF00331 PF03486 PF02452 PF01128 PF03845 PF00004 PF00571
PF00544 PF01584 PF00768 PF01636 PF01032 PF06429 PF05175 PF04616
PF00795 PF00155 PF08241 PF01183 PF08448

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF00263 PF02798 PF03739 PF08443 PF00300 PF00691 PF04203 PF01520
PF01316 PF00924

Selektierte Merkmale (t-Teststatistik)

PF00176 PF03894 PF02384 PF06969 PF02769 PF04077 PF01095 PF02452
PF05991 PF05108 PF01128 PF06182 PF03050 PF01047 PF02811 PF09560

PF00300 PF00682 PF00126 PF09397 PF10670 PF01761 PF02390 PF01263
PF02785 PF10518 PF05593 PF08443 PF03648 PF00289 PF01580 PF02782
PF02655 PF03797 PF05594 PF00953 PF01139 PF05175 PF02594 PF03609
PF01433 PF01842 PF02366 PF03613 PF08666 PF01451 PF05848 PF00590
PF09832 PF01568 PF01943 PF04413 PF03830 PF03780 PF01381 PF07477
PF07488 PF01520 PF03123 PF00768 PF01292 PF02664 PF02467 PF00460
PF10400 PF06135 PF04271 PF03725 PF07833 PF03845 PF02647 PF01643
PF04075 PF00331 PF07582 PF07693 PF00942 PF00293 PF01391 PF01522
PF00171 PF05958 PF03099 PF01106 PF00994 PF02353 PF01035 PF02581
PF04167 PF02863 PF09581 PF08447 PF02786 PF06964 PF01569 PF03577
PF01032 PF07733 PF00263 PF08238 PF02798 PF09587 PF07669 PF01244
PF00188 PF01380 PF03739 PF01618 PF09515 PF03861 PF00165 PF02096
PF01385 PF10604 PF06032 PF01554 PF01316 PF00465 PF00015 PF00083
PF01077 PF08282 PF00990 PF00529

Habitat_1

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF02743 PF01833 PF02779 PF03787 PF08242 PF01391 PF00117 PF00083
PF01526 PF00148 PF01554 PF07282 PF00376 PF00664 PF04290 PF01012
PF03480 PF02798 PF00990 PF00122 PF01979 PF00082 PF01040 PF02463
PF02195 PF00355 PF07726 PF01497 PF04389 PF01909 PF00550 PF01527
PF00553 PF02378

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF00924 PF00482 PF00425 PF01522 PF02811 PF08240 PF03537 PF03441
PF00310 PF02342

Selektierte Merkmale (t-Teststatistik)

PF10369 PF02075 PF10411 PF05762 PF02733 PF05726 PF07679 PF07517
PF00959 PF04408 PF01931 PF05239 PF01522 PF06421 PF00515 PF01904
PF03734 PF00916 PF08338 PF02436 PF00909 PF01946 PF02518 PF01969
PF03649 PF04358 PF00092 PF01483 PF00560 PF00476 PF08808 PF02342
PF02568 PF02635 PF07720 PF03480 PF05860 PF04402 PF04973 PF04066

PF00310 PF02779 PF00275 PF00201 PF00743 PF00239 PF00795 PF01051
PF00083 PF01408 PF02498 PF04039 PF03349 PF07719 PF08479 PF08548
PF03773 PF10385 PF01202 PF04909 PF01926 PF01396 PF00512 PF00043
PF01497 PF00005 PF02082 PF03807 PF00425 PF03787 PF07282 PF01636
PF00571 PF00158 PF00072 PF01656 PF04055 PF00990

Motility

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF00419 PF00102 PF10604 PF01883 PF05721 PF04015 PF00211 PF03486
PF02634 PF08376 PF06439 PF00690 PF02534 PF02885 PF00850 PF00091
PF09376 PF04471 PF00176 PF06271 PF00560 PF00753 PF01874 PF02475
PF01029 PF00669 PF01227 PF01113 PF02743 PF00112 PF00427 PF03062
PF04073 PF00082 PF07366 PF00144 PF00309 PF05222 PF02769 PF08534
PF01138 PF04185 PF00224 PF04820 PF01077 PF00905 PF01648 PF05157
PF05942 PF01979 PF01420 PF02668 PF01041 PF02355 PF04464 PF01479
PF01595 PF04892 PF00201 PF08242 PF01593 PF04069 PF04909 PF02446
PF02080 PF01522 PF00486 PF03050 PF00936 PF04607 PF04397 PF03459
PF01476 PF01814 PF02780 PF01926 PF01890 PF02662 PF01451 PF02086
PF01134 PF01934 PF00246 PF07980 PF02811 PF00571 PF09851 PF02133
PF00382

