Clusteranalyse und Phylogenomik eukaryotischer Gene prokaryotischen Ursprungs

Inaugural-Dissertation

zur Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Nicole Grünheit

aus Warstein

Januar 2010

Aus dem Institut für ökologische Pflanzenphysiologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. William Martin
Korreferent: Prof. Dr. Gerhard Steger

Tag der mündlichen Prüfung: 22.02.2010

Im Laufe dieser Arbeit wurden mit Zustimmung des Betreuers folgende Beiträge veröffentlicht:

Rezensierte Fachzeitschriften

Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10(5):203-209 (2005).

Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138(3):1723-1733 (2005).

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25(4):748-761 (2008).

Gruenheit N, Lockhart PJ, Steel M, Martin W. Difficulties in Testing for Covarion-Like Properties of Sequences Under the Confounding Influence of Changing Proportions of Variable Sites. Mol Biol Evol. 25(7):1512-1520 (2008).

Bolte K, Kawach O, Prechtl J, Gruenheit N, Nyalwidhe J, Maier UG. Complementation of a phycocyanin-bilin lyase from Synechocystis sp. PCC 6803 with a nucleomorph-encoded open reading frame from the cryptophyte Guillardia theta. BMC Plant Biol. 8:56 (2008).

Tagungsbeiträge (Poster)

Nicole Gruenheit, Tal Dagan and William Martin The covarion structure of chloroplast protein sequence data. SFB TR1 annual meeting. Munich, 2006

Nicole Gruenheit, Peter Lockhart, William Martin, Tal Dagan How does the proportion of invariable sites affect phylogeny reconstruction quality? SMBE Annual Meeting. Halifax, Canada 2007

Deusch O, Landan G, Röttger M, Grünheit N, Dagan T, Martin W Ancient phylogenies: Alignments make a difference SMBE Annual Meeting. Halifax, Canada 2007

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Zusammenfassung

In dieser Arbeit wurde die Entstehung der primären Plastiden in den Algen und Pflanzen aus Cyanobakterien und der sekundären Plastide in den Chromalveolaten aus einer Grün- oder Rotalge mittels Endosymbiose mithilfe einer vergleichenden Genomanalyse näher untersucht. Zu diesem Zweck wurden 2.825.466 kernkodierte Proteine aus 710 vollständig sequenzierten Genomen hinsichtlich ihrer Sequenzähnlichkeit und ihrer phylogenetischen Geschichte untersucht. Im Vordergrund standen die Fragen, den nächsten Verwandten der Plastiden unter den rezenten Cyanobakterien zu identifizieren, und eine Schätzung über den Anteil der cyanobakteriellen Gene in den Kerngenomen von sieben Algen und Pflanzen durchzuführen. Außerdem sollten zehn vollständig sequenzierte Chromalveolatengenome auf die Frage hin überprüft werden, ob das Verhältnis der Gene, die durch endosymbiontischen Gentransfer aus einer Grünalge oder einer Rotalge in das Genom gelangt sein könnten, eher für eine Rot- oder eine Grünalge als sekundären Endosymbionten spricht.

Um die Sequenzen in möglichst natürliche Gruppen gemäß ihren Ähnlichkeiten einteilen zu können, wurden globale Identitäten von 162 Millionen bidirektionalen besten BLAST-Treffer mithilfe des Programms `needle` berechnet, und dann verwendet, um hierarchische Clusterings nach dem Markov-Cluster-Algorithmus (`mcl`) mit einem Inflationsparameter von 2,0 und globalen Identitätsschwellenwerten von 0 % bis 70 % durchzuführen. Aus den 176.758 erstellten Clustern wurden von denjenigen, die mehr als drei Sequenzen enthielten (90.354), multiple Alignments mit dem Programm `MUSCLE` berechnet. Von jedem dieser Alignments wurde ein *neighbor-joining*-Baum basierend auf dem JTT-Modell erstellt und für jede Sequenz der Algen und Pflanzen und der Chromalveolata die kleinste monophyletische Gruppe bestimmt. Unter diesen Proteinen wurden im Fall der Algen und Pflanzen diejenigen identifiziert, die in den berechneten Bäumen in einer monophyletischen

Gruppe mit Cyanobakterien vorkamen. Dies ist ein Hinweis auf einen potenziellen cyanobakteriellen Ursprung dieses Proteins. Für die Proteine der Chromalveolaten wurden diejenigen Proteine bestimmt, die sich in einer monophyletischen Gruppe mit Proteinen der Grünalgen, Rotalge oder Pflanzen befanden, um so eine Aussage über verschiedene Hypothesen zu der Entstehung der sekundären Plastide in dieser Gruppe treffen zu können.

Der Prozentsatz aller identifizierten transferierten Gene in den verschiedenen Algen und Pflanzen lag in der konservativsten Schätzung zwischen 7 % und 8 % in den drei Pflanzen und zwischen 10 % und 11 % in den Grünalgen und der Rotalge bezogen auf alle Bäume, in denen ein Protein eines der Archaeplastida vorkam. In den Bäumen, in denen sowohl Proteine der Algen und Pflanzen als auch der Cyanobakterien vorkamen, lagen die Werte zwischen 38 % und 44 %. Bei der Analyse der Häufigkeit des Auftretens der Cyanobakterien als nächster Nachbar, wurden insbesondere die Stickstoff fixierenden Cyanobakterien aller Gruppen gefunden. Die Proteine von *Acaryochloris marina* und der Nostocales waren am häufigsten in einer monophyletischen Gruppe mit Proteinen der Algen und Pflanzen zu finden.

Durch die Analyse der Proteine der Chromalveolaten konnten Berichte zu hohen Anteilen an "grünen" Proteinen für die photosynthetischen Bacillariophyta (550, 506) und die Oomyceten (386, 367, 381) bestätigt werden. Die Ciliophora, die keine sekundäre Plastide besitzen, zeigten um die Hälfte weniger (210, 121) und die parasitischen Apicomplexa sehr wenige (41, 52, 45) "grüne" Proteine. Demgegenüber stehen 425 und 324 "rote" Gene in den Bacillariophyta, 285, 299 und 209 Gene in den Oomyceten, 153 und 123 in den Ciliophora und 123, 120, 121 in den Apicomplexa. Diese Gruppe war die Einzige, die einen höheren Anteil an "roten" als an "grünen" Proteinen aufwies. Trotz der vielen "grünen" Proteine ist die Hypothese, dass die photosynthetischen oder vormals photosynthetischen Chromalveolaten aufgrund ihrer Fähigkeit zur Photosynthese und aufgrund des sehr kleinen Rotalgengenoms viele "grüne" Proteine haben, der Hypothese einer zweiten Endosymbiose mit einer Grünalge vorzuziehen.

Die Monophylie der Chromalveolaten konnte nicht bestätigt – aber auch nicht vollständig abgelehnt – werden, da viele Gruppen entweder durch die Pflanzen und Algen oder die Pilze und Tiere gestört werden. Dies spiegelt die stark unterschiedlichen Lebensweisen der Chromalveolaten wider.

2 Abstract

Photosynthesis in eukaryotes can be traced back to several events of endosymbiosis. The primary plastids of the plants and algae stem from an endosymbiosis of a heterotrophic eukaryot with a cyanobacterium. The secondary plastid of the chromalveolates, a subgroup of the protists, is proposed to have arisen by one endosymbiosis of a heterotrophic protist with a photoautotrophic green or red alga. The question of how many genes have been transferred from the genome of the endosymbionts to the corresponding core genomes of the hosts was studied in several analyses so far. This study is the first to analyze the genomes of six plants and algae, 40 cyanobacteria, and ten chromalveolates simultanously searching for the cyanobacterial and algal genes repectively. With those genes it is possible to determine the nature of the ancestor of the primary plastid and of the ancestor of the secondary plastid respectively. The questions if the ancestor of the primary plastids was indeed most similar to the heterocyst-forming cyanobacteria of the sections IV and if the chromalveolate hypothesis postulating that the plastids of the cryptophytes, haptophytes, heterokontophytes, and alveolates go back to a single endosymbiosis event with a red alga can be supported by a whole genome analysis were of special interest.

2.825.466 proteins from 710 complete genomes were compared and clustered into protein families using 162 million bidirectional best BLAST hits and the marcov cluster algorithm (`mcl`) with a 20 % global pairwise identity cutoff and an inflation parameter of 2.0 resulting in 176.758 cluster of which 90.354 contained more than three sequences. Phylogenetic analyses were carried out for those cluster by aligning the sequences using `muscle` and building neighbor-joining trees based on the JTT model. For each protein of the plants and algae, the chromalveolates and an excavate the smallest monophyletic group was identified in the trees. To determine the nearest neighbor of the proteins those sequences that were in one

group with cyanobacteria were identified and labeled as cyanobacterial genes and the sequences of the genes of the chromalveolates that had a green or red alga as a nearest neighbor were labeled as of green or red algal origin.

If those trees in which at least one protein of a plant or alga was present were analyzed 7 % and 8 % of those proteins were found to be transferred in the plants and between 10 % and 11 % in the green and red algae. Taking only those trees into account in which at least one protein of a plant or alga and at least one protein of a cyanobacterium was found those values ranged between 38 % and 44 %. Studies that placed the ancestor of the primary plastids among the heterocyst-forming nitrogen-fixing cyanobacteria can be supported only with regard to the ability for nitrogen fixation. Among all 40 cyanobacteria *Acaryochloris marina* of section I and the nostocales of section IV were the most frequent nearest neighbors which implies that the ability to fix nitrogen is a stronger connection to the nature of the ancestor of plastids than the ability to form heterocysts.

Studies in which a high proportion of green algal genes in the genomes of the photosynthetic bacillariophyta (550, 506) and the oomycetes (386, 367, 381) were reported can be supported by the results of this analysis. In the genomes of the ciliophora that lack a secondary plastid 210 and 121 proteins of green algal origin could be identified as well as 41, 52, and 45 genes in the apicomplexan genomes. In contrast to this there were 425 und 324 genes of red algal origin found in the bacillariophyta, 285, 299 and 209 genes in the oomycetes, 153 and 123 in the ciliophora and 123, 120, 121 in the apicomplexa. The apicomplexa was the only group with a higher proportion of red algal genes compared to green algal genes. Although the presence of the green algal genes could be confirmed in this analysis the theory of a serial endosymbiosis in the chromalveolates seems to be highly unlikely. The easier explanation can be found in the ability to conduct photosynthesis. The genes related to photosynthesis can be found in all photosynthetic or formerly photosynthetic organisms.

The monophyly of the chromalveolates can neither be supported nor rejected because the groups containing proteins of chromalveolates are often disturbed by proteins of the archaeplastida, fungi and animals. These results reflect the highly different lifestyles of the chromalveolates and challenges the goal of building one reliable phylogeny by means of the comparative whole genome analyses.

