Diplomarbeit

Clustering und taxonomische Klassifizierung von Metagenom-Sequenzen

Benjamin Balluff

25. Oktober 2007

Betreuer:
Dr. Thomas Rattei
Lehrstuhl für genomorientierte Bioinformatik, TU München

Dr. Bernhard Haubold
Fakultät für Biotechnologie und Bioinformatik, FH Weihenstephan

Eidesstattliche Erklärung

Gemäß §23 Abs. 6 der Prüfungsordnung der FH Weihenstephan

Ich erkläre hiermit an Eides statt, dass die vorliegenden Arbeit von mir selbst und ohne fremde Hilfe verfasst und noch nicht anderweitig für Prüfungszwecke vorgelegt wurde.

Es wurden keine als die angegebenen Quellen oder Hilfsmittel benutzt. Wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

Freising, den 17. Oktober 2007

Inhaltsverzeichnis

1. Motivation

Mikroorganismen regieren die Welt. Sie waren die ersten Organismen auf der Erde und sie haben die Evolutionsgeschichte von Beginn an bestimmt. Auch wenn die Prokaryoten in Zeiten von Mehrzellern wie Tieren, Pflanzen und Menschen klein und unscheinbar erscheinen, ist ihr kollektiver Einfluss auf die Erde doch immens. Schätzungen gehen davon aus, dass Mikroorganismen etwa $\frac{1}{3}$ der Weltbiomasse ausmachen [17]. Durch ihre Häufigkeit und ihre Stoffwechselvielfalt sind sie die wichtigsten lebenden Bindeglieder in den globalen Kreisläufen chemischer Stoffe.

Die Mikrobiologie, die sich mit der Erforschung der Mikroorganismen befasst, durchlebt seit einigen Jahren einen starken Wandel. Technische Meilensteine in der Molekularbiologie, wie der PCR Technologie, ermöglichten neue Einblicke in die mikrobielle Vielfalt und zeigten deren ungeahnt hohe Diversität auf. Kennt man bisher etwa 18 000 prokaryotische Spezies [48], geht man mittlerweile davon aus, dass die Domäne der Bakterien mit einer geschätzten Mindestanzahl von 4 Millionen Arten [24] die artenreichste ist. Über den Großteil der Mikroorganismen auf der Erde ist somit nur wenig bekannt. Der Hauptgrund liegt darin, dass sie nicht im Labor kultiviert werden können. Dies liegt daran, dass man zu wenig über deren Physiologie weiß, um ein passendes Kulturmedium bereitstellen zu können.

Es besteht daher die Notwendigkeit, auf die genetische Information, der in einem Habitat lebenden Organismen, Zugriff zu bekommen ohne die Organismen dabei selbst kultivieren zu müssen. Die **Metagenomik** ermöglicht die Erforschung einer mikrobiellen Gemeinschaft, durch direkte Extraktion und Klonierung der gesamten genomischen DNA aus einer Umweltprobe. Die darauf folgende Anwendung moderner DNA-Analyseverfahren auf das Metagenom — der gesamten aus der Umweltprobe sequenzierten genetischen Information — erlaubt folgende Fragenstellungen anzugehen: „Welche Organismen sind da?“, „Was machen sie?“ und „Wie machen sie es?“.

In den letzten Jahren betrieben Mikrobiologen einen hohen Aufwand, um mit Hilfe der Metagenomik Habitate, wie zum Beispiel Meerwasser [31, 46], einen Biofilm in einer Eisenmine [50] oder Mammutknochen [49], erforschen zu können. Hierbei konnten viele neue Taxa anhand ihrer molekularen Signatur entdeckt werden. Die hohe phylogenetische Diversität bedeutet auch eine hohe genetische Diversität. Mehr als 1 Million möglicher neuer Gene konnten in der Meerwasserprobe aus dem Sargasso-See [31] identifiziert werden. Der resultierende Reichtum an neuen Genen und biochemischen Prozessen, der mit Hilfe der Metagenomik aus Gemeinschaften unkultivierbarer Mikroorganismen gewonnen werden kann, birgt ein enormes Potential für Forschung und Industrie.

Ein erfolgreiches Beispiel aus der Forschung ist die Erforschung des nicht kultivierbaren Bakteriums *Kuenenia stuttgartiensis*, eines Vertreters der anaeroben Ammonium Oxidierer. Dieses Bakteriums spielt eine bedeutende Rolle im Stickstoffzyklus der Meere

und ist auf Grund dessen für die Abwasserreinigung von großem Interesse.

Vor allem aber für die chemische und pharmazeutische Industrie ist die Entdeckung von neuen und effektiveren Enzymen von größter Wichtigkeit. Die Metagenomik stellt daher für die Industrie eine wichtige Quelle neuartiger bioaktiver Moleküle dar.

Das Potential der Metagenomik ist jedoch stark abhängig von Weiterentwicklungen in der Sequenziertechnik, in der Konstruktion der Klonbibliotheken, und auch in den Analysemethoden. Vor allem Fortschritte in der Bioinformatik würden das Management und die Analyse der großen Metagenomdatensätze erleichtern. Aufgrund der Größe von Metagenomdatensätzen ist es eine Erfordernis, effiziente Methoden zu entwickeln, aber auch diese Methoden an die inhaltlichen Gegebenheiten der Metagenomik anzupassen. Denn nur dann wird die Metagenomik dazu beitragen, unser Verständnis über die Welt der Mikroorganismen zu bereichern.

2. Aufgabenstellung

Ziel der Metagenomik ist die vollständige Abbildung der genetischen Information einer Umweltprobe zum Zwecke der nachfolgenden Analyse. Shotgun-Sequenzierung der DNA aus diesem Habitat ermöglicht den Zugriff auf die Genome der darin enthaltenen, nicht kultivierbaren Mikroorganismen. Aus diesem Ansatz resultieren jedoch große Datensätze meist kurzer, unklassifizierter Sequenzabschnitte aus verschiedensten Organismen, sogenannte Metagenome. Zur Strukturierung und taxonomischen Untersuchung dieser Metagenome, wurden in den letzten Jahren bereits einige Clustering- und phylogenetische Methoden entwickelt. Bis jetzt eignen sich aber die meisten dieser Methoden nicht für die Verarbeitung von Metagenomen, da sie entweder nicht automatisiert oder zu rechenintensiv sind. Angesichts der Flut an Daten aus Metagenomik-Projekten, die in naher Zukunft erwartet werden, besteht die Notwendigkeit eines vollautomatisierten Softwarepaketes zum Clustering und zur schnellen taxonomischen Studie von Metagenomen.

Ein grundlegendes Problem in der Metagenomik ist die Strukturierung der DNA-Fragmente zu Genomen oder taxonomischen Gruppen, das sogenannte *Binning*, wenn Markergene als identifizierendes Merkmal fehlen. Gegenwärtig basieren bewährte Binning-Methoden auf Merkmalen wie GC-Gehalt, Codon-Usage Profilen oder Oligonukleotidfrequenzen. Bedeutende Arbeiten zu genomischen Inhomogenitäten haben gezeigt, dass die Neigung von Genomen zu bestimmten Oligonukleotidfrequenzen ein schwaches phylogenetisches Signal in sich trägt. Daher ist es ein Ziel dieser Arbeit, eine Binning-Funktionalität auf Basis dieser Oligonukleotidfrequenzen in die Software zu integrieren, wobei das Nukleotidmuster vom Benutzer frei definiert werden kann. Das Clustering soll anhand von künstlichen Metagenomdatensätzen und verschiedenen Nukleotidmustern getestet und optimiert werden.

Während sich die Binning Implementierung auf bewährte Ansätze stützt, wurde für die taxonomische Analyse eine neuartige Methode, namens **Blast2Tree**, entwickelt. Diese soll eine schnelle und automatisierte taxonomische Klassifizierung von Metagenomendaten ermöglichen. Die Methode setzt eine vordefinierte Menge an Markergenen und einen dazugehörigen phylogenetischen Speziesbaum voraus. Dominik Lindner konnte in seiner Diplomarbeit [20] zeigen, dass es möglich war, auf diese Weise sehr schnell alle ORFs eines Metagenoms, die homolog zu einem Markergen waren, in den phylogenetischen Speziesbaum einzuordnen. Er benutzte hierbei als Referenzbaum einen Baum des Lebens, der auf Basis von 31 ausgewählten Markerproteinfamilien erstellt wurde [14].

Ziel dieser Arbeit ist es, die Blast2Tree Methodik zu implementieren und damit zu automatisieren. Mit Hilfe der Automatisierung soll die taxonomische Klassifizierung systematisch anhand des sogenannten Jack-Knife Verfahrens, validiert werden. Dazu wird jeder Knoten des Baumes einmal entfernt. Für jeden entfernten Knoten, werden alle Proteine der Taxa, die zu diesem Knoten gehören, wieder in den Baum eingeordnet.

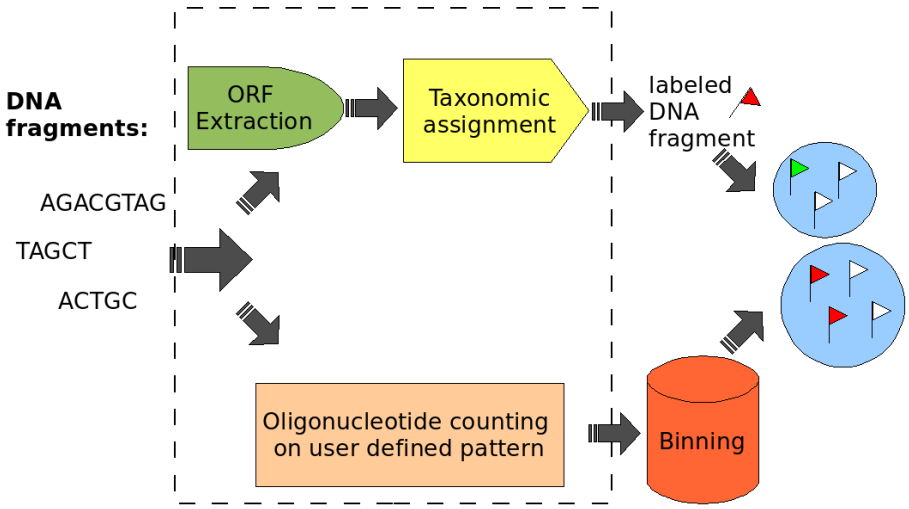


Abbildung 2.1.: Der Workflow des Software-Frameworks. Da die taxonomische Klassifizierung auf Homologie zu Markergenen basiert, werden nur die auf dem DNA-Fragment gefundenen homologen ORFs taxonomisch gekennzeichnet. Die Taxonomie kann aber auf das innehabenden Fragment übertragen werden. Das Ziel ist, am Ende Clustering der Fragmente (*Binning*) und taxonomische Analyse so miteinander zu kombinieren, so dass alle entstandenen Bins (blaue Kreise) mindestens eine zuverlässige taxonomische Bezeichnung (farbige Fähnchen) erhalten.

Durch Messen der Abweichung des gefundenen Knotens zum erwarteten Knoten, soll die Software auf algorithmischer Seite verbessert werden.

Am Ende soll ein komplettes Softwarepaket zur automatischen Analyse von Community-Genomen geschaffen werden, das Binning und taxonomische Klassifizierung zusammenführt (Abbildung 2.1). Diese Software ist als erster Schritt bei einer Analyse von Metagenomen gedacht, der einen schnellen und groben taxonomischen Überblick über das Metagenom verschaffen soll. Das Clustern der Metagenom-Sequenzen und das schnelle Zuweisen taxonomischer Marker soll auch nachfolgende Analysen, wie zum Beispiel phylogenetische oder funktionelle Untersuchungen, unterstützen und beschleunigen.

Die Software soll für wissenschaftliche Zwecke frei erhältlich sein und auch als Web-Service angeboten werden.

3. Einleitung

3.1. Systematik, Taxonomie und Phylogenie

3.1.1. Biologische Diversität und Systematik

Die Erde existiert seit ungefähr 4.5 Milliarden Jahren. Den ersten Hinweis auf Leben findet man in ca. 3.5 Milliarden alten Fossilien winziger Mikroorganismen. Seit dem hat sich die Erde mit ihrer einstigen unbewohnbaren, anoxischen und heißen Atmosphäre zu einem pulsierenden Planeten mit einer enormen biologischen Diversität entwickelt [21]. Es wurden bis heute ca. 1.8 Millionen lebende Spezies entdeckt und die Liste wird von Jahr zu Jahr länger [16].

Das Studium dieser biologischen Vielfalt im evolutionären Kontext wird als Systematik bezeichnet. Die Systematik beinhaltet sowohl die Taxonomie, also das Benennen und Klassifizieren der Arten und Artengruppen, als auch die Phylogenie, die Aufklärung der Stammesgeschichte der Arten [24].

3.1.2. Phylogenetik

Die Phylogenetik erforscht die stammesgeschichtliche Entfaltung (*Phylogenese*) der Arten und anhand von Stammbäumen [51]. Abbildung 3.1 illustriert die Terminologie eines bewurzelten Baumes, der als phylogenetischer Baum verwendet wird. Meist handelt es sich um binäre Bäume, deren innere Knoten ausgestorbene Vorfahren repräsentieren. Blattknoten hingegen repräsentieren noch nicht ausgestorbene Organismen. Im Gegensatz zum bewurzelten Baum, besitzt ein unbewurzelter Baum daher keinen gemeinsamen Vorfahren aller Blattknoten.

Um phylogenetische Stammbäume erstellen zu können, sind Informationen notwendig, die die stammesgeschichtliche Entwicklung widerspiegeln. Dazu eignen sich idealerweise homologe Merkmale, das heißt Eigenschaften, die von einem gemeinsamen Vorfahren geerbt wurden. Die Erforschung der Phylogenese erfolgte bis vor einigen Jahrzehnten noch

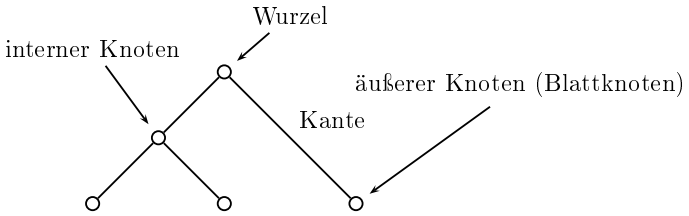


Abbildung 3.1.: Graphterminologie eines bewurzelten Baumes.

anhand morphologischer Merkmale von Fossilien und dem Vergleich morphologischer, anatomischer und physiologischer Merkmale jetztzeitiger Lebewesen [24].

Seit dem Aufkommen der Molekularbiologie ist es jedoch auch möglich, Homologien auf DNA- oder Protein-Ebene zu erklären. Diese molekulare Systematik ermöglicht es, phylogenetische Beziehungen festzustellen, die durch vergleichende Anatomie nicht erkennbar sind. Diese Methode verhalf der Mikrobiologie zur Bestimmung vieler neuer evolutionärer Verwandtschaften von Mikroorganismen.

Erstellung von phylogenetischen Bäumen

Nur bestimmte Gene und Proteine sind evolutionäre Chronometer und eignen sich deshalb zur Messung von evolutionären Verwandtschaft. Der perfekte molekulare Zeitmesser sollte für alle zu untersuchenden Organismen folgende Merkmale aufweisen ([20, 21]):

- universell verteilt, also möglichst in allen Genomen vorhanden
- funktionell homolog
- konserviert
- nicht durch horizontalen Gentransfer aufgenommen

In der Mikrobiologie haben sich vor allem ribosomale RNAs (rRNA) als evolutionäre Marker etabliert, da diese funktionell konstant und universell verteilt sind und mehrere konservierte Regionen besitzen [21]. Dazu gehören bei den Prokaryoten die rRNA-Moleküle 5S, 16S und 23S[1]. Das eukaryotische Pendant dazu ist das funktionell ähnliche 18S rRNA Molekül. Das ribosomale Datenbankprojekt (RDP) unterhält eine umfangreiche Sammlung dieser rRNA Sequenzen. Momentan (Mai 2007) sind knappe 350 000 Sequenzen online unter `http://rdp.cme.msu.edu` verfügbar.

Hat man einen idealen evolutionären Marker gefunden, so können phylogenetische Bäume nun durch ein so genanntes multiples Alignment der homologen Sequenzdaten erstellt werden. Ein Problem bei der Rekonstruktion von Bäumen ist die große Anzahl an möglichen Bäumen. So ergeben sich für den Fall, die Phylogenie von 50 Arten zu rekonstruieren, $3 * 10^{76}$ mögliche Bäume (Formel (3.1) aus [23]).

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{3.1}$$

Anzahl der bewurzelten Bäume für n Blattknoten.

Die Identifizierung des optimalen Baumes anhand verschiedener Methoden, wie der Maximum-Parsimony oder der Maximum-Likelihood-Methode, ist ein NP-hartes Problem, welches sehr viel Rechenleistung erfordert und nur durch Heuristiken erleichtert wird.

Ein weniger rechenintensiver Algorithmus zum Erstellen phylogenetischer Bäume ist der Neighbour-Joining Algorithmus, dessen Komplexität nur $O(n^3)$ bei n Taxa beträgt.

[1] „S“ steht für Svedberg

Einschränkungen beim Erstellen von phylogenetischen Bäumen

Bei jedem erstellten Stammbaum handelt es sich um eine Hypothese [24]. Wenn man eine Wahl zwischen verschiedenen Stammbäumen hat, ist jene Hypothese die optimalste, die am besten mit den verfügbaren Daten übereinstimmt. Aber Daten können auch verrauscht sein. Zum Beispiel können Proteine oder DNA-Moleküle durch horizontalen Gentransfer, Duplikation, Deletion oder multiple Substitution (Rückmutation) außerordentlich verändert worden sein. Ebenso muss man vorsichtig sein, einen Baum auf Basis von nur einem Gen zu berechnen. Es gibt Proteinfamilien, die sich in ihrer Evolutionsgeschwindigkeit stark unterscheiden, wie zum Beispiel Cytochrom c, Histone oder Hämoglobin [24]. Außerdem ist es bekannt, dass die Mutationsrate von Arten beträchtlich variieren kann, wenn deren Generationszeiten voneinander abweichen [23]. Dies alles kann zu falschen Ergebnissen führen, wenn man die falschen Markergene zur Rekonstruktion von Stammbäumen heranzieht.

Der universelle Baum des Lebens

Charles Darwin manifestierte in seiner Evolutionstheorie, dass sämtliche Organismen der Erde in einem universellen Stammbaum (*Tree of Life*) mit einem gemeinsamen Ursprung dargestellt werden können. Er glaubte somit, dass alles Leben von einem gemeinsamen Urvorfahren abstamme [24].

Ein solcher phylogenetischer Baum des Lebens, basierend auf rRNA Sequenzdaten, wurde von Carl Woese vorgeschlagen [24] und ist in Abbildung 3.2 dargestellt . Dieser zeigt deutlich die Auftrennung des Baumes in drei Hauptlinien, nämlich in die der *Bakterien*, der *Archaea* und der *Eukaryoten*. Dies sind die drei Domänen des Lebens. Die exakte Beziehung zwischen den drei Domänen war nicht immer ganz unumstritten. Heute ist diese Einteilung weitesgehend akzeptiert, da auch andere Bäume, die mit anderen Sequenzen erstellt wurden, ähnliche Ergebnisse geliefert haben.

Beispiel: Peer Bork's Baum des Lebens Erst kürzlich veröffentlichten Francesca Ciccarelli und Peer Bork (2006) eine Methode zur automatischen Erstellung eines Baum des Lebens auf Basis bestimmter Markerproteine [14] . Dabei untersuchte die Arbeitsgruppe alle *Cluster of Orthologous Groups* (COGs) der NCBI COG-Datenbank [15].

Jeder COG besteht aus individuellen, orthologen Proteinen, die in mehreren Organismen vorkommen. Für gewöhnlich ist jedem COG eine bestimmte Funktion zugewiesen, die auf konservierten Genen von mindestens drei phylogenetischen Abstammungslinien basiert. Daher eignen sich diese Proteine sowohl für eine funktionelle, als auch für eine evolutionäre Analyse. Die NCBI Datenbank enthält momentan (Mai 2007) 4872 COGs mit 192 987 Proteinen.

Wie oben beschrieben, müssen Proteine zur phylogenetischen Analyse bestimmte Kriterien erfüllen. COGs sind innerhalb ihrer Gruppe funktionell homolog und prinzipiell gut konserviert. Allerdings sind COGs nicht zwingend überall vorhanden und frei von horizontalem Gentransfer. Deswegen wurde nach COGs gesucht, die sowohl in vielen Organismen vorkommen als auch frei von horizontalem Gentransfer sind.

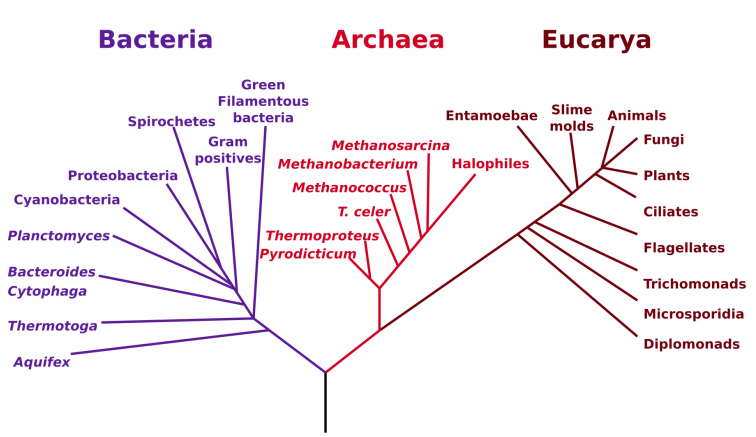


Abbildung 3.2.: Phylogenetischer Baum des Lebens nach Carl Woese, basierend auf 16S rRNA Sequenzdaten. Quelle: http://nai.nasa.gov/library/images/news_articles/274_3.jpg

Nach dieser Auswahl sind 31 orthologe Gruppen aus 191 Spezies aus allen drei Domänen übrig geblieben, die zur Erstellung eines neuen Baums des Lebens als geeignet erschienen. Mit Hilfe dieser 31 COGs wurde der — nachfolgend nur noch als Bork-Baum bezeichnete — Baum des Lebens berechnet (Abbildung 3.3). Die Einsatzmöglichkeiten des so entstandenen Baumes reichen von der klassischen Taxonomie bis hin zur Metagenomik, wo DNA Fragmente unbekannten phylogenetischen Ursprungs klassifiziert werden müssen.

3.1.3. Taxonomie

Im 18. Jahrhundert suchte der Botaniker Carl von Linné nach einer Ordnung in der Vielfalt des Lebens und war mit seinem Buch *Systema naturae* Begründer der formalen Taxonomie (aus dem Griechischen *taxis* = Ordnung; *nomos* = Gesetz). Das von Linné vorgeschlagene Modell ist eine im Fortgang späterer Forschung erweiterte Systematik zur Einteilung der Lebewesen. Dabei werden Arten hierarchisch in immer umfangreichere Organismengruppen klassifiziert. Nah verwandte Arten werden in die gleiche Gattung gestellt. Neben der Gruppierung nach Gattungen, umfasst die Taxonomie auch umfangreichere Klassifikationskategorien, wie Familie, Ordnung, Klasse, Stamm, Reich und Domäne (siehe Tabelle 3.1).

Entsprechend der von Linné eingeführten binären Nomenklatur, erhält eine Art einen zweiteiligen latinisierten Namen, der stets kursiv geschrieben wird. Der erste Teil des Namens bezeichnet die Gattung (*Genus*), zu der die Art gehört, und der zweite Teil den Artnamen (*Spezies*). Der Gattungsname wird dabei mit großem Anfangsbuchstaben geschrieben, wohingegen der Artname klein geschrieben wird. Ein Beispiel für die binäre Nomenklatur ist *Mus musculus*, zu deutsch: die Hausmaus.

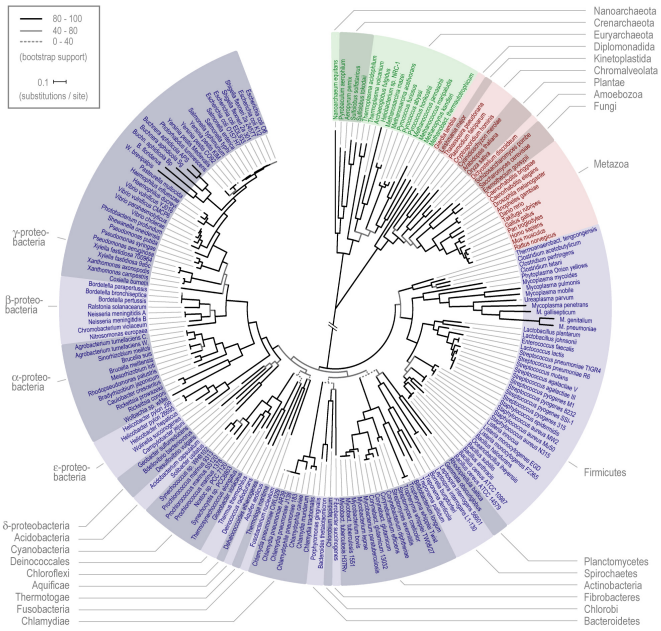


Abbildung 3.3.: Peer Borks Baum des Lebens [14]. Für eine vergrößerte Darstellung, siehe Anhang A.1.

Lateinisch	Deutsch	Englisch	Beispiel
Superregnum	Domäne	Superkingdom	Eukaryota
Regnum	Reich	Kingdom	Metazoa
Phylum	Stamm	Phylum	Chordata
Classis	Klasse	Class	Mammalia
Ordo	Ordnung	Order	Primates
Familia	Familie	Family	Hominidae
Genus	Gattung	Genus	Homo
Species	Art	Species	Homo sapiens

Tabelle 3.1.: Allgemeine taxonomische Hierarchie in aufsteigender Ordnung (von unten nach oben gesehen) mit Einordnung für *Homo sapiens*

Bakterielle Taxonomie und Taxonomie Kompendien

Die klassische Taxonomie basiert traditionell auf phänotypischen Untersuchungen. Diese phänotypischen Eigenschaften werden dazu verwendet, einen Organismus die taxonomische Leiter hinauf von der Spezies bis zur Domäne einzuordnen. Zu den Eigenschaften von taxonomischer Bedeutung gehören die Morphologie, Ernährung, Physiologie und das Habitat. Der klassische Artbegriff definiert eine Spezies als eine Gruppe von Organismen, deren Mitglieder sich kreuzen können und dabei fruchtbare Nachkommen hervorbringen. Diese Definition ist aber bei den asexuellen Organismen, wie den Prokaryoten, nicht anwendbar. Die molekulare Taxonomie benutzt daher molekulare Methoden, wie die genomische Hybridisierung oder die ribosomale rRNA Sequenzierung, um prokaryotische Spezies zu unterscheiden. Dabei betrachtet man einen Prokaryoten dann als einzigartige Spezies, wenn dessen 16S rRNA Sequenz sich um mindestens 3 % von allen anderen Arten unterscheidet [21]. Sequenzunterschiede von mehr als 5 % zu allen anderen Organismen unterstützen die Annahme, dass ein Organismus eine neue Gattung bildet. Bei der Identifizierung einer neuer Spezies ist es entscheidend, dass diese alle taxonomischen Kriterien der Ränge oberhalb ihrer Speziesbezeichnung erfüllt.

Wurden eine neue Spezies oder gar Gattung identifiziert, wird es in das *Bergey's Manual of Systematic Bacteriology* aufgenommen, dem wichtigsten taxonomischen Verzeichnis für Prokaryoten. Seit seiner Gründung im Jahre 1923 hat es weite Verbreitung in der Gemeinschaft der Mikrobiologen gefunden.

NCBI Taxonomie-Datenbank

Das *National Center for Biotechnology Information* (NCBI) hat eine Datenbank entwickelt, um Spezies taxonomisch zuordnen zu können [48]. Diese Datenbank vereint die Informationen verschiedenster primärer Datenquellen, wie dem oben beschriebenen *Bergey's Manual* oder der *RDP* Datenbank. So entstand ein einheitlicher taxonomischer Baum über alle drei Domänen, in dem jedem Taxon des Baumes eine eindeutige Kennung zugewiesen wurde. Ebenso wurden Viren in das taxonomische Verzeichnis mit aufgenommen, womit die NCBI-Taxonomie Datenbank derzeit an die 250 000 Taxa enthält (siehe Tabelle 3.2).

Beim NCBI Taxonomiebaum handelt es sich um einen bewurzelten Baum, der im Gegensatz zu den meisten phylogenetischen Bäumen nicht binär ist. Jeder Knoten besitzt eine eindeutige ID, seinen taxonomischen Rang (Spezies, Gattung, usw., siehe Tabelle 3.1) und den entsprechenden wissenschaftlichen Namen. Außerdem kennt jeder Knoten die ID seines Vaterknotens. Über diese Parent-ID wird die Struktur des Baumes festgelegt.

Ein Taxonomiebaum ist ständigen Veränderung ausgesetzt. Es kommen nicht nur neue Spezies hinzu, sondern es werden ganze Knoten gelöscht, im Baum verschoben oder verschmolzen. Daher wurde eine History eingeführt, die diese Veränderungen dokumentiert und bei Abfrage alter Daten auf diese Veränderungen verweist. Die Struktur der NCBI Taxonomie-Datenbank lässt sich somit durch drei Relationen darstellen (Abbildung 3.4).

Taxonomie Knoten	höhere Taxa	Gattung	Spezies	Niedrigere Taxa	Total
Archaea	86	99	462	59	706
Bacteria	904	1696	12323	3041	17964
Eukaryota	13924	39586	140994	10558	205061
Fungi	1031	3077	16631	857	21595
Metazoa	9996	23059	60153	5334	98542
Viridiplantae	1838	11636	58631	3784	75889
Viren	404	270	3952	20514	25140
Alle Taxa	15337	41658	161533	34199	**252726**

Tabelle 3.2.: Statistik über die in der NCBI Taxonomie-Datenbank enthaltenen Taxa (Stand: Mai 2007) [48]

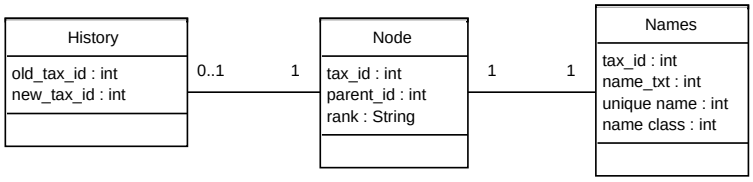


Abbildung 3.4.: Entity-Relationship-Diagramm der Architektur der NCBI Taxonomie-Datenbank.

3.2. Metagenomik

3.2.1. Was ist Metagenomik?

Ein Problem bei der Erforschung der mikrobiellen Diversität ist, dass ca. 99 % der Mikroorganismen nicht mit Standardtechniken kultiviert werden können [37]. Beispiele hierfür sind neben Mikroben, die nicht in purer Kultur gehalten werden können, auch Pathogene oder Symbionten, die nicht außerhalb ihres Wirtes existieren können, als auch fossile DNA. Es werden daher kultur-unabhängige Methoden benötigt, um die genetische Diversität, die Struktur der Gemeinschaft der Mikroorganismen und deren funktionelle ökologische Rolle zu verstehen. Es gibt eine vielversprechende Methode, die sogenannte *Metagenomik*, die das Potential hat, bei der Erforschung dieser Fragestellungen in der Mikrobiologie helfen zu können.

Die Metagenomik setzt dabei moderne Sequenziertechniken aus der Genomforschung ein, um Gemeinschaften nicht-kultivierbarer Mikroorganismen einer bestimmten ökologischen Nische zu untersuchen. Durch direkte Isolation und Klonierung der Genome der Mikroorganismen ist es möglich, auf die genetische Information des Habitats, dem *Metagenom*, zuzugreifen. Dies ermöglicht sowohl Einblicke in die Evolution, als auch auf die bisher unbekannten physiologischen Fähigkeiten von mikrobiellen Gemeinschaften, die sich auf eine ökologische Nische spezialisiert haben.

3.2.2. Ablauf eines Metagenomprojektes

Der Ablauf eines typischen Metagenomprojektes ist graphisch in der Abbildung 3.5 dargestellt. Man kann den Ablauf somit grob in drei Phasen einteilen: die DNA Extraktion, die DNA Klonierung und die Analyse.

DNA Extraktion und Klonierung

Umweltproben können DNA verschiedenster Art enthalten. Neben der DNA, der in der Probe lebenden Mikroorganismen, kann auch freiliegende DNA von bereits toten Organismen oder von Viren in der Probe vorhanden sein. Die DNA kann sich hierbei frei gelöst im Wasser befinden oder gebunden an festem Boden oder auch eingebettet in einem Biofilm. Die Extraktionsmethoden richtet sich also je nach Medium und gewünschter DNA Population. Wichtig ist, dass einem bewusst sein muss, dass man je nach Extraktionsmethode und Extraktionsprotokoll die natürliche Zusammensetzung des Metagenoms verändert [2]. Zum Beispiel werden Wasserproben typischerweise aufkonzentriert und gefiltert. Die Wahl der Siebgröße des Filters ist dabei von äußerster Wichtigkeit, da man damit je nach Wunsch eukaryotische (> 10 μm), prokaryotische ($> 0{,}1$ μm; < 10 μm) oder virale Einheiten ($< 0{,}1$ μm) anreichern kann.

Wurde die DNA extrahiert, kann sie direkt in Vektoren geklont werden. Dazu wird die DNA erst maschinell oder enzymatisch zerstückelt und dann in Vektoren eingefügt (*Ligation*). Üblicherweise werden in der Metagenomik *Bacterial Artificial Chromosomes* (BACs) als Vektoren benutzt, da sie DNA Fragmente von 80–120 kb aufnehmen können [3]. Die so entstandene Vektor-DNA wird dann zur Archivierung und Amplifizierung in einen Modellorganismus, normalerweise *E. coli*, eingebracht (*Transformation*). Zur nachfolgenden Analyse der metagenomischen Bibliothek lassen sich prinzipiell zwei verschiedene Ansätze unterscheiden, nämlich der sequenzbasierte und der funktionelle, oder aktivitätsbasierte Ansatz.

Aktivitätsbasierte Analyse

Das Ziel dieses Ansatzes ist es, Klone zu identifizieren, die eine Funktion des Metagenoms exprimieren. Diese Strategie bietet die einzige Möglichkeit, neue Klassen von Genen für neue oder bekannte Funktionen zu finden. Zum Beispiel wird zum Finden eines neuen Gens für eine bekannte Funktion die metagenomische DNA in einen Modellorganismus eingebracht, dem die gesuchte Funktion fehlt [4]. Die Wiederherstellung des für diese Funktion charakteristischen Phänotyps kann dann zur Identifizierung des Gens herangezogen werden. Der Erfolg dieses Ansatzes ist aber davon abhängig, ob die Expression der Gene überhaupt korrekt erfolgt und ausreichend stark ist [27].