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF01443 PF00460 PF01957 PF00565 PF10106 PF08666 PF01442 PF02503
PF04964 PF01610

Selektierte Merkmale (t-Teststatistik)

PF03818 PF01139 PF04752 PF00516 PF08873 PF07441 PF00144 PF00401
PF01923 PF01419 PF06353 PF09411 PF01814 PF06338 PF04961 PF08670
PF10697 PF01795 PF09862 PF06135 PF08902 PF08002 PF03899 PF05881
PF10411 PF02623 PF07552 PF10431 PF01193 PF02195 PF05223 PF02824
PF00335 PF03345 PF00350 PF03805 PF04941 PF05202 PF07858 PF01230
PF06999 PF06005 PF07521 PF03852 PF03786 PF00570 PF09137 PF05675
PF01895 PF05977 PF02335 PF10687 PF08863 PF02545 PF00662 PF06455

PF03561 PF03922 PF07130 PF09526 PF06013 PF00091 PF04964 PF00146
PF02863 PF09278 PF05721 PF06153 PF06526 PF10102 PF08883 PF02706
PF02446 PF00989 PF04854 PF06022 PF06764 PF03471 PF06314 PF08272
PF03319 PF06304 PF02976 PF00912 PF04277 PF04434 PF01120 PF01227
PF03063 PF07720 PF03772 PF03812 PF02080 PF09832 PF03953 PF07714
PF00141 PF06897 PF06192 PF03357 PF03241 PF07667 PF02643 PF02361
PF06161 PF02030 PF06415 PF04405 PF06761 PF02482 PF01039 PF05147
PF06296 PF06352 PF06039 PF03780 PF09359 PF02595 PF00435 PF09977
PF08401 PF07694 PF01727 PF03072 PF03086 PF07271 PF09644 PF01065
PF09185 PF09188 PF05220 PF03257 PF00938 PF00625 PF05658 PF04940
PF05448 PF00560 PF04385 PF08757 PF04261 PF00080 PF03372 PF10064
PF07505 PF02278 PF05109 PF07666 PF07668 PF01812 PF10094 PF08487
PF08760 PF09946 PF01619 PF01724 PF03611 PF00528 PF02625 PF01325
PF00301 PF01979 PF00112 PF03008 PF01312 PF03830 PF02108 PF08331
PF01656 PF07044 PF08479 PF10503 PF00133 PF00881 PF02668 PF01712
PF02321 PF01126 PF08406 PF03212 PF06938 PF02475 PF09992 PF01734
PF01965 PF00011 PF04472 PF04257 PF02878 PF09299 PF05732 PF00639
PF01228 PF09298 PF09574 PF03070 PF02911 PF01118 PF06744 PF01316
PF01610 PF03613 PF09948 PF08534 PF05116 PF07883 PF00547 PF00862
PF02503 PF06445 PF03646 PF05935 PF04383 PF00491 PF06719 PF01527
PF04857 PF02738 PF00036 PF02515 PF03203 PF02585 PF06857 PF03609
PF00262 PF08907 PF06196 PF05738 PF00924 PF03960 PF05159 PF08345
PF02283 PF10101 PF00588 PF04215 PF02769 PF02152 PF10091 PF08447
PF02768 PF01070 PF04328 PF09829 PF00440 PF04607 PF02409 PF00449
PF06877 PF09345 PF10070 PF00171 PF03130 PF02464 PF01113 PF03480
PF07040 PF02562 PF00005 PF09968 PF07005 PF00925 PF03120 PF08124
PF03067 PF00331 PF00580 PF00874 PF04016 PF00211 PF00128 PF01451
PF07605 PF00149 PF00145 PF00702 PF04283 PF02694 PF00015 PF08388
PF05193 PF02151 PF05985 PF02586 PF00067 PF00246 PF03062 PF01168
PF01288 PF08212 PF00905 PF04443 PF04963 PF06850 PF07071 PF00581
PF05377 PF03783 PF00691 PF02132 PF01522 PF06751 PF00565 PF04432
PF04892 PF02654 PF02880 PF06439 PF00765 PF06832 PF04552 PF07715
PF01494 PF02579 PF01443 PF08242 PF08936 PF01520 PF00089 PF00419
PF00534 PF02753 PF00460 PF01568 PF09588 PF01758 PF01883 PF00672
PF08028 PF04659 PF07683 PF00753 PF00533 PF00345 PF00085 PF00082
PF03193 PF00849 PF07186 PF00497 PF01641 PF00106 PF01510 PF00873
PF04879 PF00271 PF10110 PF00092 PF00325 PF04326 PF01917 PF01551
PF01077 PF00582 PF01479 PF01381 PF03600 PF01555 PF06429 PF00027
PF07980 PF01839 PF02879 PF04545 PF01926