3 Einleitung

3.1. Die evolutionäre Entstehung der Photosynthese

Die ersten Zellen – einfache anaerobe Bakterien – entstanden vor 3,5 bis 4 Milliarden Jahren (Hughes et al., 2004; Martin und Russell, 2003). Zu dieser Zeit unterschied sich die Atmosphäre der Erde sehr von der heutigen: Die Sauerstoffkonzentration der Luft betrug nur circa 0.002 %. Der Großteil der Luft bestand aus Stickstoff und Kohlenstoffdioxid (Des Marais, 1998). Außerdem war die Konzentration des Treibhausgases Methan sehr hoch, wodurch die Temperatur auf der Erde ungefähr der entsprach, wie wir sie heute kennen. Für circa eine Milliarde Jahre änderten sich diese Bedingungen kaum und die vorhandenen Bakterien konnten sich weiterentwickeln. Vor circa 2,5 Milliarden Jahren änderte sich die Atmosphäre, die Temperatur sowie die Anzahl und Vielfalt der lebenden Organismen drastisch. Der Sauerstoffgehalt der Luft stieg auf 20 % an, wodurch der Methananteil der Luft reduziert wurde, und die Temperatur fiel auf bis zu -50 Grad Celsius. Diese rapiden Veränderungen sind als die große Sauerstoffkatastrophe (GOE, engl. *great oxygen event*) bekannt (Anbar et al., 2007; Buick, 2008; Holland, 2006; Kasting und Howard, 2006).

Die Erzeuger des Sauerstoffs waren die Cyanobakterien (früher: Blau-Grünalgen). Diese Bakterien sind fähig oxygene Photosynthese zu betreiben, das heißt, sie können aus Kohlenstoffdioxid und Wasser mithilfe der Lichtenergie Glukose und Sauerstoff bilden. Dieser Sauerstoff gab neuen komplexeren Organismen, die sich an den höheren Sauerstoffgehalt der Luft und des Wassers angepasst hatten, die Möglichkeit sich in den Meeren auszubreiten (Kerr, 2005).

3.1.1. Cyanobakterien

Bis vor wenigen Jahrzehnten hießen Cyanobakterien "Blaualgen". Dann jedoch wurde klar, dass diese photoautotrophen Organismen weder einen Zellkern noch andere eukaryotische Merkmale besitzen, die die Zuordnung zu den Algen rechtfertigen würde. Daraufhin wurden sie in das prokaryotische Phylum der Cyanobakterien überführt (Stanier et al., 1978). Die ältesten Fossilien (Stromatolite), die vermutlich Cyanobakterien enthalten, sind circa 3,4 Milliarden Jahre alt (Allwood et al., 2006; Schopf, 1993). Sie kommen heutzutage in fast allen Lebensräumen vor. Außerdem bilden sie einen Großteil der Biomasse des Phytoplanktons und bilden so nicht nur eine wichtige Nahrungsgrundlage für Meerestiere, sondern tragen auch wesentlich zum globalen Kohlenstoff- und Stickstoffkreislauf bei, indem sie elementaren Kohlenstoff und Stickstoff fixieren und somit für andere Organismen verfügbar machen (Campbell et al., 1994). In Süßwasserseen und -flüssen werden große Cyanobakterienkolonien oftmals zu einem Problem, da viele Arten Toxine bilden, die sogar für den Menschen gefährlich sein können (Codd et al., 1999).

Cyanobakterien sind in der Lage, ein breites Spektrum des sichtbaren Sonnenlichts für die Photosynthese zu nutzen, da sie nicht nur Chlorophyll a und b als Antennenprotein verwenden, sondern auch die Phycobiline Phycoerythrin (blau) oder Phycocyanin (rot). Je nachdem wie hoch der Anteil der jeweiligen Phycobiline in einem Cyanobakterium ist, ist dieses unterschiedlich gefärbt. Durch die Nutzung des großen Farbspektrums können Cyanobakterien auch in extremen Schwachlichtbereichen überleben (Ting et al., 2002). Bisher ist nur ein Cyanobakterium, *Acaryochloris marina*, bekannt, das nicht – wie auch alle Pflanzen und Algen – Chlorophyll a und b als Antennenprotein verwendet, sondern hauptsächlich Chlorophyll d, das es in die Lage versetzt, den Infrarotbereich des Lichts zu nutzen. Hierdurch tritt *Acaryochloris marina* nicht in Konkurrenz zu anderen Cyanobakterien und kann in näherer Umgebung oder sogar in Symbiose mit anderen Cyanobakterien leben (Swingley et al., 2008).

Die verschiedenen Arten von Cyanobakterien wurden von Stanier (Rippka et al., 1979) nach morphologischen Gesichtspunkten und der Art der Teilung in fünf verschiedene Gruppen eingeteilt: Chroococcales, Pleurocapsales, Oscillatoriales, Nostocales und Stigonematales. Die Croococcales sind einzellige Bakterien, die kugel- oder auch ellipsenförmig sein können. Selten sind sie auch stäbchenförmig.

Die Teilung erfolgt binär, selten durch Knospung. Kolonien, die von manchen dieser Arten gebildet werden, werden von Schleim oder Ähnlichem zusammengehalten. Der Großteil der heute vollständig sequenzierten Cyanobakteriengenome gehört zu dieser Gruppe wie zum Beispiel alle *Prochlorococcus marinus* und *Synechococcus* sp. Arten und auch *Acaryochloris marina*.

Zur Gruppe der Pleurocapsales gehören ebenfalls einzellige Cyanobakterien, die sich jedoch durch multiple Spaltung, bei der Kolonien von bis zu 1.000 Zellen entstehen können, fortpflanzen und fähig sind Stickstoff zu fixieren. Die Kolonien werden durch eine faserige Hülle zusammengehalten. Aus dieser Gruppe wurde noch kein Genom sequenziert.

Die dritte Gruppe besteht aus filamentösen Cyanobakterien, die allerdings keine Stickstoff-fixierende Zellen und Dauersporen bilden. Die Teilung erfolgt in derselben Ebene, Verzweigungen entstehen nicht. Aus dieser Gruppe ist bisher nur das Genom von *Trichodesmium erythreum* vollständig sequenziert.

Die Nostocales bilden ebenfalls Filamente, diese können jedoch sehr weit verzweigt sein. Außerdem findet Zelldifferenzierung in Heterozysten und Akineten statt. Heterozysten werden verwendet, um unter Sauerstoffabschluss Stickstoff zu fixieren, wohingegen in den dickwandigeren Akineten Nährstoffe gespeichert werden. Aus dieser Gruppe sind bisher die Genome von drei Arten – insbesondere das von *Anabaena variabilis* – sequenziert.

Zu der fünften Gruppe gehören die komplexesten Cyanobakterien. Sie bilden Filamente und darin echte Verzweigungen. Die Teilung erfolgt in zwei Ebenen. Auch sie können Heterozysten und Akineten ausbilden. Aus dieser Gruppe ist ebenfalls noch kein Genom sequenziert.

3.1.2. Primäre Endosymbiose

Andere Organismen, die ebenfalls zur Photosynthese fähig sind, sind die eukaryotischen Algen und die Pflanzen. Sie besitzen Plastiden – Chloroplasten – in denen die Photosynthese stattfindet. Im 19ten Jahrhundert waren die Chloroplasten als “Chlorophyllkörper“ bekannt. Allerdings stellten mehrere Wissenschaftler durch Mikroskopie fest, dass diese Körper relativ eigenständig sind und meh-

rere Eigenschaften von Bakterien besitzen. Dies führte 15 Jahre später zu der Theorie von Mereschkowsky, dass sich Chloroplasten aus freilebenden Cyanobakterien entwickelt haben (Mereschkowsky, 1905). Erst 1967 wurde jedoch die Endosymbiontentheorie, die besagt, dass sowohl die Mitochondrien als auch die Chloroplasten durch eine Endosymbiose mit einem Bakterium entstanden sind, von Lynn Sagan veröffentlicht (Sagan, 1967). Allerdings postulierte sie ebenfalls, dass die Flagellen auf eine Endosymbiose mit Prokaryoten zurückgehen. In Martin (1999) zeigte William Martin erstmals, dass zwar die Mitochondrien und die Chloroplasten eindeutig auf eine Endosymbiose zurückzuführen sind, die anderen Zellkompartimente der Eukaryoten jedoch nicht.

Die erste primäre Endosymbiose erfolgte vor circa 3 Milliarden Jahren. Der Theorie nach gingen ein Archaebakterium (Wirt) und ein α-Proteobakterium (Endosymbiont) eine Endosymbiose ein. Die Wasserstoffhypothese (Martin und Müller, 1998) beschreibt ein Szenario, in dem es sich bei den beiden Organismen um ein methanogenes Archaebakterium, das auf die Zufuhr von Wasserstoff angewiesen war, als Wirt und um ein wasserstoff- und kohlenstoffdioxid-produzierendes Eubakterium als Symbionten handelt. Sowohl Wasserstoff als auch Kohlenstoffdioxid müssen also in der Umgebung, in der die beiden Organismen vorkamen, vorhanden gewesen sein. Um möglichst viel des Wasserstoffs aufnehmen zu können, versuchte der Wirt das Eubakterium weitgehend zu umschließen. Dies hatte jedoch zur Folge, dass das Eubakterium nicht mehr genug Kontakt zu der Umgebung hatte, um fermentierbare organische Substrate aufzunehmen (Glukose). Dieses Problem wurde entweder durch die Erfindung von Transportern gelöst, die die Substrate durch die Membran des Archaebakteriums importieren können, oder durch die Nutzung der in der endosymbiontischen DNA kodierten Gene für die Transporter. Durch diese Vorgänge wurden Wirt und Endosymbiont immer stärker voneinander abhängig, bis eine frühe Form der heutigen Mitochondrien erreicht und somit der erste Eukaryot entstanden war. Einer der so entstandenen heterotrophen Eukaryoten ging dann vor circa 1,2 bis 1,5 Milliarden Jahren (Kaufman et al., 2007) eine Endosymbiose mit einem Cyanobakterium ein, wodurch er autotroph und der Vorläufer der heutigen Glaucophyten, Rotalgen, Grünalgen und Pflanzen wurde.

Der heterotrophe Eukaryot hatte einen starken Selektionsvorteil, da Cyanobakterien in der Lage sind Photosynthese zu betreiben, das heißt sie können aus

Kohlenstoffdioxid und Wasser Glukose und Sauerstoff bilden. Der Eukaryot wurde also durch die Symbiose mit dem Bakterium photoautotroph. Außerdem war bioverfügbarer Stickstoff, der unter anderem benötigt wird um Aminosäuren zu synthetisieren, nicht in großen Mengen vorhanden. Einige Cyanobakterien besitzen eine Nitrogenase wodurch sie elementaren Stickstoff (N_2) fixieren können. Dieser wird zuerst in Ammoniak (NH_3) und dann in verwertbare Verbindungen umgewandelt. Diese Reaktionen benötigen sehr viel Energie, die durch die Photosynthese gewonnen wird (Falkowski und Godfrey, 2008; Fay, 1992). Allerdings müssen sie unter Sauerstoffabschluss ablaufen, da die Nitrogenase sehr sauerstoffempfindlich ist.

Für diesen Zweck haben die Cyanobakterien der Gruppen IV und V spezialisierte Zellen, die Heterozysten genannt werden und keine Photosynthese betreiben, wodurch Sauerstoff gebildet würde (Fay, 1992; Stewart, 1980). Die unizellulären Cyanobakterien trennen Photosynthese und Stickstofffixierung zeitlich. Manche betreiben bei Tageslicht Photosynthese und im Dunkeln verwenden sie die bereitgestellte Energie, um elementaren Stickstoff zu fixieren und bioverfügbar zu machen (Mullineaux et al., 1981). Andere koordinieren die Photosynthese und die Stickstofffixierung nach den Phasen des Zellzyklus (Gallon et al., 1974). Einige Arten können jedoch nur unter anaeroben Bedingungen Stickstoff fixieren (Rippka und Waterbury, 1977).