Sequenzbasierte Analyse

Die Auswahl der zu sequenzierenden Klone kann anhand eines phylogenetischen Markers geschehen oder durch zufällige Auswahl. Erstgenannte Methode eignet sich dazu, möglichst viele genomische Fragmente eines gewünschten Taxons zu sammeln und zu sequen-

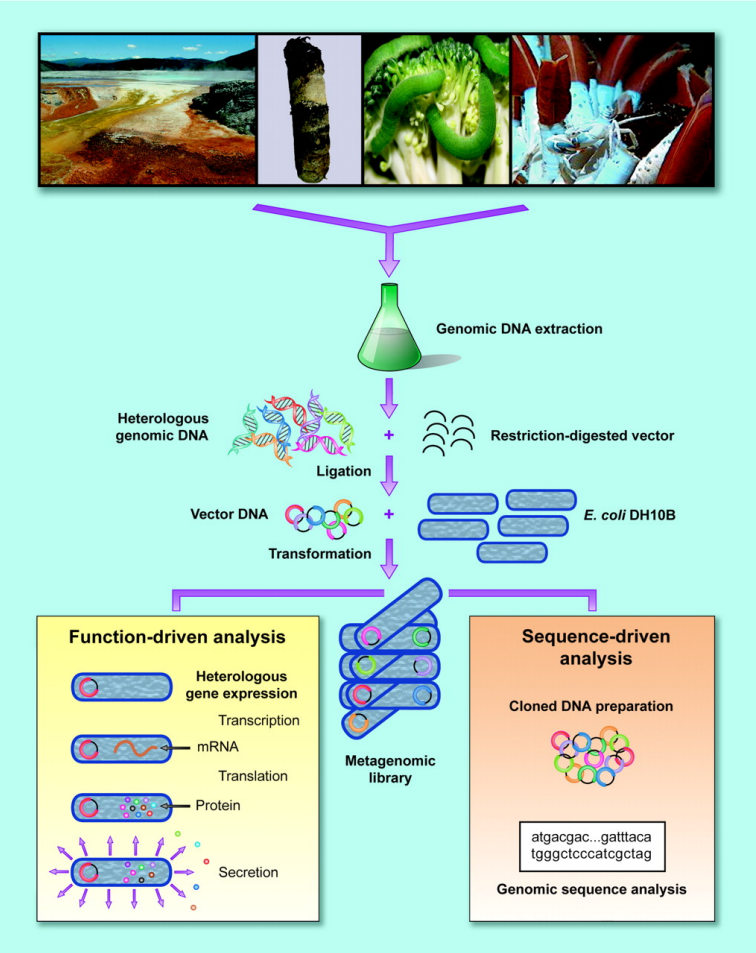


Abbildung 3.5.: Schematische Darstellung des Ablaufes eine Metagenomikprojektes, das aus Konstruktion und Analyse von Datensätzen aus einer Umweltprobe besteht. Die Bilderreihe oben zeigt dabei verschiedene Biotope, aus denen die Probe stammen könnte. Quelle: [27]

zieren. Die Möglichkeiten der Methode werden allerdings durch die begrenzte Anzahl an verlässlichen Markern eingeschränkt.

Die zufällige Sequenzierung von Klonen, als *Whole-Genome Shotgun* (WGS) Sequenzierung bezeichnet, bietet, im Gegensatz zur Marker-Methode, eine globalere Sicht auf das Metagenom. Dies schließt eine weitreichendere Untersuchung der phylogenetischen Diversität, der metabolischen Vorgänge und in einigen Fällen die Rekonstruktion von Genomen, mit ein [5]. Diese Methode hat sich vor allem durch den Fortschritt in der rechnergestützten Verarbeitung der sequenzierten Daten etablieren können. In meinen weiteren Ausführungen werde ich den Fokus auf die WGS Methodik legen.

Rekonstruktion von Genomen Eine Sequenzierreaktion liefert üblicherweise DNA Fragmente der Länge 600–1000 bp [29]. Computerprogramme können diese kurzen Fragmente bei vorhandenen Sequenzüberlappungen zu größeren kontinuierlichen Sequenzen, zu sogenannten *Contigs*, zusammensetzen (*Assemblieren*). Ein weiteres Gruppieren der Contigs zu *Scaffolds* ist durch das *Paired-End Verfahren* möglich [29]. In Kapitel 3.4 wird näher auf das Thema Rekonstruktion von Genomen anhand von Contigs und deren Merkmale eingegangen.

Da man bei dem WGS Ansatz immer den selben Primer benutzt, und die Auswahl der Klone zufällig ist, stellt sich die statistische Frage wieviel Sequenzierreaktionen man durchführen muss, um alle Nukleotide eines DNA Pools sequenziert zu haben. Dazu wurde der Begriff der *Coverage* eingeführt. Die Coverage ist in der Metagenomik von großer Bedeutung, da sich ein Metagenom aus vielen unterschiedlich vorherrschenden Arten zusammensetzt. Eine Rekonstruktion eines Genoms ist nämlich nur dann möglich, wenn es einen Großteil der Gemeinschaft ausmacht [5]. Daher ist rückblickend die Wahl der Extraktionsmethode von Wichtigkeit. Das Beibehalten der ursprüngliche Zusammensetzung des Metagenoms ist also schlecht, wenn man eine komplette Sequenzierung von bestimmten Genomen anstrebt.

Phylogenetische und funktionelle Analyse Beiden Fragestellungen ist gemein, dass man zuerst in der sequenzierten DNA nach Genen fahndet. Dies geschieht meist in einem groben Ansatz durch das Suchen von *Open-Reading-Frames* (ORFs). Ein ORF ist die Sequenz, die zwischen einem Start- und Stoppcodon liegt. Danach wird versucht, diesen ORFs eine Funktion zuzuweisen.

Diese ORFs können dann verwendet werden, um sowohl Aufschluss über das phylogenetische Spektrum als auch über das funktionelle Repertoire des Metagenoms zu bekommen. Verschiedene zur phylogenetischen Analyse benutzte Verfahren werden im Kapitel 3.3 vorgestellt.

3.2.3. Beispiele für WGS Metagenomik Projekte

Das erste großangelegte WGS basierte Umweltsequenzierprojekt untersuchte die Organismen eines extrem säurehaltigen Biofilmes in einer Eisenmine (Tyson et al. 2004). Nur einen Monat später wurde ein weitaus komplexerer Metagenomdatensatz veröffentlicht (Venter et al. 2004), der aus der Sargasso-See entnommen wurde. Im Jahre 2005 wurden

	Eisenmine	Sargasso-See	Erdboden	Walüberreste
Sequenzierte Menge	76 Mb	1350 Mb	104 Mb	78 Mb
Durchschnittl. Read-Länge	737 bp	818 bp	696 bp	673 bp
Nicht assemblierten Reads [%]	20	40	99	55
Assemblierte Genome	5	3	0	0
Annotierte ORFs	>12000	>1000000	>180000	>120000
Gefundene Spezies	5	1000	847	17
Geschätzte Anzahl an Spezies	keine Schätzung	>1 800	>3 000	25–150

Tabelle 3.3.: Statistiken über großangelegte Umweltsequenzierprojekte (Stand: Februar 2006) [25]

zwei weitere WGS Umweltdatensätze, aus sehr unterschiedlichen Habitaten, veröffentlicht. Einen Überblick über alle genannten Datensätze gibt Tabelle 3.3.

Nach dem erfolgreichen Pilotprojekt in der Sargasso See, machte sich J.C. Venter und sein Team erneut auf den Weg, die mikrobielle Vielfalt der Weltmeere zu erforschen. Die *Global Ocean Sampling* (GOS) Expedition startete im August 2003 und führte das Forschungsboot *Sorcerer II* zwei Jahre lang um die ganze Welt. Die aus dem GOS Projekt in [30] veröffentlichten Daten umfassen 7,7 Millionen Reads aus 41 verschiedenen Proben.

Um dieser Datenflut Herr zu werden, wurden daher bereits zwei Datenmanagementsysteme für Metagenomik-Projekte ins Leben gerufen. Das *Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis* Datenmanagementsystem, kurz CAMERA[2] [38] , wurde für Meeres-Metagenome, wie dem GOS Projekt, entwickelt und speichert neben den sequenzierten Metagenomen auch Metadaten, wie Ort der Probennahme, Wassertemperatur, Uhrzeit, Extraktionsprotokolle etc. Ein anderes System, das *Integrated Microbial Genomes / Metagenomics* (IMG/M) [19] legt seinen Schwerpunkt auf integrierte Analyseverfahren und weniger auf Erfassung und Speicherung der Metadaten. Aber nicht nur die Extraktionsmethoden und Protokolle variieren, sondern auch die Daten selbst.

Ein wesentlicher Faktor, der berücksichtigt werden muss, ist, dass der Artenreichtum vom Biotop abhängt. Dies zeigen die Schätzungen für die Anzahl an Spezies in den verschiedenen Metagenomen. Man geht davon aus, dass in 0,5 g Boden mehr Arten enthalten sind, als in mehreren hundert Litern Wasser [25]. Ebenso korreliert auch die Genomgröße der enthaltenen Spezies mit ihrem Habitat [3]. Genome extremer ökologische Nischen sind kleiner als Genome komplexer Umweltnischen. Für die Komplexität einer Bodenprobe ergeben sich noch zwei andere Probleme. Anhand der Assemblierungsrate, die unter 1 % liegt (Tabelle 3.3), kann man erkennen, dass kein Organismus diese ökologische Nische signifikant dominiert. Schätzungen der Autoren selbst (Tringe et al. 2005) gehen davon aus, das man ein Fünfaches sequenzieren müsste, um das dominanteste Genom vernünftig abgedecken zu können [25]. Auch konnte für ca. 47 % der ORFs kein homologer Treffer gefunden werden. Dies zeigt, dass Metagenome komplett neue Klassen von

[2]Zu erreichen unter: http://camera.calit2.net/

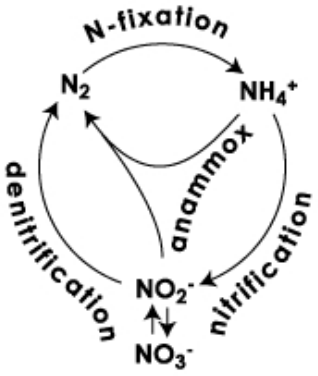


Abbildung 3.6.: Stickstoffzyklus mit den Pfaden für die anaerobe Ammonium Oxidierung. Quelle: http://www.anammox.com/

Genen enthalten und die vorhandenen Datenbanken bei weitem nicht ausreichen, um Metagenom-Gene sinnvoll zu annotieren.

Dies mag problematisch klingen. Diese neue genetische Flut beschert jedoch auch neues Potential, unbekannte Proteine und Stoffwechselwege zu entdecken und zu erforschen. Die Vorstellung des nachfolgenden Metagenoms soll hierbei als Beispiel dienen.

3.2.4. Anammox-Community Metagenom

Mitte der 1960er bemerkte Francis Richards von der University of Washington in Seattle, dass in einem anoxischem Fjord der Ammoniumgehalt unerklärbar gering war. Er vermutete, dass das Ammonium entweder anorganisch oder von Mikroorganismen anaerobisch zu Stickstoff umgewandelt werden müsse [32]. Ein Jahrzehnt später postulierte der Wiener Physikochemiker Engelbert Broda die Existenz solcher, zu dieser Zeit noch unbekannten Bakterien [34]. Wiederrum eine Dekade später führte ein Zufall in einer Hefefabrik im niederländischen Delft zur Untersuchung der anaerobischen Ammonium-Oxidierung (kurz: Anammox) durch den Wissenschaftler Gijs Kuenen [32]. Im Jahre 1999 gelang es dann holländischen Forschern um Mike Jetten und Marc Strous von der Radboud University in Nijmegen ein solches Anammox-Bakterium tatsächlich zu identifizieren [33]. Mit der Entdeckung des Bakteriums *Kuenenia stuttgartiensis*, das dem Phylum der Planctomyceten zugeordnet wurde, bewiesen sie nicht nur die Existenz eine solchen Bakteriums, sondern vor allem auch, dass es im globalen Stickstoffzyklus einen neuen Mitspieler gibt (Abbildung 3.6). Anammox nimmt in diesem Zyklus eine Abkürzung, indem es direkt von Ammonium und Nitrit zu molekularem Stickstoff übergeht. Daher ist dieses Bakterium nicht nur für Genetiker und Mikrobiologen interessant, sondern auch für Ökologen, Biotechnologen und andere. Man nimmt heute an, dass Anammox-Prozesse mit einem Anteil von bis zu 50 % an der Entfernung von fixiertem Stickstoff in den Weltmeeren beteiligt sind [1].

Zur näheren Untersuchung des Bakteriums wagte sich die Gruppe um Marc Strous

schließlich an die Sequenzierung des Genoms von *Kuenenia stuttgartiensis*. Unglücklicherweise teilen sich Anammox-Bakterien nur alle zwei Wochen und sind nicht isoliert kultivierbar. Daher wandte man die Methoden der Metagenomik an. Um eine möglichst große Kultur an *Kuenenia stuttgartiensis* zu züchten, setzte man einen Bioreaktor an. Nach 15 Generationen (ca. 1 Jahr) konnte *K. stuttgartiensis* mit einem Anteil von 73±5 % im Reaktor angereichert werden. Aus der nachfolgenden Sequenzierung des Metagenoms konnte das 4,2 Mb große Genom zu 98 % geschlossen werden und daraus die biochemischen Pfade des Anammox-Prozesses abgeleitet werden [1]. Eine phylogenetische Analyse des Metagenoms ergab außerdem, dass sich mindestens weitere 29 taxonomische Einheiten im Anammox-Community Genom befinden. Eine Diversitätsabschätzung ergab, dass sich weitere 70 bis 100 Taxa im Metagenom befinden.

Dieses Metagenom gibt also noch genügend Anlass zur weiteren Untersuchung.

3.3. Taxonomische Analyse von Metagenomen

Die Ergründung der in einem Metagenom vorhandenen taxonomischen Einheiten und deren phylogenetische Einordnung und Verwandtschaft zueinander, ist eines der wichtigsten Ziele in der mikrobiellen Analyse eines Metagenoms. Die Metagenomik versucht, neben der Frage nach dem „Wer ist da?“ , auch die Frage nach deren quantitativen Anteile im Metagenom zu beantworten. In den folgenden drei Unterkapiteln werden drei verschiedene Ansätze vorgestellt. Auf letztem basiert die Aufgabenstellung dieser Diplomarbeit.

3.3.1. SSU rRNA Methode

Dieser Ansatz hat sich als Methode der Wahl zur phylogenetischen Einordnung von Organismen in der Mikrobiologie bewährt. Sie basiert auf der Betrachtung der Homologien von Markermolekülen auf DNA- oder Protein-Ebene.

Wie bereits im Kapitel zur Phylogenetik angeführt, eignen sich nur bestimmte Gene und Proteine zur Messung phylogenetischer Verwandtschaft. Man hat viele Gene und Proteine als evolutionäre Zeitmesser vorgeschlagen. Gene, die für ribosomale RNAs codieren, sind die in der Mikrobiologie am häufigsten verwendeten Zeitmesser, da sie einige hervorragenden Eigenschaften besitzen. Sie sind funktionell konstant, universell verteilt und enthalten mehrere Regionen, in denen die Nukleotidsequenz konserviert ist. 5S, 16S und 23S rRNAs bezeichnen Moleküle, die Teile der Untereinheiten eines prokaryotischen Ribosoms sind [21]. Da 16S ein Teil der kleinen Untereinheit eines Ribosoms ist, ist die Abkürzung SSU (engl. small subunit) rRNA ein Synonym für 16S Sequenzierung.

Mit Hilfe dieser SSU rRNA Sequenz kann man einen bekannten Organismus in einem Metagenom identifizieren oder einen unbekannten in seinen evolutionären Kontext platzieren. Es ist allerdings zu beachten, dass eine Aussage für einen Read nur dann gemacht werden kann, wenn dieser auch einen Marker beherbergt. Daher wird diese Methode auch meist in der Metagenomik dazu verwendet, einen Einblick in die breite Diversität einer mikrobiellen Gemeinschaft zu bekommen.

Dazu wird die DNA direkt aus der Umweltprobe extrahiert (Abbildung 3.3.1). Die 16S Gene in der Probe werden mittels *Polymerasekettenreaktion* (PCR) vermehrt. Diese

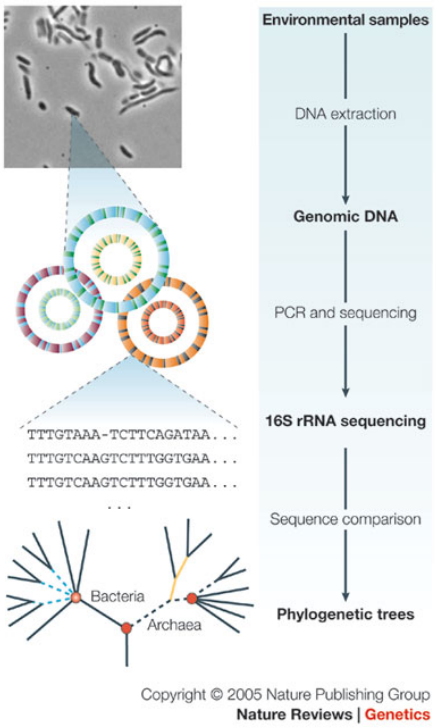


Abbildung 3.7.: Verlauf einer 16S ribosomalen Analyse einer mikrobiellen Gemeinschaft. Quelle: [2]

Vorgehensweise ist sowohl schnell als auch spezifisch, da man zur gezielten Amplifizierung der konservierten Sequenzen universelle Primer benutzen kann. Die PCR Produkte werden sequenziert und stehen für Computeranalysen bereit.

Ein Vergleich dieser Sequenzen mit Datenbanken von 16S rRNA Genen, erlaubt eine phylogenetische Klassifizierung. Zeigt keiner der Vergleiche für eine 16S rRNA eine Ähnlichkeit von mehr als 97 %, so entstammt diese Sequenz per Definition einer neuen Spezies [21]. Multiple Alignments bewerkstelligen diese Vergleiche. Aus den Ergebnissen kann mit Hilfe eines Algorithmus ein phylogenetischer Baum erstellt werden , der am besten die in den Sequenzen evolutionäre Information beschreibt. Die Erstellung eines solchen Baumes kann jedoch in Abhängigkeit des verwendeten Algorithmus ein sehr rechenintensives Unterfangen (Gleichung (3.1)) sein. Am Ende steht jedoch eine verlässliche taxonomische und phylogenetische Einordnung eines Organismus und seiner Markersequenz.

Die Häufigkeit von bestimmten SSU rRNAs lassen einen groben Schluss über die Struktur der mikrobiellen Gemeinschaft zu, da 16S Sequenzen von dominanten Organismen öfter vorhanden sein sollten. Eine solche quantitative Analyse mit Hilfe verschiedenster Marker, wurde von J.C. Venter für das Sargasso Metagenom durchgeführt. Die Ergebnisse können im Anhang A.3 eingesehen werden.

Zusammenfassend lässt sich sagen, dass dieser Ansatz die zuverlässigsten Ergebnisse

aller Methoden liefert. Er ist aber auch, aufgrund der rechenintensiven Konstruktion der Bäume, der aufwändigste. Zudem lässt sich eine taxonomische oder phylogenetische Aussage über eine Sequenz nur dann machen, wenn diese ein Markergen besitzt. Der Anteil an Markergenen im einem Genom beträgt üblichweise weniger als 5 % [8].

3.3.2. Best-Blast-Hit Methode

Die wohl naheliegenste Methode, einer Sequenz eine taxonomische Zuordnung zu geben, besteht darin, die besten Homologen in öffentlichen Sequenzdatenbanken zu suchen, da man davon ausgehen kann, dass Homologien auf Verwandtschaftsverhältnisse hindeuten. Dieser Ansatz umgeht das Problem, dass ein phylogenetischer Marker vorhanden sein muss. Es lässt sich somit zu fast allen unbekannten Sequenzen, die ein Homologes in einer Datenbank haben, eine taxonomische Aussage machen. Der Vergleich einer unbekannten Sequenz mit dem Inhalt einer Sequenzdatenbank, geschieht üblicherweise mit einem Alignment-Tool wie BLAST [35] oder FASTA [36]. Zu dem Vorteil, dass dieser Ansatz pro Sequenz zudem auch schneller ist als die Berechnung eines phylogenetisches Baumes, gesellen sich aber auch Nachteile.

Ergebnisse einer Best-Blast-Hit Analyse sind nämlich nur dann zuverlässig, wenn Verwandte in der Datenbank vorhanden sind. Sie sind aber gänzlich nutzlos, wenn kein Verwandter vorhanden ist [2]. Die Resultate hängen aber nicht nur von der inhaltlichen Ausrichtung der Datenbank ab, sondern auch davon, ob die zu untersuchende Sequenz von horizontalem Gentransfer beeinflusst wurde. Ersterem Nachteil kann abgeholfen werden, indem man eine möglichst universelle und große Datenbank benutzt, wobei zu beachten ist, dass dann der Aufwand der Suche linear mit der Größe der Datenbank steigt. Für besonders große Datensätze können dann die Sequenzvergleiche zum Engpass der Methode werden.

Ein weiterer großer Nachteil der Methode ist, dass man nur eine bestimmte Anzahl an Homologen betrachtet, z.B. nur den Besten oder die zwei Besten etc., und damit die Information zu den schlechteren Treffern ignoriert [20].

MEGAN

Vor kurzem veröffentlichte Daniel Huson von der Universität Tübingen die Software MEGAN (Metagenome Analyzer), ein neues Programm auf Basis der Best-Blast-Hit Methode zur taxonomischen Analyse von Sequenzdatensätzen [18].

In einem vorverarbeitendem Schritt muss ein Sequenzvergleich für alle Reads des Datensatzes mit einer geeigneten Referenzdatenbank vollzogen werden. Für gewöhnlich wird dazu BlastN oder BlastX zur Suche in der NCBI-nt[3] oder NCBI-nr[4] Sequenzdatenbank benutzt.

[3] NCBI Nukleotidsequenzdatenbank bestehend aus allen GenBank+RefSeq Nukleotiden+EMBL+DDBJ+PDB Einträgen

[4] NCBI Peptidsequenzdatenbank bestehend aus allen nicht-redundanten GenBank CDS Translationen+RefSeq Proteinen+PDB+SwissProt+PIR+PRF

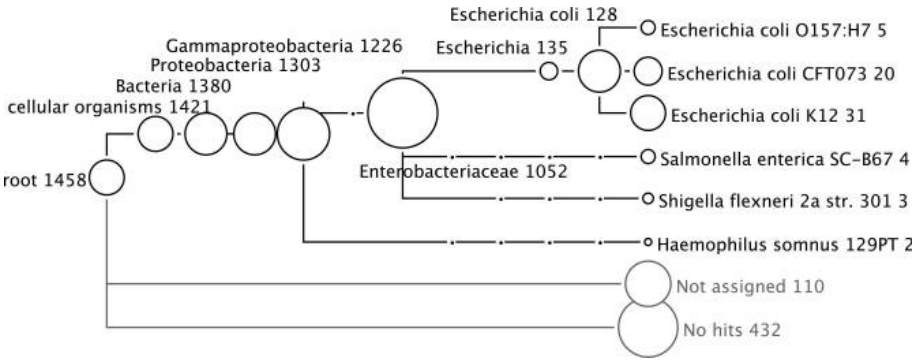


Abbildung 3.8.: MEGAN Analyse von 2000 Reads aus E. coli K12, basierend auf BlastX und der NCBI-nr Datenbank. Quelle: [18]

Die Sequenzvergleiche werden nach MEGAN importiert und MEGAN benutzt dann diese, um den taxonomischen Inhalt des Datensatzes zu berechnen und zu erforschen. Hierzu wird beim Programmstart zunächst der NCBI-Taxonomie Baum geladen. Ein einfacher Algorithmus ermittelt nun für jeden Read den letzten gemeinsamen Vorfahren für die Menge an Taxa, die im Sequenzvergleich für diesen Read gefunden wurden. Die Ergebnisse können durch einige Parameter beeinflusst werden. Ein Min-Score Filter setzt den Mindestwert, den ein Alignment haben muss, um von dem Algorithmus berücksichtigt zu werden. Dies ist ein sehr wichtiger Parameter, da er direkten Einfluss auf zu Zuweisung eines Taxons hat. Wählt man den Score zu niedrig, werden auch schwache, weit entfernte Homologien berücksichtigt, was dazu führt, dass diese Sequenz zu weit oben im Taxonomiebaum eingeordnet wird. Der Min-Support Filter wird dazu benutzt Falsch-Positive zu verwerfen, indem er einen Schwellenwert für die minimale Anzahl an zugewiesenen Reads für ein Taxon im Baum setzt.

Das Resultat der Berechnung wird dem Benutzer graphisch in Form einer Baumstruktur präsentiert (Abbildung 3.8). Der Radius eines Kreises im Baum ist der Menge an Reads proportional, die diesem Taxon zugewiesen wurden. Ebenso erlaubt die Software die Ergebnisse der Berechnung einzusehen und zu exportieren. Von Interesse sind hierbei vor allem die Liste an Reads, die einem Taxon zugeordnet worden sind und umgekehrt das Taxon, das einem Read zugewiesen worden ist.

Die Entwickler konnten außerdem zeigen, dass die Methode dazu fähig ist, auch noch Fragmente mit einer Kürze von 100 bp mit ausreichender Korrektheit zu klassifizieren, was sie für die Auswertung von Datensätze aus modernen Sequenzierverfahren, wie der Pyrosequenzierung, attraktiv macht.

3.3.3. Blast2Tree

Blast2Tree ist eine phylogenetische Methode zur taxonomischen Klassifizierung, die Dominik Lindner in seiner Diplomarbeit entwickelt hat [20]. Sie ist deshalb von Vorteil, da in einem Metagenom meist eine Vielzahl von noch nicht sequenzierten Genomen im Ver-

borgenen liegen und man daher nur Verwandtschaftsaussagen machen kann. Da aber die Rekonstruktion von phylogenetischen Bäumen sehr aufwendig ist, besteht die Idee darin, eine Sequenz in einen bereits vorhandenen phylogenetischen Speziesbaum einzuordnen. Eine unbekannte Sequenz kann jedoch nur dann taxonomisch zugeordnet werden, wenn sie ein Markergen beherbergt, das zu den Markergenen homolog ist, aus denen der Speziesbaum erstellt wurde. Man kann somit folgenden Voraussetzungen zusammenfassen:

- phylogenetischer Speziesbaum mit Angabe der Zweiglängen
- Datenbank von Markergenen, aus dem Speziesbaum, mit folgenden Eigenschaften:
  - universell verteilt
  - gut konserviert
  - frei von horizontalem Gentransfer

Dominik Lindner verwendete für seine Zwecke den Baum des Lebens der Bork Gruppe [14] , der auf 31 Gruppen orthologer Proteine basiert (Siehe Kapitel 3.1.2).

Vorbereitung des Speziesbaumes

Zu jedem Knoten im Baum wird ein Distanzvektor aus den gegebenen Zweiglängen berechnet, der die Distanzen zu allen Blattknoten des Baumes angibt (Abbildung 3.9). Aus Gleichung (3.2) folgt die Gesamtanzahl an Distanzvektoren für einen binären, gewurzelten Baum in Abhängigkeit von dessen Anzahl an Blattknoten.

$$N_{total} = 2 * N_{Blatt} - 1 \tag{3.2}$$

N_{total}: Gesamtanzahl an Knoten für einen binären, gewurzelten Baum; N_{Blatt}: Anzahl an Blattknoten

Zusätzlich zur Berechnung der Distanzvektoren, muss jeder Knoten eine taxonomische Bezeichnung besitzen. Da Blattknoten bereits Spezies repräsentieren, ist es kein Problem, ihnen eine taxonomische Bezeichnung zu geben. Inneren Knoten, falls sie noch keine taxonomische Benennung besitzen, wird eine NCBI Taxonomie-ID nach Algorithmus 1 zugeteilt .

Algorithmus 1

Für jeden inneren Knoten K_{innen} ***:***

1. *Bestimme die Menge* M *aller Blattknoten, die unterhalb von* K_{innen} *liegen.*
2. *Starte im NCBI-Baum bei einem Blattknoten aus* M *und steige in der taxonomischen Hierarchie soweit auf, bis ein NCBI-Taxon* T *alle Taxa aus* M *enthält.*
3. *Weise dem Knoten* K_{innen} *das Taxon* T *zu.*

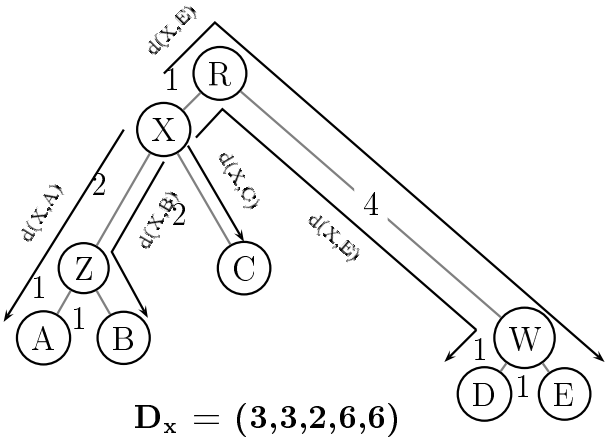


Abbildung 3.9.: Beispiel: Berechnung des Distanzvektors für Knoten X. Die Distanzen ergeben sich aus der Summe der einzelnen Zweiglängen zu den Blattknoten. Quelle: [20]

Des Weiteren muss für jeden Knoten die durchschnittliche Distanz zu seinen Blattknoten ermittelt werden. Diese entspricht der durchschnittlichen evolutionären Distanz, die eine neue und unbekannte Spezies von eben diesem Vorfahren — repräsentiert durch den betrachteten Knoten — entfernt ist.

Nach der Vorbereitung muss für jeden Knoten des Baumes folgendes vorhanden sein:

- Individueller Distanzvektor mit den Distanzen zu allen Blattknoten des Baumes
- Taxonomische Bezeichnung durch eine NCBI Taxonomie-ID
- Durchschnittliche Distanz zu den Blattknoten, die unterhalb des betrachteten Knotens liegen

Genvorhersage

Wie bereits erwähnt, kann eine Sequenz nur dann klassifiziert werden, wenn sie ein homologes Gen zu den Markergenen des phylogenetischen Baumes enthält. Dazu müssen erstmal alle Gene auf der Sequenz identifiziert werden und auf Homologie getestet werden.

Eine notwendige Bedingung, dass eine DNA-Sequenz für ein Protein kodiert, ist, dass die Sequenz einen Open-Reading-Frame, kurz ORF, bildet . Ein ORF ist eine Region innerhalb einer DNA-Sequenz, die mit einem Startcodon beginnt (z.B ATG) und einem Stoppcodon (z.B. TAA) endet [29]. Bakterien unterscheiden sich aber häufig darin, welches Startcodon sie benutzen. Daher verwendet man bei der Suche nach ORFs in Bakterien oft die Definition, dass ein ORF zwischen zwei Stoppcodons liegt. Es ist anzunehmen, dass sich in Metagenomen hauptsächlich Bakterien befinden werden, und da in Bakterien nahezu das gesamte Genom für Gene kodiert und es somit kaum nicht kodierende DNA-Abschnitte gibt [7], ist diese Definition für einen ORF durchaus verwendbar. Zu beachten ist aber, dass ein Gen bei der Sequenzierung durchtrennt worden

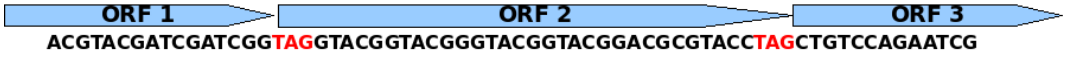


Abbildung 3.10.: Eine DNA-Sequenz mit drei gefundenen ORFs für einen Leserahmen. ORF 2 liegt nach Definition zwischen zwei Stoppcodons. ORF 1 und 3 sind partielle ORFs, da deren Anfang bzw. Ende nicht auf diesem Contig liegen.

sein kann, das heißt, nur partiell auf dem Contig vorhanden ist und somit nicht zwischen zwei Stoppcodons liegen kann (Abbildung 3.10). Dieser Fall gilt als partieller ORF und wird bei der späteren Klassifizierung gesondert behandelt .

Ein weiterer wichtiger Parameter bei der Genvorhersage ist die minimale Länge eines ORFs. Einerseits will man kein kleines biologisch relevantes Gen übersehen, andererseits kann ein zu kurzer ORF auch zufallsbedingt auftreten [7]. Eine sinnvolle Mindestlänge, die bei kompletten Genomen zur Suche von ORFs verwendet wird ist 150 bis 200 Nukleotide. Da aber im Falle von einzelnen Contigs partielle ORFs auftreten, die durchschnittlich kürzer als komplette ORFs sind, aber dennoch biologisch relevant sein können, verwendete Dominik Lindner eine Mindestlänge von 100 Nukleotiden [20].

Die Suche nach kompletten und partiellen ORFs auf der gegebenen DNA-Sequenz muss auf allen sechs Leserahmen durchgeführt werden. Eine mögliche Homologie eines gefundenen ORFs zu den Genen aus dem Markergensatz kann schnell durch Programme wie BLAST [35] oder FASTA [36] festgestellt werden.

Berechnung der ORF-Distanzen

Durch die die Genvorhersage liegen jetzt nur noch die interessanten (zu Markergenen homologe) ORFs vor. Anhand der Homologien zu den Markergenen des Baumes kann für jeden ORF ein Distanzvektor *do* berechnet werden:

$$\text{Distanzvektor } do = (do_0, do_1, do_2, ..., do_n)$$

$$do_i = -\begin{cases} ln(\frac{S}{S_s}) & \text{für: partiellen ORF} \\ ln(\frac{\frac{S}{S_s}+\frac{S}{S_{sm}}}{2}) & \text{für: kompletten ORF} \end{cases} \tag{3.3}$$

n: Anzahl der Spezies/Blattknoten im Baum; do_i: evolutionäre Distanz des ORFs zu Spezies i; S: Smith-Waterman Alignment Score zwischen ORF und Markergen der Spezies i; S_s: Selfscore des ORFs; S_{sm}: Selfscore des Markergens der Spezies i

Dazu wird der ORF mit den Markergenen aller Spezies (Blattknoten) des Baumes aligniert. Die Alignment-Scores werden dann nach Gleichung (3.3) in Distanzen umgewandelt [20]. Gibt es mehrere homologe Gene pro Spezies und ORF, wird das mit dem höchsten Score verwendet. So erhält man für jeden ORF einen Vektor, der die evolutionären Distanzen des ORFs und dem Organismus aus dem er stammt, zu den Spezies im Baum wiederspiegelt.