Oxygen_Req

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF01541 PF01434 PF08239 PF05345 PF04972 PF04069 PF01075 PF00145
PF03459 PF01074 PF00581 PF00684 PF00580 PF01943 PF00733 PF03372
PF01464 PF01938 PF08668 PF00041 PF02719 PF00702 PF01494 PF04205
PF01170 PF00586 PF00691 PF07963 PF02872 PF00378 PF02687 PF02416
PF00294 PF01867 PF07007 PF00293 PF07969 PF04471 PF01844 PF04616
PF01832

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF02803 PF04095 PF03595 PF05656 PF01578 PF08666 PF00037 PF02837
PF00550 PF08245

Selektierte Merkmale (t-Teststatistik)

PF07278 PF00850 PF03916 PF09686 PF07513 PF08900 PF00437 PF07511
PF05662 PF06146 PF04245 PF03382 PF07942 PF08013 PF06458 PF05656
PF09899 PF06710 PF02104 PF09423 PF04267 PF04095 PF10418 PF00358
PF00171 PF05922 PF08707 PF01184 PF09347 PF08379 PF02771 PF00117
PF05425 PF00733 PF00892 PF07883 PF00033 PF01909 PF01011 PF04940
PF00990 PF01925 PF02096 PF00227 PF00873 PF02080 PF02629 PF00578
PF02016 PF00703 PF00891 PF00343 PF02770 PF03595 PF05552 PF04171
PF06628 PF02195 PF09829 PF08323 PF07690 PF07497 PF00455 PF07922
PF04235 PF00486 PF00857 PF03592 PF03358 PF03880 PF02357 PF02837
PF03672 PF04069 PF05175 PF03838 PF00199 PF00294 PF01541 PF00293
PF01039 PF00361 PF00586 PF03572 PF00668 PF09313 PF08803 PF10423
PF00702 PF01075 PF00684 PF01839 PF01381 PF00037 PF08666 PF02836
PF08722 PF07721 PF01970 PF00108 PF08238 PF04972 PF00165 PF01022
PF01497 PF01527 PF02574 PF03918 PF00593 PF00196 PF00529 PF07992
PF10604 PF04055 PF00128 PF02803 PF08241 PF02518

Super_Kingdom

Eliminierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF00535 PF01979

Selektierte Merkmale: Schnittmenge (Absolutbetrag)

PF04542

Selektierte Merkmale: Vereiningungsmenge (Absolutbetrag)

PF04542 PF00753 PF01978 PF01472 PF00486 PF07724 PF02875 PF02775 PF00571 PF01163

Selektierte Merkmale (t-Teststatistik)

PF02778 PF04895 PF01287 PF01020 PF00935 PF01015 PF01157 PF01194 PF01655 PF09249

Anhang C

Domänen unbekannter Funktion (DUF)

Auflistung der Domänen unbekannter Funktion („Domains of Unknown Function", DUF), die unter den selektierten Merkmalen sind. Die Merkmalsmenge wird über die t-Teststatistik ermittelt. Es handelt es um die Merkmale, an Hand derer in dem Eliminierungsschritt mit maximaler Klassifikationsrate klassifiziert wird.