3.1.3. Sekundäre Endosymbiose

Nicht alle photosynthetischen Eukaryoten haben ihre Plastide direkt über die primäre Endosymbiose erhalten. Innerhalb der Gruppe der photosynthetischen Protisten sind mindestens zwei Endosymbiosen eines dieser Protisten als Wirt mit einer Grün- oder Rotalge als Endosymbionten bekannt. So besitzt zum Beispiel *Euglena gracilis* eine Plastide, die sehr stark der Plastide der Grünalgen ähnelt, aber ansonsten keinerlei Ähnlichkeit mit dieser Gruppe aufweist (Gibbs, 1978; Taylor, 1974).

Diese Entdeckung zeigte, dass es möglich ist, dass Eukaryoten Plastiden von einem anderen Eukaryoten erhalten. Im Fall von Euglena war es ein nicht-photosynthetischer Vorfahr der Eugleniden, der eine Symbiose mit einer Grünalge eingegangen ist. Dieser Vorgang wird sekundäre Endosymbiose genannt. Im Laufe

der Evolution ging der Zellkern der Grünalge verloren, allerdings ist die sogenannte sekundäre Plastide bei den Eugleniden von drei, bei den Chromalveolaten von vier anstatt nur zwei Membranen umgeben. Eine andere Gruppe, die eine sekundäre Plastide aufweist, sind die Chlorarachniophyten, bei denen jedoch immer noch der Zellkern der Grünalge mit einem stark reduzierten Genom – ein Nucleomorph – vorhanden ist (Archibald, 2007).

Die Oomyceten sind Pflanzenparasiten, die nicht in der Lage sind Photosynthese zu betreiben, da die sekundäre Plastide in dieser taxonomische Gruppe verloren gegangen ist. Die Ciliophora oder Wimperntierchen gehören ebenfalls zu den Chromalveolaten. Diese taxonomische Gruppe enthält nicht-photosynthetische Einzeller, die die sekundäre Plastide, wenn sie sie jemals besessen haben, vollständig verloren haben (Archibald und Keeling, 2002). Zu dieser Gruppe gehören zum Beispiel *Paramecium tetraurelia* und *Tetrahymena thermophila*.

Eine weitere Gruppe der Chromalveolaten – die Apicomplexa – setzt sich aus nicht-photosynthetischen Parasiten der Tiere zusammen. Diese Parasiten, zu denen zum Beispiel der Malariaerreger *Plasmodium falciparum*, der Dünndarm-Parasit *Cryptosporidium parvum* und der Rinderparasit *Theileria parva* zählen, enthalten eine sekundäre Plastide, den Apicoplasten. Sie sind jedoch nicht in der Lage Photosynthese zu betreiben, da alle Gene, die hierfür benötigt werden, nicht mehr in dem Genom des Apicoplasten kodiert sind (Foth und McFadden, 2003; McFadden und Waller, 1997).

Im Gegensatz zu den Eugleniden und den Chlorarachniophyten ist der Vorfahr der Chromalveolaten eine Symbiose mit einer Rotalge eingegangen (Fast et al., 2001; Harper und Keeling, 2003; Keeling, 2009; Patron et al., 2004). Allerdings ist noch nicht geklärt, wie oft bei diesen Gruppen die sekundäre Endosymbiose vorkam. Thomas Cavalier-Smith hat 1999 die Chromalveolatenhypothese formuliert (Cavalier-Smith, 1999). Er verglich die Anzahl der Membranen in den Gruppen der Cryptophyten, Haptophyten, Heterokonta, und Alveolata (Dinoflagellaten, Apicomplexa und Ciliaten) und die Struktur der Plastiden.

Weil der Prozess der sekundären Endosymbiose sehr komplex ist – es musste ein Importmechanismus über drei oder vier Membranen erfunden werden – ist es eher unwahrscheinlich, dass er sehr häufig vorkam.

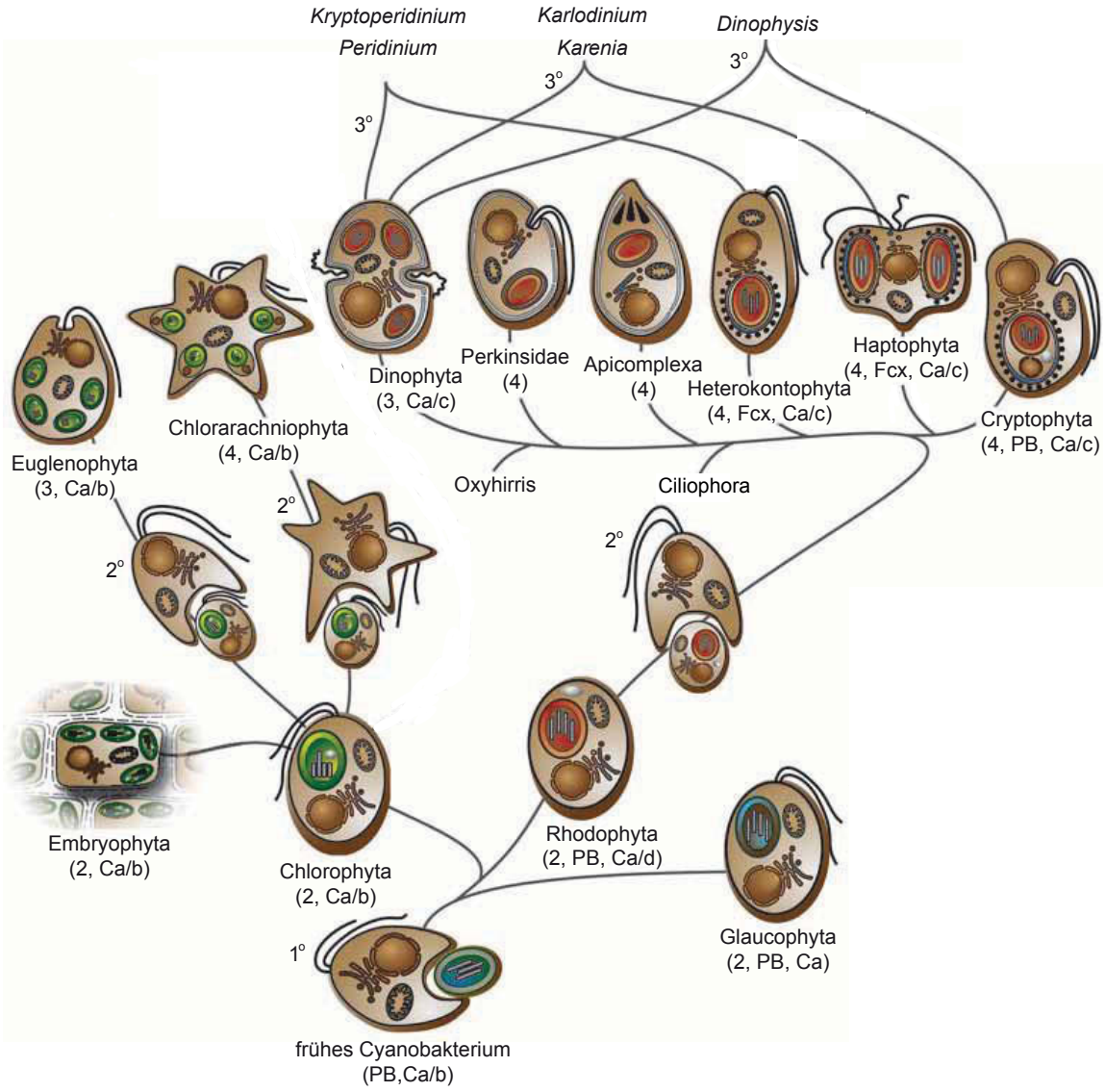

Abbildung 3.1.: Schematische Beschreibung der primären und sekundären Endosymbiose verändert nach Gould et al. (2008). Als erstes ging ein heterotropher Eukaryot eine Symbiose mit einem Cyanobakterium ein (1°), das im Laufe der Evolution zu einer Plastide reduziert wurde. Aus diesem photoautotrophen Eukaryoten entstanden die Chlorophyten (und aus diesen die Pflanzen), Rhodophyten und Glaukophyten. Später gingen jeweils ein Vorfahr der Eugleniden und ein Vorfahr der Chlorarachniophyten eine Endosymbiose mit einer Grünalge ein, wodurch sie sekundäre Plastiden erhielten (2°). Nach der Chromalveolatenhypothese ging außerdem der Vorgänger der Chromalveolaten eine Symbiose mit einer Rotalge ein. Die so entstandene rote sekundäre Plastide ging jedoch in der Linie der Ciliaten verloren. Einige andere Linien (zum Beispiel einige Dinoflagellaten) erhielten ihre Plastide über eine tertiäre Endosymbiose (3°), das heißt über die Endosymbiose eines Organimus mit sekundärer Plastide. In Klammern sind jeweils die Anzahl der Membranen um die Plastiden (2,3,4), die Art des verwendeten Chlorophylls (Chlorophyll a, b, c oder d) und das Vorhandensein von Phycobilinproteinen (PB) oder Fucoxanthin (Fcx) verzeichnet.

Deswegen schlug Cavalier-Smith 2001 vor, dass ein gemeinsamer Vorfahr dieser Gruppen eine Endosymbiose mit einer Rotalge eingegangen ist (Cavalier-Smith, 1999). Da jedoch zum Beispiel die Ciliaten nicht photosynthetisch sind, muss in dem Vorfahr dieser Gruppe die sekundäre Plastide wieder verloren gegangen sein. Von den Dinoflagellaten ist nur ungefähr die Hälfte der Organismen fähig Photosynthese zu betreiben. Sie besitzen entweder eine von drei Membranen umschlossene sekundäre Plastide oder eine tertiäre Plastide, die durch eine Symbiose mit einem Organismus, der eine sekundäre Plastide aufweist, entstanden ist.

Die Chromalveolatenhypothese wird immer noch stark diskutiert, da zwar in manchen Analysen die Monophylie der Chromalveolaten bestätigt werden konnte (Bachvaroff et al., 2005; Fast et al., 2001; Harper und Keeling, 2003; Patron et al., 2004; Yoon et al., 2002), in anderen diese jedoch abgelehnt werden musste (Arisue et al., 2002; Baldauf et al., 2000; Bodyl, 2005; Harper et al., 2005; Simpson et al., 2006). Allerdings konnte die enge Verwandtschaft in einigen Teilgruppen der Chromalveolaten bestätigt werden. So zeigen die Stramenopilen, Alveolata und zusätzlich dazu die Rhizaria eine enge Verwandtschaft (Patron et al., 2007). Diese Gruppe wurde in Burki et al. (2007) die SAR-Gruppe genannt.

Auch die Haptophyten und die Cryptophyten bilden eine eigene Gruppe (Hackett et al., 2007). Obwohl inzwischen allgemein anerkannt ist, dass der sekundäre Endosymbiont der Chromalveolaten eine Rotalge war, weisen sowohl die ersten Genomanalysen als auch eine Analyse von Moustafa et al. (2009) darauf hin, dass es einen relativ hohen Anteil an von Grünalgen und Pflanzen stammenden Genen in den Kerngenomen der Bacillariophyta *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* gibt (Armbrust et al., 2004; Bowler et al., 2008). Auch die Oomyceten, die eng mit den Bacillariophyta verwandt sind und zusammen mit ihnen die Gruppe der Stramenopilen bilden, zeigen einen hohen Anteil an "grünen" Genen (im Weiteren auch "grüne" beziehungsweise "rote" Signale genannt) (Tyler et al., 2006).