Taxonomische Zuordnung

Eine taxonomische Aussage wird durch die Einordnung des ORFs anhand seines Distanzvektors zu einem der Knoten des Baum möglich. Dazu wird der Distanzvektor eines ORFs mit den Distanzvektoren der Baumknoten verglichen. Der Knoten mit dem ähnlichsten Distanzvektor wird ausgewählt und das Taxon dieses Knotens dem ORF zugewiesen.

Bevor die Distanzvektoren jedoch miteinander verglichen werden können, muss der Distanzvektor des ORFs jedem Knoten des Baumes mathematisch angepasst werden. Dies hat allem voran zwei Gründe. Zum einen ist es sehr wahrscheinlich, dass die Spezies, aus dem der ORF stammt, nicht im Baum vorhanden ist. Daher ist sicher anzunehmen, dass die neue Spezies einen Vorfahr besitzt, der durch einen inneren Knoten im Baum repräsentiert wird. Die bereits für jeden Knoten bestimmte durchschnittliche evolutionäre Distanz zu den Blattknoten gibt an, wie weit diese neue Spezies wahrscheinlich von dem betrachteten Vorfahr entfernt ist. Diese Tatsache fließt in die mathematische Skalierung des Distanzvektors durch einen Summanden s ein.

Zum anderen besitzen die Distanzvektoren aus ORF und Knoten verschiedene Wertebereiche. Die Werte des ORFs werden daher durch einen Faktor f auf das Niveau des Knotenvektors k gebracht (Formel (3.4)) [20]. Dies kann nur unter der Annahme geschehen, dass die Summen aller Astlängen in Genbaum und Referenzbaum gleich sind. Die Anpassung des ORF-Vektors muss für jeden Knoten des Baumes separat geschehen, da jeder Knoten eine andere durchschnittliche evolutionäre Distanz zu seinen Blattknoten besitzt.

$$f_k = \frac{\sum_{i=0}^{n} dk_i + s_k}{\sum_{i=0}^{n} do_i} \tag{3.4}$$

f_k: Multiplikator; n: Anzahl der Spezies/Blattknoten im Baum; dk_i: evolutionäre Distanz des Knotens k zu Spezies i; s_k: durchschnittliche evolutionäre Distanz des Knotens k zu seinen Blattknoten; do_i: evolutionäre Distanz des ORFs zu Spezies i; [5]

$$do' = (do_0 * f_k - s_k, do_1 * f_k - s_k, do_2 * f_k - s_k, ..., do_n * f_k - s_k) \tag{3.5}$$

do': skalierter ORF Distanzvektor; n: Anzahl der Spezies/Blattknoten im Baum; f_k: Multiplikator; s_k: durchschnittliche evolutionäre Distanz des Knotens k zu seinen Blattknoten; do_i: evolutionäre Distanz des ORFs zu Spezies i; [6]

Nach Anwendung der Parameter f_k und s_k nach Formel (3.5) auf die Rohdistanzen des ORF-Vektors, ist dieser mit dem Wertevektor des jeweiligen Knotens k vergleichbar [20]. Nach Wiederholung der Skalierung für alle Baumknoten, kann der ähnlichste Knoten über die Summe der Fehlerquadrate bestimmt werden (Formel (3.6)). Die Knoten mit der kleinsten Summe — mit dem kleinsten Fehler e — sind dem ORF am ähnlichsten.

[5] die Formel aus Quelle [20] wurde korrigiert
[6] die Formel wurde gegenüber der aus Quelle [20] korrigiert

$$e = \sum_{i=0}^{n} (dk_i - do'_i)^2 \tag{3.6}$$

e: Fehler zwischen ORF- und Knotenvektor; n: Anzahl der Spezies/Blattknoten im Baum; do'_i: skalierte evolutionäre Distanz des ORFs zu Spezies i; dk_i: evolutionäre Distanz des Knotens k zu Spezies i

Die gesamte Prozedur ist in Algorithmus 2 noch einmal kurz dargestellt.

Algorithmus 2

Allgemein gegeben: ORF-Distanzvektor do

Für jeden Knoten** k**:

- *Gegeben: durchschnittliche evolutionäre Distanz s_k des Knotens k zu seinen Blattknoten*
- *Gesucht: skalierter ORF-Distanzvektor do' zu Knoten k*

1. *Berechne Faktor f_k nach Formel* (3.4)
2. *Skaliere do auf Niveau von k nach Formel* (3.5)
3. *Messe Ähnlichkeit/Fehler e_k zwischen do' und dk nach Formel* (3.6)

Weise dem ORF das Taxon von Knoten** k **zu, wo** e_k **minimal ist

Eine Rechenbeispiel kann im Anhang A.4 vorgefunden werden.

3.3.4. Vergleich der Methoden

An dieser Stelle soll ein kurzer Vergleich zwischen den vorgestellten Methoden zur taxonomischen Analyse gemacht werden.

Eine phylogenetische Analyse durch Rekonstruktion von Bäumen anhand von Markergenen liefert mit Sicherheit die verlässlichsten Ergebnisse. Der Aufwand für die Berechnung der Bäume ist aber sehr hoch, was bei großen Datensätzen unpraktikabel wird.

Eine weitaus schnellere Methode ist die hier vorgestellte Blast2Tree Methode, die ebenfalls eine phylogenetische Zuordnung eines ORFs erlaubt; aber auch nur dann, wenn auf dem ORF ein Markergen vorhanden ist.

Die Best-BLAST-Hit Methode ermöglicht dagegen zu (fast) jedem ORF eine taxonomische Aussage zu machen. Da es aber kein phylogenetisches Verfahren ist, ist eine Zuordnung zu einem inneren Knoten eines Baumes nicht möglich. Nur die Software MEGAN ist durch den Lowest-Common-Ancestor Algorithmus dazu fähig [18], Sequenzen auch inneren Knoten zuzuordnen. Der Einfluss von horizontalem Gentransfer kann aber dazu führen, dass der Algorithmus die Sequenz zu weit oben im Baum platziert.

Sowohl die Blast2Tree Methode, als auch der Best-Blast-Hit Ansatz, wurden von Dominik Lindner in ersten Versuchen auf das Anammox Metagenom angewandt und die Ergebnisse miteinander verglichen [20]. Zusammenfassend lässt sich sagen, dass beide

Methoden aufgrund ihres unterschiedlichen Ansatzes nicht direkt vergleichbar sind. Je nach Fragestellung und Bedingungen bietet sich die eine oder andere an, oder sie können sich durchaus auch ergänzen.

3.4. Rekonstruktion von Genomen — Binning

3.4.1. Einleitung

Eines der wohl wichtigsten Ziele in der Metagenomik ist die Gewinnung von kompletten Genomen der in der Umweltprobe enthaltenen Mikroorganismen. Ein grundlegendes Problem ist dabei die Zuordnung der sequenzierten DNA Fragmente zu bestimmten Genomen oder taxonomischen Einheiten — das sogenannte Binning —, wenn Markergene als identifizierendes Merkmal nicht vorhanden sind . Wenn man eine durchschnittliche Genomgröße von 4 Mb annimmt, würden nur etwa 5–10 % der Klone ein phylogenetisches Markergen, wie zum Beispiel dem für Prokaryoten anerkannten 16S rRNA Molekül, beherbergen [8]. Daher besteht in Abwesenheit solcher Markergene ein klares Bedürfnis nach anderen Möglichkeiten, um zu zeigen, dass zwei Fragmente zum gleichen Organismus gehören. Ein traditionell erster Schritt besteht darin, die DNA Sequenzen durch Überlappung zu verschmelzen und , in einem zweiten Schritt, durch Bildung von Scaffolds in spezies-spezifische „bins“ zu packen [5] . Weitere sequenz-basierte Merkmale, die verwendet werden können, um zu beurteilen, ob zwei unverknüpfte DNA Sequenzen zum gleichen Organismus gehören, sind genomische Fingerabdrücke, wie das GC-Verhältnis, Neigungen des Genoms zum Gebrauch bestimmter Codons (Codon Usage) und Oligonukleotidfrequenzen. Ebenso können Best-Blast Treffer wertvolle Hinweise auf den Ursprung des Sequenz liefern.

3.4.2. Genomische Signaturen

Eine Eigenschaft, aus der man viele Schlüsse über die Taxonomie eines Organismus ziehen kann, ist das GC-Verhältnis der genomischen DNA . Dieses wird als Prozentsatz von Guanin plus Cytosin in der DNA des eines Organismus definiert. Die GC-Werte haben eine große Schwankungsbreite, wobei die Werte von 20 % bis zu fast 80 % bei den Prokaryoten reichen [21]. Doch das GC-Verhältnis kann nicht nur zwischen den Organismen schwanken, sondern es schwankt auch innerhalb prokaryotischer Genome beträchtlich [8]. Da in der Metagenomik vor allem fragmentierte DNA prokaryotischen Ursprungs zusammengefasst werden sollen, führt diese Tatsache dazu, dass das GC-Verhältnis kein starkes phylogenetisches Signal in sich trägt und sich damit nur bedingt zum Binning eignet.

Betrachtet man den Gengehalt eines typischen Fosmid Inserts mit einer Länge von ca. 40 Kb, so enthält dieses im Durchschnitt um die 40 Gene. Bei einer Suche in einer öffentlichen Gendatenbank wie der NCBI Protein Datenbank mittels BLAST, liefern meist nur die Hälfte der Gene signifikante Treffer [8] . Diese Treffer sind als kritisch anzusehen, da viele Datenbanken einfach keinen nahen Verwandten zu der betrachteten Spezies besitzen.

Codon Usage, auch Codon Bias, beschreibt das Phänomen, dass Varianten des universellen genetischen Codes von verschiedenen Spezies unterschiedlich verwendet werden. Dabei werden bestimmte Codons des degenerierten Codes bevorzugt benutzt. Allerdings gilt auch hier, dass Codon Vorlieben innerhalb eines Genoms deutlich variieren und eher mit dem Genexpressionslevel korrelieren, als ein phylogenetisches Signal beinhalten [28].

Im Gegensatz dazu kann ein phylogenetisches Signal in den Frequenzen von Oligonukleotiden in einer genomischen Sequenz beobachtet werden.

3.4.3. Oligonukleotidfrequenzen als phylogenetisches Merkmal

Unterschiede von Genomen in deren Oligonukleotidfrequenzen sind schon seit geraumer Zeit bekannt. Statistische Maßzahlen zum Messen und Interpretieren von Heterogenität in Oligonukleotid-Zusammensetzungen zwischen und innerhalb von Genomen, wurden erstmals von Karlin und Cardon im Jahre 1994 vorgeschlagen [10]. Sie stellten sich die Frage, ob sich die relativen Häufigkeiten von Oligomeren zwischen Genomen verschiedener Organismen, aber auch innerhalb von Genomen, signifikant unterscheiden. Es konnte für Oligomerlängen von zwei bis drei Nukleotiden bewiesen werden, dass deren relative Häufigkeiten ein spezies-spezifisches Signal darstellen. Dieses Phänomen konnte zumindest für Dinukleotidfrequenzen eingehender geklärt werden [10].

In darauf folgenden Publikationen von Rocha [12], Pride [11] und Teeling [8] wurde daher das diskriminatorische Potential von längeren Oligonukleotidfrequenzen (4–8 Nukleotide) und deren phylogenetische Signale untersucht. In einem Kompromiss zwischen minimaler Wortlänge und maximalem Informationsgewinn, wurde vor allem für Tetranukleotide gezeigt, dass deren Frequenzen eine schwache phylogenetische Signatur besitzt [11]. Tetranukleotidfrequenzen bieten somit vor allem bei vergleichenden Analysen großer Datensätze eine gute Balance zwischen Rechenanforderung und annehmbarer Auflösung. Es konnte außerdem bewiesen werden, dass phylogentische Bäume, die auf Basis von Tetranukleotiden gerechnet wurden, eine Kongruenz mit den allgemein anerkannten 16S rRNA Bäumen besitzen [11]. Dies lässt den Schluss zu, dass Binning nicht nur auf Spezies-Ebene möglich ist, sondern auch für höheren Ordnungen.

3.4.4. Statistische Maßzahlen für Oligonukleotidfrequenzen

Sei f_x die Frequenz eines Nukleotides X in einer Sequenz S. Will man Sequenzen von verschiedenen Organismen, Chromosomen oder Strängen miteinander vergleichen, muss man die komplementäre antiparallele Struktur der doppelsträngigen DNA berücksichtigen [10]. Daher kombiniert man die Sequenz S mit ihrer komplementären Inversen S^T zu $S + S^T = S^*$. In S^* sind die Frequenzen der Mononukleotide Adenin f_A^* und Thymin f_T^* somit $f_A^* = f_T^* = (f_A + f_T)/2$. Entsprechend gilt auch für die Frequenzen von Guanin und Cytosin $f_G^* = f_C^* = (f_G + f_C)/2$. Dies lässt sich auch auf alle Oligonukleotide erweitern (Formel (3.7)).

$$f_W^* = \frac{f_W + f_{\bar{W}}}{2} \tag{3.7}$$

f^*_W: erwartete Frequenz des Oligonukleotides W in S^*; f_W: erwartete Frequenz des Oligonukleotides W in S; $\bar{W}$: komplementäres Inverses des Oligonukleotides W

Zum Beispiel errechnet sich die Frequenz des Dinukleotid Guanin-Thymin f^*_{GT} dann als $f^*_{GT} = (f_{GT} + f_{AC})/2$, usw für alle weiteren Oligonukleotide. Es ist gebräuchlich die Frequenzen auf deren erwartete Anzahl und die Länge der Sequenz zu normalisieren.

Eine anerkannte Methode zur Normalisierung von Oligonukleotidfrequenzen ist das Markov-Ketten-Modell [12]. Die Begründung hierfür ist, dass man Frequenzen eines Oligomers von Effekten bereinigen sollte, die auf dem Aufkommen kleinerer Wörter beruhen, die innerhalb des betrachteten Oligomers vorkommen. Jedoch zeigt die Arbeit von Pride [11], dass durch das Entfernen dieser Effekte Vergleiche von Spezies schwieriger werden. Diese Effekte tragen nämlich zur Entwicklung von organismus-spezifischen Oligomusterfrequenzen bei. Pride schlägt daher eine andere Methode vor, die auf der Markov-Kette nullter Ordnung basiert. Diese bereinigt den Vergleich von Oligonukleotidfrequenzen nur von der Verteilung ihrer Mononukleotidfrequenzen. Hierzu sei nun $O(W)$ die beobachtete Frequenz eines Wortes W in S^*. Die erwartete Anzahl an Oligonukleotiden $E(W)$ berechnet sich nach Formel (3.8).

$$E(W) = \left[(f^*_A)^a * (f^*_C)^c * (f^*_G)^g * (f^*_T)^t\right] * N \tag{3.8}$$

$E(W)$: erwartete Anzahl des Wortes W; f^*_X: erwartete Frequenz des Nukleotides X in S^*; a, c, g, t: Anzahl der Nukleotide A,C,G,T in W; N: Länge von S^*

Die bereinigte Frequenz $F(W)$ für ein Oligonukleotid errechnet sich dann als Quotient zwischen $O(W)$ und $E(W)$ (Formel (3.9)).

$$F(W) = \frac{O(W)}{E(W)} \tag{3.9}$$

$F(W)$: normalisierte Frequenz des Wortes W in S^*; $O(W)$: beobachtetes Vorkommen des Wortes W in S^*; $E(W)$: erwartete Anzahl des Wortes W in S^*

Ein Distanzmaß zur Messung ähnlicher Muster zweier genomischer Fragmente in ihren Oligonukleotidfrequenzen, wurde von Karlin et al. vorgeschlagen [10] und ist in Formel (3.10) dargestellt.

$$D(S_1, S_2) = \frac{1}{4^n} * \sum_{w=0}^{4^n} \mid F(w)_1 - F(w)_2 \mid \tag{3.10}$$

D: Distanz zwischen zwei Sequenzen S_1 und S_2; $F(w)_i$: normalisierte Frequenz des Wortes W (dargestellt durch Index w) für eine Sequenz i; n: Länge des Oligonukleotidwortes W

Andere Ansätze verwenden als Ähnlichkeitsmaß den Korrelationskoeffizienten auf Basis von standardisierten Oligonukleotidfrequenzen [8]. Dieser Ansatz setzt jedoch normalverteilte Daten voraus und hängt daher kritisch von der Stichprobenanzahl, in diesem Fall von der Länge der Sequenz, ab. Daher muss man bei der Wahl des Distanzmaßes, als auch bei der Wahl der Wortlänge die Probleme der Methodik und die damit verbundenen Einschränkungen kennen.

3.4.5. Probleme und Einschränkungen

Unabhängig vom benutzten statistischen Modell, variiert die Signifikanz des Unterschiedes zwischen $O(W)$ und $E(W)$ mit der Anzahl von $O(W)$, da ein statistischer Test Abweichungen besser erkennen kann, wenn die Anzahl von $O(W)$ größer ist. Dies hat zwei entscheidende Effekte. Einerseits wird, bei einer gegebenen Wortlänge, die Trennung besser je länger die durchschnittliche Länge der Sequenzen ist. Umgekehrt wird bei einer gegebenen durchschnittlichen Sequenzlänge das diskriminatorische Potential bei Verwendung kürzerer Oligomere mächtiger. Das heißt aber nicht, dass kürzere Wortlängen generell besser sind. Kürzere Oligonukleotide besitzen weniger Permutationen und bieten der Diskriminanzfunktion somit weniger Dimensionen zur Trennung. Aus diesem Grund kann das Trennungspotential durch die Wahl eines zu kurzen Wortes eingeschränkt werden.

Eine Wahl hat man meist bei der durchschnittlichen Länge der zu untersuchenden DNA Fragmente nicht. Somit stellt sich die Frage, bei welchen Sequenzlängen welche Oligomere ihr diskriminatorisches Optimum erreichen oder an ihre Grenzen stoßen. Für Analysen mit Dinukleotiden wird eine Länge von 5 kb als ausreichend angegeben [10]. Für Vergleiche anhand von Tetranukleotidneigungen konnte gezeigt werden, dass in den meisten Fällen Fragmente von 10 kb und auch selten noch von 1 kb korrekt klassifiziert werden konnten [8].

Die Trennungsmöglichkeiten anhand von bestimmten Oligonukleotidneigungen hängt aber nicht nur von der Länge der DNA Sequenzen, sondern auch vom Artenreichtum des betrachteten Sequenzpooles, ab; im Fall der Metagenomik, vom Artenreichtum des Metagenoms. Es ist wichtig zu erwähnen, dass das diskriminatorische Potential dieser Methode mit zunehmender Anzahl an vorhandenen Spezies abnimmt [8]. Für Habitate, in denen die natürliche Diversität aufgrund von extremen Bedingungen reduziert ist, oder von einer Spezies stark dominiert wird, wird die Analyse möglicherweise gute Ergebnisse erzielen können. Für Proben aus artenreicheren ökologische Nischen, wie Proben aus dem Erdboden, ist diese Methode ungeeignet. Zusätzlich zu den genannten Schwierigkeiten, gibt es Einschränkungen in der Methodik, die sozusagen in der Natur der Sache liegen. Zum Beispiel können Fragmente, die viele Gene durch lateralen Gentransfer aufgenommen haben, nicht richtig zugewiesen werden, da sie ein für ihr Genom atypisches Oligonukleotid-Frequenzmuster aufzeigen. Ebenso kann ein hoher Grad an Polymorphismen zu einer falschen Zuordnung führen.

Insgesamt lassen sich somit folgende negative Faktoren zusammenfassen:

- kurze DNA Sequenzen
- artenreiches Metagenom
- starker Polymorphismus
- lateraler Gentransfer

Den letzten beiden natürlichen Einflüssen ist eigen, dass sie jegliche molekulare phylogenetische Signatur verwischen, so dass auch andere molekulare Methoden unwirksam werden.

3.4.6. Anwendung

Mit Hilfe von Mustern in Oligonukleotidneigungen können genügend lange genomische Fragmente von der selben Spezies mit vernünftiger Wahrscheinlichkeit einander zugeordnet werden. Daher hat dieser Ansatz bereits Anwendung in der Zuweisung von genomischen Fragmenten in der Metagenomik gefunden [8]. Es gibt einen Webserver „Tetra“ der Korrelationskoeffizienten zwischen Sequenzen auf Basis von Tetranukleotidfrequenzen berechnet [9].

Ein konkreter Fall in dem Oligonukleotidneigungen benutzt werden, ist die Trennung von Sequenzen von pathogenen Organismen und deren Wirtsorganismus [15]. Man wählte hier die Tripletnukleotidfrequenzen als Merkmale zur Klassifizierung aus zwei Gründen. Zum einen, da die traditionelle Methode der Suche von homologen Genen in öffentlichen Datenbanken nur dann funktioniert, wenn auch verwandte Homologe in der Datenbank vorhanden sind. Zum anderen, da deren erster Ansatz – Klassifizierung anhand der Codon Usage – den Nachteil hatte, dass man hierzu erst kodierende Regionen voraussagen musste. Dies führte dazu, dass Fragmente, die aus nicht kodierenden Regionen bestehen, nicht klassifiziert werden konnten. Aus dem Projekt ist ein Onlineservice hervorgegangen, der EST (Expressed Sequence Tags) Sequenzen mittels Tripletfrequenzen klassifiziert.

3.5. Zusammenfassung

Die Einleitung gab eine Einführung in die Entwicklung des Lebens auf unserer Erde und den damit verbundenen Wissenschaften der Phylogenetik und der Taxonomie. Eine wichtige Technologie, die das Tor zu der enormen biologischen Diversität aufgestoßen hat, ist die Metagenomik, welche in Form mehrerer von Forschern durchgeführter Projekte näher vorgestellt wurde. Dabei wurde auch auf die Problematiken der Metagenomik eingegangen.

Es wurden Techniken zur Analyse von Metagenomen vorgestellt, darunter Methoden sowohl zur phylogenetischen/taxonomischen Analyse als auch zur Rekonstruktion der Genome, die sich fragmentiert in den Metagenom-Datensätzen befinden (Binning).

Die eingehend dargestellte Blast2Tree Methode soll im weiteren Verlauf der Diplomarbeit implementiert, evaluiert und eingesetzt werden. Ebenso soll das Binning anhand der Oligonukleotidfrequenzen näher untersucht und schließlich im selben Zuge mit Blast2Tree angewandt werden.

4. Technische Realisierung

4.1. Zielsetzungen

Ziel dieser Arbeit ist es ein komplettes Softwarepaket zur automatischen Analyse von Metagenomen zu schaffen, das sowohl Binning als auch die taxonomische Klassifizierung miteinander vereint.

Während die Binning Prozedur auf bewährten Ansätzen und vorhandenen Methoden basiert, soll die taxonomische Klassifizierung anhand des Blast2Tree Verfahrens vollständig neu implementiert werden.

Dabei wurden sowohl technische als auch inhaltliche Zielsetzungen formuliert, die bei der Erstellung der Software erfüllt werden sollten.

Technische Zielsetzungen:

- Schnelle und schlanke Implementierung:
  Metagenomdatensätze sind für gewöhnlich sehr umfangreich und die wichtigste Anforderung an die Software ist, einen schnellen und zuverlässigen Überblick über die Taxa im Metagenom zu erhalten. Daher ist eine schnelle und vor allem effiziente Programmiertechnik von Nöten. Das impliziert auch eine schlanke Implementierung, da ein Overhead die Laufzeit nur unnötigweise verlangsamen würde.

- Multi-Threading:
  Die Tatsache, dass einzelne Sequenzen auf gleiche Weise untersucht werden müssen, sich sozusagen die gleichen Objekte im Hauptspeicher teilen, legt die Verwendung von Leichtgewichtsprozessen nahe. Auch in Hinblick auf die zunehmende Verbreitung von Mehrkern-Prozessoren, sollte die Programmiersprache Multi-Threading unterstützen.

- Leichte Verständlichkeit und übersichtliche Architektur des Programms (Objekt Orientierte Programmierung):
  Da die Implementierung dieser Software im Rahmen einer Diplomarbeit abläuft, ist es wahrscheinlich, dass weitere nachfolgende Studenten an der Software weiterarbeiten werden. Um ihnen eine schnelle Einarbeitung zu ermöglichen, ist es wichtig die Software verständlich zu gestalten und übersichtlich zu gliedern.

- Leichte Portierbarkeit:
  Die Zielgruppe dieser Software sind vorwiegend Wissenschaftler aus den Bereichen Biologie, Mikrobiologie und Bioinformatik. Eine leichte Portierbarkeit der Software, würde es jedem ermöglichen, die sie benutzen zu können, unabhängig von dessen Betriebssystem.

- Leichte Erstellung eines Web-Services:
  Eine weitere Anforderung an das Projekt ist Einrichtung eines Web-Interfaces, der kleine Sequenzdatensätze taxonomisch analysieren soll. Die Software sollte somit nur durch leichte Veränderung Web-fähig gemacht werden können.

- Integration von etablierten Programmen:
  Es sollte nichts implementiert werden, das nicht schon effizient realisiert wurde und zur freien Verfügung steht. Allerdings sollten dadurch nicht die zuvor genannten Anforderungen verletzt werden.

Aufgrund der angeführten technischen Anforderungen, fiel die Entscheidung der zu verwendenden Programmiersprache auf JAVA. JAVA ist eine objekt-orientierte Programmiersprache, mit der man plattformunabhängige Applikationen entwickeln kann. Es unterstützt Multi-Threading und eignet sich durch die JAVA Enterprise Komponenten JSP (Java Server Pages) und Servlets gut zur Webentwicklung. Eine Integration der BioJava Tools entfiel, da die Überschneidung zwischen den von BioJava angebotenen Funktionen und unseren Anforderungen zu gering war.

Inhaltliche Zielsetzungen:

- Benutzerdefinierte Oligonukleotidmuster zum Binning:
  Der Benutzer soll angeben können nach welchem Muster die Oligonukleotide gezählt werden. Üblich ist das Zählen von Di-, Tri oder Tetranukleotiden.

- Austauschbarkeit von Referenzbaum und Markergenset:
  Die Implementierung soll unabhängig vom verwendeten Speziesbaum und dessen Markergene sein. Dies bedeutet, dass Parser integriert werden müssen, die neue Bäume und deren Markergene einlesen können.

- Parameter vom Benutzer definierbar:
  Wie bereits erwähnt, werden die Benutzer der Software vor allem Wissenschaftler sein. Um ihnen für ihre Forschung viele Möglichkeiten zu geben, sollten die wichtigsten Parameter vom Benutzer einstellbar sein.

- Unterstützung der in der Bioinformatik üblichen Standardformate:
  In Anlehnung an die Anforderung der Austauschbarkeit, soll die Software die in der Bioinformatik am weitest verbreitetsten Formate unterstützen, sowohl in der Eingabe, als auch in Ausgabe von Daten.

4.2. Vorstellung der Software

4.2.1. Unterstützte Dateiformate

Multi-FASTA Format Das FASTA Format ist ein textbasiertes Format, um Nukleotid- oder Peptidsequenzen darzustellen. Da es das populärste Dateiformat zur Speicherung

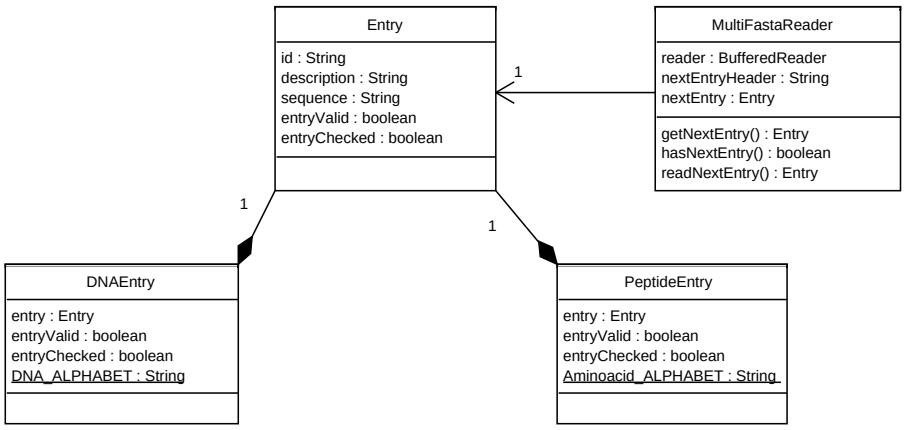


Abbildung 4.1.: Klassendiagramm der Objekte `Entry` und deren durch Komposition erbende Klassen `DNAEntry` und `PeptideEntry`.

von Sequenzdaten in der Bioinformatik ist, war es erforderlich einen Parser für das FASTA Format in die Software zu integrieren. Hierzu wurde eine Klasse `MulitFastaReader` implementiert, die sequentiell Einträge aus einer FASTA Datei liest, die Einträge auf ihre syntaktische Gültigkeit prüft und diese dann in Objekte des Typs `Entry` überführt (Abbildung 4.1).

Die von `Entry` erbenden Klassen `DNAEntry` und `PeptideEntry` überprüfen, ob deren Sequenz eine nach [40] gültige Sequenz ist.

Newick Format Das Newick Format ist ein Standard, um Baumstrukturen durch verschachtelte Klammerausdrücke darzustellen. Im Newick Format sind die einzelnen Knoten eines Baumes durch Komma getrennt und Knoten mit demselben Elternknoten durch Klammern zusammengefasst. Die Zweiglänge (Distanz zum Elternknoten) kann durch einen Doppelpunkt getrennt, hinter dem Knotennamen angegeben werden. Namen interner Knoten folgen auf die rechte Klammer des Klammernpaares derer Kindsknoten. Beispiel für den Baum aus Abbildung 3.9 :

$$((A{:}1,B{:}1)Z{:}2,C{:}2)X{:}1,(D{:}1,E{:}1)W{:}4;$$

Auch hier wurde eigens dafür ein Parser `NewickParser` implementiert, der einen Newick-Baum rekursiv parst und in eine JAVA Datenstruktur umwandelt (Anhang B.1). Üblicherweise kreiert man dazu Objekte der Klasse `TreeNode`. Die Baumstruktur wird abgebildet, indem jeder Knoten jeweils seinen Vaterknoten und/oder seine Kindknoten kennt (Abbildung 4.2). Die übrigen Instanzvariablen ergeben sich aus den Anforderungen des Blast2Tree Algorithmus (Siehe: Kapitel 3.3.3). Die fehlende Instanzvariable des Distanzvektors erbt `TreeNode` von der Klasse `TreeVector`.

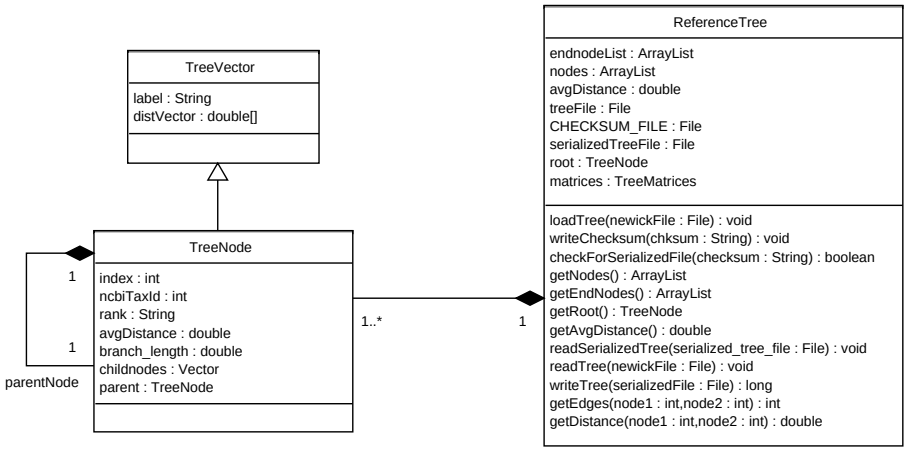


Abbildung 4.2.: Klassendiagramm der Objekte `TreeVector` und deren erbende Klasse `TreeNode`, sowie der Klasse `ReferenceTree`, die die `TreeNode` Objekte speichert.

4.2.2. Workflow

Die Methodik aus Dominik Lindners Diplomarbeit wurde weitesgehend beibehalten, dennoch aber an einigen Stellen verändert und voll automatisiert. Der Workflow der Software ist in Abbildung 4.3 dargestellt und wird nun im Folgenden näher erläutert.

Workflow: 1) Laden der Markergene und des Referenzbaumes

Das Markergenset wird aus einer Datei geladen, die im Multiple-FASTA Format vorliegen muss. Handelt es sich um mehrere Genfamilien, muss im Header jeder Sequenz auch die Bezeichnung der Genfamilie angegeben werden. Generell muss jeder Eintrag der FASTA Datei also folgende Information enthalten:

- NCBI Taxonomie des Organismus, aus dem die Sequenz stammt
- Eindeutige Bezeichnung der Sequenz
- Bezeichnung der Genfamilie
- Aminosäure Sequenz

Der `MultiFastaReader` parst die Einträge der Datei, überführt sie in eine Datenstruktur des Typs `MarkerProtein` und sammelt sie in einem `MarkerProteinSet` (Abbildung 4.4). Somit werden die Markerproteine im Hauptspeicher gehalten.