Endspores

PF09925	DUF2157	Predicted membrane protein (DUF2157)
PF02583	DUF156	Uncharacterised BCR, COG1937
PF01796	DUF35	Domain of unknown function DUF35
PF04456	DUF503	Protein of unknown function (DUF503)
PF06835	DUF1239	Protein of unknown function (DUF1239)
PF02589	DUF162	Uncharacterised ACR, YkgG family COG1556
PF05127	DUF699	Putative ATPase (DUF699)
PF02649	DUF198	Uncharacterized ACR, COG1469
PF09855	DUF2082	Nucleic-acid-binding protein containing Zn-ribbon domain (DUF2082)
PF02588	DUF161	Uncharacterized BCR, YitT family COG1284
PF07241	DUF1429	Protein of unknown function (DUF1429)
PF06993	DUF1304	Protein of unknown function (DUF1304)
PF02656	DUF202	Domain of unknown function DUF
PF04365	DUF497	Protein of unknown function (DUF497)
PF07561	DUF1540	Domain of Unknown Function (DUF1540)
PF05949	DUF881	Bacterial protein of unknown function (DUF881)
PF10035	DUF2179	Uncharacterized protein conserved in bacteria (DUF2179)
PF02405	DUF140	Domain of unknown function DUF140

PF02659	DUF204	Domain of unknown function DUF

Gram_Stain

PF05991	DUF901	Protein of unknown function (DUF901)
PF05108	DUF690	Protein of unknown function (DUF690)
PF06182	DUF990	Protein of unknown function (DUF990)
PF02594	DUF167	Uncharacterised ACR, YggU family COG1872
PF09832	DUF2059	Uncharacterized protein conserved in bacteria (DUF2059)
PF03780	DUF322	Protein of unknown function (DUF322)
PF06135	DUF965	Bacterial protein of unknown function (DUF965)
PF02647	DUF196	Uncharacterized ACR, COG1343
PF04075	DUF385	Domain of unknown function (DUF385)
PF04167	DUF402	Protein of unknown function (DUF402)
PF06032	DUF917	Protein of unknown function (DUF917)

Habitat_1

PF01931	DUF84	Protein of unknown function DUF84
PF01904	DUF72	Protein of unknown function DUF72
PF08338	DUF1731	Domain of unknown function (DUF1731)
PF01969	DUF111	Protein of unknown function DUF111
PF04402	DUF541	Protein of unknown function (DUF541)
PF03773	DUF318	Predicted permease

Motility

PF08873	DUF1834	Domain of unknown function (DUF1834)
PF06353	DUF1062	Protein of unknown function (DUF1062)
PF10697	DUF2502	Protein of unknown function (DUF2502)
PF09862	DUF2089	Uncharacterized protein conserved in bacteria (DUF2089)
PF06135	DUF965	Bacterial protein of unknown function (DUF965)
PF08902	DUF1848	Domain of unknown function (DUF1848)
PF08002	DUF1697	Protein of unknown function (DUF1697)
PF06005	DUF904	Protein of unknown function (DUF904)
PF05675	DUF817	Protein of unknown function (DUF817)

PF05977	DUF894	Bacterial protein of unknown function (DUF894)
PF10687	DUF2495	Protein of unknown function (DUF2495)
PF09526	DUF2387	Probable metal-binding protein (DUF2387)
PF06153	DUF970	Protein of unknown function (DUF970)
PF06526	DUF1107	Protein of unknown function (DUF1107)
PF10102	DUF2341	Domain of unknown function (DUF2341)
PF04854	DUF624	Protein of unknown function, DUF624
PF06764	DUF1223	Protein of unknown function (DUF1223)
PF06304	DUF1048	Protein of unknown function (DUF1048)
PF09832	DUF2059	Uncharacterized protein conserved in bacteria (DUF2059)
PF06897	DUF1269	Protein of unknown function (DUF1269)
PF07667	DUF1600	Protein of unknown function (DUF1600)
PF02643	DUF192	Uncharacterized ACR, COG1430
PF06161	DUF975	Protein of unknown function (DUF975)
PF06296	DUF1044	Protein of unknown function (DUF1044)
PF06352	DUF1061	Protein of unknown function (DUF1061)
PF03780	DUF322	Protein of unknown function (DUF322)
PF09977	DUF2134	Predicted membrane protein (DUF2134)
PF08401	DUF1738	Domain of unknown function (DUF1738)
PF01727	DUF30	Domain of unknown function DUF30
PF03072	DUF237	MG032/MG096/MG288 family 1
PF03086	DUF240	MG032/MG096/MG288 family 2
PF09185	DUF1948	Domain of unknown function (DUF1948)
PF09188	DUF1951	Domain of unknown function (DUF1951)
PF10064	DUF2302	Uncharacterized conserved protein (DUF2302)
PF10094	DUF2332	Uncharacterized protein conserved in bacteria (DUF2332)
PF08760	DUF1793	Domain of unknown function (DUF1793)
PF09946	DUF2178	Predicted membrane protein (DUF2178)
PF01724	DUF29	Domain of unknown function DUF29
PF03008	DUF234	Archaea bacterial proteins of unknown function
PF08331	DUF1730	Domain of unknown function (DUF1730)
PF07044	DUF1329	Protein of unknown function (DUF1329)
PF06938	DUF1285	Protein of unknown function (DUF1285)
PF09992	DUF2233	Predicted periplasmic protein (DUF2233)
PF04472	DUF552	Protein of unknown function (DUF552)
PF09298	DUF1969	Domain of unknown function (DUF1969)
PF09574	DUF2374	Protein of unknown function (Duf2374)
PF06744	DUF1215	Protein of unknown function (DUF1215)
PF09948	DUF2182	Predicted metal-binding integral membrane protein (DUF2182)
PF08907	DUF1853	Domain of unknown function (DUF1853)
PF06196	DUF997	Protein of unknown function (DUF997)