Moustafa et al. (2009) postulieren dass der Vorfahr der Chromalveolaten zuerst eine Endosymbiose mit einer Grünalge eingegangen ist. Dieser Endosymbiont ging dann jedoch verloren und wurde durch eine Rotalge als Endosymbionten ersetzt. Eine andere Theorie besagt, dass die Chromalveolaten eine Schwestergruppe zu den Archaeplastida sind (Hampl et al., 2009; Nozaki et al., 2007, 2009). Dies würde

bedeuten, dass die grünen Gene bereits in dem Vorfahren der Archaeplastida und Chromalveolata präsent waren und nicht durch endosymbiontischen Gentransfer (siehe Abschnitt 3.1.4) in das Kerngenom der Protisten eingefügt wurden.

Eine dritte Gruppe der photosynthetischen Eukaryoten erwarb die Fähigkeit zur Photosynthese über eine Endosymbiose mit einem Protisten, der selber eine sekundäre Plastide besaß. Dies wird tertiäre Endosymbiose genannt und ist bisher nur in den Dinoflagellaten und einigen Ciliaten bekannt (Imanian und Keeling, 2007; Kim et al., 2007; Vesteg et al., 2009).

3.1.4. Endosymbiontischer Gentransfer

Die Veränderungen, die der jeweilige Endosymbiont sowohl bei der primären als auch bei der sekundären Endosymbiose im Laufe der Evolution durchgemacht hat, sind enorm. Viele Gene, die in dem Genom des Vorläufers der Cyanobakterien kodiert waren, gingen verloren oder wurden durch endosymbiontischen Gentransfer (EGT) in das Kerngenom des Wirts übertragen (Martin et al., 1993). Dies setzt jedoch voraus, dass von dem Wirt eine neue Importmaschinerie in die Plastiden entwickelt werden musste, so dass die Proteine, die zum Beispiel für die Photosynthese in den Chloroplasten gebraucht werden, aus dem Cytoplasma importiert werden können. Andere endosymbiontische Gene haben sogar ursprüngliche Gene des Wirts oder mitochondriale Gene ersetzt (Allen, 2003; Martin und Herrmann, 1998; Martin und Schnarrenberger, 1997).

Im Fall der sekundären Endosymbiose sind die Möglichkeiten für endosymbiontischen Gentransfer sogar noch komplexer. Hier wurden sowohl Gene aus dem Kerngenom des Endosymbionten als auch Gene aus der Plastide des Endosymbionten in das Kerngenom des Wirts übertragen. Außerdem ist die Importmaschinerie über drei bzw. vier Membranen weitaus komplizierter und zu großen Teilen noch nicht vollständig erforscht (Chaal und Green, 2005; Hempel et al., 2009; Ishida et al., 2000). Das Kerngenom des Endosymbionten ging in den meisten Fällen vollständig verloren. Nur die Endosymbionten der Cryptophyceen und Chlorarachniophyceen weisen noch einen Nucleomorph auf, der zwischen den äußeren beiden und den inneren beiden Membranen lokalisiert ist, allerdings ein stark reduziertes Genom aufweist.

Der Gentransfer vom Endosymbionten zum Wirt und der Genverlust waren so stark, dass nur noch ein Bruchteil der Gene, die in dem frühen Cyanobakterium kodiert waren in dem Plastidengenom zu finden sind. In heutigen primären Plastiden finden sich zwischen 60 und 200 Genen (Martin et al., 2002), während heutige Cyanobakterien zwischen 1.800 und 7.400 Gene besitzen. Analysen, in denen die genaue Zahl der in das Kerngenom von Pflanzen und Grünalgen transferierten Gene bestimmt wurde, zeigen, dass bis zu 18 % des Wirtsgenoms aus cyanobakteriellen Genen besteht (Martin et al., 2002). In Deusch et al. (2008) wurde diese Analyse mit einer größeren Datenbank und außerdem mit *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae* wiederholt. Auch hier fanden sich Werte zwischen 12,7 % und 18 %. Für den Glaucophyten *Cyanophora paradoxa*, für den allerdings bisher nur eine EST-Datenbank existiert, wurde gezeigt, dass 11 % des Kerngenoms cyanobakteriellen Ursprungs sind (Reyes-Prieto et al., 2006).

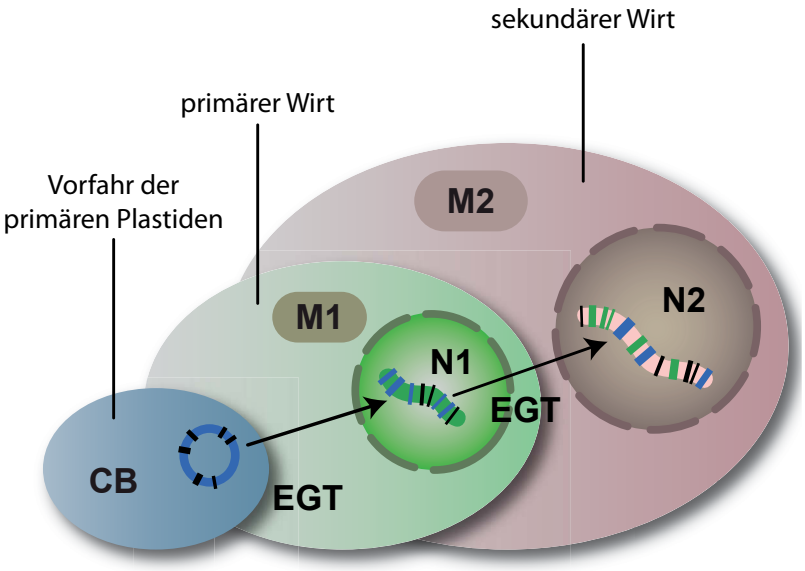

Abbildung 3.2.: Schematische Darstellung des endosymbiontischen Gentransfers verändert nach Elias und Archibald (2009). Gene aus dem Genom des primären Plastiden wurden im Laufe der Evolution in das Kerngenom des primären Wirts transferiert. Das Genom der Plastide wurde stark reduziert. Nach der Etablierung der sekundären Endosymbiose wurden Gene aus dem Kerngenom des Endosymbionten (Grün- oder Rotalge) in das Kerngenom des sekundären Wirts übertragen. Bei den meisten Protisten mit sekundärer Plastide ging das Kerngenom des Endosymbionten vollständig verloren, aus dem Genom der sekundären Plastide wurden ebenfalls Gene in das Kerngenom des Wirts transferiert.

Moustafa et al. (2009) haben 2.533 und 2.423 Proteine in dem Kerngenom der photosynthetischen Protisten *Thalassiosira pseudonana* und *Phaeodactylum tricornutum*

identifiziert, für die in phylogenetischen Bäumen ein Ursprung aus Grünalgen, Pflanzen oder Rotalgen abgeleitet werden kann. Mehr als 1.700 dieser Proteine zeigen in dieser Analyse eine Verbindung zu Grünalgen und Pflanzen.

3.2. Vom Protein zur Proteinfamilie

Eine Analyse, die die Verteilung und Verwandtschaft von Proteinen in verschiedenen Organismen untersuchen soll, beruht darauf, dass diese homologen Gene beziehungsweise Proteine in verschiedenen Orgnanismen gefunden und verglichen werden. Je nachdem wie ähnlich sich die Proteine aus verschiedenen Organismen sind, können daraus Rückschlüsse über die Verwandtschaft zwischen den jeweiligen Organismen geschlossen werden und auch darüber, ob ein Gen sich ursprünglich in dem Organismus entwickelt hat oder über endosymbiontischen Gentransfer in das Genom gelangt ist. Ein Protein, dass eine höhere Verwandtschaft zu einer anderen Organismengruppe aufweist als zu der eigenen, ist mit hoher Wahrscheinlichkeit durch Gentransfer in das Genom des Organismus gelangt.

Um diese Entscheidung treffen zu können, müssen jedoch die Homologen aus verschiedenen Organismen verlässlich erkannt werden. Im Folgenden wird der Weg von einem einzelnen Protein zu der zugehörigen Proteinfamilie, bestehend aus den Homologen, und die weitere phylogenetische Analyse beschrieben.

3.2.1. Graphen

Eine häufig verwendete Darstellungsmethode für die Verbindungen zwischen Genen, Proteinen oder auch Organismen ist der Graph. Um mit diesen mathematisch gut beschriebenen Konstrukten arbeiten zu können sind einige Definitionen notwendig:

Sei V eine endliche Menge an Elementen im Weiteren **Vertices** oder **Knoten** genannt und E eine endliche Menge an **Kanten**, die jeweils zwei Vertices verbinden.

1. Ein gewichteter Graph G bestehend aus Paaren (V,E), ist ein Graph, der eine Funktion w enthält, die jeder der n Kanten $e_1, \ldots, e_n$ einen Wert (eine Gewichtung) aus $\mathbb{R}_{\geq 0}$ zuweist. Die Größe der Gewichtung gibt an, wie stark die zwei Vertices miteinander verbunden sind.

a) Enthält w sowohl einen Wert für $V_1 \longmapsto V_2$ als auch für $V_2 \longmapsto V_1$, so ist w **symmetrisch** und G **ungerichtet**. Ansonsten ist w **unsymmetrisch** und G **gerichtet**.

b) Enthält G keine Schleifen, das heißt $w(v,v) = 0, \forall\, v \in V$, dann wird der Graph **irreflexiv** genannt, **reflexiv** sonst.

2. Ein **Ähnlichkeitsraum** besteht aus einer symmetrischen Funkion s und einer endlichen Menge an Vertices V, wobei s folgende Bedingung erfüllt:
$s(u,v) = \infty \Longleftrightarrow u = v$.
s wird Ähnlichkeitsmaß oder Ähnlichkeitskoeffizient genannt.

In diesem Zusammenhang können z.B. Organismen als Vertices in einem Graphen betrachtet werden. Die Kanten zwischen den Vertices entsprechen dann den Verwandtschaftsbeziehungen zwischen diesen Organismen. Ein sogenannter **phylogenetischer Stammbaum** ist also ein Graph, der die Evolution der darin enthaltenen Organismen graphisch darstellt. Ist dieser Graph gerichtet, so wird er gewurzelt genannt und ein Knoten wird als Wurzel bezeichnet, ist der Graph ungerichtet so ist er ungewurzelt.

Definition 3.1 ***Der Grad eines Vertex***
*In einem ungerichteten Graph kann jedem Vertex $v \in V$ ein **Grad** zugeteilt werden. Dieser gibt die Anzahl der Kanten an, die diesen Vertex mit einem anderen verbinden. In einem gerichteten Graphen wird der Grad des Vertex durch den **Eingans-** (Anzahl der eingehenden Kanten) und einen **Ausgangsgrad** (Anzahl der ausgehenden Kanten) beschrieben.*

Ein phylogenetischer Stammbaum kann aus drei verschiedenen Arten von Vertices bestehen:

- Die **Wurzel** hat den Eingangsgrad 0 und den Ausgangsgrad 2 und kommt nur in einem gewurzelten Baum vor.

- Die inneren Vertices haben den Eingangsgrad 1 und den Ausgangsgrad 2. Sie werden auch **HTUs** (engl. *hypothetical taxonomic unit*) genannt.

- Die äußeren Vertices (Blätter) haben den Eingangsgrad 1 und den Augangsgrad 0. Sie werden als **OTUs** (engl. *operational taxonomic unit*) bezeichnet und repräsentieren die enthaltenen Organismen.