Der Baum liegt mit Angabe der Zweiglängen im Newick Format vor. Der `NewickParser` erstellt daraus die `TreeNode` Objekte, die in der Klasse `ReferenceTree` gespeichert werden (Abbildung 4.2). Nun werden die Knoten des Referenzbaumes rekursiv durchlaufen. Dabei werden sowohl die Distanzvektoren, als auch die durchschnittlichen Distanzen jedes

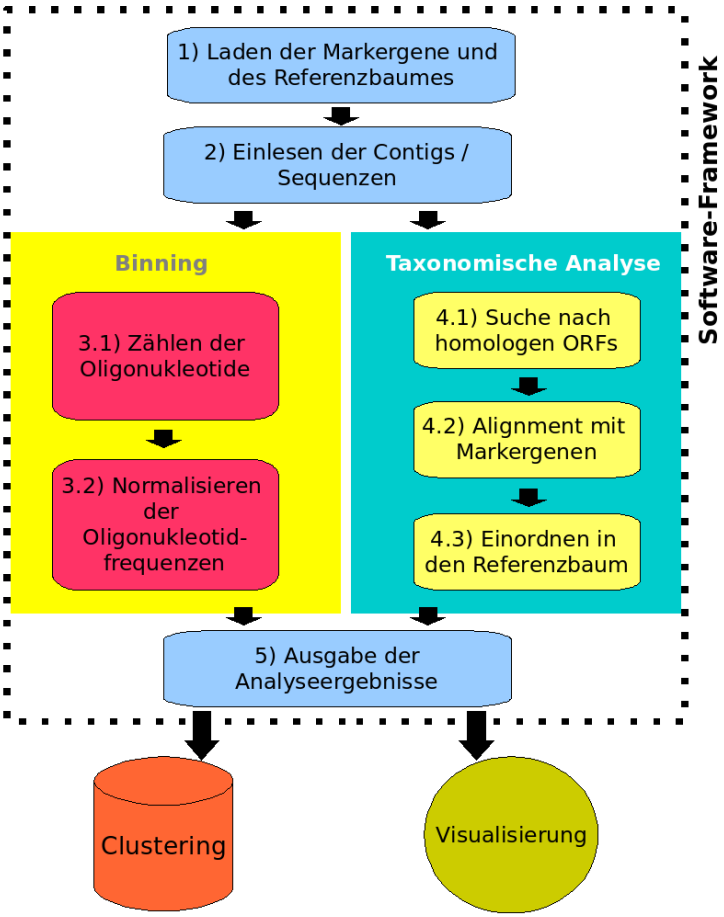


Abbildung 4.3.: Grober Workflow der Software. Der gestrichelte Rahmen stellt das Aufgabengebiet der zu programmierenden Software dar. Sowohl Binning, also das Clustering der Contigs, als auch die Visualisierung, sollen durch externe Programme geschehen.

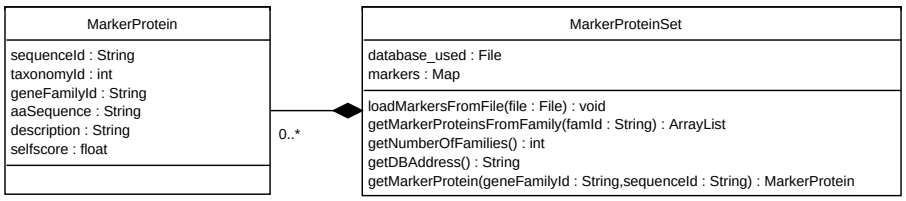


Abbildung 4.4.: Klassendiagramm der Objekte `MarkerProtein` und der agglomerierenden Klasse `MarkerProteinSet`.

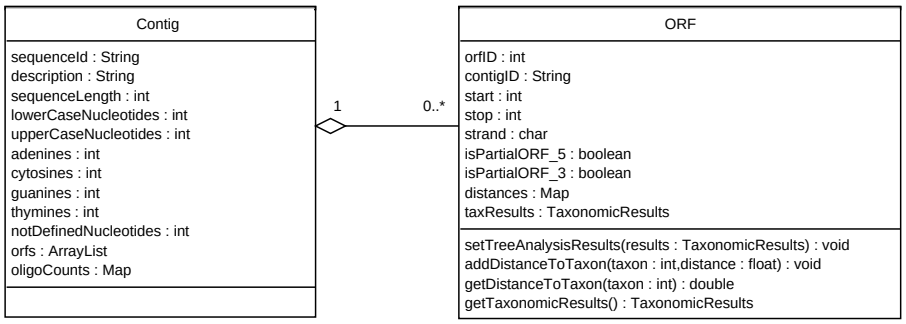


Abbildung 4.5.: Klassendiagramm der Objekte `Contig` und `ORF`. Wird ein oder mehrere ORFs auf der Sequenz des Contigs gefunden, wird für jeden ORF eine `ORF` Instanz generiert und im jeweiligen `Contig` Objekt abgespeichert.

Knotens zu seinen Blattknoten berechnet. Zusätzlich zur Berechnung der Distanzvektoren, muss jedem Knoten die entsprechende NCBI Taxonomie zugewiesen werden. Dazu wird der NCBI-Baum in den Speicher geladen (Abbildung 3.4). Jedem Knoten des Referenzbaumes wird dann, anhand des in Kapitel 3.3.3 beschriebenen Algorithmus 1, die eindeutige NCBI-Taxonomie ID, der taxonomischen Rang (Spezies, Gattung, usw., siehe Tabelle 3.1) und der entsprechende wissenschaftliche Name zugeteilt.

Der so prozessierte `ReferenceTree` wird serialisiert und in eine binäre Datei geschrieben. Bei einem erneuten Lauf verhindert eine Checksummen-Überprüfung ein erneutes Prozessieren, wenn der Baum bereits als serialisierte Datei vorhanden ist. Somit kann der Baum in kürzester Zeit direkt aus der Datei heraus geladen werden.

Workflow: 2) Einlesen der Contigs/Sequenzen

Auch die zu analysierenden Sequenzen müssen im Multiple-FASTA Format vorliegen. Man hat häufig mit mehreren tausend Einträgen zu tun. Das Halten aller `Entry` Objekte samt Sequenzdaten im Speicher würde den Hauptspeicher der meisten Computer sprengen. Daher müssen die Sequenzen transient behandelt und deren Metainformation in einer anderen Datenstruktur gespeichert werden. Hierzu wurde die Klasse `Contig` geschaffen (Abbildung 4.5).

Die Metadaten, wie Länge der Sequenz, Anzahl der einzelnen Nukleotide usw. werden von einer Klasse `CharacterCounter` gesammelt und dem `Contig` übergeben. Die letzten beiden Feldvariablen `orfs` und `oligoCounts` zeigen auch schon auf, was noch getan werden muss.

Workflow: 3.1) Zählen der Oligonukleotide

Zur Suche der Oligonukleotide wurde eine Klasse `OligoNucleotideCounter` geschrieben, die zur Aufgabe hat, die Anzahl aller Oligonukleotide für ein bestimmtes Muster zu suchen. Muster werden hierbei aus folgenden Zeichen konstruiert: ein „N“ gibt das Vor-

handensein eines Nukleotides an, ein „x“ das Vorhandensein irgendeines Zeichens. Übliche Muster sind „NN“ für Di-, „NNN“ für Tri- und „NNNN“ für Tetranukleotide. „NNxNN“ entspricht dem Betrachten zweier nacheinander folgender Codons, mit dem Vernachlässigen der 3.($\equiv$ x) und 6. Base. Der Benutzer kann bei Start des Programms ein beliebiges Muster angeben. Der `OligoNucleotideCounter` zählt auch die Oligonukleotide auf dem reversen komplementären Strang. Man erhält somit die Frequenz f_W^* für ein Wort W nach Formel (3.7). Das Vorkommen eines Wortes W wird in einem Array gespeichert, wobei der Index des Arrays das Wort W kodiert. Das Array mit den Frequenzen wird dann mit dem Oligomuster als Schlüssel in einer Hashmap des Contigs gespeichert.

Workflow: 3.2) Normalisieren der Nukleotidfrequenzen

Die beobachteten Oligonukleotidfrequenzen eines Contigs können nun anhand der in der Einleitung vorgestellten Formeln (3.8) und (3.9) normalisiert werden. Die Vorkommen der einzelnen Nukleotide sind durch den `CharacterCounter` bekannt.

Workflow: 4.1) Suche nach homologen ORFs

Ziel ist es, Gene auf dem aktuell zu analysierenden Contig zu finden, die homolog zu den verwendeten Markerproteinfamilien sind. Daher besteht diese Aufgabe eigentlich aus zwei Teilen, zum einen aus der Genvorhersage, und zum anderen aus der Prüfung, ob ein gefundenes Gen homolog zu einer der Markerproteinfamilien ist.

Die Genvorhersage besteht, wie in Kapitel 3.3.3 beschrieben, aus der Identifizierung von *Open-Reading-Frames* auf einem Contig. Eine Klasse `ORF_Finder` sucht auf allen sechs Leserahmen der Nukleotidsequenz nach ORFs. Dabei wird eine Sequenz als ORF identifiziert, wenn diese zwischen zwei Stoppcodons liegt, oder partiell ist. Das heißt, dass das zweite Stoppcodon schon nicht mehr auf dem Contig liegt (Abbildung 3.10). Außerdem muss der gefundene ORF die angegebene Mindestlänge haben.

Der `ORF_Finder` arbeitet als endlicher Automat, der jedes Zeichen der Contigsequenz liest und dadurch seinen Zustand verändert bis er auf ein Stoppcodon trifft (Endzustand) (Anhang B.2). Ein gefundener ORF wird mit den Angaben seiner Lage in der Instanz seines Contigs gespeichert (Abbildung 4.5).

Nachdem alle ORFs eines Contigs ermittelt wurden, wird jeder ORF in einem eigenen Thread analysiert. Als erstes wird die Nukleotidsequenz des ORFs mit Hilfe der Klasse `DNA_Translator`, die auf Basis eines endlichen Automaten funktioniert (Anhang B.3), in eine Peptidsequenz übersetzt. Die ORF Peptidsequenz wird dann mittels eines externen BLAST Prozesses gegen die Markerprotein-Datenbank geblastet. Der BLAST dient hier als Filter, um alle ORFs mit einem zu hohen E-Value für die weitere Analyse auszuschließen, da von diesen angenommen werden kann, dass sie keine Markerproteine sind. Der E-Value ist standardmäßig auf 10^{-5} eingestellt. Dieser Wert lässt sich aber vom Benutzer verändern. Die BLAST Ausgabe wird analysiert , und — wenn vorhanden — das am signifikantesten homologe Markerprotein zurückgegeben.

Workflow: 4.2) Alignment mit Markergenen

Falls ein homologes Markerprotein gefunden wurde, wird die Peptidsequenz des ORFs mit allen Proteinen der entsprechenden Markerproteinfamilie aligniert. Dazu wird auf den JALigner[1] zurückgegriffen, einer Open-Source JAVA Implementierung des Smith-Waterman Algorithmus. Das Alignment verwendet dabei folgende Standardparameter, die aber jederzeit verändert werden können:

- Gap-Opening: -10
- Gap-Extension: -2
- Substitutionsmatrix: BLOSUM 62

Die erhaltenen Scores werden nach Formel (3.3) in Distanzen zu den jeweiligen Taxa der Markerproteine umgerechnet. Jede Distanz wird in der Hashmap des betreffenden `ORF` Objektes mit der NCBI TaxID des Markerproteins als Schlüssel gespeichert. Somit erhält der ORF einen Distanzvektor zu den Taxa der homologen Markergruppe und somit auch einen Distanzvektor zu den Blattknoten des Referenzbaumes.

```
protected void alignORFwithProteins(String homologFamily) {
      // Gets the members of the homolog marker protein family
      ArrayList<MarkerProtein> markerGroup
            = marker_db.getMarkerProteinFamily(homologFamily);
      Sequence orfSeq = new Sequence(orfSequence);
      // For each marker protein do the SW alignment
      for (MarkerProtein protein : markerGroup) {
            Sequence markerProtein =
                  new Sequence(protein.getSequence());
            Alignment sw_alignment
                  = SmithWatermanGotoh.align(orfSeq, markerProtein,
                  substMatrix, SW_ALIGNMENT_OPEN,
                   SW_ALIGNMENT_EXTEND);
            float score = sw_alignment.getScore();
            // Calculate the distance
            float distance = calculateDistance(score, protein);
            // Save the distance in the ORF instance
            orf.addDistanceToTaxon(markerProtein.getTaxon(),
                  distance);
      }
}
```

Workflow: 4.3) Einordnen in den Referenzbaum

Der ORF wird nun anhand seines Distanzvektors in den Referenzbaum eingeordnet. Der in Kapitel 3.3.3 vorgestellte Algorithmus 2 wurde in der Klasse `TreeMapping` verwirklicht. Der ORF Distanzvektor wird für jeden Knoten des Referenzbaumes nach den Formeln

[1] Erhältlich unter: http://jaligner.sourceforge.net/

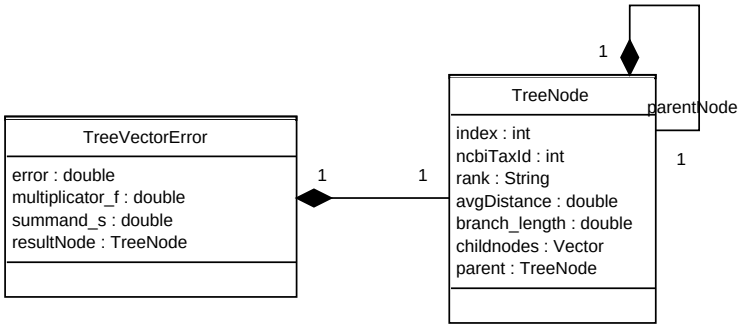


Abbildung 4.6.: Klassendiagramm der Objekte `TreeVectorError` und `TreeNode`. Für einen homologen ORF wird für jeden Knoten des Baumes ein Fehler errechnet, der in dann in einer `TreeVectorError` Instanz gespeichert wird. Dem Knoten mit dem kleinsten Fehler ist der ORF dann am ähnlichsten.

(3.4) und (3.5) skaliert und der jeweilige Fehler (Formel (3.6)) bestimmt. Für jeden dieser Vergleiche wird ein `TreeVectorError` Objekt generiert, das den betrachteten Knoten und den zugehörigen Fehler speichert (Abbildung 4.6).

Alle `TreeVectorError` Objekte werden in einer Liste gespeichert, die am Ende nach Größe der Fehler sortiert wird. Die jeweils x besten Vergleiche (die `TreeVectorError` Objekte mit den kleinsten Fehlern) werden dem ORF als Ergebnis übergeben. Somit können die Taxa der Ergebnisknoten auf den ORF übertragen werden, und weiter auf den Contig des ORFs.

```
// Mapps ORF to reference tree and stores results in the ORF object
TreeMapping refTreeAnalyser = new TreeMapping(refTree);
List<TreeVectorError> resultList =  refTreeAnalyser.allocateORFintoTree(this.orf);
TaxonomicResults results =
      new TaxonomicResults(resultList, homologFamily, refTree);
orf.setTreeAnalysisResults(results);
```

4.2.3. Ausgabe und Analyse der Ergebnisse

Die Ausgabe der Analyseergebnisse erfolgt für taxonomische Analyse und Binning separat.

Binning Da es bereits sehr schnelle und effiziente Clustering-Programme gibt, entschieden wir uns, durch eine geeignete Dateiausgabe eine Schnittstelle zu diesen Programmen herzustellen. Unterstützt wird derzeit die Ausgabe in den Formaten für Cluster 3.0[2] und CLUTO[3] [53], in der Version 2.1.2. Dazu wird für jeden Contig der normalisierte Vektor der Oligonukleotidfrequenzen im gewünschten Format ausgegeben. Die Daten können

[2]Erhältlich unter: http://bonsai.ims.u-tokyo.ac.jp/ mdehoon/software/cluster
[3]Erhältlich unter: http://glaros.dtc.umn.edu/gkhome/views/cluto

seqID	length	GC	orfID	partial	length	markerFam	node	label	error	homogenity
seq1	1133	0.33	seq1_3	false	669	COG92	[6]	Proteobacteria (1224)	0.15	0.08
seq1	1133	0.33	seq1_3	false	669	COG92	[92]	Alphaproteobacteria (28211)	0.16	0.08
seq1	1133	0.33	seq1_3	false	669	COG92	[110]	Rickettsiaceae (775)	0.17	0.08
seq2	924	0.53								
seq3	3427	0.36	seq3_11	true	924	COG52	[231]	Firmicutes (1239)	0.37	0.35
seq3	3427	0.36	seq3_11	true	924	COG52	[248]	Onion yellows phyto. (100379)	0.39	0.35
seq3	3427	0.36	seq3_11	true	924	COG52	[247]	M. mycoides subsp. (44101)	0.39	0.35

Tabelle 4.1.: Beispielausgabe der Software mit taxonomischer Untersuchung ohne internes Clustering. Es wurden drei Sequenzen (seq1, seq2 und seq3) untersucht, von denen zwei (seq1, seq3) ein homologen ORF besitzen und für die somit eine taxonomische Aussage gemacht werden kann. Es werden die jeweils drei ähnlichsten Knoten des Baumes als Ergebnis betrachtet. Zu seq2 kann daher nur die Minimalinformation ausgegeben werden. Die Homogenität gibt an wie homogen die Ergebnisgruppe für diesen ORF ist. Dieser Wert kann dazu benutzt werden, um die Qualität einer Aussage zu einem ORF zu beurteilen. Mehr dazu im Paragraph zur Beurteilung der Ergebnisse.

nun, je nach Wunsch und Informationsstand, mit verschiedenen Clusteralgorithmen gebinnt werden.

Es wird jedoch auch ein programminternes k-means Clustering angeboten, das durch das Weka-Statistik-Paket[4] [52] miteingebunden wurde. Es ist aber im Vergleich zu vorherig genannten Programmen deutlich langsamer und eignet sich damit nur für kleine Datensätze und zu Testzwecken. Die Ergebnisse des internen Clusterings — falls durchgeführt — werden in der Standardausgabe, mit der sich der nächste Paragraph befasst, ausgegeben.

Ausgabe der taxonomische Analyse Diese Ausgabe erweitert die Standardausgabe um die Ergebnisse der taxonomische Analyse. Die Standardausgabe besteht aus der Minimalinformation zu den eingelesenen Sequenzen, wie Länge und GC-Gehalt. Das heißt man kann das Programm auch ohne taxonomische Analyse und Binningfunktion laufen lassen, bzw. wahlweise auch nur mit jeweils einer Funktion oder mit beiden. In Tabelle 4.1 ist als Beispiel die Ausgabe eines normalen Durchlaufes mit taxonomischer Untersuchung dargestellt.

Beurteilung der Ergebnisse Eine weitere Zielsetzung, die während der Evaluation und Entwicklung der Software aufkam, war dem Benutzer ein Mittel zu geben, um seine Ergebnisse beurteilen und visualisieren zu können.

Eine Beurteilung der einzelnen Ergebnisse ist wichtig, um mögliche unsinnige oder falsche Aussagen herausfiltern zu können. Eine Möglichkeit die Qualität einer taxonomischen Aussage zu bestimmen, ist über die Homogenität der Einzelaussagen; das heißt, zu überprüfen wie stimmig die Einzelaussagen zur Herkunft eines ORFs sind. Dazu sollten die besten x Ergebnisknoten im Baum möglichst nah beieinanderliegen. Im Idealfall sollten die Ergebnisknoten benachbart sein. Der Wert der Homogenität wird also anhand der

[4] Erhältlich unter: http://www.cs.waikato.ac.nz/ml/weka/

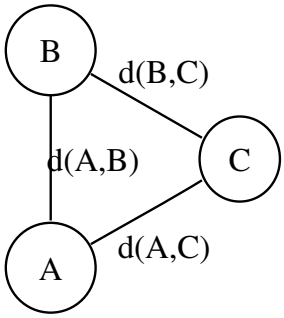


Abbildung 4.7.: Gegeben seien drei Ergebnisknoten A,B und C für einen ORF. Dieser Graph stellt nicht deren Struktur im Referenzbaum nach, sondern gibt die Anzahl der möglichen Beziehungen wieder. Nach Formel (4.3) ergeben sich für drei Knoten, drei Beziehungen. Anhand der Summe der Distanzen für jede Knotenbeziehung, kann man eine Aussage über die Homogenität der Knotenclique machen. Je kleiner die Summe desto verwandter sind die Knoten untereinander, und desto sicherer ist die taxonomische Aussage zu dem ORF.

Summe der Distanzen errechnet, die die Ergebnisknoten untereinander besitzen (Abbildung 4.7). Die Summe wird sowohl auf die Anzahl der Distanzen (Formel (4.3)), als auch auf die durchschnittliche Distanz im gesamten Baum (Formel (4.2)), normiert (Formel (4.1)).

$$h = \frac{\sum_{i=0}^{m} \sum_{j=i+1}^{m} d(k_i, k_j)}{o^{(4.3)} * t^{(4.2)}} \tag{4.1}$$

h: Homogenität; m: Anzahl der Ergebnisknoten pro ORF; k_i: Knoten i; o: Anzahl der Distanzen zwischen m Knoten; t: durchschnittliche Distanz über alle Knoten des gesamten Baum

$$t = \frac{\sum_{i=0}^{n} \sum_{j=i+1}^{n} d(k_i, k_j)}{\sum_{i=1}^{n} (n - i)} \tag{4.2}$$

t: durchschnittliche Distanz über alle Knoten des gesamten Baum; n: Anzahl der Knoten im Referenzbaum; k_i: Knoten i

$$o = \sum_{i=1}^{m} (m - i) \tag{4.3}$$

o: Anzahl der Distanzen zwischen m Knoten; m: Anzahl der Ergebnisknoten pro ORF

Der Benutzer kann die Homogenitätsberechnung beeinflussen, indem er die Anzahl der besten Ergebnisknoten einstellt. Ein konkreter maximaler Wert zur Filterung der Ergebnisse kann nicht genannt werden. Man muss die Ergebnisse von Fall zu Fall durchschauen

und anhand derer einen maximalen Homogenitätswert festlegen. Zum Beispiel kann man in der Ausgabe aus Tabelle 4.1 erkennen, dass die Ergebnisse für seq3 sehr inhomogen sind. Ein möglicher maximal zulässiger Homogenitätswert, nach dem hier gefiltert werden könnte, ist daher 0.3.

Visualisierung Ziel ist es einen Überblick über die taxonomische Zusammensetzung eines Metagenoms zu bekommen. Da aber die Ausgabedatei sehr groß und damit unübersichtlich sein kann, sollten die Ergebnisse der Ausgabe visualisierbar sein.

Eine Möglichkeit zur Visualisierung, die hier als Beispiel vorgestellt werden soll, ist die graphische Darstellung anhand von *iTOL*. iTOL (Interactive Tree Of Life) ist ein mächtiges Online-Tool zur Visualisierung von Bäumen, das erst kürzlich von Ivica Letunic veröffentlicht [43] wurde. iTOL erlaubt Informationen auf einem selbst definiertem Baum graphisch darzustellen. Zum Beispiel wurde der Baum des Lebens von Peer Bork mit iTOL dargestellt (Anhang A.1).

Um einen Überblick über die Verteilung der phylogenetischen Einheiten im untersuchten Metagenom zu bekommen, könnte man jeden Knoten des Baumes mit einem Häufigkeitswert versehen, der proportional zur Häufigkeit des Taxons im Metagenom ist. Eine Möglichkeit das Vorkommen eines Taxons zu bestimmen, ist durch einfaches Abzählen des Auftretens des entsprechenden Knotens in der Ergebnisausgabe. Eine ausgeklügeltere Variante berücksichtigt auch den Fehler der jeweiligen Zuweisung. Je kleiner der Fehler, desto verlässlicher die Aussage. Für jeden Knoten werden die reziproken Fehler aller Zuweisungen summiert (Formel (4.4)). Je größer der Wert, desto häufiger das Aufkommen des Knotens/Taxons.

$$i(k) = \sum_{i=0}^{n} \frac{1}{e_i} \tag{4.4}$$

$i(k)$: Intensität des Knotens/Taxons k; n: Anzahl der Zuweisungen für Knoten k; e_i: Fehler einer Zuweisung i zu Knoten k

Somit kann man schnell einen quantitativen Querschnitt der taxonomischen Zusammensetzung des Metagenoms erhalten. Abbildung 4.8 zeigt eine derartige Beispielvisualisierung.

4.2.4. Software-Architektur

Um den Überblick über die Programmstruktur zu bewahren, wird das System in größeren Einheiten, sogenannten Paketen, organisiert. Paketdiagramme geben eine abstrakte Gesamtsicht auf das System. Folgendes Paketdiagramm (Abbildung 4.9) zeigt die Gliederung des Programms dieser Diplomarbeit. Dabei wurden die wichtigsten Klassen, die bereits vorgestellt wurden, in das Paketdiagramm integriert.

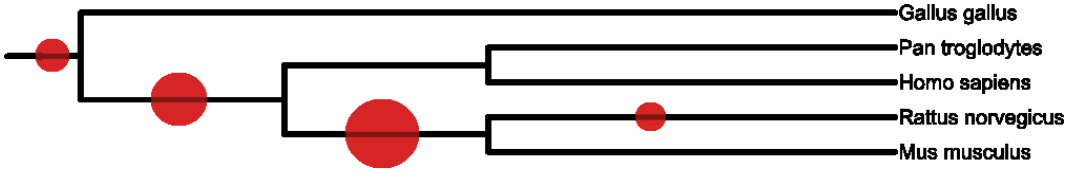


Abbildung 4.8.: Die Ergebnisse einer fiktiven taxonomischen Analyse werden in diesem Beispielbaum graphisch aufgezeigt. Die Größe der roten Kreise ist proportional zur Intensität bzw. zum Vorkommen des jeweiligen Taxons im untersuchten Datensatz. Man kann zum Beispiel erkennen, dass im Datensatz hauptsächlich Spezies aus der Klasse Säugetiere, genauer aus der Ordnung Nagetiere, vorkommen. Die Markierung von internen Knoten, also höheren Taxa, bedeutet, dass die Sequenzen aus Spezies stammen, die nicht im Baum vorhanden sind, aber vom gleichen Vorfahren abstammen. Die Intensität der Kreise kann durch verschiedene Ansätze — einer ist in Formel (4.4) dargestellt — errechnet werden.

4.3. Zusammenfassung

Die Software wurde unter Berücksichtigung der Anforderungen implementiert und steht als vollständiges Programm zur Verfügung (Anhang F). Die Vollautomatisierung der Blast2Tree Methode erlaubt nun nicht nur die komfortable Analyse von großen Metagenomdatensätzen, sondern ermöglicht auch das systematische Verbessern dieser Methode anhand geeigneter Evaluierungsverfahren. Außerdem ist es möglich, das Programm direkt mit den anderen vorgestellten Methoden zur taxonomischen Analyse zu vergleichen, um zum Beispiel Laufzeitverhalten, Weiterprozessierbarkeit der Ergebnisse usw. zu untersuchen.

Ziel ist es nach der Diplomarbeit die Implementierung auch als Web-Service für kleine Metagenome ($<$ 1000 Sequenzen) anzubieten.

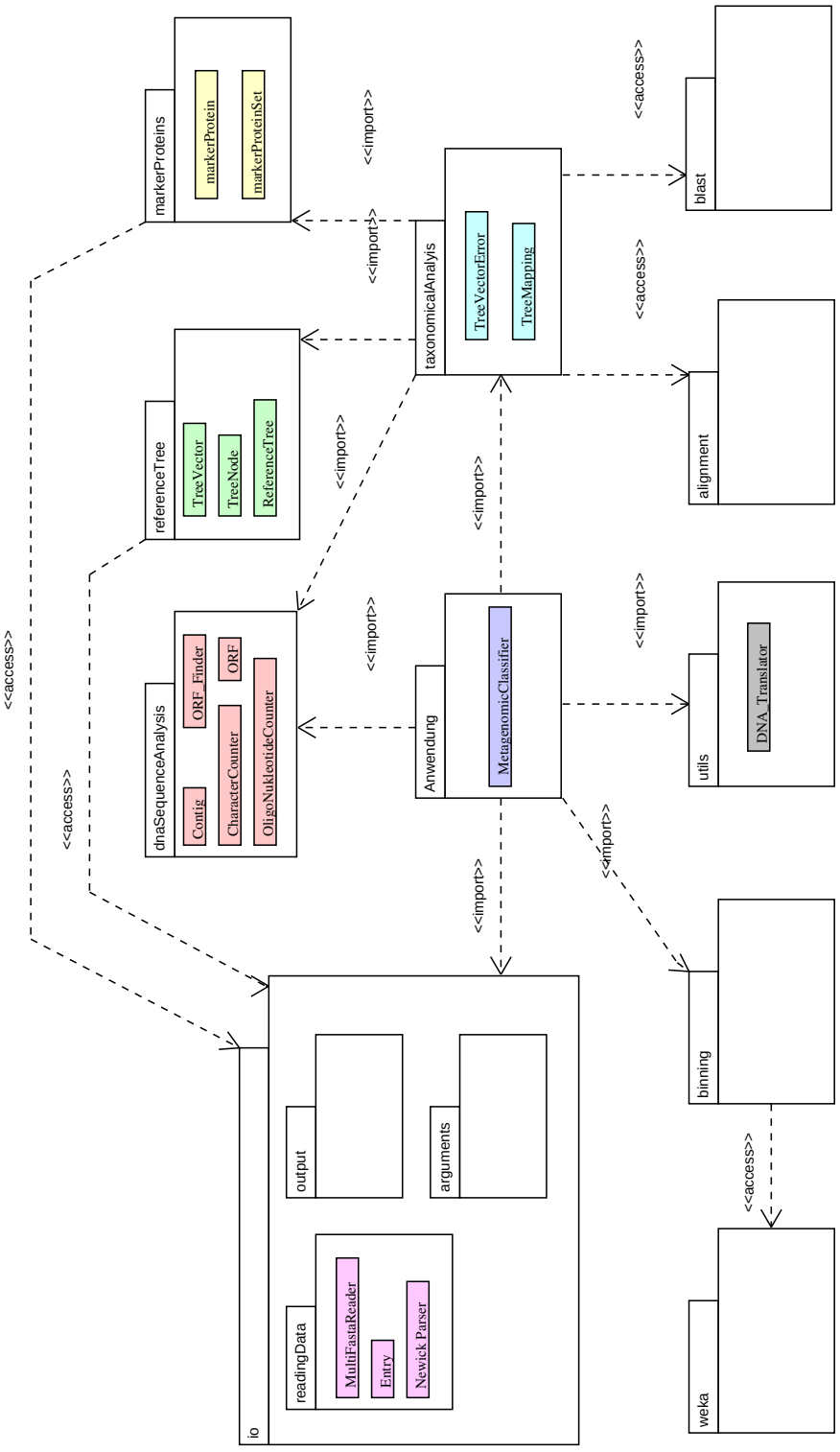


Abbildung 4.9.: Die Programmstruktur wird hier an einem Paketdiagramm (nach UML 2 Standard) aufgezeigt. Klassen wurden dabei farblich hervorgehoben.

5. Validierung der Blast2Tree Methode

5.1. Jack-Knife Prinzip

Mit Hilfe der Implementierung wurde die Blast2Tree Methode voll automatisiert. Dies erlaubt nun die taxonomische Klassifizierung durch das sogenannte Jack-Knife Verfahren systematisch zu evaluieren und die Methodik somit iterativ zu verbessern. Dazu wird jeder Knoten des Baumes einmal entfernt (Abbildung 5.1). Für jeden entfernten Knoten, werden alle Proteine der Taxa, die zu diesem Knoten gehören, wieder in den Baum eingeordnet. Dabei ist zu beachten, dass bei der Wiedereinordnung eines Proteins der Teilbaum, der unterhalb des gelöschten Knotens liegt, nicht mehr verfügbar ist. Als ideales Ergebnis würde man also die Zuordnung eines Proteins zu dem Vaterknoten des gelöschten Knotens erwarten (Abbildung 5.2). Anhand der Abweichung vom Idealergebnis kann die Methodik auf algorithmischer Seite verbessert werden.

5.2. Validierung

5.2.1. Verwendeter Baum und Markersatz

Als Referenzbaum dient der Baum des Lebens von Ciccarelli und Peer Bork [2], der in Kapitel 3.1.2 vorgestellt wurde. Dieser basiert auf der Verknüpfung von 31 Gruppen orthologer Proteine (COGs) aus 191 komplett sequenzierten Genomen und wurde von Effekten des horizontalen Gentransfers bereinigt. Die 31 relevanten COGs und deren Proteine wurden der STRING Datenbank [39] entnommen. In Version 6.3 sind dies 6

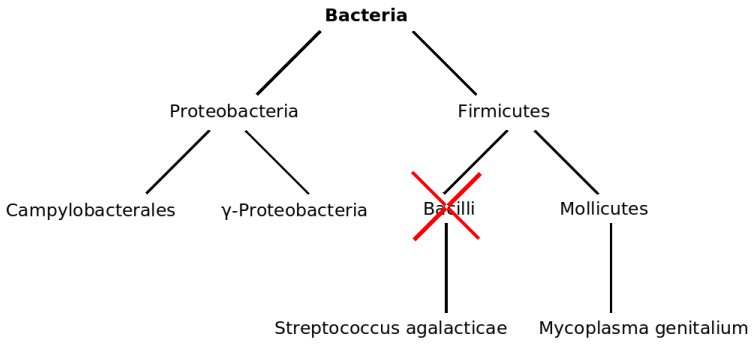


Abbildung 5.1.: Im ersten Schritt des Jack-Knife Verfahrens wird ein Knoten aus dem Baum gelöscht.

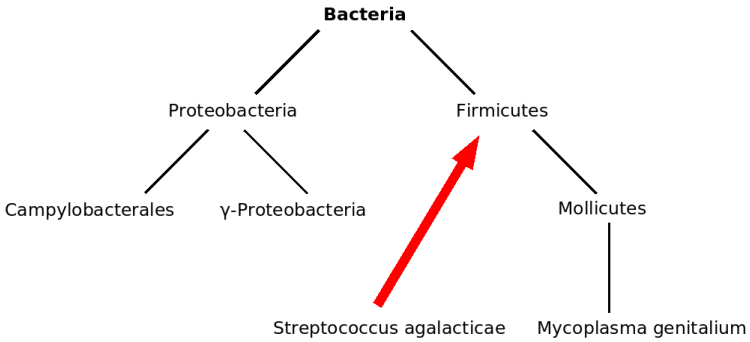


Abbildung 5.2.: Im zweiten Schritt wird versucht die Proteine der Taxa, die zu dem gelöschten Knoten gehören, wieder in den Baum zu mappen. Idealerweise wird der Vaterknoten des entfernten Knotens als Ergebnis gewünscht.

537 Proteine. Der Referenzbaum liegt mit Angabe der Zweiglängen im Newick Format vor. Das Programm liest den Baum ein, weist den internen Knoten NCBI Taxonomie Nummern zu und berechnet für alle Knoten (interne als auch Blattknoten) die Distanzen zu den Blattknoten des Baumes, so dass für jeden Knoten ein Distanzvektor vorliegt. Dieser Distanzvektor kann auch als graphisches Profil eines Knotens dargestellt werden (Abbildung 5.3).