PF10101	DUF2339	Predicted membrane protein (DUF2339)
PF10091	DUF2329	Uncharacterized protein conserved in bacteria (DUF2329)
PF04328	DUF466	Protein of unknown function (DUF466)
PF09829	DUF2057	Uncharacterized protein conserved in bacteria (DUF2057)
PF06877	DUF1260	Protein of unknown function (DUF1260)
PF09345	DUF1987	Domain of unknown function (DUF1987)
PF10070	DUF2309	Uncharacterized protein conserved in bacteria (DUF2309)
PF07040	DUF1326	Protein of unknown function (DUF1326)
PF09968	DUF2202	Uncharacterized protein conserved in archaea (DUF2202)
PF07005	DUF1537	Protein of unknown function, DUF1537
PF04016	DUF364	Domain of unknown function (DUF364)
PF07605	DUF1568	Protein of unknown function (DUF1568)
PF04283	DUF439	Protein of unknown function (DUF439)
PF02586	DUF159	Uncharacterised ACR, COG2135
PF07071	DUF1341	Protein of unknown function (DUF1341)
PF06439	DUF1080	Domain of Unknown Function (DUF1080)
PF01883	DUF59	Domain of unknown function DUF59
PF03193	DUF258	Protein of unknown function, DUF258
PF06429	DUF1078	Domain of unknown function (DUF1078)

Oxygen_Req

PF07278	DUF1441	Protein of unknown function (DUF1441)
PF07513	DUF1527	Protein of unknown function (DUF1527)
PF08900	DUF1845	Domain of unknown function (DUF1845)
PF07511	DUF1525	Protein of unknown function (DUF1525)
PF03382	DUF285	Mycoplasma protein of unknown function, DUF285
PF05656	DUF805	Protein of unknown function (DUF805)
PF09899	DUF2126	Uncharacterized protein conserved in bacteria (DUF2126)
PF06710	DUF1197	Protein of unknown function (DUF1197)
PF09347	DUF1989	Domain of unknown function (DUF1989)
PF00892	DUF6	Integral membrane protein DUF6
PF01925	DUF81	Domain of unknown function DUF81
PF04171	DUF405	Protein of unknown function (DUF405)
PF09829	DUF2057	Uncharacterized protein conserved in bacteria (DUF2057)
PF04235	DUF418	Protein of unknown function (DUF418)
PF09313	DUF1971	Domain of unknown function (DUF1971)
PF01970	DUF112	Integral membrane protein DUF112

Super_Kingdom

PF04895 DUF651 Archaeal protein of unknown function (DUF651)

Literaturverzeichnis

[1] Aizerman M., Braverman E., Rozonoer L. (1964) Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, **25**(6):821-837.

[2] Bardy S. L., Ng S. Y. M., J. K. F (2003) Prokaryotic motility structures. *Microbiology*, **149**:295-304.

[3] Boser B.E., Guyon I., Vapnik V.N. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, 5. ACM Pittsburgh, S. 144-152.

[4] Cabeen M. T., Jacobs-Wagner C. (2005) Bacterial cell shape. *Nat. Rev. Microbiol.*, **3**(8): 601-610.

[5] Cristianini N., Shawe-Taylor J. (2000) An Introduction to Support Vector Machines. Cambridge University Press, S. 93ff.