Ein gewurzelter phylogenetischer Stammbaum enthält eine Wurzel und macht so Annahmen über den gemeinsamen Vorläufer aller darin enthaltenen Organismen. Von dieser Wurzel ausgehend haben sich alle HTUs und schlussendlich die OTUs entwickelt. Der Graph ist demnach gerichtet. Ein ungewurzelter Baum verzichtet auf diese Annahme. Als Ähnlichkeitsmaß s zwischen den Organismen können sowohl morphologische Kriterien als auch Unterschiede in den Gen- oder Proteinsequenzen dienen.

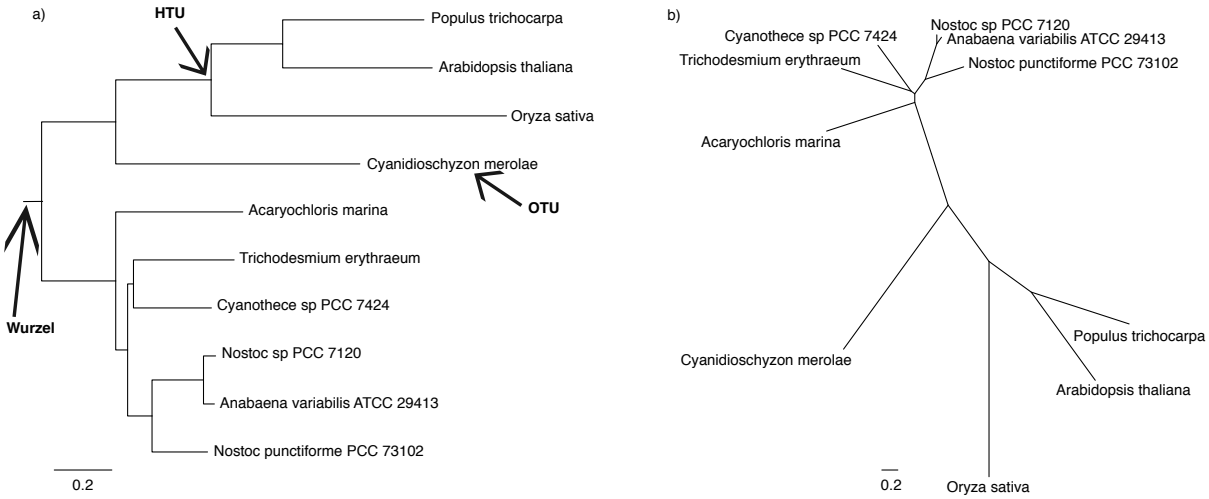

Abbildung 3.3.: a) Beispiel eines gewurzelten Baums mit 10 Taxa. Die Wurzel, die OTUs und die HTUs sind gekennzeichnet. b) Beispiel eines ungewurzelten Baums mit 10 Taxa.

Definition 3.2 ***Ein Pfad in einem Graphen***

1. *Ein **Pfad** in einem Graphen beschreibt eine Folge von Vertices $v_1, \ldots, v_n$, die einen Vertex mit einem anderen verbinden.*
2. *Die Länge L eines Pfades ist die Anzahl der Kanten auf dem Pfad.*

Der Pfad zwischen zwei OTUs in einem phylogenetischen Baum kann als die evolutionäre Distanz zwischen den beiden Organismen interpretiert werden, indem die Werte des Ähnlichkeitsmaßes s für jede Kante, die auf dem Pfad liegt, aufsummiert wird.

Nicht nur die Organismen selber sondern auch die Proteine der Organismen und ihre Ähnlichkeit können in einem Graphen abgebildet werden. Um herauszufinden, welche Proteine aus welchem Organismus zu einer Proteinfamilie gehören, müssen

diese aufgrund ihrer Ähnlichkeit in Gruppen eingeteilt werden. Dieses Verfahren wird Clustering genannt.

Definition 3.3 ***Partition oder Clustering eines Graphen***
Als Partition oder ***Clustering*** *der Vertices V eines Graphen wird die Menge aller paarweise disjunkten Mengen $V_1, \ldots, V_d$ bezeichnet, wobei jede Menge V_i eine Teilmenge von V bildet und die Vereinigung $\cup_{i=1,\ldots,d} V_i$ V entspricht. Die Mengen V_i werden* ***Cluster*** *genannt.*

Das Erstellen eines Clusterings, in dem die Vertices in einem Cluster signifikant ähnlicher zueinander als zu den Vertices in anderen Clustern sind, ist nicht trivial lösbar und es gibt inzwischen viele verschiedene Algorithmen für dieses Problem. Einige Algorithmen werden im nächsten Kapitel vorgestellt.

3.2.2. Hierarchische Clusteranalyse

Das Ziel einer Clusteranalyse ist es, viele Elemente nach ihren Eigenschaften in Gruppen einzuteilen. Innerhalb einer Gruppe sollten die Elemente möglichst ähnlich sein, die Gruppen selber sollten jedoch möglichst verschieden sein. So ist das Berechnen eines phylogenetischen Baumes durch das Clustern der zugrundeliegenden Merkmale (z.B. Sequenzähnlichkeit) der Organismen zu lösen. Dies ist jedoch nur möglich, wenn man bereits weiß, welche Gene oder Proteine zu einer "Familie" gehören und auf das Gen oder Protein eines gemeinsamen Vorläufers zurückzuführen sind. Um diese Familien zu finden, muss die Menge aller Proteine aller Organismen, für die eine Analyse druchgeführt werden soll auf ihre Ähnlichkeit untereinander untersucht und daraufhin in Gruppen eingeteilt werden. Es gibt viele verschiedene Ansätze, um dieses Problem der Einteilung in Gruppen zu lösen.

In der Biologie wird häufig ein agglomeratives hierarchisches Clusterverfahren verwendet. Hierbei werden die einzelnen Elemente so in die Cluster eingeteilt, dass ein gewurzelter, binärer Baum (Dendrogramm) entsteht.

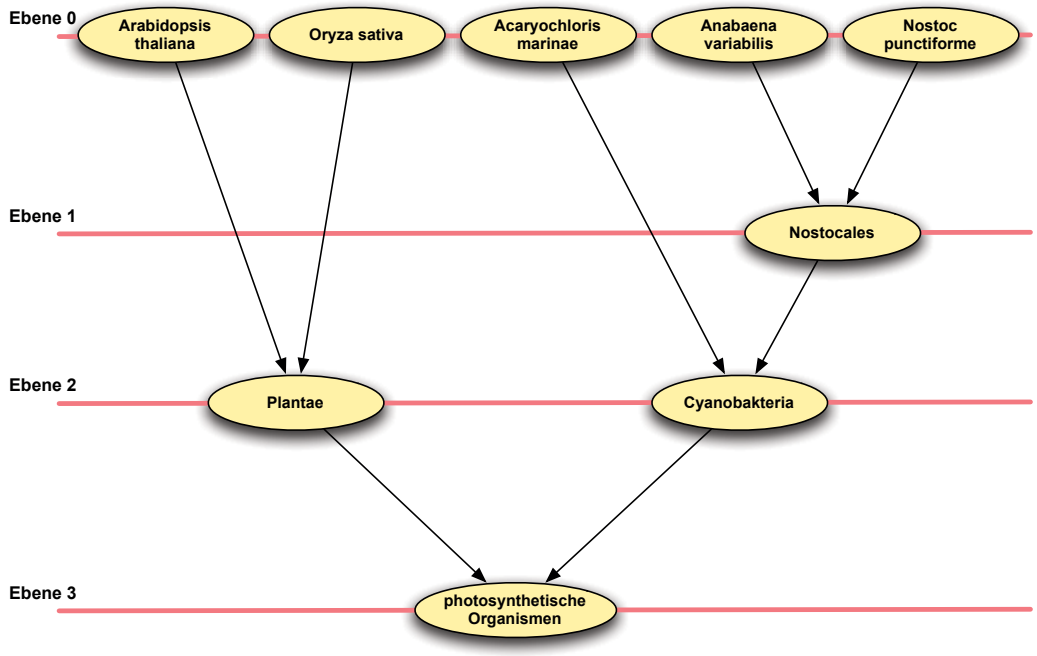

Abbildung 3.4.: Dendrogramm des Clusterings eines Proteins der Organismen *Arabidopsis thaliana*, *Oryza stiva*, *Nostoc punctiforme*, *Anabaena variabilis* und *Acaryochloris marina*. Zuerst werden die beiden Pflanzenproteine zu einem Cluster Plantae und zwei der cyanobakteriellen Proteine zu einem Cluster Nostocales zusammengefasst. Zu dieser Gruppe wird dann das dritte cyanobakterielle Protein hinzugefügt, wodurch ein Cluster entsteht, der alle cyanobakteriellen Proteine enthält. Zuletzt werden alle Proteine in einem Cluster vereint. Die verschiedenen Ebenen in diesem Dendrogramm entsprechen unterschiedlichen Clustergranularitäten.

Die Blätter des Baumes repräsentieren die einzelnen Elemente, also Cluster der Größe 1, die auch Waisen (engl. *orphans*) genannt werden. Diese werden dann auf jeder Baumebene zu größeren Gruppen zusammengefasst, indem diejenigen Cluster, deren Abstand minimal ist, zu einem Cluster vereint werden. Dies erfordert auf jeder Ebene die Berechnung der Distanz zwischen zwei Clustern und die Definition einer Distanzfunktion. Generell gibt es drei verschiedene Ansätze, die Distanz zwischen zwei Clustern zu berechnen:

- *single-linkage*
- *average-linkage*
- *complete-linkage*

Bei dem *single-linkage*-Verfahren (Florek et al., 1951), das auch als Nächste-Nachbarn Methode bekannt ist (vgl. Abbildung 3.5a), entspricht die Distanz zwischen zwei

Clustern X und Y der minimalen Distanz zwischen zwei Elementen x und y der Cluster:

$$dist(X,Y) = \min_{x \in X, y \in Y}(x,y)$$

Dieses Verfahren ist jedoch sehr anfällig für Ausreißer, da immer nur ein Element pro Cluster betrachtet wird. Außerdem werden hierdurch eher "kettenförmige" Cluster erstellt, da es nur wenige Elemente benötigt, um zwei Cluster zu vereinen (Jain und Dubes, 1988).

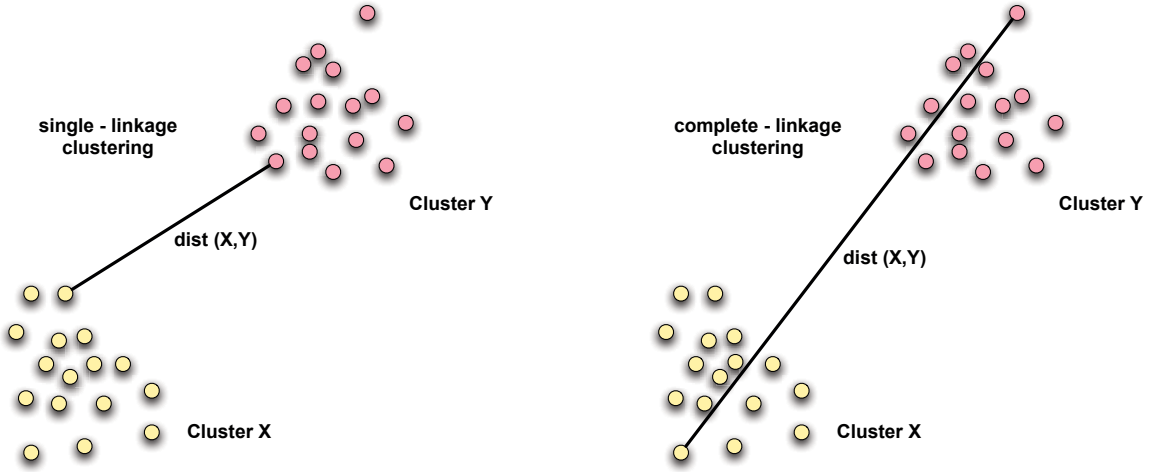

Abbildung 3.5.: Graphische Darstellung des *single-* und *complete-linkage-*Clusteringverfahrens. Zwei Cluster X und Y werden aufgrund der minimalen bzw. maximalen Distanz zweier Elemente der Cluster vereint.