Da der verwendete Baum 191 Blattknoten besitzt, gibt es insgesamt 381 Knoten (Formel (3.2)). Eine Printversion des so prozessierten Baumes liegt zur Übersicht im Anhang C.1 bei. Jeder dieser 381 Knoten wird nun einmal gelöscht und die Proteine der Spezies, die darunter liegen, wieder in den Baum platziert. Insgesamt müssen für diese 381 Knoten 69 675 mal Proteine in den Baum eingeordnet werden. Um die Qualität des Verfahrens beurteilen zu können, muss ein Bewertungsschema für die Mappings eingeführt werden.

5.2.2. Bewertungsschema

Das Bewertungsschema misst die Abweichung zwischen gefundenem Knoten z und erwartetem Knoten y. Eine korrekte Zuweisung ist dann erfolgt, wenn der erwartete Knoten y gleich dem Vaterknoten des gelöschten Knotens ist. Die Abweichung zwischen z und y kann taxonomisch und phylogenetisch gemessen werden:

- taxonomisch: Rangdifferenz(z,y)
- phylogenetisch:
  - Kantendifferenz(z,y)
  - Distanzdifferenz(z,y)

Für jedes Mapping eines Proteins werden alle drei Werte ermittelt.

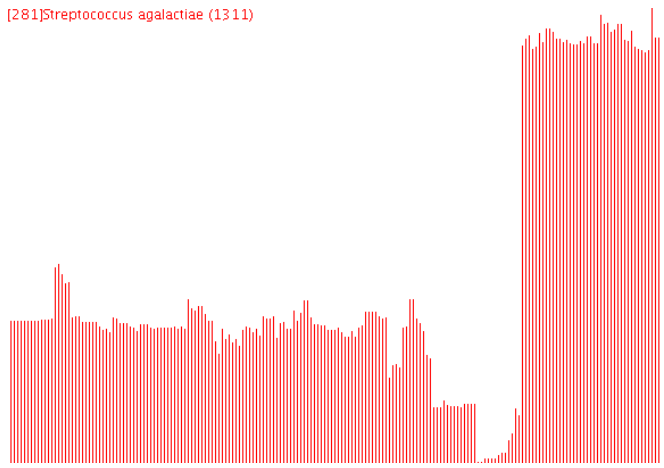


Abbildung 5.3.: Die Distanzen eines Knotens zu den Blattknoten des Baumes lassen sich auch graphisch als Profil darstellen. Hier ist als Beispiel das Distanzprofil des Speziesknotens von *Streptococcus agalactiae* dargestellt. Jeder Balken repräsentiert dabei die Entfernung des Knotens zu allen 191 Spezies des verwendeten Referenzbaumes [14]. In der Mitte des Bildes sind besonders niedrige Werte zu erkennen. Diese stellen die Distanzen zu den nächsten Verwandten im Phylum der Klasse Bacilli dar. Links davon befinden sich die Entfernungen zu den restlichen Bacteria, wohingegen der rechte Teil der Graphik die Distanzen zu den Eukaryota und Archaea darstellt.

5.2.3. Verbesserung der Blast2Tree Methode

Um die Aussagekraft der Ergebnisse zu verbessern, gibt es auf algorithmischer Seite verschiedene Ansatzpunkte:

- Verwenden einer anderen Gleichung zur Umrechnung von Alignment-Scores zu Distanzen (Formel (3.3))
- Alternativen zur Kleinst-Fehlerquadrate-Berechnung zur Auswahl des besten Knotens (Formel (3.6))

Da keine andere Gleichung zur Distanzberechnung bekannt ist, konzentriert sich der Rest der Arbeit mit der Verbesserung zur Auswahl des ähnlichsten Knotens. Dabei werden Alternativen für die verwendete Kleinst-Fehlerquadrate-Berechnung gesucht.

Distanzmaße

Sieht man die Distanzvektoren des Knotens und des einzuordnenden Proteins als Vektoren in einem euklidischen Raum mit n Dimensionen, wobei n die Anzahl an Blattknoten im Baum ist, so bieten sich verschiedene Distanzmaße an. Eine allgemeine Formel zur Berechnung von Distanzen liefert die Minkowski-Norm (5.1).

$$d = \left(\sum_{i=0}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{5.1}$$

d: Distanz zwischen x und y; n: Anzahl der Dimensionen im Raum; p: Grad der Distanz; x: Punkt x; y: Punkt y.

Aus der Minkowski-Norm lassen sich bekannte Distanzmaße ableiten, die für diesen Fall interessant sind:

- Manhatten Distanz ($p = 1$)
- Euklidische Distanz ($p = 2$)
- Chebyshev Distanz ($p = \infty$)

Die Chebyshev Distanz betrachtet nur die maximale Distanz eines Wertepaares zweier Vektoren und vernachlässigt somit den Rest der Information. Sie ist daher gänzlich ungeeignet für die Bestimmung des ähnlichsten Knotens. Die Euklidische Distanz ist bereits in leicht abgewandelter Form in der Kleinst-Fehlerquadrate-Berechnung enthalten. Es gilt nun zu überprüfen, ob die Manhattan Distanz bessere Ergebnisse liefert als das euklidische Distanzmaß.

Eine Maßzahl, die sich ebenso eignet den Zusammenhang zwischen zwei Wertevektoren zu erfassen, ist der Pearson'sche Korrelationskoeffizient. Voraussetzung ist allerdings, dass ein linearer Zusammenhang besteht. Dies ist der Fall, da man für ein Wertepaar im

Idealfall gleiche Distanzen hätte. Allerdings setzt der Korrelationskoeffizient Normalverteilung der Werte in beiden Vektoren voraus, wovon man in diesem Fall nicht ausgehen kann. Daher wird auf eine Verwendung dieser Maßzahl verzichtet.

Das Jack-Knife Verfahren wurde für beide Distanzmaße durchgeführt und die jeweils 69 675 Mappings nach oben ausgeführtem Bewertungsschema analysiert. Die Ergebnisse für beide Distanzmaße wurden zum direkten Vergleich graphisch dargestellt (Abbildung 5.4). Aus allen drei Vergleichskategorien geht eindeutig hervor, dass das Manhattan Distanzmaß zu korrekteren Ergebnissen führt, als das euklidische.

Gewichtung von Alignments

Um die Klassifikationsgenauigkeit weiter zu verbessern, wurden schlechte Zuordnungen näher analysiert. Ein bildliches Beispiel einer falschen Zuweisung und dem eigentlichen richtigen Knoten zeigen die Graphen in Abbildung 5.5. Der Algorithmus begeht hier einen Klassifikationsfehler, da er ein eindeutiges Merkmal, nämlich die niedrigen Distanzen zu gewissen Spezies knoten, übersieht. Eine Möglichkeit besteht nun darin, Alignments mit höherem Score stärker zu gewichten als Alignments mit niedrigerem Score. Dies macht biologisch durchaus Sinn, da Alignments an Aussagekraft verlieren, wenn die zwei beteiligten Spezies phylogenetisch sehr weit voneinander entfernt sind.

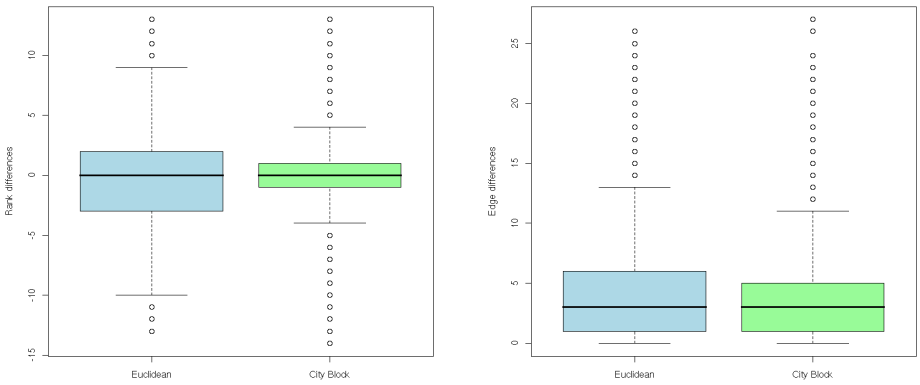
Dazu werden ein maximaler und ein minimaler Wert festgelegt, zwischen denen ein linearer Gradient verläuft. Anhand der Lage eines Distanzwertes in diesem Gradienten wird ein Faktor w ermittelt, der dann die Distanzdifferenz zwischen Knoten und Protein gewichtet (Formel (5.2)).

Drei mathematische Ansätze wurden entwickelt, implementiert und untersucht (Gleichungen (5.3) bis (5.5)). Gleichung (5.3) skaliert die Rohdistanzen auf das Maximum der Rohdistanzen, beurteilt also die Alignments unabhängig vom betrachteten Knoten. Gleichung (5.4) geht dagegen vom Knoten aus und favorisiert so Proteine, deren Scores dem des Knotens ähnlich sind. Gleichung (5.5) betrachtet wiederum die angepassten Proteindistanzen zur Berechnung des Gewichtes und ist damit sowohl vom Knoten abhängig, als auch vom ursprünglichen Distanzvektor.

$$e = \sum_{i=0}^{n} (w(i) * |dk(i) - do'(i)|) \tag{5.2}$$

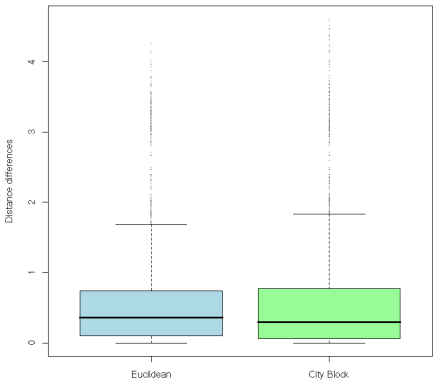
e: Fehler zwischen Protein- und Knotenvektor; dk: Distanzvektor des betrachteten Knotens; do': Skalierter Distanzvektor des Proteins; $w(i)$: Faktor zur Gewichtung der Distanzdifferenz zwischen Protein und Knoten bezogen auf Blattknoten i; n: Anzahl der Blattknoten im Baum.

$$\begin{aligned} max &= max(do) \\ min &= 0 \\ w(i) &= \frac{(max - do(i))}{(max - min)} \end{aligned} \tag{5.3}$$



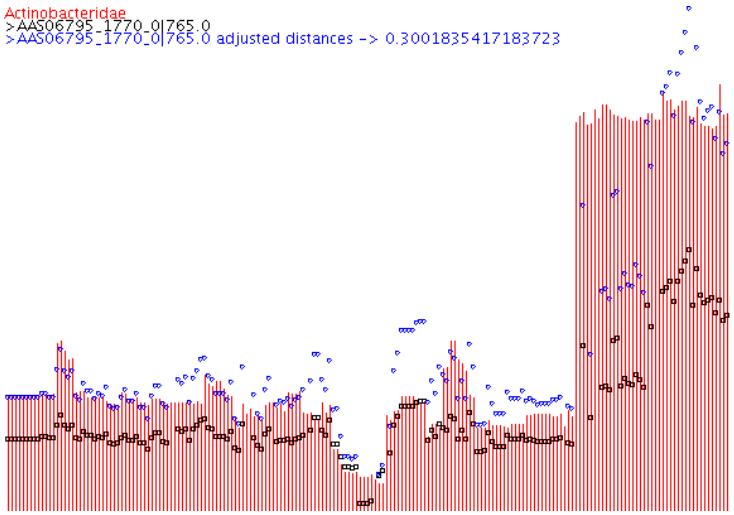
(a) Rangdifferenzen

(b) Kantendifferenzen

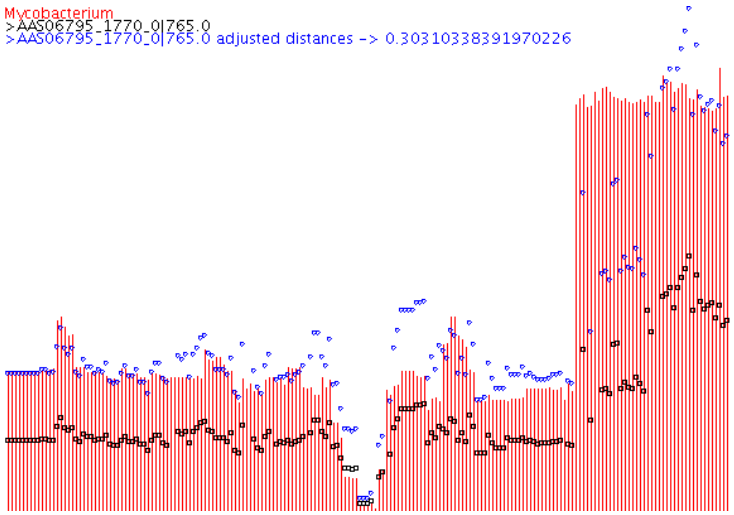


(c) Distanzdifferenzen

Abbildung 5.4.: Die Jack-Knifing Prozedur wurde auf jeden Knoten des Bork-Baumes, einmal unter Verwendung der euklidischen (hellblau) Distanz und einmal der Manhattan Distanz (grün), angewandt. Die Ergebnisse werden in drei Boxplots miteinander verglichen. Dabei werden in Plot (a) die Rangdifferenzen, in Plot (b) die Kantendifferenzen und die Plot (c) die Distanzdifferenzen zwischen erwarteten und erhaltenen Knoten aller Mappings gegenübergestellt. Je näher die Quartilen der Boxplots am Nullpunkt sind, desto besser ist das Gesamtmapping. Rangdifferenzen können dabei auch einen negativen Werte x annehmen. Dies bedeutet, dass der gefundene Knoten einen um x Ränge zu hohen Rang gegenüber dem erwarteten Knoten besitzt.



(a) Gefundener Knoten



(b) Erwarteter Knoten

Abbildung 5.5.: Der Spezies knoten *Mycobacterium avium* wurde im Rahmen der Jack-Knife Validierung entfernt und dessen Proteine wieder in den Baum eingeordnet. Für die Klassifikation des COG102-Proteins wurde der Knoten gefunden, dessen Distanzprofil in Graph (a) dargestellt ist. Die Rohdistanzen sind als kleine schwarze Quadrate im Profil mitaufgetragen, ebenso wie die skalierten Distanzen als blaue Dreiecke. Der Fehler, der sich aus den angepassten Distanzen errechnet, befindet sich oben mittig im Bild. Graph (b) zeigt dagegen den gewünschten Knoten.

do: Distanzvektor des Proteins; $w(i)$: Faktor zur Gewichtung der Distanzdifferenz zwischen Protein und Knoten bezogen auf Blattknoten i.

$$\begin{aligned} max &= max(dk) \\ min &= min(dk) \\ w(i) &= \frac{(max - dk(i))}{(max - min)} \end{aligned} \tag{5.4}$$

dk: Distanzvektor des betrachteten Knotens; $w(i)$: Faktor zur Gewichtung der Distanzdifferenz zwischen Protein und Knoten bezogen auf Blattknoten i.

$$\begin{aligned} max &= max(do') \\ min &= min(do') \\ w(i) &= \frac{(max - do'(i))}{(max - min)} \end{aligned} \tag{5.5}$$

do': Skalierter Distanzvektor des Proteins; $w(i)$: Faktor zur Gewichtung der Distanzdifferenz zwischen Protein und Knoten bezogen auf Blattknoten i.

Die Wirkung dieser Modifikationen wurden mittels Jack-Knifings verglichen. Bevor jedoch auf die Ergebnisse eingegangen wird, soll noch eine weitere Überlegung vorgestellt werden, die zur Verbesserung der Identifizierung des richtigen Knotens dienen soll.

Betrachtung des Profilverlaufes

Ein weiterer Ansatz ist die Betrachtung des Profilverlaufes. Für das menschliche Auge ist es auf Anhieb einfach, den richtigen Knoten zu identifizieren (siehe wieder Abbildung 5.5). Dabei stellen wir uns die Werte von angepassten Distanzen und den Distanzen des Knotens als Kurven vor und vergleichen deren Verläufe. Sind die Verläufe ähnlich, besteht wahrscheinlich auch eine Ähnlichkeit zwischen Knoten und Protein.

Das Verhalten einer Kurve lässt sich mit der Ableitung beschreiben. Die erste Ableitung gibt die Steigung m wieder (Formel (5.6)).

$$m = \frac{\triangle x}{\triangle y} \tag{5.6}$$

m: Steigungskoeffizient; x: Wert auf der x-Achse; y: der zu x entsprechende Wert auf der y-Achse.

Da wir es hier mit diskreten Werten in einem Vektor zu tun haben, ist $\triangle y = 1 = konst.$ und $\triangle x$ die Differenz zwischen einem Wert $x(i)$ und seinem Nachfolgendenwert $x(i+1)$. Zwei Distanzvektoren lassen sich vergleichen, indem man die Steigungen an alle Punkten miteinander vergleicht. Ein konkreter Wert kann durch Aufsummieren aller

Steigungsdifferenzen zwischen den Punktsteigungen beider Vektoren ermittelt werden (Formel (5.7)).

$$e_m = \sum_{i=0}^{n} |(do'(i+1) - do'(i)) - (dk(i+1) - dk(i))| \tag{5.7}$$

e_m: Gesamtfehler der Steigungen zwischen do' und dk; n: Anzahl der Blattknoten im Baum; dk: Distanzvektor des betrachteten Knotens; do': Skalierter Distanzvektor des Proteins.

Dieser Wert e_m kann zum normalen Fehler e addiert werden. Damit erhält man einen Gesamtfehler e_g, der als Auswahlkriteriums für die Suche nach dem dem ähnlichsten Knoten für ein Protein herangezogen werden kann. Auch der Einsatz dieses Verfahrens soll evaluiert werden.

5.2.4. Ergebnisse

Evaluierung der Ergebnisse

Es wurde untersucht, wie sich die oben angeführten Gewichtungsfunktionen und die Anwendung des Steigungsfehlers auf die über 69 000 Mappings der Jack-Knife Methode auswirken. Als Basis-Distanzmaß wurde dabei die Manhattan Distanz verwendet, die sich ja bereits gegenüber der euklidischen als besser erwiesen hat. Die Ergebnisse wurden graphisch in Form von Boxplots ausgearbeitet (Abbildung 5.6).

Es zeigt sich, dass die Gewichtung der Distanzunterschiede durch die Formeln (5.3) und (5.5) eine bessere Zuordnung der Proteine ermöglicht. Auch die Berücksichtigung der Summe der Steigungsunterschiede bringt eine geringe Verbesserung mit sich.

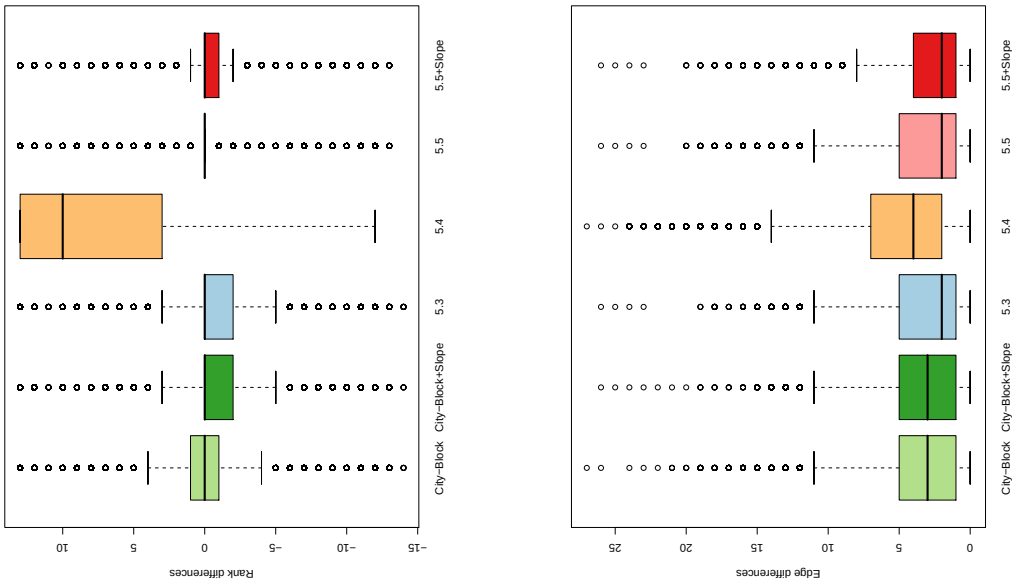
Eine differenziertere Bewertung der Resultate im Anhang C.2 nach folgenden Gesichtspunkten

- exakter Treffer (gefundener Knoten = erwarteter Knoten)
- taxonomischer Treffer (taxon. Beschriftung des gefundenen Knotens = taxon. Beschriftung des erwarteten Knotens)
- auf Stammeslinie (gefundener Knoten liegt auf der Stammeslinie des erwarteten Knotens)
- nicht auf Stammeslinie (negativ zu bewerten)

macht noch einmal deutlich, dass die Formel (5.5) mit Einberechnung der Steigungsunterschiede die besten Ergebnisse erzielt. Auf Basis genannter Kriterien soll für zukünftige Evaluationen ein neues und differenzierteres Bewertungsschema geschaffen werden.

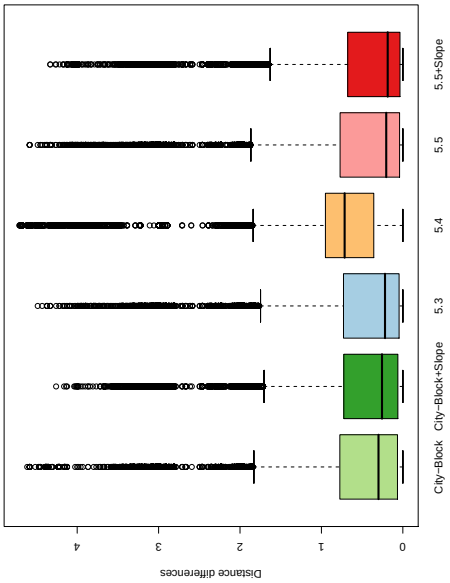
Weitere Schlussfolgerungen

Die Jack-Knife Methode kann nicht nur dazu dienen, Erkenntnisse methodischer Art zu gewinnen, sondern auch inhaltlicher Art.



(a) Rangdifferenzen

(b) Kantendifferenzen



(c) Distanzdifferenzen

Abbildung 5.6.: Die Zuordnungsqualität der 6 537 Proteine aus den 31 COGs des verwendeten Referenzbaumes wurden in 69 000 Mappings anhand des Jack-Knife Verfahrens untersucht. Dabei sollte der Einfluss folgender Parameter verglichen werden (v.l.): Manhattan Distanz (hellgrün), Manhattan Distanz mit Berücksichtigung der Steigungsunterschiede nach Formel (5.7) (dunkelgrün), Gewichtung der Distanzdifferenzen nach Formel (5.3) (hellblau), Gewichtung nach Formel (5.4) (orange), Gewichtung nach Formel (5.5) (rosa) und Gewichtung nach Formel (5.5) mit Berücksichtigung der Steigungsunterschiede nach Formel (5.7) (rot). Die Ergebnisse für Rang- (a), Kanten- (b) und Distanzdifferenzen (c) sind in den drei Boxplots dargestellt.

COG Performanz Der verwendete Referenzbaum hat die Eigenschaft, dass er aus 31 orthologen Gruppen erstellt wurde und damit einen gemittelten Baum mehrerer Genbäume darstellt. Daher interessiert es, wie stark die Genbäume dieser Gruppen vom Bork-Baum abweichen und inwiefern sich damit Unterschiede in der Zuordnungsqualität in Abhängigkeit von der verwendeten Markerproteinfamilie (COG) ergeben.

Eine punktebasierte Auswertung, die die Ergebnisse in Abhängigkeit der einzelnen COGs untersuchte, ergab, dass es sehr wohl Unterschiede in der Qualität der Zuordnung gibt (Anhang C.3). Dies lässt darauf schließen, dass die Genbäume von besser arbeitenden COGs dem gemittelten Bork-Baum ähnlicher sind als schlechter arbeitende. Eine Möglichkeit zur Verbesserung der Zuweisungsqualität besteht eventuell darin, für jeden COG einen eigenen Baum bereitzustellen.

Zuordnungsqualität auf Phylum-Ebene Eine weitere Fragestellung lautet, ob die Zuordnungsqualität innerhalb des Baumes variiert. Als taxonomische Vergleichsebene wurden die Phyla gewählt. Das heisst, es wurde untersucht, ob die Qualität sich zwischen den Phyla unterscheidet. Dazu wurden alle Mappings nach ihrer Zugehörigkeit zu dem jeweiligen Phylum geordnet und nach bekanntem Schema ausgewertet (Anhang C.4: Abbildungen 1–3).

Es zeigt sich auch hier, dass die Qualität unterschiedlich ist. Zum einen kann hier wiederum der gemittelte Baum verantwortlich sein, zum anderen sind die verschiedenen Phyla unterschiedlich repräsentiert. Insgesamt enthält der Bork-Baum 191 Arten. Davon fällt ein Drittel allein schon auf das Phylum der Proteobacteria (63 Spezies), gefolgt vom Phylum Firmicutes (39) und den Eukaryota (23). Viele Phyla sind deutlich durch weniger Spezies repräsentiert, wie die zum Beispiel das Phylum Planctomycetes (2) oder die Chlorobien Gruppe (3). Eine Sequenz, die sehr viele schlechte Alignments besitzt, könnte dann per Zufall eher den stark repräsentierten Phyla zugeordnet werden.

5.3. Zusammenfassung

Durch das sogenannte Jack-Knife Verfahren wurde der Einfluss verschiedener Parameter auf die Zuordnungsqualität der Blast2Tree Methode untersucht. Dabei wurden sowohl verschiedene Distanzmaße, als auch unterschiedliche Funktionen zur Gewichtung von Distanzunterschieden verglichen. Es zeigte sich, dass die Anwendung des Manhattan Distanzmaßes und einer dazugehörigen Gewichtung der Distanzunterschiede nach Formel (5.5) die besten Ergebnisse erzielte. Aufgrund dieser Evaluierung wird diese Kombination als Standardverfahren zur Berechnung des Fehlers in das Programm integriert.

Zudem ließ das Evaluierungsverfahren auch Schlüsse über die Zuordnungsqualität in Abhängigkeit der Baumeigenschaften zu. So wurde für den aus 31 orthologen Gruppen (COG) erstellten Referenzbaum festgestellt, dass die Qualität für die Einordnung eines Proteins in diesen Baum von der homologen orthologen Gruppe abhängt. Es wird angenommen, dass die jeweiligen Genbäume der einzelnen COGs vom Baum divergieren. Es konnte eine Rangliste erstellt werden, die die Ähnlichkeit des Genbaumes eines COGs mit dem Referenzbaum angibt und damit auch die Zuordnungsqualität.

Unabhängig von verwendeten COG, variiert aber auch die Zuordnungsqualität innerhalb des Baumes, was auf Phylum Ebene bestätigt wurde. Dies sind Erkenntnisse, die man bei der Weiterentwicklung und Verbesserung der Blast2Tree Methodik beachten muss.

Es gibt bereits mehrere Ansatzpunkte, um die Aussagekraft der Ergebnisse zu verbessern:

- Genvorhersage
  - Verwendung einer echten Genvorhersage anstatt „nur“ der ORF-Suche
  - Filterung der gefundenen ORFs nach Frameshift Mutationen
- Algorithmus
  - Einsatz multipler Alignments anstelle paarweiser Alignments zur Ermittlung der Distanzen, da die verwendeten Bäume auch über multiple Alignments erstellt wurden
- Weitere Validierung
  - Testen von partiellen homologen Proteinen, d.h. was ist die Mindestlänge von korrekt zuordbaren Homologen
  - Testen verschiedener Bäume: für jeden COG einen Baum, Verwendung eines 16s rRNA Baumes anstelle der COG Bäume

6. Binning von Metagenomen

Ziel des Binnings ist das Zusammenfassen von DNA-Fragmenten zu deren Ursprungsgenomen. Kapitel 3.4 hat die wichtigsten Merkmale von DNA Sequenzen vorgestellt, anhand derer solch ein Clustering durchgeführt werden kann. Wie bereits berichtet, haben bedeutende Arbeiten gezeigt, dass genomweite Oligonukleotidfrequenzen einem phylogenetischen Fingerabdruck gleichkommen [8, 9, 10, 11, 12, 13]. Jedoch hat sich keines dieser Veröffentlichungen mit den für die Metagenomik üblichen kurzen Sequenzlängen befasst.

Daher widmet sich dieser Teil der Diplomarbeit der Anwendung des Binnings auf Metagenom-Sequenzen mithilfe von Oligonukleotidfrequenzen. Das Clustering soll anhand von künstlichen Metagenomdatensätzen und verschiedenen Nukleotidmustern getestet und optimiert werden.

6.1. Parameter

Es gibt eine Reihe von Parametern, die auf die Qualität des Binnings Einfluss nehmen. Dabei kann zwischen unbeeinflussbaren und beeinflussbaren unterschieden werden:

6.1.1. Unbeeinflussbare Parameter

In dem Moment, in dem die DNA des Metagenoms bereits fertig sequenziert wurde, sind folgenden Parameter nicht mehr beeinflussbar:

Zusammensetzung des Metagenoms

Das diskriminatorische Potential nimmt mit zunehmender Anzahl an vorhandenen Spezies im Metagenom ab. Im Kapitel 3.2.2 wurde beschrieben, dass, um die Diversität des Metagenoms zu erniedrigen bzw. bestimmte Spezies anzureichern, Filter benutzt werden. Zu beachten ist aber, dass ähnliche Spezies nicht nur eine ähnliche Physiologie haben und daher durch Filter schlecht trennbar sind, sondern auch ein ähnliches Genom besitzen. Daher sind sie auch durch molekulare Fingerabdrücke, wie den Oligonukleotidfrequenzen, nur bedingt voneinander trennbar.

Durchschnittliche Sequenzlänge

In den Fällen, bei denen man ähnliche Spezies/Genome im Metagenom hat, ist die durchschnittliche Länge der Reads von entscheidender Bedeutung für die Trennleistung des Binnings. Aber gerade die Probendiversität hat wiederum Einfluss auf die Assemblierungsrate und damit auf die Länge der Contigs. Daher kann man hier nur durch geeignete Klonierungs- und Sequenziertechniken Einfluss nehmen.

6.1.2. Beeinflussbare Parameter

Im Gegensatz zu den unbeeinflussbaren Parametern, sind nachfolgenden Parameter direkt vom Benutzer der Software beeinflussbar und ergebnisbestimmend:

Oligomuster

Zur Auswahl stehen neben den klassischen Di-,Tri-,Tetranukleotiden auch verschiedenste Muster, wie „NNxNN“ oder auch Hexa-, Hepta-, und Oktanukleotide. Das diskriminatorische Potential ist höher, je länger das Oligomuster, da es mehr „Eigenschaften“ der Sequenz betrachtet. Dies trifft jedoch nur dann zu, wenn die zu untersuchende Sequenz lang genug ist, um ihre charakteristische Verteilung an Oligonukleotiden zu erhalten[1]. Mit langen Sequenzen kann aber in der Metagenomik mit den neuen Sequenziertechnologien[2] nicht gerechnet werden. Zu kurze Oligomuster eignen sich aber aufgrund ihrer niedrigen Trennleistung nicht für komplexe mikrobielle Gemeinschaften.

Außerdem werden zu große Muster, aufgrund der möglichen Permutationen multipliziert mit der Anzahl an zu untersuchenden Sequenzen, für das Clusteringprogramm bzw. den Computer schnell zum Speicherproblem. Der Speicherbedarf und das Laufzeitverhalten in Abhängigkeit des Oligomusters sind in Abbildung 6.1 beispielhaft dargestellt. Mit einer Anzahl an Sequenzen, die typisch für ein Metagenomik-Projekt ist (∼800.000 für das Sargasso-See Metagenom) stößt man daher sehr schnell an die Grenzen des Hauptspeichers und der Laufzeit.

Es ist daher wichtig bei der Wahl des optimalen Musters, den Spagat zwischen biologischer Diversität, durchschnittlicher Sequenzlänge und zur Verfügung stehender Hardware zu schaffen.

Repräsentation der Werte

Die Frequenzwerte können zum einen roh zum weiteren Clustering hergenommen werden, oder aber auch normalisiert werden. Wobei eine Normalisierung sicherlich üblich ist, zumindest auf die Länge der Sequenz. Meist wird jedoch eine beobachtete Nukleotidfrequenz auf ihren Erwartungswert standardisiert. Kapitel 3.4.4 hat sich intensiv der Normalisierung der Werte anhand ihrer erwarteten Mononukleotidfrequenzen gewidmet.

Distanzfunktion und Clustermethode

Distanzfunktion und Clustermethode sind hier zusammengefasst, da diese meist im verwendeten Clusteringprogramm ausgewählt werden können.

Distanzfunktionen errechnen einen Wert, um zu messen, wie ähnlich oder unähnlich zwei Sequenzen anhand ihrer Oligofrequenzen sind. Übliche Funktionen zur Messung der Unähnlichkeit ist die euklidische Distanz oder die Manhattan-Distanz, welche der Formel

[1] Extremes Negativbeispiel: das Oligomuster ist genauso lang wie die Sequenz.

[2] Die Pyrosequencing-Technologie der Firma 454 Life Sciences erreicht derzeit Längen von 100–300 bp pro Read.

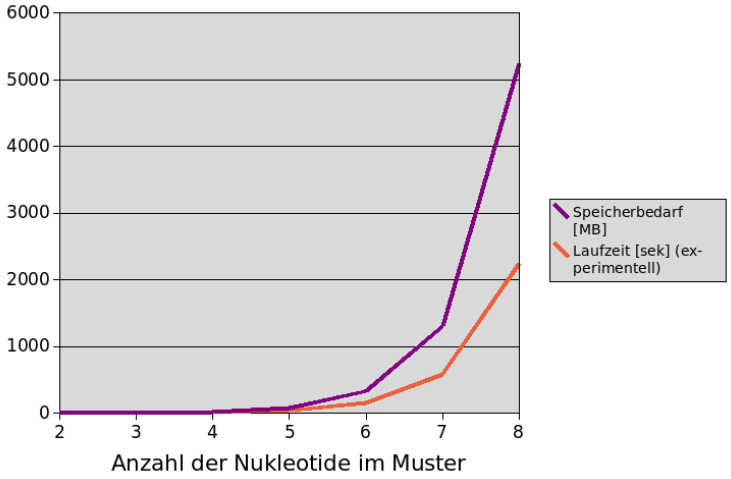


Abbildung 6.1.: Der Speicherbedarf und die Clusteringlaufzeit wird in Abhängigkeit der Länge des Oligomusters untersucht. Als Testdatensatz dienen die ersten 10 000 Contigs des Sargasso-See Metagenoms, die eine durchschnittliche Länge von 900 Residuen aufweisen. Das Clusteringprogramm CLUTO [53] hat eine Speicherplatzkomplexität von $O(N * M)$, wobei N die Anzahl der Objekte und M die Anzahl an Dimensionen ist. Speichert man die normalisierten Frequenzwerte in einem `double` Datentypen der Größe 8 Byte, so errechnet sich der Hauptspeicherbedarf mit einem Muster der Länge m nach folgender Formel: $10000 * 8\text{Byte} * 4^m$. Daher erscheint der exponentielle Verlauf des Speicherbedarfs logisch. Die Laufzeiten wurden für das Clustering der 10 000 Contigs mit der k-means Implementierung von CLUTO und k = 20 experimentell bestimmt. Auch hier ist der Verlauf exponentiell von der Länge der Oligomere abhängig.