[6] Dulyaninova N. G., Podlepa E. M., Toulokhonova L. V., Bykhovsky V. Y. (2000) Salvage pathway for NAD biosynthesis in Brevibacterium ammoniagenes: regulatory properties of triphosphate-dependent nicotinate phosphoribosyltransferase. *Biochim. Biophys. Acta*,**1478**(2):211-220.

[7] Eddy S. R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**(9):755-763.

[8] Eichenberger P., Jensen S. T., Conlon E. M., van Ooij C., Silvaggi J., González-Pastor J. E., Fujita M., Ben-Yehuda S., Stragier P., Liu J. S. Losick, R. (2003) The sigma(Epsilon) regulon and the identification of additional sporulation genes in Bacillus subtilis. *J. Mol. Biol.*, **327**(5):945-972.

[9] Errington J. (2003) Regulation of endospore formation in *Bacillus subtilis*. *Nat. Rev. Micro.*, **1**(2):1740-1526.

[10] Finn R. D., Tate J., Mistry J., Coggill P. C., Sammut S. J., Hotz H.-R., Ceric G., Forslund K., Eddy S. R., Sonnhammer E. L. L., *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**(Database issue): D281-D288.

[11] Gianoulis T. A., Raes J., Patel P. V., Bjornson R., Korbel J. O., Letunic I., Yamada T., Paccanaro A., Jensen L. J., Snyder M. *et al.* (2009) Quantifiying environmental adaption. *PNAS*, **106**(5):1374-1379.

[12] Gruber T. M., Gross C. A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.*, **57**(1):441-466.

[13] Guyon I., Weston J., Barnhill S., Vapnik V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, **46**(1):389-422.

[14] Handelsman J. (2004) Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**(4):669-685.

[15] Henriques A. O., Moran C. P. (2007) Structure, Assembly, and Function of the Spore Surface Layers. *Annu. Rev. Microbiol.*, **61**(1):555-588.

[16] Hoa N. T., Brannigan J. A., Cutting S. M. (2002) The Bacillus subtilis Signaling Protein SpoIVB Defines a New Family of Serine Peptidases. *J. Bacteriol.*, **184**(1):191-199.

[17] Iyer L. M., Koonin E. V., Aravind L. (2001) Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. *Proteins*, **43**(2):134-144.

[18] Kastenmüller G., Gasteiger J., Mewes J.-W. (2008) An environmental perspective on large-scale genome clustering based on metabolic capabilities. *Bioinformatics*, **24**:i56-i62.

[19] Kunin V., Raes J., Harris J. K., Spear J. R., Walker J. J., Ivanova N., von Mering C., Bebout B. M., Pace N. R., Bork P. *et al.* (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol.*, **4**:198.

[20] Langer D., Hain J., Thuriaux P., Zillig W. (1995) Transcription in Archaea: Similarity to that in Eucarya. *Proc. Natl. Acad. Sci. USA*, **92**(13): 5768-5772.

[21] Meinicke P. UFO: a web server for ultra-fast functional profiling of whole genome protein sequences, to be published by *BMC Genomics*.

[22] Meyer Y., Sialaa W., Bashandya T., Riondeta C., Vignolsa F., Reichhelda J. P. (2008) Glutaredoxins and thioredoxins in plants. *Biochim. Biophys. Acta*,**1783**(4):589-600.

[23] Munk K. (Hrsg.) (2001) Grundstudium Biologie Genetik, 1. Auflage. Spektrum Akademischer Verlag Heidelberg, S. 7-2.

[24] Nagy Z., Chandler, M. (2004) Regulation of transposition in bacteria. *Res. Microbiol.*, **155**(5):387-398.

[25] Navarre W. W., Schneewind O. (1999) Surface Proteins of Gram-Positive Bacteria and Mechanisms of Their Targeting to the Cell Wall Envelope. *Microbiol. Mol. Biol. Rev.*, **63**(1):174-229.

[26] NCBI (Hrsg.): Prokaryotic Genome Project. *http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi* (Abruf 18.7.2009)

[27] Paredes C. J., Alsaker K. V., Papoutsakis E. T. (2005) A comparative genomic view of clostridial sporulation and physiology. *Nat. Rev. Microbiol.*, **3**(12):969-978.