Das *complete-linkage*-Verfahren versucht diesen Effekt zu verhindern, indem der Durchmesser der Cluster minimiert wird (Sorensen, 1948). Somit ist hier die Distanz zwischen zwei Clustern X und Y die maximale Distanz zwischen zwei Elementen x und y der Cluster:

$$dist(X,Y) = \max_{x \in X, y \in Y}(x,y)$$

Dies resultiert darin, dass eher viele, relativ kleine Cluster gebildet werden, die einen ähnlichen Durchmesser aufweisen. Ein weiteres häufig verwendetes Verfahren ist das *average-linkage*-Verfahren, das eher unter dem Namen UPGMA (engl. *unweighted pair-group method using the average approach*) bekannt ist (Sokal und Michener, 1958). Hierbei bestimmt nicht nur jeweils ein Element des Clusters den Abstand zu einem anderen, sondern die mittlere Distanz aller Elemente eines

Clusters:

$$dist(X,Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} dist(x,y)$$

Hierdurch werden Cluster erzeugt, die in ihrer Struktur zwischen den beiden anderen Verfahren liegen. Weder sind sie kettenförmig noch sehr klein. Allerdings ist dieses Verfahren sehr rechenintensiv, da alle Abstände zwischen allen Elementen einbezogen werden. Für sehr große Datensätze bedeutet dies, dass die Berechnungen sehr viel Arbeitsspeicher benötigen.

Allerdings haben die hierarchischen Clusteringverfahren folgende Nachteile:

- Es muss eine Distanzfunktion zwischen den Clustern definiert werden, die das Clustering sehr stark beeinflusst.
- Das Clustering muss noch aus dem erhaltenen Dendrogramm extrahiert werden, das heißt es muss ein Schwellenwert s festgesetzt werden, der bestimmt ob eine Gruppe in den einen oder anderen Cluster gehört.

Um diesen Nachteilen zu entgehen, wurde das sogenannte Markov-Clustering entwickelt, das im Folgenden vorgestellt wird.

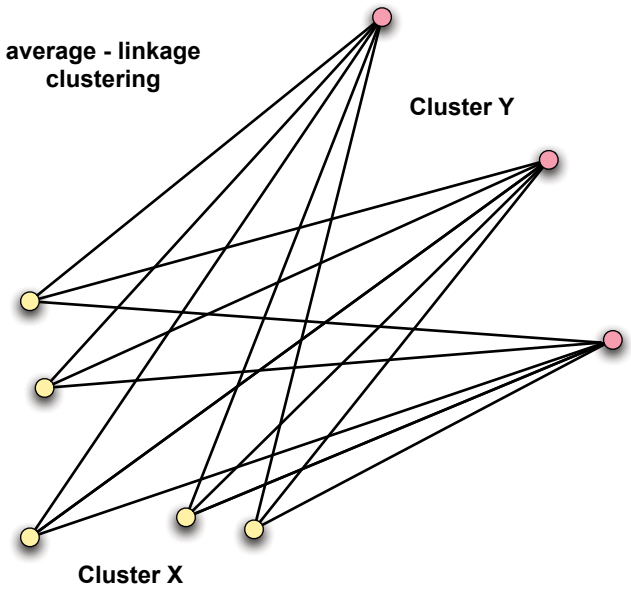

Abbildung 3.6.: Graphische Darstellung des *average-linkage*-Clusteringverfahrens. Zwei Cluster X und Y werden aufgrund der durchschnittlichen Distanz aller Elemente der Cluster vereint.

3.2.3. Das Markov-Clustering

Das Markov-Clustering, das 2000 von Stijn van Dongen entwickelt wurde (van Dongen, 2000*b*), basiert auf der Markov-Kette, einem diskreten stochastischen Prozess.

Definition 3.4 ***Die Markov-Kette***
Eine Markov-Kette (nach Andrei Andrejewitsch Markov, 1856 - 1922) ist ein diskreter stochastischer Prozess, der der ***Markoveigenschaft*** *genügt:*
$\mathbb{P}(X_{t+1} = i+1 | X_t = i_t, X_{t-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{t+1} = i+1 | X_t = i_t)$
Dies bedeutet, dass der Zustand der Markov-Kette zum Zeitpunkt $t+1$ nur von dem aktuellen Zustand zum Zeitpunkt t abhängt, nicht jedoch von den Zuständen davor. Die Markov-Kette wird auch ***Irrfahrt*** *genannt.*

Das Markov-Clustering wird vorrangig verwendet, um Gene oder Proteine aufgrund ihrer Sequenzähnlichkeit in Gen- beziehungsweise Proteinfamilien einzuteilen. Jedes Protein repräsentiert einen Vertex des ungerichteten Graphen G, der die Grundlage für das Clustering bildet. Die Kanten zwischen den Proteinen werden durch Homologiesuchen (z.B. reziproker BLAST Abschnitt 4.3.5) bestimmt und die paarweisen Sequenzidentitäten bilden das Ähnlichkeitsmaß s. Auf diese Art bilden sich innerhalb des Graphen bereits Regionen aus, in denen die Vertices stärker untereinander verbunden sind und Regionen, in denen sehr wenige Kanten vorkommen.

Definition 3.5 ***Die Zusammenhangskomponente***

1. *Gibt es in einem gerichteten Graphen sowohl einen Pfad von v nach w als auch einen von w nach v bzw. in einem ungerichteten Graphen einen Pfad zwischen v und w, so sind die beiden Vertices zusammenhängend oder stark verbunden.*
2. *Sind alle Vertices in einem Graphen G zusammenhängend, so ist G zusammenhängend.*
3. *Ein Teilgraph G' eines Graphen G heißt Zusammenhangskomponente, wenn alle Vertices von G' zusammenhängend sind.*
4. *G heißt k-zusammenhängend, wenn es k Zusammenhangskomponenten gibt.*

Eine Irrfahrt auf einem Graphen kann wie folgt beschrieben werden: Die Irrfahrt startet zum Zeitpunkt 0 an einem zufällig gewählten Vertex V. Der nächste Vertex,

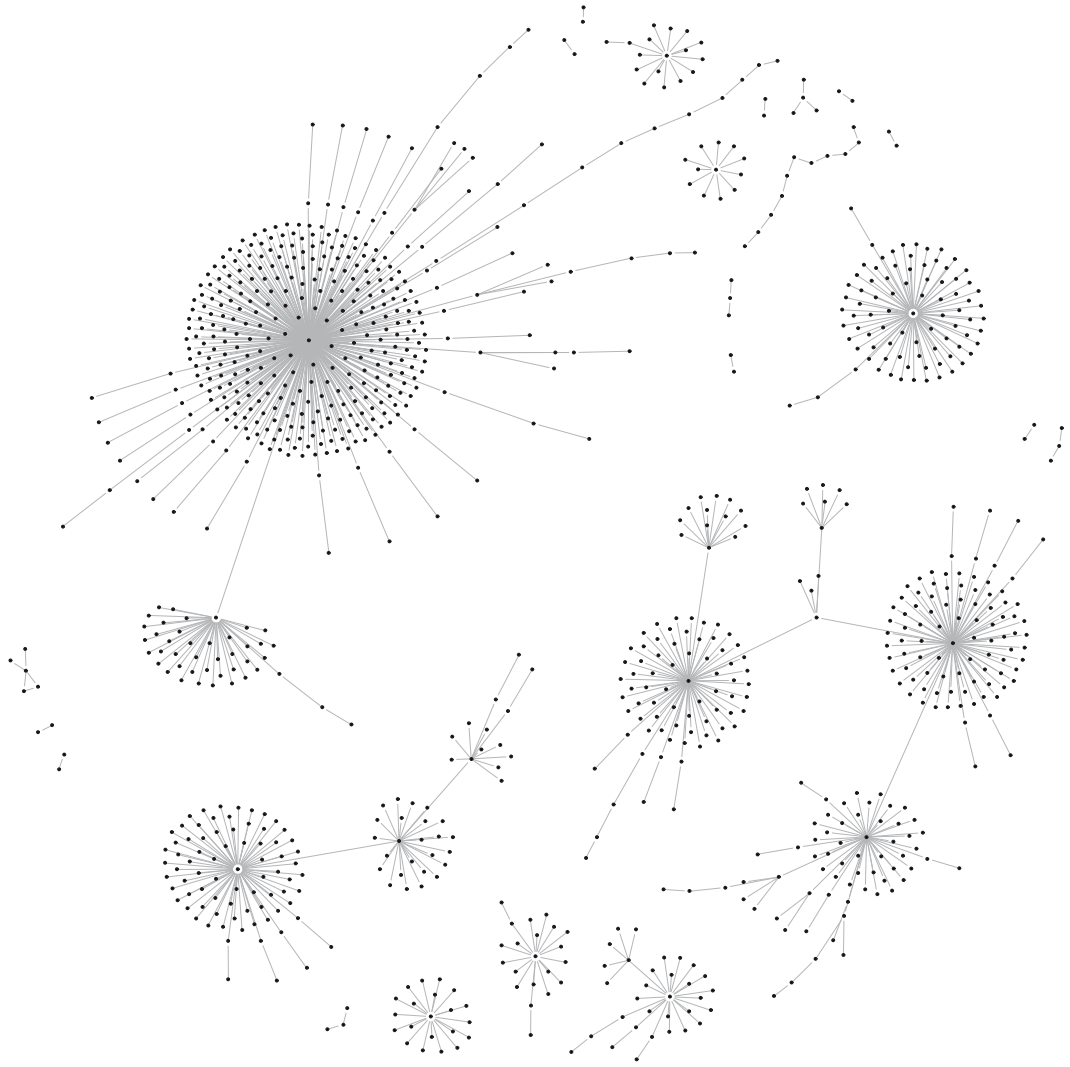

Abbildung 3.7.: Beispiel eines ungerichteten Graphen mit einer eindeutigen Clusterstruktur. Die Vertices (schwarze Punkte) sind mit Kanten (graue Linien) verbunden. Je nachdem wie stark die Vertices untereinander verbunden sind, können sie zu Clustern zusammengefasst werden. Manche Vertices haben nur sehr wenig Verbindungen zu anderen und können bei einem Clustering als sogenannte Waisen nicht in einen Cluster eingeordnet werden.

der besucht wird, ist nur abhängig von den Gewichtungen der von diesem Vertex ausgehenden Kanten. Das heißt die Irrfahrt geht am wahrscheinlichsten zu dem Vertex *W* weiter, der zu diesem Vertex am ähnlichsten ist. Ist die Irrfahrt an diesem Vertex angekommen, werden wieder nur die Übergangswahrscheinlichkeiten zu den Vertices betrachtet, die zu dem Vertex *W* adjazent sind. Welcher Vertex als nächstes – zum Zeitpunkt 2 – besucht wird ist also nur abhängig von dem Vertex, in dem sich die Irrfahrt gerade befindet und nicht mehr von dem Vertex, in dem sie sich vorher befunden hat.