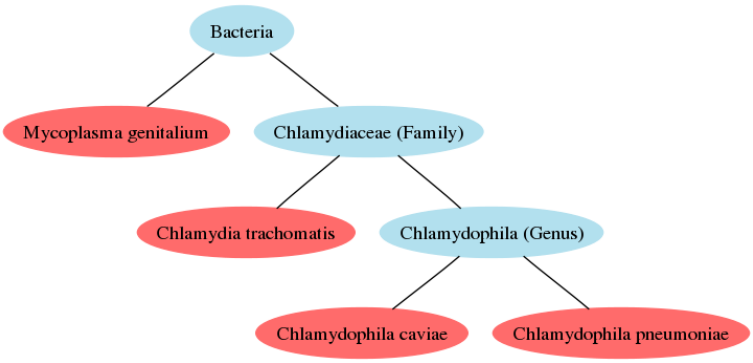


Abbildung 6.2.: Graphische Darstellung der taxonomischen Verwandtschaftsverhältnisse der am künstlichen Metagenom beteiligten Spezies (rot markiert).

(3.10) entspricht. Eine Distanzfunktion zur Messung der Ähnlichkeit, die im Bereich Binning ebenfalls gebräuchlich ist, ist der Pearson'sche Korrelationskoeffizient [8].

Für das Clustering selber stehen verschiedene Algorithmen zur Wahl. Weiß man ungefähr mit wievielen phylogenetischen Einheiten man rechnen kann, bietet sich ein partitionierender Algorithmus, wie der k-means Algorithmus, an. Ebenso kann jedoch ein agglomerativer Ansatz gewählt werden, anhand dessen man in einem Dendrogramm die Anzahl der phylogenetischen Einheiten abschätzen könnte.

Es gibt auch unüberwachte Clusteringprogramme, die die Anzahl der Cluster automatisch bestimmen. Jedoch gibt es auch bei diesen Programmen einen Parameter, der die Körnigkeit des Ergebnisses beeinflusst.

6.2. Untersuchung der Parameter an einem künstlichen Metagenom

6.2.1. Erstellung des künstlichen Metagenoms

Auswahl der Genome

Da die Genomlänge keine Rolle spielt (die Genome werden ja später fragmentiert), wurden kurze Genome ausgewählt, um die Analysezeiten kurz zu halten.

Es wurde der Organismus *Chlamydophila caviae* als Bezugsorganismus gewählt. Um zu untersuchen, wie gut das Trennungspotential der Methode bei verschiedenen Verwandtschaftsgraden funktioniert, wurden drei weitere Organismen — mit steigendem Verwandtschaftsgrad zur Bezugsspezies — hinzugezogen; zum einen *Mycoplasma genitalium*, als weit entfernten „Verwandten", dann eine Spezies des gleichen Genus, *Chlamydophila pneumoniae*, und eine der gleichen Familie, *Chlamydia trachomatis* (Abbildung 6.2). Alle Genome wurden von der NCBI Genbank Datenbank[3] heruntergeladen.

[3] http://www.ncbi.nlm.nih.gov/Genbank/

Erstellung des künstlichen Metagenom-Datensatzes

Metagenom-Daten werden meist mittels Shotgun-Sequenzierung erhoben. Um eine Shotgun-Sequenzierung zu simulieren, wurden die oben genannten Genome mithilfe der Software *Sequencer* [44] fragmentiert und die so erhaltenen Fragmente zufällig vermischt.

Da das Diskriminierungspotential der Oligonukleotidmethode kritisch von der Länge der Fragmente abhängt, wurde die durchschnittliche Länge der Fragmentebildung in den Schritten 1500 bp, 5000 bp, und 10 000 bp variiert. Somit kann die Qualität des Clusterings in Abhängigkeit der Größe der Fragmente getestet werden.

Ebenso war es von großem Interesse zu sehen, wie gut die Performanz der Methode bei Datensätze verschiedener phylogenetischer Komplexität ist. Die Komplexität steigt mit dem Grad an Verwandtschaft und der Menge an Spezies im Metagenom. Daher wurden die Metagenome in unterschiedlichen Zusammensetzungen untersucht:

- *Chlamydophila caviae*, *Chlamydophila pneumoniae*
- *Chlamydophila caviae*, *Chlamydia trachomatis*
- *Chlamydophila caviae*, *Mycoplasma genitalium*
- *Chlamydophila pneumoniae*, *Mycoplasma genitalium*
- *Chlamydophila caviae*, *Chlamydia trachomatis*, *Mycoplasma genitalium*
- *Chlamydophila caviae*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae*
- *Chlamydophila caviae*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae*, *Mycoplasma genitalium*

Insgesamt wurden somit 21 Metagenom-Datensätze geschaffen, die nun untersucht werden sollten.

6.2.2. Clustering

Wahl des Oligomusters

In einem Kompromiss zwischen maximalem diskriminatorischem Potential und minimaler Oligonukleotidlänge, sollten Clusterings mit folgenden Mustern durchgeführt werden:

- NNN (Trinukleotid) [10]
- NNNN (Tetranukleotid) [8]
- NNxNN (entspricht dem betrachten zweier nacheinander folgender Codons, mit dem Vernachlässigen der 3. ($\equiv$ x) und 6. Base)

Repräsentation der Werte

Die Werte wurden auf ihre erwarteten Mononukleotidfrequenzen standardisiert (Kapitel 3.4.4).

Metagenom-zusammensetzung	Kenn-größe	1500			5000			10000		
		NNN	NNNN	NNxNN	NNN	NNNN	NNxNN	NNN	NNNN	NNxNN
C. caviae,	Spez	0,99	0,99	0,99	1	1	1	1	1	1
M. genitalium	Sens	0,99	0,99	0,99	1	1	1	1	1	1
C. caviae,	Spez	0,59	0,62	0,41	0,78	0,89	0,93	0,91	0,97	0,99
C. trachomatis	Sens	0,59	0,62	0,61	0,78	0,89	0,93	0,91	0,97	0,99
C. caviae,	Spez	0,61	0,61	0,56	0,83	0,84	0,84	0,91	0,93	0,92
C. pneumoniae	Sens	0,62	0,62	0,63	0,83	0,84	0,84	0,91	0,93	0,92
C. pneumoniae,	Spez	0,99	0,99	0,99	1	1	1	1	1	1
M. genitalium	Sens	0,99	0,99	0,99	1	1	1	1	1	1
C. caviae,	Spez	0,82	0,83	0,75	0,8	0,92	0,97	0,7	0,92	0,99
C. trachomatis,	Sens	0,72	0,74	0,73	0,85	0,92	0,95	0,89	0,96	0,99
M. genitalium										
C. caviae,	Spez	0,65	0,65	0,66	0,7	0,74	0,77	0,74	0,86	0,93
C. trachomatis,	Sens	0,48	0,49	0,46	0,62	0,65	0,65	0,72	0,81	0,86
C. pneumoniae										
C. caviae,	Spez	0,78	0,77	0,79	0,77	0,77	0,77	0,79	0,8	0,91
C. trachomatis,	Sens	0,6	0,6	0,59	0,68	0,69	0,68	0,74	0,79	0,87
C. pneumoniae,										
M. genitalium										

Tabelle 6.1.: Die Kombinationen aller Parameter und deren Clusteringergebnisse sind in dieser Tabelle zusammengefasst. Dabei stellen die Werte jedes Postens die Mittelwerte des Clusterings nach 500 maliger Wiederholung dar. „Spez“ steht für Spezifität und „Sens“ für Sensitivität.

Clusteringmethode

Es wurde ein partitionierender Clusteringalgorithmus gewählt, da hier die Anzahl an enthaltenen Spezies in den einzelnen Metagenomen bekannt war. Geclustert wurde mithilfe der k-means Implementierung des bereits vorgestellten Clustering Tools CLUTO [53]. Der Parameter k richtete sich jeweils nach der Anzahl an Organismen im Datensatz.

Der k-means Algorithmus definiert seine Startmengen üblicherweise zufällig. Das Ergebnis eines Clusterings ist daher meist nicht wiederholbar. Um ein repräsentatives Ergebnis zu bekommen, wurde das Clustering für jedes der Metagenome, jedes Muster und jede Fragmentlänge 500 mal durchgeführt. Es mussten somit 31 500 Clusteringläufe durchgeführt werden.

6.2.3. Ergebnisse und Bewertung

Die Fragmente der mannigfaltigen Metagenome wurden anhand ihres molekularen Fingerabdruckes geclustert und sollten im Idealfall ihrem Ursprungsgenom wieder zugewiesen werden.

Da es sich bei dem Binning um ein Klassifikationsproblem handelt, wurden die dafür üblichen Kennwerte Sensitivität und Spezifität berechnet, um die Qualität eines Clusterings messen zu können. Wie man diese Kenngrößen in diesem speziellen Fall errechnet, wird im Anhang D.1 erklärt. Die Kombinationen aller Freiheitsgrade und deren Clusteringergebnisse sind in Tabelle 6.1 dargestellt. Die zweite Tabelle 6.2 mittelt die Werte nach Länge und Oligomuster.

Wie man den Ergebnissen entnehmen kann, wurden einige Vermutungen bestätigt. So

Metagenomzusammensetzung	Kenngröße	NNN	NNNN	NNxNN	1500	5000	10000
Chlamydophila caviae,	Spez	1	1	1	0,99	1	1
Mycoplasma genitalium	Sens	1	1	1	0,99	1	1
Chlamydophila caviae,	Spez	0,76	0,83	0,78	0,54	0,87	0,95
Chlamydia trachomatis	Sens	0,76	0,83	0,84	0,61	0,87	0,95
Chlamydophila caviae,	Spez	0,78	0,79	0,77	0,59	0,84	0,92
Chlamydophila pneumoniae	Sens	0,78	0,8	0,8	0,62	0,84	0,92
Chlamydophila pneumoniae,	Spez	1	1	1	0,99	1	1
Mycoplasma genitalium	Sens	1	1	1	0,99	1	1
Chlamydophila caviae,	Spez	0,77	0,89	0,9	0,8	0,89	0,87
Chlamydia trachomatis,	Sens	0,82	0,88	0,89	0,73	0,91	0,95
Mycoplasma genitalium							
Chlamydophila caviae,	Spez	0,7	0,75	0,79	0,65	0,74	0,84
Chlamydia trachomatis,	Sens	0,61	0,65	0,66	0,47	0,64	0,8
Chlamydophila pneumoniae							
Chlamydophila caviae,	Spez	0,78	0,78	0,82	0,78	0,77	0,84
Chlamydia trachomatis,	Sens	0,67	0,7	0,71	0,6	0,68	0,8
Chlamydophila pneumoniae,							
Mycoplasma genitalium							

Tabelle 6.2.: Gemittelte Werte aus Tabelle 6.1 nach Länge und Muster.

nimmt generell die Qualität des Clusterings mit zunehmender Fragmentlänge zu und mit zunehmeder Verwandtschaft zwischen den Organismen ab (Tabelle 6.2). Nur bei sehr weit entfernten Organismen scheint auch die Länge nicht mehr der entscheidende Faktor zu sein.

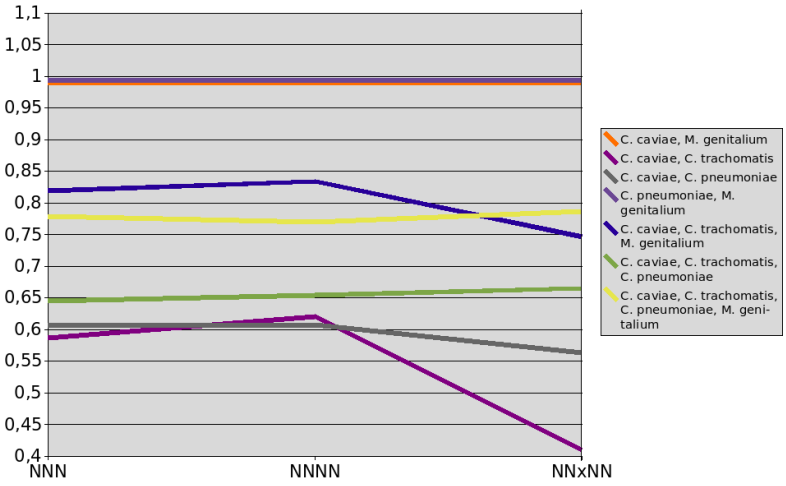
Betrachtet man speziell die Qualität der Trennung nach verwendetem Oligomuster (Abbildungen 6.3 und 6.4), so ist zu erkennen, dass jedes Muster ein Optimum für eine bestimmte Sequenzlänge und Zusammensetzung des Metagenoms hat. Besonderer Aufmerksamkeit bedürfen hier die Ergebnisse der beiden Muster „NNNN“ und „NNxNN“, da beide gleich viel Speicher benötigen. So hat das Muster „NNNN“ sein Optimum bei 1500 bp und artenarmer Metagenome, wohingegen „NNxNN“ bei längeren Sequenzen eine höhere Trennleistung besitzt.

Es ist anzunehmen, dass wenn man es mit kurzen Sequenzen ($<$ 1500 bp) zu tun hat, sich ein kürzeres Muster, wie das Trinukleotid, empfiehlt; aber auch nur dann, wenn die Komplexität gering ist.

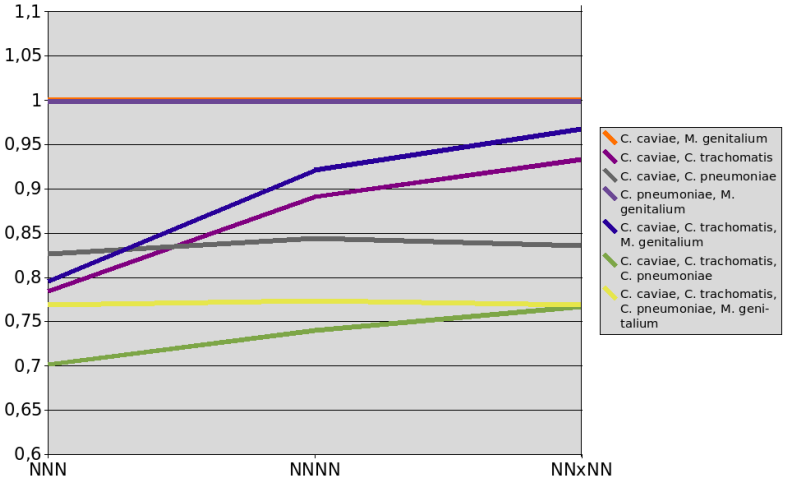
6.3. Zusammenfassung

In diesem Kapitel wurde auf die Binningbedingungen und deren Auswirkungen auf die Clusteringqualität eingegangen. Dazu wurden Clusteringexperimente in Abhängigkeit der Sequenzlängen und des Oligomusters an künstlichen Metagenomen in unterschiedlicher phylogenetischer Zusammensetzung durchgeführt.

Dieser Versuch zeigt sehr deutlich den phylogenetischen Charakter der Oligonukleotidfrequenzen auf. An den teilweise schlechten Ergebnissen kann man erkennen, dass es sehr schwer ist, nah verwandte Fragmente richtig von einander zu trennen, wenn die

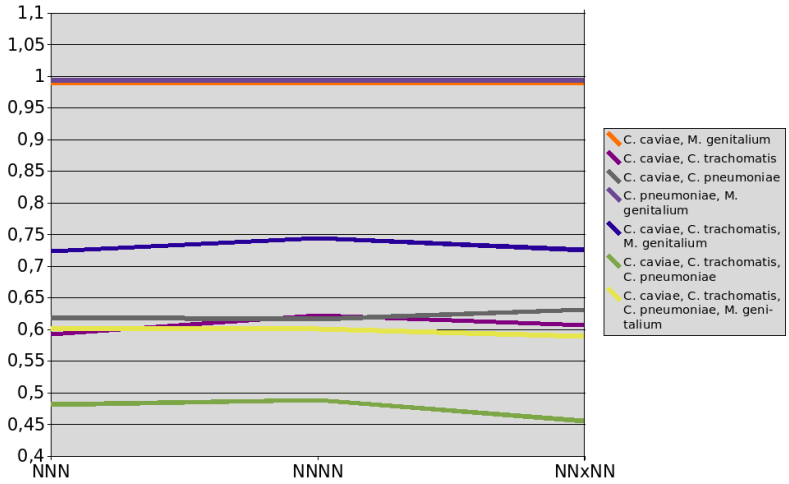


(a) Spezifität, øSequenzlänge = 1500 bp

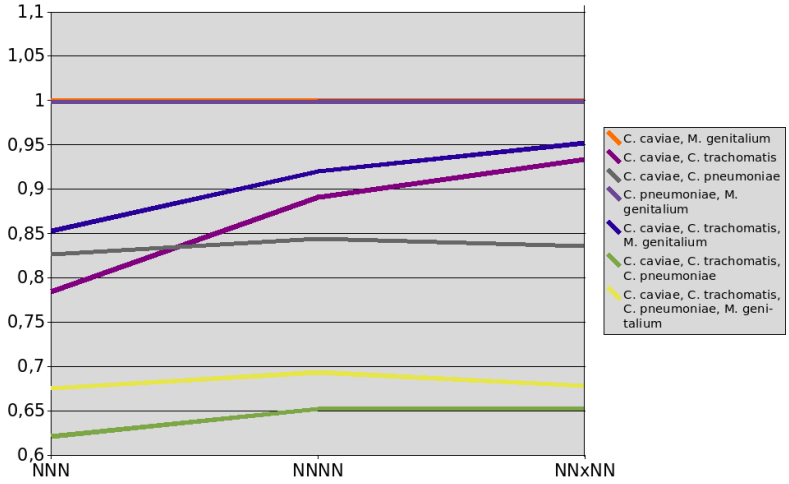


(b) Spezifität, øSequenzlänge = 5000 bp

Abbildung 6.3.: DNA Fragmente ungleicher künstlicher Metagenome, bestehend aus den Organismen *Chlamydophila caviae*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae* und *Mycoplasma genitalium*, wurden anhand verschiedener Oligonukleotidmuster unter Variation der Sequenzlänge geclustert. Die Prozedur wurde für jedes dieser Clusterings 500 mal wiederholt und die Ergebnisse gemittelt. Hier dargestellt ist die Abhängigkeit der Spezifität (y-Achse) vom verwendeten Oligomuster (x-Achse) und der Diversität des Metagenoms für die Sequenzlängen 1500 bp (a) und 5000 bp (b).



(a) Sensitivität, øSequenzlänge = 1500 bp



(b) Sensitivität, øSequenzlänge = 5000 bp

Abbildung 6.4.: DNA Fragmente ungleicher künstlicher Metagenome, bestehend aus den Organismen *Chlamydophila caviae*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae* und *Mycoplasma genitalium*, wurden anhand verschiedener Oligonukleotidmuster unter Variation der Sequenzlänge geclustert. Die Prozedur wurde für jedes dieser Clusterings 500 mal wiederholt und die Ergebnisse gemittelt. Hier dargestellt ist die Abhängigkeit der Sensitivität (y-Achse) vom verwendeten Oligomuster (x-Achse) und der Diversität des Metagenoms für die Sequenzlängen 1500 bp (a) und 5000 bp (b).

Sequenzen zu kurz sind. Die Methode ist aber sehr wohl in der Lage, auch bei kurzen Sequenzen, phylogenetische Gruppen auf höherer Ebene voneinander zu trennen. Für die Rekonstruktion von Genomen ist dies allerdings eine ernüchternde Tatsache. Man will ja hier auf Speziesebene clustern und dies erscheint bei Betrachten der üblichen Sequenzlängen in der Metagenomik sehr utopisch.

Für den hier einzigen zu beeinflussbaren Parameter, nämlich das Oligonukleotidmuster, kann zusammenfassend gesagt werden, dass das optimale Muster von den Ausgangsbedingungen des Metagenoms, wie der durchschnittlichen Sequenzlänge und der biologischen Diversität, abhängt. Es muss somit von Fall zu Fall entschieden werden, welches das optimale Muster ist.

7. Anwendung auf reale Metagenome

7.1. Sargasso-See Metagenom

7.1.1. Datensatzbeschreibung

Die Sargasso-See ist eine subtropische Region, die sich im östlichen Nordatlantik befindet. Da es ein sehr nährstoffarmes Gebiet ist, ist das Wasser aufgrund des geringen Planktonwachstums extrem klar [24]. Diese Tatsache und die Nähe zur Ostküste der USA macht dieses marine Ökosystem zu einem begehrten Forschungsobjekt.

So führte auch im Jahre 2004 die Gruppe um J.C. Venter dort ein Metagenomprojekt durch [31]. Ziel war es die DNA mikrobieller Gemeinschaften aus dem Oberflächenwasser der Sargasso-See zu sequenzieren. Aus der extrahierten DNA entstanden so sieben unabhängige Datensätze mit einer Gesamtmenge von 1.3 Gigabasen. Über 1 Million Gene konnten in dem Metagenom gefunden werden, wobei die meisten der vorhergesagten Genen nicht mit ausreichender Bestimmtheit einer phylogenetischen Gruppe zugewiesen werden konnten. Abschätzungen gehen von 1800 phylogenetischen Gruppen im Metagenom aus.

Auch nach drei Jahren ist dieser Datensatz immer noch der zweitgrößte — nur J.C. Venter hat sich mit dem *Global Ocean Sampling Projekt* selber übertroffen — und der meist zitierte, wenn es um Metagenomikprojekte geht.

Daher habe auch ich mich entschieden, den Datensatz aus der Sargasso-See mit unserer Software zu untersuchen und mit anderen veröffentlichten Ergebnissen zu vergleichen.

Fakten zum Datensatz

Aus den 1.3 Gigabasen an qualitätsgefilterten Rohdaten konnten durch Assemblieren etwa 800 000 Contigs mit einer Gesamtmenge von etwa 817 Millionen Nukleotiden erstellt werden (Anhang E.1). Diese Contigs sind über die Homepage des J.C. Venter Institutes[1] und über GenBank frei verfügbar. Das Metagenom wurde mit der Shotgun-Methode sequenziert. Wie zu erwarten war, liegt die Sequenzlänge deshalb hauptsächlich zwischen 700 und 1200 Basenpaaren (Abbildung 7.1).

7.1.2. Taxonomische Analyse

Laufzeitvergleich

Da die Größe zukünftiger Metagenome wohl der des Sargasso-See Datensatz ähnlich sein werden, haben wir an diesem eine Laufzeitanalyse durchgeführt. Dabei war es von In-

[1] https://research.venterinstitute.org/sargasso/

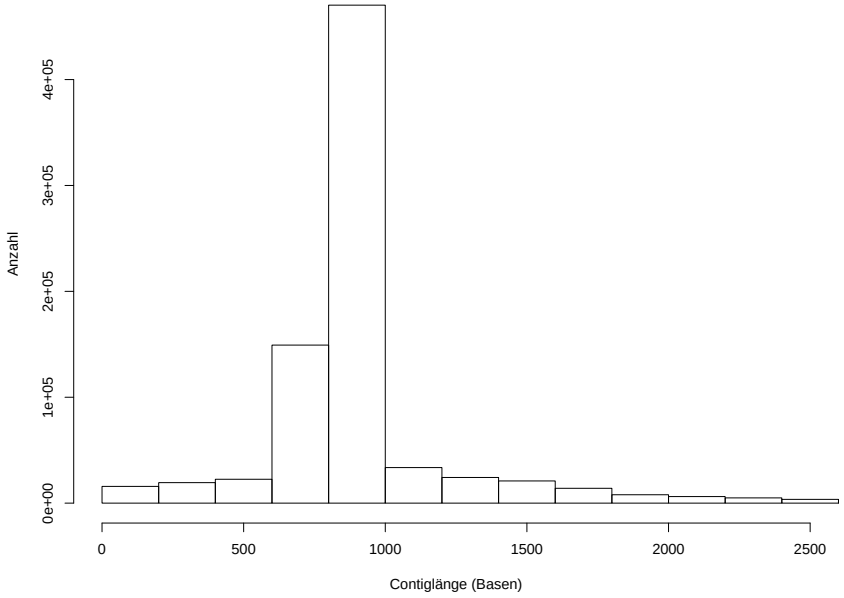


Abbildung 7.1.: Längenverteilung der Contigs aus dem Sargasso-See Metagenom.

teresse, unsere Methode mit den Methoden aus Kapitel 3.3 in ihrem Laufzeitverhalten zu vergleichen. Die Best-Blast-Hit Methode wurde mit dem Programm MEGAN durchgeführt, während die phylogenetischen Bäume mit dem Neighbour-Joining Algorithmus des PHYLIP Paketes berechnet wurden. Für den Vergleich wurde die durchschnittliche Dauer für die Analyse eines Contigs ermittelt, unter der Annahme dass jeder Contig ein Markergen beherbergt.

Abbildung 7.2 zeigt, dass der Blast2Tree Ansatz um den Faktor 100 schneller ist als der Blast-Hit Ansatz, und um den Faktor 3000 schneller als die Rekonstruktion des Baumes anhand multipler Alignments. Die Blast2Tree Methode bietet somit die schnellste Möglichkeit einen Überblick über die Zusammensetzung eines großen Metagenoms zu bekommen. Anhand nachfolgender Analyse soll durch den Vergleich mit bereits bekannten Ergebnissen untersucht werden, ob die Ergebnisse von Blast2Tree auch zuverlässig sind.

Genvorhersage

Bei einer Mindestlänge von 100 Nukleotiden wurden 9 604 084 ORFs gefunden, wovon etwa 70 % eine Länge von 100 bis 200 Nukleotiden haben (Abbildung 7.3). Wie man erkennen kann, sind die meisten dieser kurzen ORFs nicht partiell. Es wird vermutet, dass ein Großteil dieser ORFs aufgrund ihrer Kürze keine biologische Relevanz im Sinne eines Genes haben.

Homologensuche

34 094 (=0,36 %) ORFs wiesen eine Homologie zu den verwendeten COG Proteinen auf, womit zu 3,3 % aller Contigs eine taxonomische Aussage gemacht werden kann.

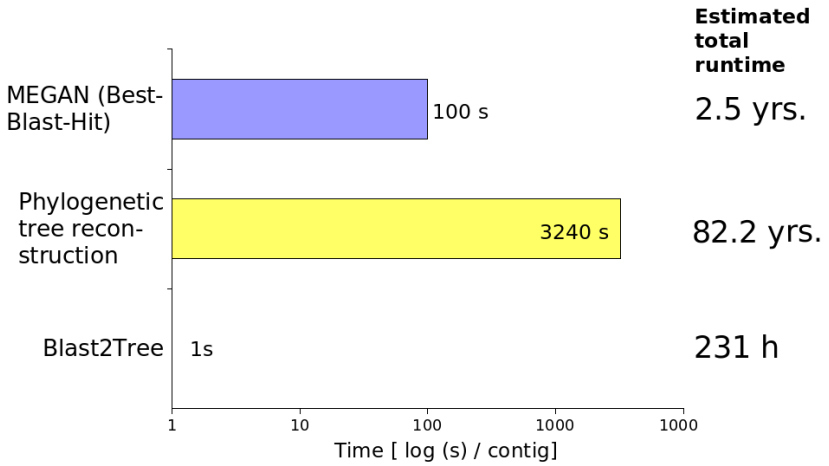


Abbildung 7.2.: Laufzeitvergleich verschiedener Ansätze zur taxonomischen oder phylogenetischen Einordnung eines Contigs. Die Ergebnisse sind auf logarithmischer Skala aufgetragen. Die Gesamtlaufzeiten beziehen sich auf das Sargasso-See Metagenom, wenn auf jedem der 800 000 Contigs ein Markergen wäre.

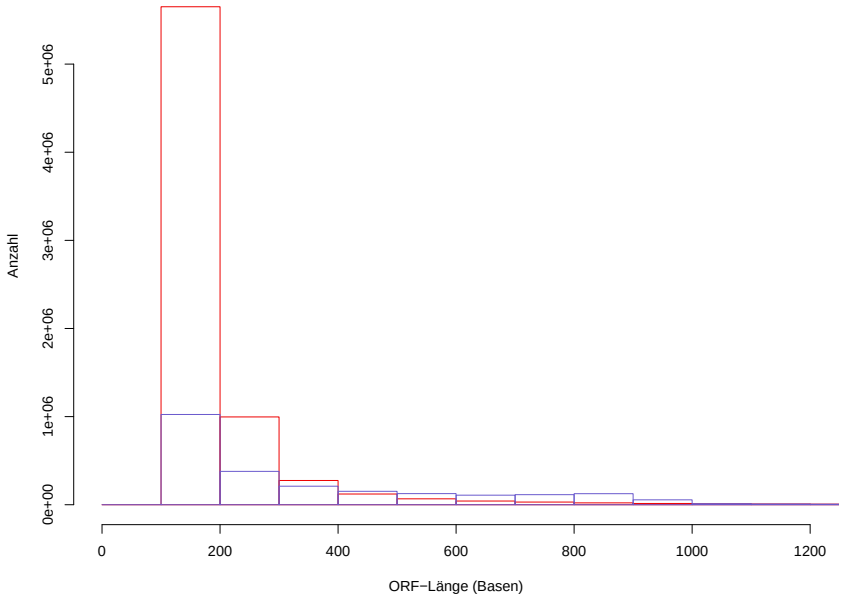


Abbildung 7.3.: Längenverteilung der gefundenen ORFs mit einer gegebenen Mindestlänge von 100 Nukleotiden. Der rote Verlauf stellt die nicht partiellen ORFs dar, der blaue die partiellen.

	øLänge [bp]	Anzahl	Anteil	Homologe	Homologenrate
Gesamt	223,7	9604084		34094	0,004
Nicht partielle ORFs	183,3	7266978	0,76	10629	0,001
Partielle ORFs, gesamt	349,5	2337106	0,24	23465	0,010
Partiell...					
an beiden Enden	714,8	24253	0,00	608	0,025
an einem Ende	345,6	2312853	0,24	22857	0,010

Tabelle 7.1.: Statistik zu den ORFs, die in den Contigs bei einer gegebenen Mindestlänge von 100 Nukleotiden identifiziert wurden.

Die Ergebnisse zeigen außerdem (Tabelle 7.1), dass die nicht partiellen ORFs zwar 76 % der gesamten ORFs darstellen, aber nur 31 % der Homologen stellen. Die Homologenrate ist bei den nicht partiellen ORFs damit um den Faktor 10 geringer als bei den partiellen. Dies ist vor allem auf die durschschnittliche Länge zurückzuführen. Dieser Effekt ist auch zwischen partiellen ORFs an beiden Enden und partiellen ORFs an einem Ende zu beobachten. Kurzum bestätigt dies die These, dass viele der kurzen ORFs keine kodierende Regionen sind und möglicherweise parallel zum Strang einer kodierenden Region als sogenannter „Shadow ORF“ liegen.

Ergebnisse

Die gefundenen homologen ORFs wurden mit dem Blast2Tree Algorithmus taxonomisch beschriftet. Dazu wurden jeweils die drei besten Treffer für einen ORF berücksichtigt. Die Ergebnisse wurden mit einem maximalen Homogenitätswert von 0,4 gefiltert, so dass 18 946 ORFs zur weiteren Betrachtung übrig blieben.

Die Ermittlung eines sinnvollen Homogenitätswertes kann dadurch geschehen, dass man die Menge an unspezifischer taxonomischer Information, die dadurch herausgefiltert wird, in Abhängigkeit vom Homogenitätswert untersucht. Dies wurde für drei verschiedene Werte (0,3 , 0,4 und 0,5) durchgeführt (Tabelle 7.2). In einem Kompromiss zwischen größtmöglicher Minimierung an unspezifischer taxonomischen Information und minimalem Verlust an spezifischer Information, ergab sich hier ein maximaler Homogenitätswert von 0,4 als sinnvoll.

Zur Visualisierung wurde für jeden Knoten des Bork-Baumes ein Häufigkeitswert nach Formel (4.4) berechnet und die Ergebnisse ins iTOL-Format gebracht. Die Daten wurden dann auf der Onlineplattform von iTOL visualisiert (Abbildung 7.4). Dies vermittelt sehr übersichtlich einen Überblick über die phylogenetische Zusammensetzung des Metagenoms. Da die Knoten taxonomisch beschriftet sind, kann man so eine Aussage über vorherrschende oder abwesende Taxons machen. Es ist zum Beispiel ersichtlich, dass Spezies des Phylums Proteobacteria das Metagenom dominieren.

Die phylogenetische Diversität der Sargasso-See Sequenzen wurde bereits von den Autoren untersucht [34]. Sie benutzten sechs verschiedene phylogenetische Markergene, um die Häufigkeit der darin enthaltenen Phylotypen zu bestimmen. Die Ergebnisse dieser

Abbildung 7.4.: Quantitative Darstellung der vorkommenden Taxa im Sargasso-See Metagenom, nach Filterung der Ergebnisse mit einem Homogenitätswert von 0,4 unter Berücksichtigung der drei besten Treffer pro homologem ORF. Die Teile des Baumes sind dabei farbkodiert: Rosa steht für Eukaryota, Grün für Archaea, Orange für Proteobacteria und Cyanblau für den Rest an Bacteria.

	H≥0,5	**H<0,5**		**H<0,4**		**H<0,3**		**Ziel**
Gesamt	34094	31011	(90 %)	18946	(55 %)	18742	(55 %)	-
Cellular Organisms	15484	12421	(80 %)	429	(3 %)	377	(2 %)	minimieren
Bacteria	9457	9450	(100 %)	9406	(99 %)	9405	(99 %)	neutral
Spezifischere Taxa	9153	9140	(100 %)	9111	(99 %)	8960	(98 %)	maximieren

Tabelle 7.2.: Menge an gefilterten Markern für verschiedene Taxa in Abhängigkeit unterschiedlicher maximaler Homogenitätswerte H. Bei der Auswahl eines sinnvollen Wertes für H ist es das Ziel die Menge an unspezifischen Markern (hier repräsentiert durch die Taxa Cellular Organisms und Bacteria) zu minimieren und die Menge an spezifischen Markern zu erhalten. Der größte Verlust an unspezifischer Information findet hier bei einem Wert von 0,4 statt, bei fast gleichbleibender Informationsmenge für die spezifischen Taxa.