[28] Parkhill J., Sebaihia M., Preston A., Murphy L. D., Thomson N., Harris D. E., Holden M. T. G., Churcher C. M., Bentley S. D., Mungall K. L. *et al.* (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat. Genet.*, **35**(1):32-40.

[29] Pei J. Grishin N. V. (2001) GGDEF domain ist homologous to adenylyl cyclase. *Proteins*, **42**(2):210-216.

[30] Ponting C. P., Russel R. R. (2002) The Natural History of Protein Domains. *Annu. Rev. Biophys. Biomol. Struct.*, **31**(1): 45-71.

[31] Potts M. (1994) Desiccation tolerance of prokaryotes. *Microbiol. Mol. Biol. Rev.*, **58**(4):755-805.

[32] Robinson A. J., Overy C., Kunji E. R. S. (2008) The mechanism of transport by mitochondrial carriers based on analysis of symmetry. *PNAS*, **105**(46):17766-17771.

[33] Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.*, **5**(3):e77.

[34] Sandhya S., Rani S. S., Pankaj B., Govind M. K., Offmann B., Srinivasan N., Sowdhamini R. (2009) Length Variations amongst Protein Domain Superfamilies and Consequences on Structure and Function. *PLoS One*, **4**(3):e4981.

[35] Springer T. A. (1997) Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain. *Proc. Natl. Acad. Sci. USA*, **94**(1):65-72.

[36] Steil L., Serrano M., Henriques A. O., Volker U. (2005) Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of Bacillus subtilis. *Microbiology*, **151**:399-420.

[37] Turnbaugh P. J., Ley R. E., Hamady M., Fraser-Liggett C. M., Knight R., Gordon J. I. (2007) The Human Microbiome Project. *Nature*, **449**(7164):804-810.

[38] White J. R., Nagarajan N., Pop M. (2009) Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS*, **5**(4):e1000352.

[39] Woese C. R., Kandler O., Wheelis M. L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *PNAS*, **87**(12):4576-4579.

[40] Woyke T., Xie G., Copeland A., Gonzáles J. M., Han C., Kiss H., Saw J. H., Senin P., Yang C., Chatterji S., *et al.* (2009) Assembling the Marine Metagenome, One Cell at a Time. *PLoS One*, **4**(4):e5299.

Danksagung

An dieser Stelle bleibt nur noch eines zu sagen: Danke!

Als erstes möchte ich mich bei Dr. Peter Meinicke und Dr. Maike Tech bedanken, die mir die Arbeit ermöglicht haben, am Fortschritt interessiert waren und immer ein offenes Ohr für Fragen hatten. Vielen Dank für all die konstruktiven Anmerkungen und das Korrekturlesen. Ein großes Dankeschön geht an meinen „Büromitbewohner“ Christian, der mir das SVN – soweit es ging – näher gebracht hat, Anne, die mir Latex durch Beispiele schmackhaft gemacht hat und Rasmus für die technische Unterstützung.

Vielen Dank an alle, die mich während des Studiums unterstützt haben, für mich da waren und sind. Und die mich besonders in der letzten, „heißen“ Phase daran erinnert haben, dass ich auch die Bachelorarbeit schaffen werde. Allen voran danke ich meiner Familie. Claudia, Du meine allerliebste Schwester, ohne Dich gäbe es noch mehr grammatikalische Absonderlichkeiten in dieser Arbeit. Du hast mich zuverlässig mit Lebensweisheiten versorgt („Denk an die Vitamine! “) und mich damit mehr als einmal zum Schmunzeln gebracht.

Fabian, ohne Deine Zeit und Geduld, hilfreiche Diskussionen, kritische Anmerkungen („Department of Redundancy Department“) und die eine oder andere Tasse Tee, die wir zusammen geschlürft haben, wäre diese Zeit nicht halb so schön gewesen. Danke, dass Du Dich zu nachtschlafender Zeit aus dem Bett gequält hast (und damit bewiesen hast, dass es Dich auch vor zehn Uhr gibt!), um mit mir Fahrrad zu fahren.

Ein großes Dankeschön an Rike und Daggi für ihre Freundschaft und all die schönen Spieleabende. Sei es die „ bombige Palme“ oder andere Assoziationen: Macht das Leben lebenswerter. Nicht vergessen möchte ich all die Menschen, die mit mir die Musik, die schöne Seite des Lebens, teilen: Katharina und Unicante.

Fabian, Papa, Danke für’s Korrekturlesen!

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Göttingen, den 19. August 2009