Ausgehend von diesen Überlegungen hat Stijn van Dongen in seiner Doktorarbeit (van Dongen, 2000*b*) und in dem Technischen Report (van Dongen, 2000*a*) das Markov-Clustering entwickelt und beschrieben. Der Algorithmus wird in dieser Arbeit in Abschnitt 4.3.8 näher erläutert, hier soll nur die dem Clustering zugrundeliegende Idee erklärt werden.

Definition 3.6 ***Das Graph Clustering Paradigma***

1. *Die Anzahl der längeren Pfade zwischen zwei Vertices ist groß, wenn sich diese in demselben stark verbundenen Cluster bzw. einer Zusammenhangskomponente befinden. Gehören die Vertices zu unterschiedlichen Clustern oder Zusammenhangskomponenten, so gibt es nur wenige längere Pfade zwischen ihnen.*
2. *Eine Irrfahrt, die in einer Zusammenhangskomponente startet, wird diese nicht verlassen, bevor nicht viele der zu der Komponente gehörenden Vertices besucht wurden.*
3. *Betrachtet man die Menge aller kürzesten Pfade in G, so werden die Kanten, die zwei Zusammenhangskomponenten verbinden, Teil vieler dieser kürzesten Pfade sein.*

Dies bedeutet, dass zum Beispiel eine endliche Irrfahrt, die bei einem Vertex innerhalb einer beliebigen Zusammenhangskomponente startet, diese sehr wahrscheinlich gar nicht verlässt oder erst fast alle anderen Vertices in dieser Komponente besucht. Also findet man innerhalb der Komponenten sehr viele sehr lange Pfade, zwischen den Komponenten jedoch nur sehr wenige und wenn dann nur sehr kurze Pfade. Für eine ausführliche Beschreibung ist hier die Lektüre von van Dongen (2000*b*) empfohlen.

3.3. Vergleich der Proteine in einer Proteinfamilie

In einer phylogenetischen Analyse sollen homologe Proteine, also Proteine, die in verschiedenen Spezies und dem gemeinsamen Vorfahren dieser Spezies vorkommen und dieselbe Funktion haben, auf ihre Ähnlichkeit untersucht werden. Aus der Ähnlichkeit dieser Proteine soll dann die Verwandtschaftsbeziehung zwischen den Spezies geklärt werden. Der erste Schritt ist also ein Vergleich aller Proteine aller Spezies mit allen Proteinen aller anderen Spezies (siehe Abschnitt 4.3.5), dann müssen diese Proteine in Proteinfamilien eingeteilt werden (siehe Abschnitt 4.3.8). Wurden diese Familien durch ein Clusteringverfahren oder eine Sternsuche, bei der man von nur einer Spezies ausgeht anstatt alle mit allen zu vergleichen, erstellt, so müssen die Proteine in eine Form gebracht werden, in der gleiche Positionen identifiziert und verglichen werden können. Für diesen Zweck werden paarweise oder multiple Alignments von diesen Proteinen erstellt.

3.3.1. Alignments

Bei der Erstellung eines Alignments wird versucht, homologe Positionen in Proteinen, die zu derselben Proteinfamilie gehören, verschiedener Organismen zu identifizieren, wodurch Punktmutationen, Insertionen und Deletionen erkannt werden können. Bei einem paarweisen Alignment werden nur zwei Sequenzen miteinander verglichen (siehe Abschnitt 4.3.6.1), bei einem multiplen Alignment werden mehrere Sequenzen aligniert (Abschnitt 4.3.6.2). Das Zeichen "–" wird in eine Sequenz eingefügt, wenn eine Deletion in dieser oder eine Insertion in der anderen Sequenz stattgefunden hat.

Es wurden viele veschiedene Algorithmen entwickelt, um Alignments zu erstellen, die die Verwandtschaft zwischen den Sequenzen akkurat wiedergeben (Edgar, 2004*a*; Needleman und Wunsch, 1970; Notredame et al., 2000; Thompson et al., 1994). Aus einem Alignment kann dann unter anderem die Ähnlichkeit und die Identität der Sequenzen berechnet werden.

3.3.2. Evolutionäre Distanzen

Nachdem ein multiples Alignment berechnet wurde, kann aus diesem auf die Distanz zwischen zwei Sequenzen geschlossen werden und auf diese Weise eine Distanzmatrix erstellt werden. Hierfür gibt es viele Algorithmen und Program-

me. Die einfachste Distanz ist die p-Distanz oder Hammingdistanz (Hamming, 1950). Hierbei werden die Unterschiede zwischen zwei Sequenzen gezählt, aufsummiert und durch die Anzahl der verglichenen Positionen geteilt. Auf Gene oder Proteine bezogen hat dieses Modell hat den Nachteil, dass es weder mehrfache Mutationen noch Rückmutationen mit einbezieht und die Mutationsraten zwischen verschiedenen Paaren von Aminosäuren in Wirklichkeit nicht gleich sind. So ist die Wahrscheinlichkeit, dass eine Aminosäure in eine ähnliche Aminosäure mutiert höher als in eine, die ganz andere chemische Eigenschaften besitzt.

Aus diesem Grund entwickelten erst Jukes und Cantor in Jukes und Cantor (1969) ein stochastisches Markov-Modell für DNA-Sequenzen und später Margaret Dayhoff (Dayhoff et al., 1972, 1978) ein Modell für Proteinsequenzen sowie eine Substitutionsmatrix, für die die Übergangswahrscheinlichkeiten zwischen den einzelnen Aminosäuren anhand ihres Auftretens und der Häufigkeit der Austausche in Alignments von Sequenzen, die mindestens zu 85 % identisch sind, berechnet wurden. 1992 entwickelten Jones et al. (1992) eine Substitutionsmatrix anhand einer größeren und neueren Datenbank.

Die Werte in diesen Matrizen geben an, wie ähnlich sich zwei Aminosäuren sind. Ein höherer Wert bedeutet, dass sich zwei Aminosäuren sehr ähnlich sind, zwei unähnliche Aminosäuren haben einen sehr kleinen beziehungsweise sogar negativen Eintrag in der Matrix. Mit diesen Übergangswahrscheinlichkeiten können zum Beispiel mithilfe eines *Maximum-Likelihood*-Schätzers die phylogenetischen Distanzen zwischen den Sequenzen ermittelt werden (siehe Abschnitt 4.3.9.1). Einen Überblick über verschiedene Modelle gibt zum Beispiel (Lió und Goldman, 1998).

3.3.3. Phylogenetische Bäume

Es gibt viele verschiedene Modelle, Algorithmen und Programme um phylogenetische Bäume aus Sequenzdaten zu erstellen. Eine Methode, die nicht sehr rechenintensiv ist und deswegen für große Datensätze oft verwendet wird, ist die in Abschnitt 3.3.2 beschriebene Distanzmethode. Aus der Distanzmatrix kann mithilfe des UPGMA- oder des *neighbor-joining* (siehe Abschnitt 4.3.10.1) Algorithmus ein Baum erstellt werden. Beide Algorithmen beruhen auf Clusteringalgorithmen (3.3), wobei UPGMA dem *average*- und *neighbor-joining* dem *single-linkage* Verfahren

entspricht.

Eine Methode zur Erstellung von phylogenetischen Bäumen, die im Allgemeinen als überlegen zu der *neighbor-joining* Methode angesehen wird, ist die *Maximum-Likelihood*-Methode. Dieser sehr rechenintensive Algorithmus wurde im Jahr 1964 in Cavalli-Sforza et al. (1964) das erste Mal erwähnt. Er basiert auf der Idee, dass für jede Position überprüft wird, wie wahrscheinlich ein Baum ist gegeben ein Modell und die Daten (ein multiples Alignment). Die Wahrscheinlichkeiten für das komplette Alignment werden aufsummiert und es wird versucht den Baum zu finden, der die Wahrscheinlichkeit maximiert. Da es $\frac{(2n-5)!}{2^{n-3}(n-3)!}$ ungewurzelte Bäume mit n Taxa gibt, ist es nicht möglich, die Wahrscheinlichkeit für alle möglichen Bäume zu berechnen. Deswegen hat Joseph Felsenstein 1981 eine Vereinfachung des Algorithmus mithilfe einer Heuristik entwickelt (Felsenstein, 1981).

3.3.4. Supernetzwerke

Während eines Vergleichs ganzer Genome werden mehrere Tausend Proteinfamilien verglichen und phylogenetische Bäume für diese berechnet. Für die Auswertung dieser Bäume kann zum einen betrachtet werden, welche Organismen häufiger in einer Gruppe beobachtet werden als andere. Diese Daten sind jedoch nicht einfach darstellbar. Ein anderer Ansatz ist die Erstellung eines Consensusbaums (Adams III, 1972; Margush und McMorris, 1981), eines Superbaums (Binindа-Emonds et al., 1999; Creevey und McInerney, 2005; Semple und Steel, 2000) oder eines Supernetzwerks (Holland et al., 2007, 2008; Huson und Bryant, 2006; Whitfield et al., 2008).

Bei der Erstellung eines Consensusbaums wird versucht herauszufinden, welche Phylogenie von einem Großteil der Daten unterstützt wird. Allerdings müssen in jedem Baum, der zu dieser Berechnung verwendet wird, dieselben Taxa präsent sein. Dann wird für jede mögliche Kante über eine 50 % Mehrheitsregel entschieden, ob sie gut unterstützt ist, oder nur in einem Teil der Daten vorkommt. Ein Superbaum kann berechnet werden, wenn in unterschiedlichen Bäumen auch unterschiedliche Organismen vorkommen können. Auch kann die Anzahl der Taxa variieren. Die einfachste Methode ist die Erstellung einer Supermatrix, in die alle Gruppen aller Bäume mit 0 und 1 kodiert eingetragen werden. Aus dieser Matrix

wird über ein Mehrheitskriterium ein Baum berechnet. Hierbei gehen allerdings viele Informationen, zum Beispiel über widersprüchliche Signale in den Daten, verloren, die nur in einem Supernetzwerk angezeigt werden können.

In einem Netzwerk kann es mehr als eine Verbindung (engl. *split*) von einem Taxon zu einem anderen geben. Dies wird graphisch durch mehrere parallele Linien anstatt der einen Linie eines Astes realisiert. Hat ein Netzwerk eine Baumstruktur, also sind sehr wenige parallele Linien vorhanden, so gibt es wenig Konflikte in den Daten. Anderenfalls gibt ein Teil der Daten ein anderes phylogenetisches Signal als ein anderer Teil der Daten.

3.4. Zielsetzung

Die heutigen photosynthetischen Eukaryoten erhielten ihre Fähigkeit zur Photosynthese, indem sie entweder eine Endosymbiose mit einem Cyanobakterium (primäre Plastide, Algen und Pflanzen) oder mit einem anderen photosynthetischen Eukaryoten (sekundäre Plastide, Protisten) eingingen. Im Laufe der Evolution wurden viele Gene des Endosymbionten in das Genom des jeweiligen Wirts übertragen oder gingen verloren. Im Falle der primären Endosymbiose wurde endosymbiontischer Gentransfer von dem Genom des Cyanobakteriums in das Kerngenom des Eukaryoten und in das Genom der Mitochondrien nachgewiesen. Der Gentransfer bei der Etablierung der sekundären Endosymbiose ist komplizierter: Gene wurden aus dem Kerngenom des Endosymbionten (Grünalge oder Rotalge) und auch aus dem Plastidengenom in das Kerngenom des Wirts übertragen.