Analyse sind im Anhang A.3 dargestellt. Die phylogenetische Verteilung wurde mit den Ergebnissen der Blast2Tree Methode verglichen. Dazu wurden die sechs Werte für jedes Taxon gemittelt und meinen Ergebnissen in Abbildung 7.5 direkt gegenübergestellt.

Es ist offensichtlich, dass bei vielen Taxa beide Analysen übereinstimmen, wie zum Beispiel bei den Proteobacteria, den Archaea oder den Cyanobacteria. Für eine Vielzahl von Taxa konnten aber grobe Abweichungen festgestellt werden. Das Taxon Firmicutes ist im Vergleich überproportional vorhanden. Der Grund hierfür ist wahrscheinlich die Güte einzelner COG Proteine. Wie im Kapitel 5.2.4 gezeigt wurde, weichen einige COG Proteine vom verwendeten Baum ab und führen damit zu fehlerhaften Platzierungen, vor allem im Bereich der Firmicutes.

Darüber hinaus habe ich auch die Ergebnisse der restlichen Phyla mit in die Graphik integriert. Hier fällt vor allem der hohe Anteil an Acidobacteria auf, die eigentlich im Erdreich heimisch sind. Es ist unwahrscheinlich, dass die Wasserproben vom Meeresboden kontaminiert worden sind, da nur Oberflächenwasser entnommen wurde. Erst kürzlich veröffentlichte Arbeiten werfen jedoch ein neues Licht auf das Phylum der Acidobacteria. Zum einen scheint die Diversität innerhalb dieses Phylums höher als bisher angenommen [45], zum anderen konnte ein ebenfalls hoher Anteil an Acidobacteria auch in einem Pazifik-ähnlichen[2] marinen Habitat nachgewiesen werden [46]. Die Aussage über die Häufigkeit an Acidobacteria in meinen Ergebnissen ist somit nicht grundsätzlich falsch, sondern bedarf einer genaueren Untersuchung. Das Vorkommen von eukaryotischer DNA ist dagegen nicht so verwunderlich, da Meerwasser gerade an der Oberfläche Lebensraum für viele eukaryotische Spezies bietet.

Zusammenfassend lässt sich sagen, dass die Ergebnisse durchaus mit denen der Venter-Gruppe vergleichbar sind und damit in diesem Fall das Ziel erfüllt haben, schnell einen groben Überblick über das Metagenom zu verschaffen.

[2]Die Zusammensetzung an Taxa war der einer aphotischen Region des Pazifiks sehr ähnlich.

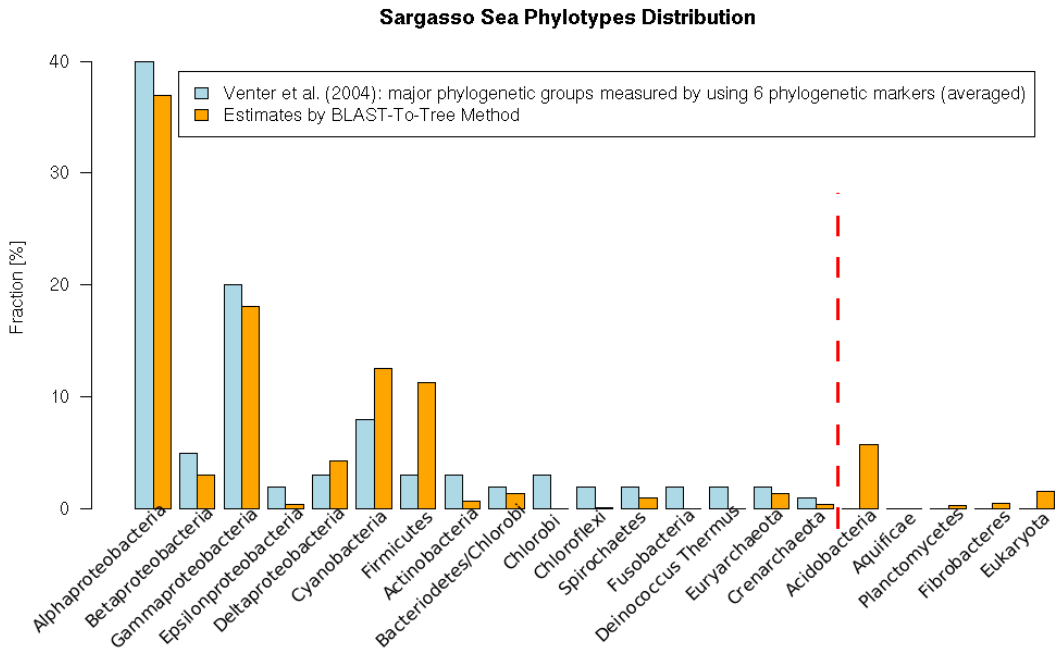


Abbildung 7.5.: Vergleich der Analyseergebnisse zur phylogenetischen Diversität des Sargasso-See Metagenoms. Die ursprüngliche phylogenetische Analyse der Autoren (gelbe Balken) basiert auf sechs phylogenetischen Markern, deren Werte in dieser Darstellung für jedes Taxon gemittelt wurden (Anhang A.3). Die Werte der Blast2Tree Methode entstanden durch das Aufsummieren aller Häufigkeitswerte der Knoten, die unterhalb oder auf dem betrachteten Taxon liegen. Somit wurden die Ergebnisse höherer Taxons nicht mit betrachtet, da sie hierfür zu unspezifisch sind. Die phylogenetische Verteilung wurde außerdem um die Taxa der Blast2Tree-Ergebnisse erweitert, welche rechts durch eine rote Linie getrennt dargestellt sind.

7.1.3. Binning

Binningparameter

Nach einem ersten Lauf stellte sich heraus, dass der Datensatz von 800 000 Contigs zu groß und speicherintensiv für die zur Verfügung stehende Hardware[3] ist, um ihn zu clustern. Daher wurden $\frac{1}{4}$ der kürzesten Contigs aus dem Datensatz entfernt. Ein Contig wurde aber nur dann vom Binning ausgeschlossen, wenn sich auf ihm keiner der 31 COG Marker befand. Nach der Filterung blieben etwa 612 000 Sequenzen übrig.

Die Oligonukleotidfrequenzen jedes Contigs wurden nach dem Muster „NNxNN“ gezählt und nach den Formeln (3.8) und (3.9) normalisiert. Die Wahl dieses Musters wird mit der vermuteten hohen Diversität der Probe begründet.

Das Programm CLUTO v2.1.1 [53] wurde eingesetzt, um die Contigs mithilfe des k-means Verfahrens in Bins zu gruppieren. Da vermutet wird, dass sich über 1800 Spezies im Metagenom befinden [31], wurde als Parameter $k = 2000$ gewählt. Die Dauer des Clusterings betrug 2 Stunden und 45 Minuten.

Ergebnisse

Nach dem Clustering wurden folgende Eckdaten erfasst:

- Durchschnittliche Größe eines Clusters: 306 Contigs
- Größter/kleinster Cluster (Contigs): 1771 (1364 Contigs)/ 1352 (1 Contig)
- Größter/kleinster Cluster (Basen): 1001 (1,8 Mb) / 1352 (860 bp)

Eine Beurteilung der Richtigkeit des Binnings kann nur anhand der taxonomischen Information geschehen, die für einen Bin bekannt ist. Von den 2000 Clustern haben 1907 mindestens einen Contig mit taxonomischer Beschriftung. Das heißt, dass jeder dieser 1907 Bins im Durchschnitt 10 Marker mit taxonomischer Aussage enthält (bei einem Homogenitätswert kleiner 0,4). Abbildung 7.6 zeigt aber, dass die Größen der Cluster nicht gleichverteilt sind, so dass für jeden Bin somit proportional weniger oder mehr taxonomische Information vorhanden ist.

Als Beispiel werde ich jeweils zwei homogene und inhomogene Cluster vorstellen. Bei der Suche nach homogenen Clustern kann die Ausgabe von CLUTO dienen. CLUTO gibt zu jedem Cluster einen Wert für die Ähnlichkeit der einzelnen Objekte innerhalb des Clusters aus. Nach diesem kann man sich bei der Suche richten. Je zwei Beispiele sind in den Abbildungen 7.7 und 7.8 dargestellt.

Man sieht, dass das Binning nur bedingt zur Klassifikation von Genomfragmenten geeignet ist. Allerdings sind die Ergebnisse schlecht beurteilbar, da keine wirkliche Überprüfung der Daten stattfinden kann. Klar ist nur, dass das Binnen in Anbetracht der bekannten Schwierigkeiten, wie HGT, zu kurzen Sequenzen oder ein hoher Grad an Diversität, bei weitem nicht perfekt funktionieren kann.

[3] Systemkonfiguration: Intel(R) Xeon(R) CPU 5140 @ 2.33 GHz, 8 GB Arbeitsspeicher.

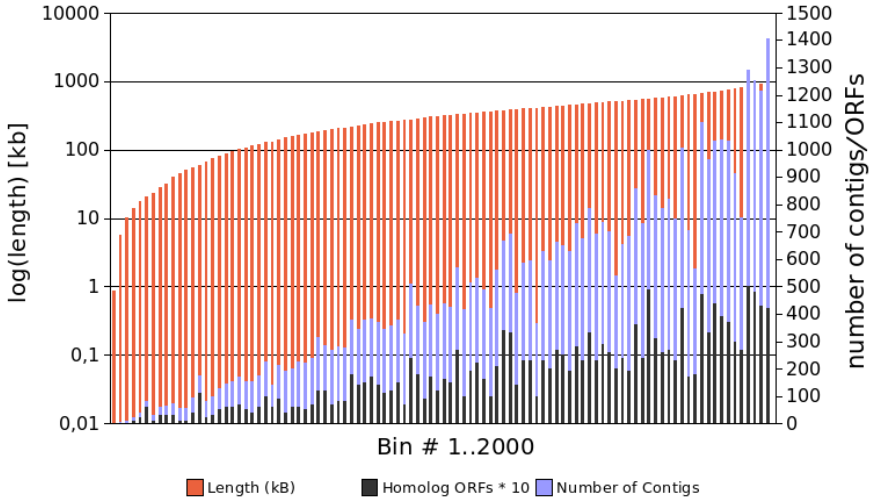
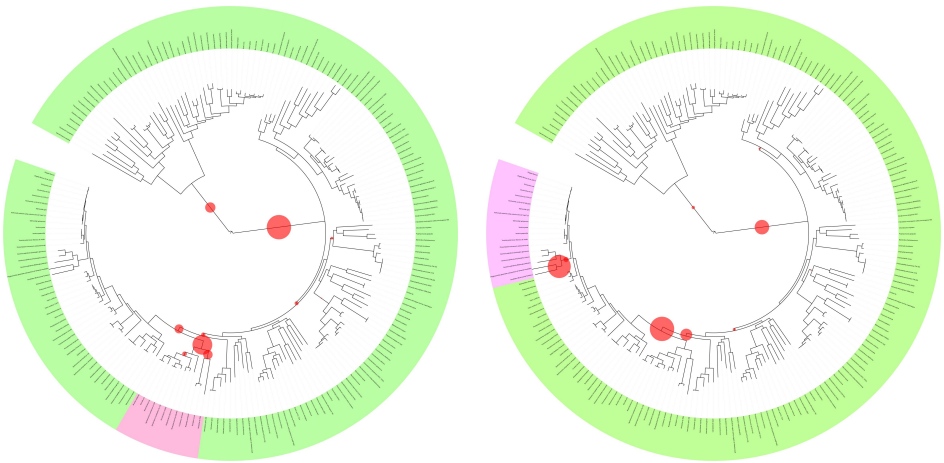
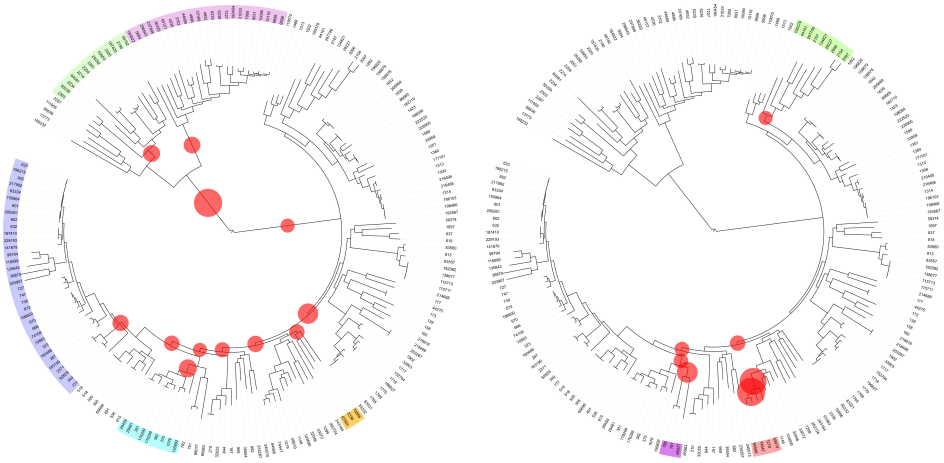


Abbildung 7.6.: Dargestellt ist die Verteilung der Contigs, der homologen ORFs und Basenpaaren auf die einzelnen Cluster. Dabei wurden die Bins nach ihrer Basenpaarmenge aufwärts sortiert. Zur Verbesserung der Lesbarkeit wurde, sowohl die Skala für die Sequenzmenge logarithmiert, als auch die Anzahl an homologen ORFs mit 10 multipliziert. Wie man erkennen kann, sind die Bins in ihrer Menge an Contigs nicht gleichverteilt. Dies führt dazu, dass für einen Bin verschieden viel taxonomische Information zu Verfügung steht, wie man unschwer an der Verteilung der homologen ORFs erkennen kann.



(a) Cluster 24: 861 Contigs, 26 Marker (b) Cluster 678: 305 Contigs, 17 Marker

Abbildung 7.7.: Visualisierung der taxonomischen Information für die beiden Cluster 24 (a) und 678 (b) in iTOL. Die Marker klassifizieren Cluster 24 sehr klar zur Klasse der Alphaproteobacteria, während Cluster 678 vorwiegend Marker aus der Familie der Enterobacteriaceae enthält.



(a) Cluster 1482: 384 Contigs, 17 Marker (b) Cluster 1550: 461 Contigs, 7 Marker

Abbildung 7.8.: Visualisierung der taxonomischen Information für die beiden Cluster 1482 (a) und 1550 (b) in iTOL. Die Marker sind sich in beiden Fällen uneins über die Zuordnung der Bins.

7.2. Anammox-Community Metagenom

Dieser Bereich ist nicht öffentlich, da die Forschungsarbeiten noch nicht abgeschlossen sind.

7.3. Zusammenfassung

Die in dieser Diplomarbeit vorgestellten Methodiken zur taxonomischen Untersuchung und Binning wurde auf zwei reale Metagenome angewandt: dem Metagenom aus der Sargasso-See und dem Anammox-Community Metagenom. Diese Metagenome war deshalb interessant, da sie sich hinsichtlich ihrer Größe, der biologischen Diversität und bereits vorhandener Information zum Inhalt des Metagenoms unterscheiden.

Aufgrund dieser verschiedenen Voraussetzungen mussten für beide Fälle unterschiedliche Verfahrensweisen in Analyse und Bewertung der Ergebnisse angewandt werden. Konnte man die Resultate aus der taxonomischen Untersuchung der Sargasso-See Contigs mit bereits vorhandenen Ergebnissen vergleichen, wusste man bei dem Anammox-Metagenom nur um die starke Präsenz von Anammox-Bakterien. Allerdings hatte man hier Zugang zu anderen Informationen, wie den Supercontigs aus dem *Scaffolding*, was für die Bestimmung optimaler Binningparameter nützlich war.

Das für das Sargasso-See Metagenom gefundenen taxonomische Profil konnte gut mit den bereits veröffentlichten Ergebnissen in Einklang gebracht werden. Zudem konnte die Blast2Tree Methode neue Erkenntnisse über vorhandene Phylotypen im Metagenom liefern. Die Ergebnisse für das Anammox-Community Metagenom konnten konnten sich zumindest mit dem decken, was man über die phylogenetische Einordnung dieser Bak-

terien weiß. Zusammenfassend kann man daher sagen, dass die Blast2Tree Methode ihr Ziel erreicht hat, schnell eine grobe Übersicht über die taxonomische Zusammensetzung eines Metagenoms vermitteln zu können.

Ebenso diente dieses Kapitel dem Vergleich der Methoden (MEGAN, Blast2Tree, CLUTO) in Bezug auf Laufzeitverhalten, Resultate und Komplementierfähigkeit; das heißt, wie sehr sich die Methodiken in ihren Ergebnissen ergänzen konnten. MEGAN konnte die Ergebnisse aus der Blast2Tree Analyse für das Anammox-Metagenom vervollständigen und somit bei der Interpretation helfen. Die Best-Blast-Hit Methode konnte aber angesichts der langen Laufzeit nicht auf das Sargasso-See Metagenom angewandt werden.

Auch galt es die implementierte Blast2Tree Methode in der Praxis zu erproben. Wichtig war dabei zu sehen, wie der Benutzer seine Ergebnisse beeinflussen, bewerten und interpretieren kann. Möglichkeiten wurden anhand des Homogenitätswertes und der Visualisierung auf der Online-Plattform iTOL vorgestellt.

Anhand der für beide Metagenom präsentierten Ergebnisse zum Binning, kann man erkennen, dass die Methodik nur teilweise fähig war, vernünftige Aussagen zu machen. Zu beachten sei aber der Fakt, dass die durchschnittliche Länge der Sequenzen (zwischen 1 kb und 2 kb) weit unter der liegt, wie sie in den Veröffentlichungen zu diesem Thema vorkommen (Teeling [8]: 40 kb, Karlin [10]: 5 kb). Und der Erfolg der Methode hängt kritisch von der Länge der Sequenzen ab, die gebinnt werden sollen.

Während der Betrachtung der taxonomischen Kennzeichnung inhomogener Bins kam jedoch der Gedanke auf, ob eine Spezies nicht mehrere phylogenetische Signale an unterschiedlichen Stellen im Baum verursachen könnte. Es ist zwar bekannt, dass die verwendeten Markergene auf lateralen Gentransfer hin untersucht wurden, aber auszuschließen ist nicht, dass viele der unbekannten Organismen im Metagenom nicht doch Markergene ausgetauscht haben könnten. Dies erschwert die ohnehin schwierige Beurteilung der Qualität der Binningergebnisse.

8. Zusammenfassung

Die Metagenomik wird sich in naher Zukunft als eines der wichtigsten Verfahren in der modernen Biologie etabliert haben und eine Unmenge von Daten liefern. Um die zukünftige Datenflut auswerten zu können, bedarf es neuer Methoden und Programme.

Daher war es das Ziel im Rahmen dieser Diplomarbeit, ein Softwarepaket zur schnellen und automatischen taxonomischen Analyse von Metagenomen zu schaffen. Als Basis hierfür wurde die von Dominik Lindner [20] entwickelte Blast2Tree Methode verwendet, die zur taxonomischen Zuweisung Markergene benutzt. Im Laufe der Arbeit wurde diese anhand des Jack-Knife Verfahrens validiert und optimiert.

In einer Anwendung auf zwei reale Metagenome, nämlich dem Anammox-Community- und dem Sargasso-See Metagenom, wurde die Software auch im Praxisfall getestet. Dabei wurde die Software mit anderen Verfahren und deren Ergebnissen verglichen. Blast2Tree war dabei nicht nur die schnellste aller Methoden, um ein taxonomisches Profil des Metagenoms zu erstellen, sondern lieferte auch ähnliche Ergebnisse. Diese Methode kann somit als erster Schritt in einer phylogenetischen Analyse eines Metagenoms dienen und eine spätere phylogenetische Rekonstruktion unterstützen.

Ein Ziel war es auch das Binning, also das Clustern von DNA Fragmenten, in den Workflow zu integrieren. Das Binning wurde anhand von Oligonukleotidfrequenzen am Beispiel von künstlichen Metagenomen evaluiert und der Einfluss wichtiger Parameter, wie der Diversität des Metagenoms, der durchschnittlichen Länge der Sequenzen und das verwendete Oligomuster, auf die Trennleistung hin untersucht. Da das optimale Oligomuster, als einziger beeinflussbarer Parameter, von Fall zu Fall variieren kann, gibt die Software dem Benutzer die Möglichkeit, das Muster selbst zu definieren.

Das Binning wurde ebenfalls auf die oben genannten realen Metagenome in einem Zug mit Blast2Tree angewandt. Zum Clustern der DNA Fragmente wurde das k-means Verfahren gewählt, wobei der Parameter k vom jeweiligen geschätzten Artenreichtum abhing. Am Ende konnte das Ziel erreicht werden, dass die meisten Bins auch einen taxonomischen Marker enthielten. Eine Bewertung der Binningresultate fiel aber schwer, da zum einen die Binningqualität in Anbetracht der bekannten Schwierigkeiten, vor allem der kurzen Sequenzen, gelitten hat, und zum anderen aber keine konkreten Überprüfungmöglichkeiten vorhanden waren.

Die wichtigsten Eckdaten der Software lassen sich stichpunktartig zusammenfassen:

- Blast2Tree ermöglicht schnell einen groben Überblick über die taxonomische / phylogenetische Zusammensetzung des Metagenoms zu bekommen
- Die Software erlaubt das Definieren eigener Oligomuster zum optimalen Binning verschiedenster Metagenome anhand von Oligonukleotidfrequenzen

- Dem Benutzer wurden Mittel gegeben und Wege aufgezeigt, um seine Ergebnisse filtern und darstellen zu können
- Die Software erlaubt das Auswechseln des Baumes und der dazugehörigen Markergene

Die Software bietet dem Benutzer somit schon jetzt die Möglichkeit, ein Metagenom schnell und zuverlässig zu untersuchen. Daher soll die Software im Herbst diesen Jahres als Web-Service veröffentlicht werden und ich hoffe, dass diese Software Anwendung und zufriedene Benutzer finden wird.

Literaturverzeichnis

[1] Strous M, Wagner M et al. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature, 440:790-4

[2] Tringe S, Rubin E (2005) Metagenomics: DNA Sequencing of environmental samples. Nat Rev Genet, 6:805-14

[3] Xu J (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Molecular Ecology, 15:1713-31

[4] Shelswell KJ (2004) Metagenomics: the science of biological diversity. The Science Creative Quarterly, http://www.scq.ubc.ca/?p=509

[5] Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol, 1:106-12

[6] Lorenz P (2006) Metagenomics für die weiße Biotechnologie. Chemie Ingenieur Technik, 78:461-468

[7] Binnewies T et al. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct Integr Genomics, 6:165-85

[8] Teeling H et al. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol, 6:938-47

[9] Teeling H et al. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics, 5:163

[10] Karlin S, Cardon L (1994) Computational DNA sequence analysis. Annu Rev Microbiol, 48:619-54

[11] Pride D et al. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Research, 13:145-158

[12] Rocha E, Viari A, Danchin A (1998) Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucleic Acids Research, 26:2971-2980

[13] Emmersen J, Rudd S, Mewes HW, Tetko IV (2007) Separation of sequences from host-pathogen interface using triplet nucleotide frequencies. Fungal Genet Biol, 44:231-41

[14] Ciccarelli F, Bork P et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science, 311:1283-7

[15] Tatusov R et al. (1997) A genomic perspective on protein families. Science, 278:631-7

[16] Kimball JW (2006) Kimball's Biology Pages, http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/

[17] von Mering C, Bork P et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science, 315:1126-30

[18] Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research, 17:377-86

[19] Markowitz VM et al. (2006) An experimental metagenome data management and analysis system. Bioinformatics, 22:e359-e367

[20] Lindner D (2006) Entwicklung von Werkzeugen zur taxonomischen Analyse von Proteinhomologen und deren Anwendung im Anammox-Metagenomics Projekt. Diplomarbeit an der FH Weihenstephan und TU München

[21] Madigan MT, Martinko JM (2006) Brock Mikrobiologie (11. deutsche Auflage). Pearson Studium, ISBN:978-3-8273-7187-4

[22] Lecointre G, Le Guyader H (2001) Biosystematik (1. deutsche Auflage). Springer, ISBN:975-3-540-24037-2

[23] Kramer et al., Universität Heidelberg: Phylogenetische Bäume (Unterrichtsunterlagen)

[24] Campbell NA, Reece JB (2003) Biologie (6. deutsche Auflage). Spektrum Akademischer Verlag, ISBN:3-8274-1352-4

[25] Foerstner KU, von Mering C, Bork P (2006) Comparative analysis of environmental sequences: potential and challenges. Philos Trans R Soc Lond B Biol Sci, 61:519-23

[26] Langer M, Gabor EM, Liebeton K, Meurer G, Niehaus F, Schulze R, Eck J, Lorenz P (2006) Metagenomics: an inexhaustible access to nature's diversity. Biotechnol J, 1:815-21

[27] Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev, 68:669-85

[28] Karlin S, Campbell A M, Mrázek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet, 32:185-225

[29] Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R (2004) Molecular Biology of the Gene (5. englischsprachige Ausgabe). Pearson Education Benjamin Cummings, ISBN: 0-321-22368-3

[30] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007), The Sorcerer II Global Ocean Sampling Expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol., 5:e77

[31] Venter JC et. al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science, 304:66-74

[32] Pilcher H (2005) Pipe dreams. Nature, 437:1227-28

[33] Strous M, Fuerst JA, Kramer EH, Logemann S, Muyzer G, van de Pas-Schoonen KT, Webb R, Kuenen JG, Jetten MS (1999) Missing lithotroph identified as new planctomycete. Nature, 400:446-449

[34] Broda E (1977) Two kinds of lithotrophs missing in nature. Zeitschrift für allgemeine Mikrobiologie, 17:491-3

[35] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol., 215:403-410

[36] Lipman D, Pearson W (1985) Rapid and sensitive protein similarity searches. Science, 227:1435-41

[37] Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol., 3:reviews0003.1–reviews0003.8

[38] Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol., 5:e75

[39] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res, 33:433-7

[40] International Union Of Pure And Applied Chemistry (IUPAC), www.chem.qmul.ac.uk/iupac/

[41] Floyd RW (1962) Algorithm 97: Shortest Path. Communications of the ACM, 5:345

[42] Rattei T, Arnold R, Tischler P, Lindner D, Stümpflen V, Mewes HW (2006) SIMAP: the similarity matrix of proteins. Nucleic Acids Research, 34:252-6

[43] Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23:127-8

[44] Haubold B (2004) Sequencer, a software to simulate shotgun sequencing (version 1.4). http://adenine.biz.fh-weihenstephan.de/homePage/

[45] Quaiser A, Ochsenreiter T, Lanz C, Schuster SC, Treusch AH, et al. (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. Mol Microbiol., 50:563–575

[46] Martín-Cuadrado AB, López-García P, Alba JC, Moreira D, Monticelli L, Strittmatter A, Gottschalk G, Rodríguez-Valera F (2007) Metagenomics of the deep mediterranean, a warm bathypelagic habitat. PLoS ONE, 2:e914.

[47] van Niftrik LA, Fuerst JA, Sinninghe Damsté JS, Kuenen JG, Jetten MS, Strous M (2004) The anammoxosome: an intracytoplasmic compartment in anammox bacteria. FEMS Microbiol Lett, 233:7-13.

[48] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 28:10-4

[49] Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science, 311:392-4

[50] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 428:37-43

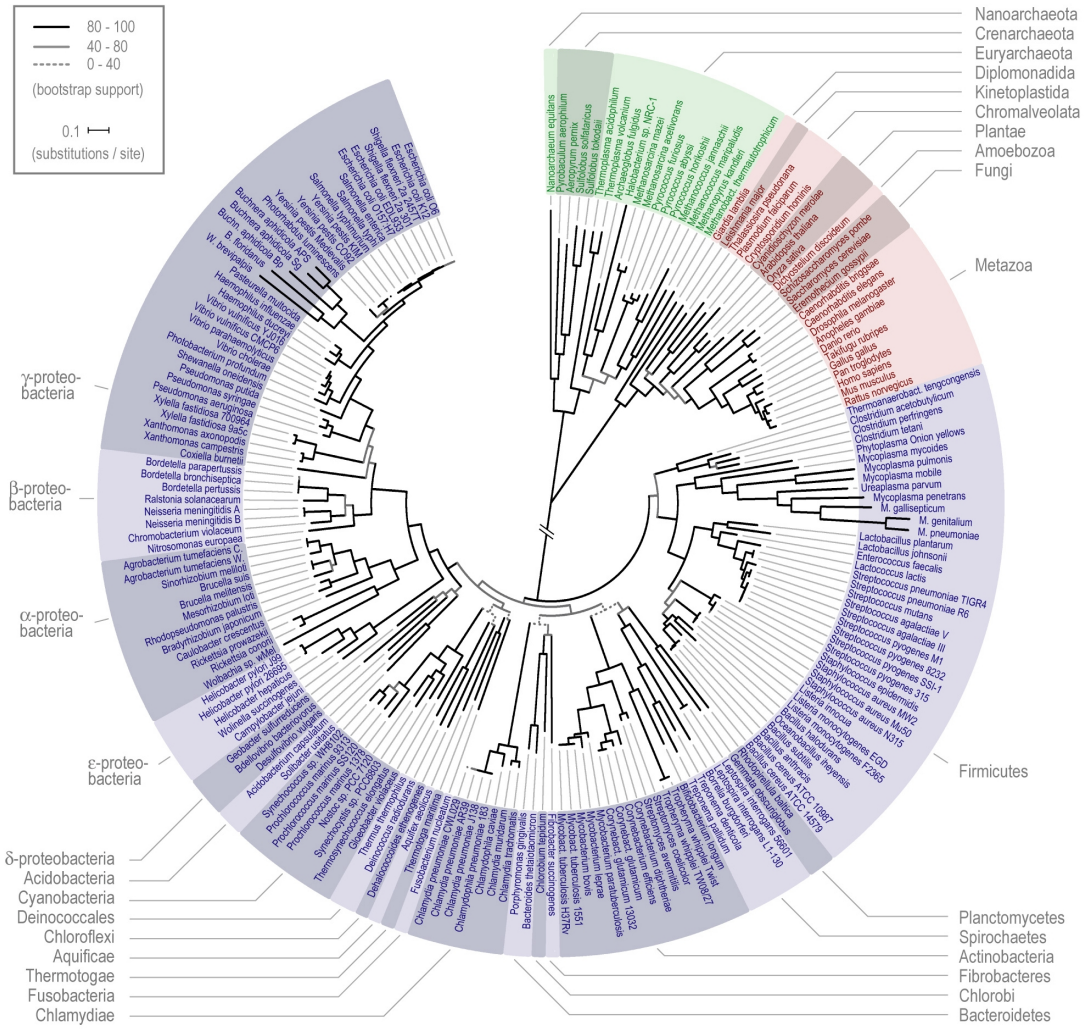
[51] Felsenstein J (2004) Inferring Phylogenies. Sinauer Assoc., ISBN:0-8789-3177-5

[52] Witten IH and Frank E (2005) Data Mining: Practical machine learning tools and techniques (2nd Edition). Morgan Kaufmann, San Francisco, ISBN: 0-1208-8407-0

[53] Karypis G (2002) CLUTO: Software for Clustering High-Dimensional Datasets. http://glaros.dtc.umn.edu/gkhome/views/cluto/

A. Einleitung

A.1. Baum des Lebens von Ciccarelli, Bork et al.



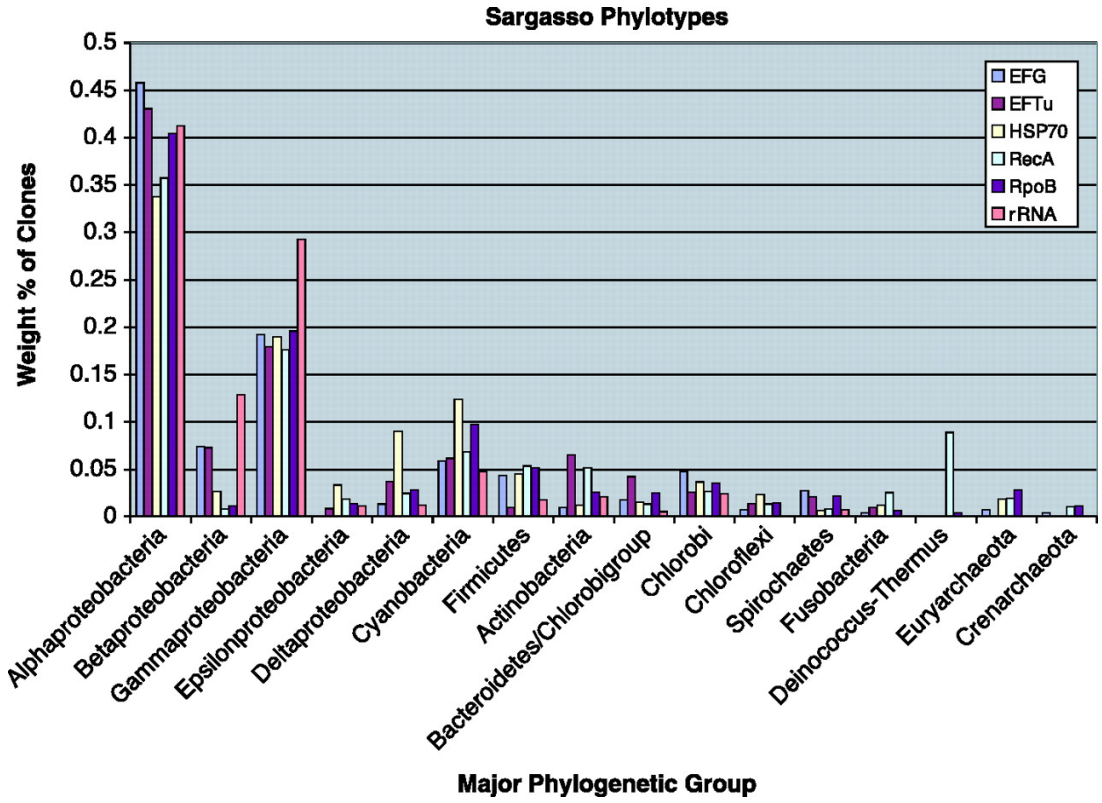
Anhang A.1.1: Baum des Lebens von Francesca Ciccarelli, Peer Bork et al. [14]. Die drei Domänen werden durch die Farben blau (Bacteria), grün (Archaea) und rot (Eukaryota) veranschaulicht.

A.2. Universelle, nicht-HGT COGs

Orthologe Gruppe	Annotation	Gene in Prokaryoten	Gene in Eukaryoten
COG0012	Predicted GTPase, probable translation factor	171	30
COG0016	Phenylalanine-tRNA synthethase alpha subunit	168	42
COG0048	Ribosomal protein S12	168	48
COG0049	Ribosomal protein S7	169	41
COG0052	Ribosomal protein S2	168	79
COG0080	Ribosomal protein L11	170	61
COG0081	Ribosomal protein L1	168	61
COG0087	Ribosomal protein L3	168	54
COG0091	Ribosomal protein L22	168	75
COG0092	Ribosomal protein S3	168	30
COG0093	Ribosomal protein L14	168	41
COG0094	Ribosomal protein L5	169	36
COG0096	Ribosomal protein S8	168	55
COG0097	Ribosomal protein L6P/L9E	168	65
COG0098	Ribosomal protein S5	168	110
COG0099	Ribosomal protein S13	168	49
COG0100	Ribosomal protein S11	169	51
COG0102	Ribosomal protein L13	168	54
COG0103	Ribosomal protein S9	168	52
COG0172	Seryl-tRNA synthetase	177	37
COG0184	Ribosomal protein S15P/S13E	168	41
COG0186	Ribosomal protein S17	170	46
COG0197	Ribosomal protein L16/L10E	168	54
COG0200	Ribosomal protein L15	168	70
COG0201	Preprotein translocase subunit SecY	178	37
COG0202	DNA-directed RNA polymerase, alpha subunit	171	45
COG0256	Ribosomal protein L18	168	50
COG0495	Leucyl-tRNA synthetase	172	43
COG0522	Ribosomal protein S4 and related proteins	174	46
COG0525	Valyl-tRNA synthetase	169	37
COG0533	Metal-dependent proteases with chaperone activity	168	35

Anhang A.2.1: Universell verteile, nicht-lateral transferierte COGs. Quelle: Supporting Online Material aus [14].

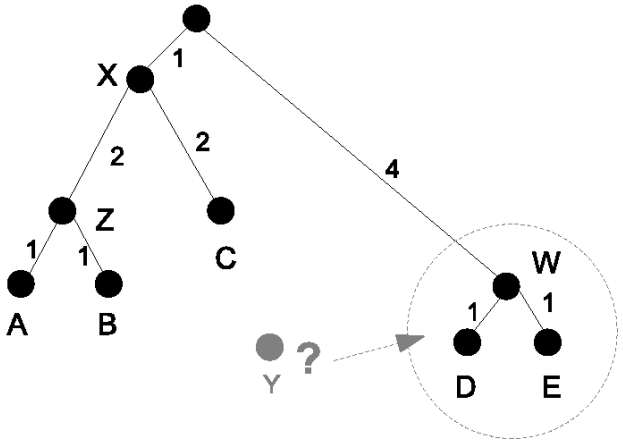
A.3. Sargasso-See Phylotypen Verteilung



Anhang A.3.1: Überblick über die phylogenetische Diversität der Sargasso-See Probe. Der relative Anteil bedeutender Stämme am Metagenom wurde mit mehreren phylogenetischen Markern gemessen: 16S rRNA, RecA, EF-Tu, EF-G, HSP70, and RNA polymerase B (RpoB). Quelle: [31].

A.4. Bsp: Taxonomische Zuordnung eines ORFs durch die Blast2Tree Methode

Ein zu den Markergenen homologer ORF Y mit dem Distanzvektor $do_y = (4.1, 3.9, 3.5, 0.4, 0.6)$ soll in den Speziesbaum aus Abbildung A.4 eingeordnet werden. Jeder Knoten k des Baumes besitzt einen Distanzvektor dk_k und eine durchschnittliche Distanzen s_k zu seinen Blattknoten:



Anhang A.4.1: Beispielbaum, in den der Distanzvektor des ORFs Y eingeordnet werden soll.

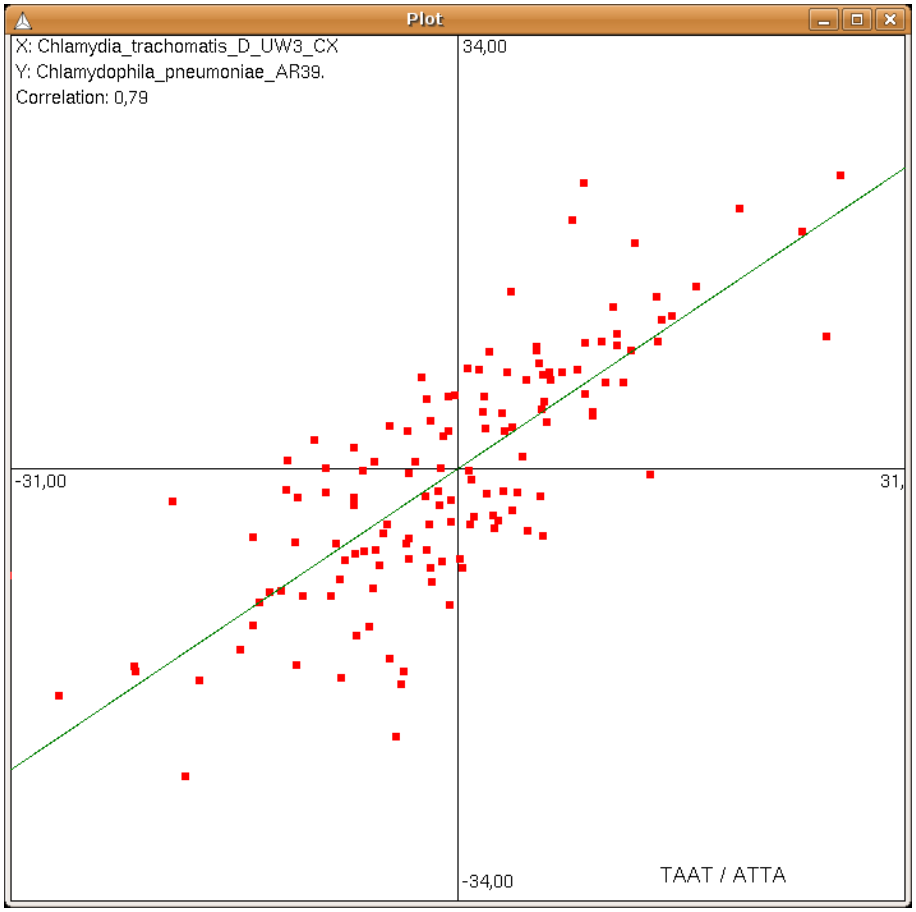
$$dk_A = (0, 2, 5, 9, 9); s_A = 0$$
$$dk_B = (2, 0, 5, 9, 9); s_B = 0$$
$$dk_C = (5, 5, 0, 8, 8); s_C = 0$$
$$dk_D = (9, 9, 8, 0, 2); s_D = 0$$
$$dk_E = (9, 9, 8, 2, 0); s_E = 0$$
$$dk_Z = (1, 1, 4, 8, 8); s_Z = 1$$
$$dk_X = (3, 3, 2, 6, 6); s_X = 2.67$$
$$dk_W = (8, 8, 7, 1, 1); s_W = 1$$

Jeder Knoten des Baumes wird nun durchlaufen, wobei der Distanzvektor do_y an jeden Knoten angepasst wird und dann der Fehler berechnet wird:

- Knoten A: $f = 2.0, s = 0.0 \Rightarrow do'_y = (8.2, 7.8, 7.0, 0.8, 1.2) \Rightarrow e = 232.96$
- Knoten B: $f = 2.0, s = 0.0 \Rightarrow do'_y = (8.2, 7.8, 7.0, 0.8, 1.2) \Rightarrow e = 231.36$
- Knoten C: $f = 2.08, s = 0.0 \Rightarrow do'_y = (8.53, 8.11, 7.28, 0.83, 1.25) \Rightarrow e = 172.10$
- Knoten D: $f = 2.24, s = 0.0 \Rightarrow do'_y = (9.18, 8.74, 7.84, 0.90, 1.34) \Rightarrow e = 1.36$
- Knoten E: $f = 2.24, s = 0.0 \Rightarrow do'_y = (9.18, 8.74, 7.84, 0.90, 1.34) \Rightarrow e = 3.15$
- Knoten Z: $f = 2.16, s = 1.0 \Rightarrow do'_y = (7.86, 7.42, 6.56, 0.0, 0.30) \Rightarrow e = 218.17$
- Knoten X: $f = 2.668, s = 2.67 \Rightarrow do'_y = (8.27, 7.74, 6.67, 0.0, 0.0) \Rightarrow e = 143.97$
- Knoten W: $f = 2.4, s = 1.0 \Rightarrow do'_y = (8.84, 8.36, 7.4, 0.0, 0.44) \Rightarrow e = 2.30$

Die Fehler werden nach aufsteigender Reihenfolge sortiert und dem ORF kann der Knoten mit dem kleinsten Fehler (hier: Knoten D mit $e = 1.36$) zugewiesen werden.

A.5. Tetra: Analyse und Vergleich von Tetranukleotidfrequenzen in DNA Sequenzen



Anhang A.5.1: Dieser Screenshot aus der Stand-alone Version von TETRA [9] zeigt einen Dotplot auf Basis von Tetranukleotidfrequenzen zweier Sequenzen aus den Organismen *Chlamyida trachomatis* und *Chlamydophila pneumoniae*. Dabei entspricht jeder Punkt einem Tetranukleotid (z.B. ATTA). Die Lage des Punktes auf x- und y- Achse wird durch die standardisierten Frequenzen dieses Tetranukleotides in beiden Sequenzen festgelegt. Es wird sowohl eine Regressionsgerade als auch der dazugehörige Korrelationskoeffizient berechnet, der als Maß für die Ähnlichkeit zweier Sequenzen verwendet wird (in diesem Fall: 0,79).

B. Technische Realisierung

B.1. Grammatik des NewickParser

$$
\begin{array}{rcl}
tree & \Longrightarrow & descendant\_list\ [root\_label]\ [:\ branch\_length]\ ; \\
descendant\_list & \Longrightarrow & (\ subtree\ \{\ ,\ subtree\ \}\ ) \\
subtree & \Longrightarrow & descendant\_list\ [internal\_node\_label]\ [:\ branch\_length] \\
 & \Longrightarrow & leaf\_label\ [:\ branch\_length] \\
root\_label & \Longrightarrow & String\ value \\
internal\_node\_label & \Longrightarrow & String\ value \\
leaf\_label & \Longrightarrow & String\ value \\
branch\_length & \Longrightarrow & double\ value
\end{array}
$$

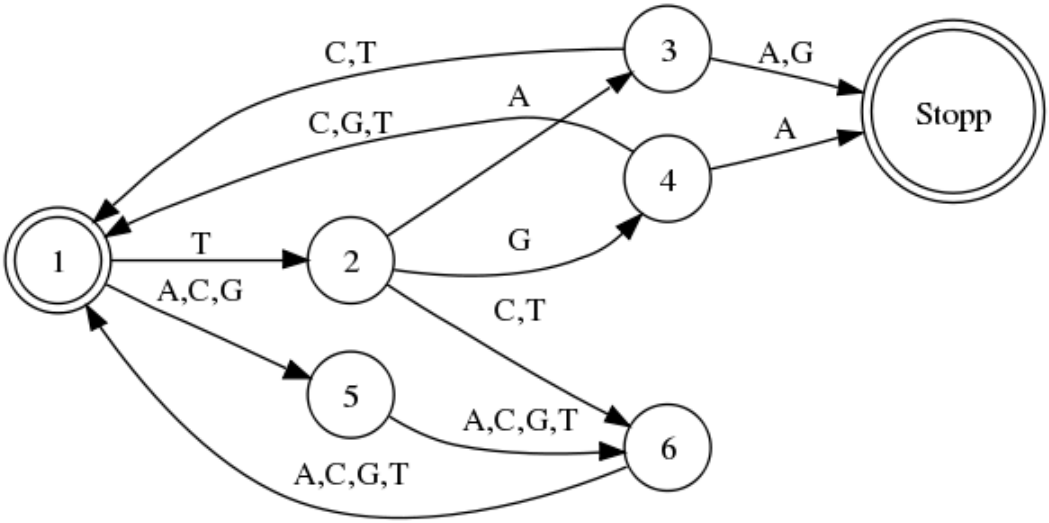
Anhang B.1.1: Die Produktionsregeln des `NewickParser`.

Legende:

- { } : kann keinmal oder mehrmals auftreten
- [ ] : optional, tritt einmal oder keinmal auf
- Alle anderen Zeichen (Klammer, Strichpunkt, Doppelpunkt) werden vom Newick Format selbst benutzt

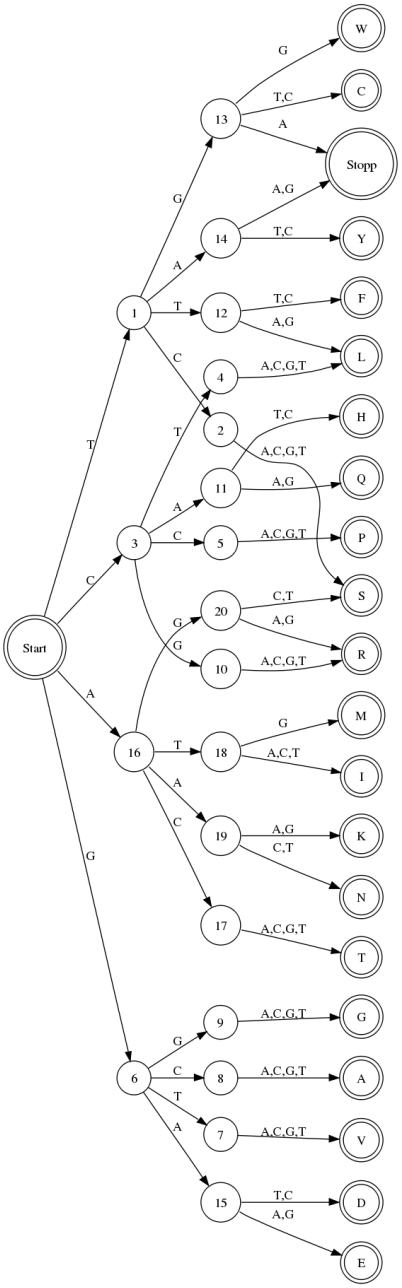
Quelle: http://evolution.genetics.washington.edu/phylip/newick_doc.html

B.2. Zustandsautomat des ORF_Finder



Anhang B.2.1: Der Zustandsautomat mit dem der `ORF_Finder` Stoppcodons in einer Sequenz erkennt. Zustand 1 ist der Startzustand und Stopp der Endzustand, wo eines der Stoppcodons (TAA,TAG,TGA) entdeckt wurde.

B.3. Zustandsautomat des DNA_Translator



Anhang B.3.1: Der Zustandsautomat mit dem der `DNA_Translator` Nukleotidtripletts in eine Aminosäure übersetzt. Der Zustand Stopp steht für ein Stoppcodon. Nicht im Graphen gezeigt sind die Fälle, in denen ein unbekanntes Zeichen gelesen wurde. In der aktuellen Implementierung wird das unbekannte Codon in der Peptidsequenz mit einem X annotiert.

C. Validierung der Blast2Tree Methode

C.1. Taxonomisch beschrifteter Bork-Baum

Auf dem beiliegendem Faltblatt ist der in dieser Diplomarbeit verwendete Referenzbaum dargestellt. Dieser basiert auf Francesca Ciccarellis und Peer Borks Baum des Lebens [14]. Die inneren Knoten wurden nach Algorithmus 1 taxonomisch benannt.

Dieser Baum wurde dann sowohl zur Validierung der Blast2Tree Methode als auch zur Untersuchung realer Metagenome (Kapitel 7) benutzt.

C.2. Vergleich der Resultate verschiedener Gewichtungsfunktionen

Funktion	**Exact Hit**	**Tax. Hit**	**On lineage**	**Not on lineage**
Euklid	7607	20572	38330	31345
Manhattan	10061	23422	39629	30046
Func. (5.3)	11953	26448	45327	24348
Func. (5.4)	5308	8425	9751	59924
Func. (5.5)	12582	26859	43195	26480
Func. (5.5) + slope	13106	27727	44743	24932

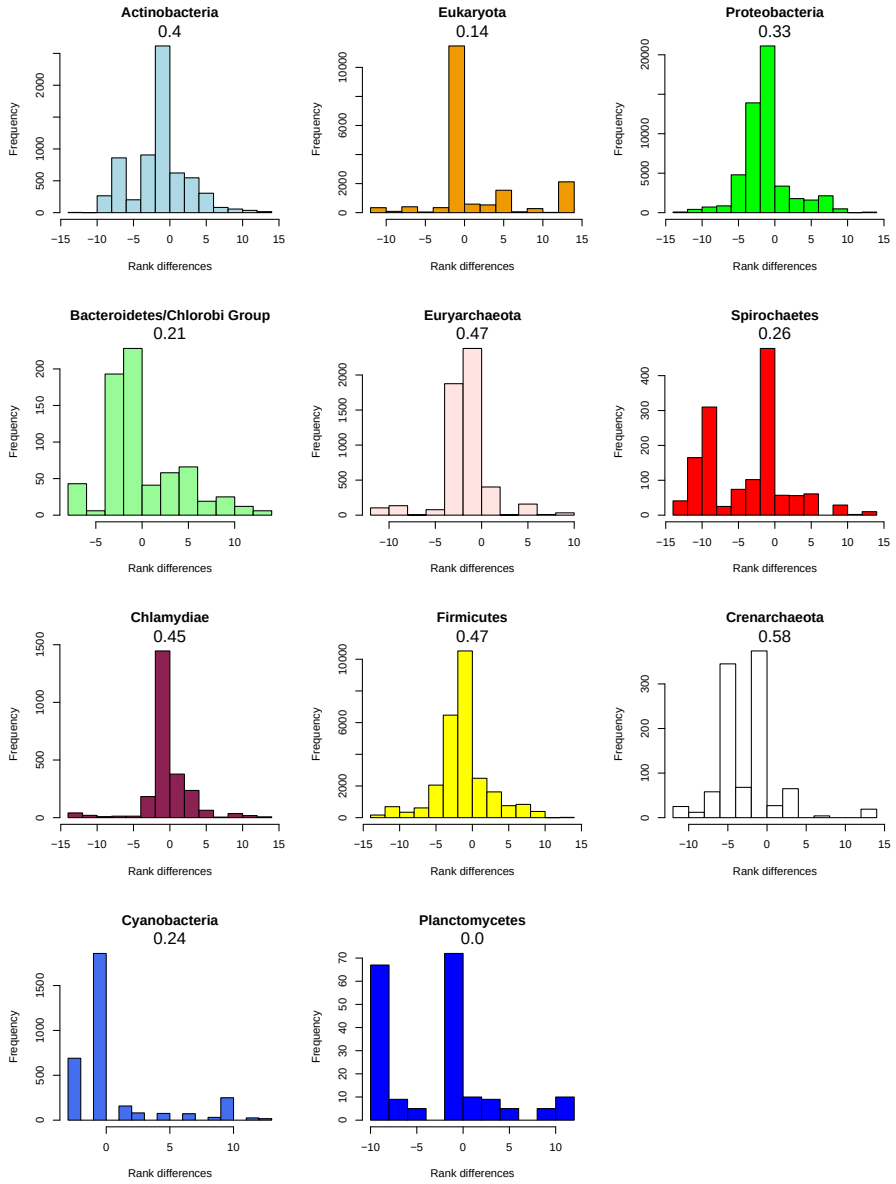
Anhang C.2.1: Die Resultate des Jack-Knifings bei Anwendung verschiedener Gewichtungsfunktionen werden miteinander verglichen. Die Zahlen in der Tabelle stellen die Anzahl an Mappings dar. Allen Gewichtungsfunktionen liegt dabei das Manhattan Distanzmaß zu Grunde.

C.3. COG Klassifizierungsqualität

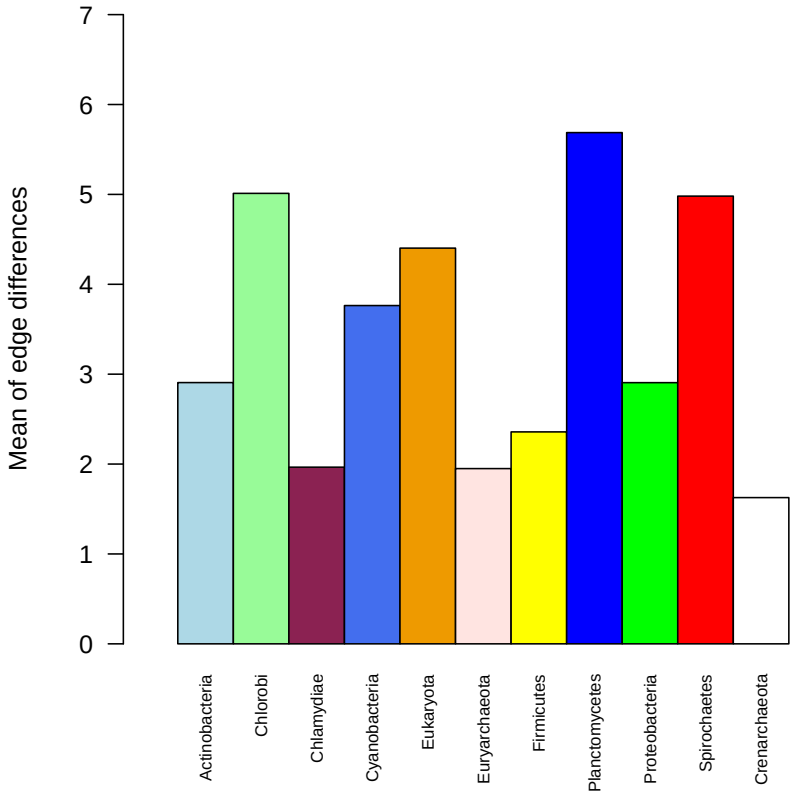
COG	Anzahl an Mappings	Erster Treffer = Zielknoten	Anteil [%]	Zielknoten unter ersten 3 Treffern	Anteil [%]	Score
COG0522	2190	655	0.299	1168	0.533	-367
COG0092	1980	523	0.264	994	0.502	-463
COG0087	2266	646	0.285	1107	0.488	-513
COG0080	2390	536	0.224	1189	0.497	-665
COG0100	2236	513	0.229	1058	0.473	-665
COG0096	2290	532	0.232	1053	0.459	-705
COG0256	2254	522	0.231	1011	0.448	-721
COG0103	2238	482	0.215	1038	0.463	-718
COG0081	2361	534	0.226	1072	0.454	-755
COG0093	2104	485	0.230	842	0.400	-777
COG0012	2013	379	0.188	849	0.421	-785
COG0016	2119	390	0.184	881	0.415	-848
COG0495	2160	401	0.185	887	0.410	-872
COG0200	2475	538	0.217	1045	0.422	-892
COG0186	2157	410	0.190	815	0.377	-932
COG0202	2154	376	0.174	817	0.379	-961
COG0049	2112	390	0.184	754	0.357	-968
COG0102	2251	434	0.192	826	0.366	-991
COG0091	2569	452	0.175	1074	0.418	-1043
COG0094	2074	324	0.156	684	0.329	-1066
COG0201	2150	338	0.157	681	0.316	-1131
COG0533	2009	192	0.095	680	0.338	-1137
COG0099	2244	331	0.147	761	0.339	-1152
COG0097	2448	370	0.151	815	0.332	-1263
COG0172	2118	171	0.080	594	0.280	-1353
COG0184	2089	244	0.116	484	0.231	-1361
COG0052	2630	385	0.146	858	0.326	-1387
COG0098	3054	482	0.157	1030	0.337	-1542
COG0525	2067	74	0.035	444	0.214	-1549
COG0048	2196	240	0.109	407	0.185	-1549
COG0197	2277	233	0.102	494	0.216	-1550

Anhang C.3.1: Die Zuordnungsgüte aller 31 COGs wurde untersucht. Dazu wurde eine Punktgebung für jedes Mapping eingeführt. Für jeden exakten Treffer, das heißt der erste Treffer ist auch der erwartet Knoten, gibt es 1 Punkt. War der gewünschte Knoten unter den ersten 3 Treffern, gibt es 0 Punkte. Einen Punkt Abzug gibt es in jedem anderen Fall. Die COGs wurden dann anhand ihrer Gesamtpunktzahl aufsteigend geordnet.

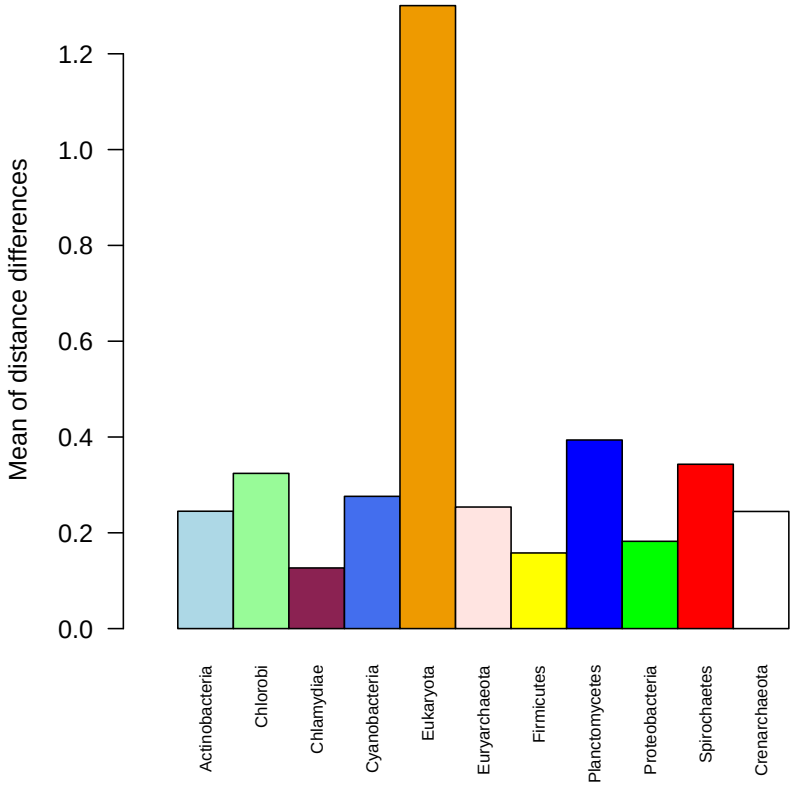
C.4. Zuordnungsqualität auf Phylum-Ebene



Anhang C.4.1: Die Rangdifferenzen sind hier für die 11 wichtigsten Phyla als Histogramme dargestellt. Die Zahlen, die unter den Namen der Phyla stehen, geben Auskunft über den Anteil an exakten Treffer an den Mappings, die eine Rangdifferenz von Null haben. Für diese Auswertung wurde nur jeweils der erste Treffer eines gemappten Proteins berücksichtigt.



Anhang C.4.2: Dieses Balkendiagramm zeigt die durchschnittliche Kantendifferenz für die 11 dominierenden Phyla.



Anhang C.4.3: Dieses Balkendiagramm zeigt die durchschnittliche Distanzdifferenz für die 11 dominierenden Phyla. Hier fällt vor allem der Balken, der zu den Eukaryota gehört, im Vergleich zu den anderen Balken durch seinen hohen Mittelwert auf. Dies ist möglicherweise auf die generell höhere Kantenlänge im Teilbaum der Eukaryoten zurückzuführen.

D. Binning

D.1. Berechnung der Kenngrößen Sensitivität und Spezifität

Da bei jedem Clustering eine Sequenz einem Cluster zugewiesen wird und man auch dessen Ursprungsgenom kennt, kann eine Tabelle mit den Clusteringergebnissen aufgestellt werden. Hier eine Beispieltabelle:

	Cluster 1	Cluster 2	Cluster 3	Sequenzen im Genom
Genom A	**800**	100	100	1000
Genom B	100	**300**	100	500
Genom C	**100**	50	50	200
Sequenzen in Cluster	1000	450	250	

Um die Kennwerte berechnen zu können, müssen erst die Klassifikationsziele definiert werden. Dazu wird jedem Genom ein bevorzugter Cluster zugewiesen und zwar ist es der Cluster, in dem die meisten Sequenzen des betrachteten Genoms liegen (in Fettschrift in obiger Tabelle gekennzeichnet). Im Beispiel wäre dies der Cluster 1 für Genom A, Cluster 2 für Genom B und ebenfalls Cluster 1 für Genom C. Hier ist ein Konflikt zwischen Genom A und C um Cluster 1 entbrannt. In diesem Fall besagt die Klassifikationsregel, dass das Genom das Vorrecht für den Cluster erhält, das den höchsten Anteil an Fragmenten in diesem Cluster stellt. Hier wäre dies Genom A mit 800 Sequenzen. Genom C kann auch nicht Cluster 2 zugewiesen werden, da auch hier Genom B mehr Anteile hält. Also wird für die Sequenzen aus Genom C Cluster 3 als Klassifikationsziel festgelegt.

Nun kann für jedes Genom x mit Klassifikationsziel k eine Wahrheitstabelle aufgestellt werden:

	in Cluster k	in anderen Clustern
Sequenzen aus Genom x	Richtig positive (= Klassifikationsziel) (**a**)	Richtig negative (kumuliert über restliche Cluster) (**d**)
Sequenzen anderer Genome	Falsch positive (kumuliert über restliche Genome) (**b**)	Falsch negative (kumuliert über restliche Genome und Cluster) (**c**)

Die Sensitivität $a/(a+c)$ ist die Wahrscheinlichkeit, dass eine Sequenz richtig klassifiziert wird. Die Spezifität $d/(b+d)$ gibt die Wahrscheinlichkeit an, mit der eine Sequenz nicht falsch klassifiziert wird.

Als Rechenbeispiel soll Genom A dienen. Hier ergeben sich folgende Werte: (a) = 800; (b) = 200; (c) = 500; (d) = 200. Damit errechnet sich eine Spezifität von 200/(200+200) = 0,5 und eine Sensitivität von $800/(800 + 500) = 0,62$.

Die Kenngrößen werden für alle Genome berechnet und am Schluss gemittelt. Dieser gemittelte Wert repräsentiert dann am Schluss die Güte dieses Binnings.

E. Anwendung

E.1. Sargasso-See Metagenom Datensatz

Anzahl an Contigs:	811372		
Gesamtmenge an Basen:	817 Mb		
	Länge	**GC-Gehalt**	**Gaps**
Maximum:	978 kb	0,96	0,01
Minimum:	100 bp	0	0
Durchschnitt:	1006,4 bp	0,38	0

Anhang E.1.1: Statistik zu den Contigs, die aus den Rohdaten des Sargasso-See Metagenoms assembliert wurden.

F. Quellcode

F.1. CD Inhalt

Auf der beiliegenden CD befindet sich der Quellcode, der im Rahmen der Diplomarbeit erstellt wurde. Auch wurde eine begleitende JavaDoc erstellt.

Die Datei `README.txt` gibt alle weiteren Informationen zum Inhalt der CD und zum Starten des Blast2Tree Programms.

F.2. Programmparameter

```
Blast2Tree - command line version
USAGE: java -cp <classpath variables> B2T [OPTIONS...]

HELP OPTIONS:
 -?,--usage   Prints out this help.
 -h,--help    Prints out this help.

MANDATORY ARGUMENTS:
 -b,--blast <path>              Specifies the path to the BLAST executable.
 -d,--markerDB <filename>       Specifies the location of the file that
                                contains the marker genes. This file must
                                be in FASTA format.
 -f,--inputFile <filename>      Specifies the filename of the metagenomic
                                sequences file. The file must be in FASTA format.

FACULTATIVE OPTIONS:
Taxonomical analysis options:
 -e,--eValue <double>           Specifies BLAST maximum e-value for marker
                                homologs finding. (default: 1.0E-4)
 -a,--taxAnalysis <boolean>     Specifies if taxonomical analysis should be
                                performed. (default: true)
 -l,--length <integer>          Sets the minimal nucleotide length of an ORF
                                (default: 100)
 -m,--matrix <BLOSUMx|PAMy>     Sets the alignment scoring matrix. Be aware
                                that BLAST will use the specified matrix.
                                Possible matrices are BLOSUM45, BLOSUM50, BLOSUM62,
                                BLOSUM 80, PAM30 and PAM70. (default: BLOSUM62)
```

```
 -n,--ncbiTaxFile <filename>   Specifies the location of the NCBI taxonomy
                               Zip file, as it can be downloaded from the
                               NCBI FTP site:
                               ftp://ftp.ncbi.nih.gov/pub/taxonomy/
                               Only necessary if a new reference tree has been
                               specified.
                               (default: referenceTree/source/taxdmp.zip)
 -r,--maxError <integer>       Specifies the maximal error for an
                               assignment of a taxon to an ORF. (default: no limit)
 -t,--tree <filename>          Specifies the reference species tree file
                               in Newick format. (default:
                               referenceTree/source/tree_Feb15_midpoint_rooted.txt)
 -x,--maxTaxa <integer>        Specifies the maximal number of assigned
                               taxa per ORF. (default: 3)
Clustering options:
 -k,--kMeans <integer>         Setting this argument > 1 will perform an
                               internal k-means clustering on the data according
                               to the oligonucleotide pattern and the GC-content
                               of the contigs. Because k-means algorithm is
                               implemented by the WEKA package and therefore
                               working slowly, it is recommended to use this
                               feature only for small data sets. (default: 0)
 -p,--pattern <string>         Specifies the oligonucleotide pattern to be used
                               for clustering: N = any nucleotide, x = omitted
                               nucleotide (maximum 15 characters)(default: NNxNN).
                               You can write an ouput file for clustering by
                               specifying parameter c. Activate internal k-means
                               clustering by setting parameter k > 1.
 -c,--cluster <prefix>         Setting this argument will create a set of
                               data files containing the oligonucleotide usage for
                               each contig. This file can be used for clustering
                               with CLUTO 2.x, a clustering toolkit. (default:null)
 -g,--rawCounts <boolean>      Choosing this argument will use the raw counts
                               of the nucleotide pattern usage for clustering the
                               data (only normalised to the length of the sequence).
                               (default: false)
Output options:
 -i,--iTOL <directory>         By setting this argument the necessary files for
                               displaying the results in iTOL (http://itol.embl.de)
                               will be written to that directory. (default: null)
 -o,--out <filename>           Setting this argument will write the output of
                               the analysis to the selected file. (default: STDOUT)
```