Für *Arabidopsis thaliana* wurde in früheren Studien ein Anteil von 12,7 % beziehungsweise 18 % berechnet, für *Cyanidioschyzon merolae* ein Anteil von 17,1 %. Werte für *Oryza sativa* (13,6 %), *Chlamydomonas reinhardtii* (14,2 %) und *Cyanophora paradoxa* (11 %) wurden ebenfalls in anderen Analysen abgeleitet (Archibald, 2006; Deusch et al., 2008; Martin et al., 2002; Reyes-Prieto et al., 2006). Darüberhinaus zeigten frühere Analysen, dass die Fähigkeit zur Stickstofffixierung des Cyanobakteriums eine grundlegende Rolle bei der Etablierung der Endosymbiose gespielt haben könnte (Deusch et al., 2008).

Die Chromalveolatentheorie von Cavalier-Smith besagt, dass die sekundären Plastiden in den Cryptophyten, Haptophyten, Heterokonta, und Alveolata (Dinoflagellaten, Apicomplexa und Ciliaten) aus einer Endosymbiose mit einer Rotalge hervorgingen (Cavalier-Smith, 1999). Phylogenetische Analysen der plastidären Proteine und der Aufbau der Plastiden können diese Hypothese zumindest teilweise bestätigen (Iida et al., 2007; Khan et al., 2007; Li et al., 2006; Nosenko et al., 2006; Oudot-Le Secq et al., 2007; Sanchez-Puerta et al., 2007). In einigen Analysen wurden jedoch viele Gene in dem Kerngenom der Protisten identifiziert, die aus Grünalgen stammen könnten (Moustafa et al., 2009). Diese Erkenntnisse werden in Dagan und Martin (2009) diskutiert.

Vor diesem Hintergrund war es das Ziel der vorliegenden Arbeit, durch einen Vergleich von sieben vollständig sequenzierten Kerngenomen von Algen und

Pflanzen mit den Genomen von 40 vollständig sequenzierten Cyanobakterien diese Gene zu identifizieren und abzuschätzen, in welchem Maße Gentransfer bei der Etablierung der primären Endosymbiose stattgefunden hat. Ferner sollte aus der phylogenetischen Analyse dieser Gene dasjenige Cyanobakterium identifiziert werden, das dem freilebenden Vorfahren der primären Plastiden am Ähnlichsten ist.

Darüber hinaus sollten diejenigen Gene identifiziert und quantifiziert werden, die in dem Kerngenom der elf Repräsentanten der Chromalveolaten und Excavata präsent sind und aus Grün- oder Rotalgen stammen. Auf diese Weise kann der Vorfahr des sekundären Endosymbionten näher charakterisiert und die Chromalveolatentheorie getesten werden.

4 Material und Methoden

4.1. Daten

Für diese Analyse wurden 2.825.466 Translationen der kernkodierten offenen Leseraster von 710 vollständig sequenzierten Genomen aus den in Abbildung 4.1 aufgeführten Organismengruppen verwendet. Im Folgenden werden diese als Proteine bezeichnet. Alle Bakterienproteine und die Proteine der Tiere, Pilze, *Phythophtora infestans* und *Oryza sativa* wurden im Februar 2008 von dem FTP - Server des NCBI[1] heruntergeladen. Die Sequenzen von *Thalassiosira pseudonana* wurden in der Version 1.0 aus der JGI[2] Datenbank extrahiert, ebenso wie die Daten für *Populus trichocarpa* in der Version 1.1, *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Chlamydomonas reinhardtii* in der Version 3.1, *Phaedactylum tricornutum* in der version 2.0 und *Phytophtora ramorum* in der Version 1.0 und *Phytophtora sojae* in der Version 1.1. Das Proteom von *Arabidopsis thaliana* wurde von der TAIR[3] Internetseite erhalten. Die Proteine von *Cyanidioschyzon merolae* wurden von der Homepage des Genomprojektes[4] heruntergeladen. Proteine für *Trichomonas vaginalis*, *Entamoeba histolytica* und *Tetrahymena thermophila* wurden aus der TIGR Datenbank extrahiert. Für die Protisten *Plasmodium falciparum*[5], *Cryptosporidium parvum*[6] und *Dictyostelium tricornutum*[7] existieren eigene Genomprojekte, von denen die Daten erhalten wurden. Eine detaillierte Aufstellung der einzelnen Organismen und der Anzahl der Proteine pro Genom kann in Tabelle A.1 gefunden werden.

1 http://www.ncbi.nlm.nih.gov/
2 http://www.jgi.doe.gov/
3 http://www.arabidopsis.org/
4 http://merolae.biol.s.u-tokyo.ac.jp/
5 http://plasmodb.org/plasmo/
6 http://cryptodb.org/cryptodb/
7 http://dictybase.org/

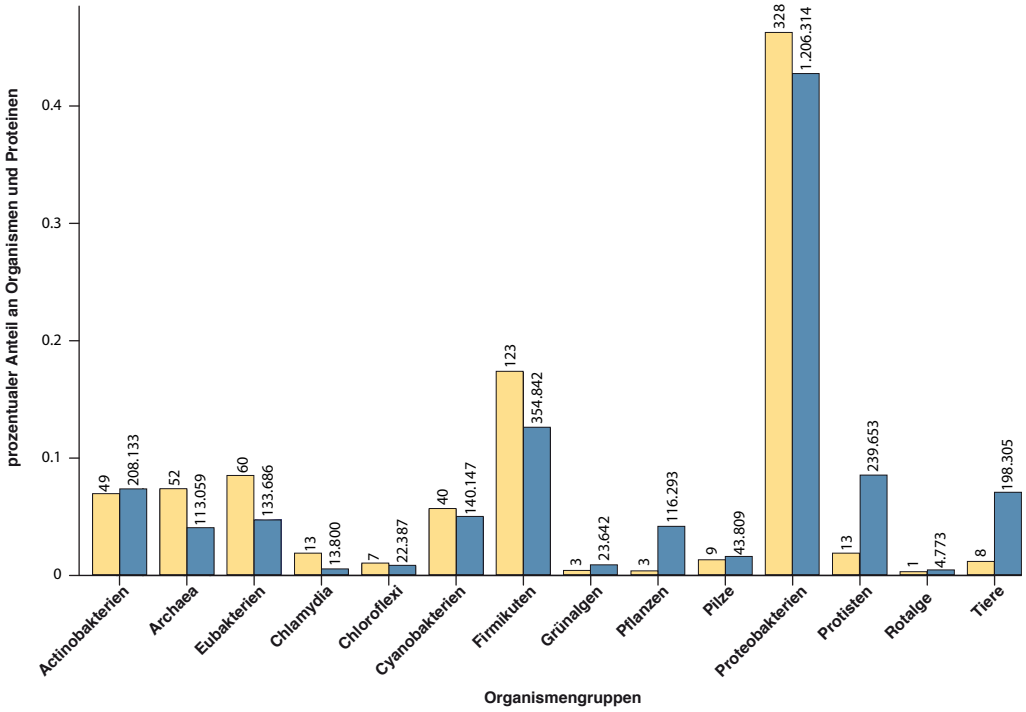

Abbildung 4.1.: Zusammensetzung der Datenbank. Die gelben Balken zeigen den prozentualen Anteil der jeweiligen Organismengruppe an der Anzahl der Organismen (710). Die blauen Balken symbolisieren den Anteil an der Gesamtzahl der Proteine (2.825.466). Über den Balken sind die absoluten Zahlen vermerkt.

4.2. Verwendete Computer

Aufgrund der großen Anzahl an Proteinen und notwendigen Berechnungen wurden mehrere Computer und Großrechner verwendet: Der Großteil der Berechnungen wurde auf dem Opteron-Cluster des Zentrums für Informations- und Medientechnologie (ZIM) durchgeführt. Er besteht aus drei Servern mit jeweils vier Dual-Core-Prozessoren, wodurch bis zu 24 Prozesse gleichzeitig gestartet werden konnten. Zusätzlich dazu wurden die Server "Jukes" und "Horst" des Instituts für Ökologische Pflanzenphysiologie für das Clustering mit `mcl` verwendet, da nur auf diesen Rechnern 32 GB Arbeitsspeicher zur Verfügung standen. Bei den Servern handelt es sich um das Modell ProLiant DL360 der Firma Hewlett-Packard. Beide sind mit zwei Quad-Core-Prozessoren in der Lage 8 Prozesse gleichzeitig auszuführen. Ein MacBook Pro mit dem Betriebssystem MacOS X

wurde als Arbeitsplatzrechner verwendet. Für die Erstellung von Abbildungen wurde das Statistikprogramm R und der Adobe Illustrator verwendet und für Textkalkulationen das Programm Excel aus dem Microsoft Office Paket.

4.3. Programme und Algorithmen

4.3.1. Perl und Bioperl

Für die Bearbeitung der Dateien und Durchführung vieler Stapelaufgaben, bei denen für viele Dateien derselbe Programmaufruf ausgeführt werden musste, wurden Perl[8]-Skripte geschrieben. Die verwendete Perl-Version ist 5.8.8. Außerdem wurden zusätzlich dazu Bioperl[9]-Module in der Version 1.6 verwendet, mit denen zum Beispiel ein phylogenetischer Baum im Newick-Format gelesen und analysiert werden kann.

4.3.2. R

Die Statistiksoftware `R`[10] wurde verwendet, um die Daten zu analysieren und Abbildungen zu erstellen. `R` ist kostenfrei erhältlich, umfasst jedoch dieselben Funktionen wie eine Bezahlsoftware. Durch die ausschließliche Bedienung über die Kommandozeile kann das Paket große Datenmengen gut verwalten und Berechnungen können im Stapelbetrieb ausgeführt werden.

4.3.3. vi und Textwrangler

Die Perlskripte und andere Dateien wurden mit dem Kommandozeilen basierten Texteditor `vi`[11] geschrieben. Mithilfe dieses Editors kann zum Beispiel mit regulären Ausdrücken gesucht und ersetzt werden. Sein größter Vorteil liegt darin, dass er keine GUI besitzt und auf jedem Server verwendet werden kann. Auf dem lokalen Rechner wurde der Texteditor `Textwrangler`[12] verwendet. Auch hier können reguläre Ausdrücke verwendet und auch Ersetzungen in mehrere Dateien gleichzeitig durchgeführt werden.

8 www.perl.org/
9 http://www.bioperl.org/
10 http://www.r-project.org/
11 http://macvim.org/OSX/index.php
12 http://www.barebones.com/products/TextWrangler/

4.3.4. MySQL

Die Datenbanksoftware MySQL[13] kann große Datenmengen effizient speichern und gibt so die Möglichkeit, schnelle Abfragen durchzuführen. In dieser Arbeit wurden sowohl die Informationen zu den einzelnen Clustern als auch der Proteinindex und die nächsten Nachbarn der Proteine gespeichert. Außerdem wurden mehrere Parameter der Alignments und Bäume in diese Datenbank integriert, so dass für jeden Cluster ein Datensatz dieser Parameter einfach abgefragt werden konnte.

4.3.5. BLAST

Das ***B**asic **L**ocal **A**lignment **T**ool* (BLAST) (Altschul et al., 1990, 1997) wurde entwickelt, um Sequenzähnlichkeiten einer Sequenz zu einer großen Sequenzdatenbank zu bestimmen. Da BLAST eine Heuristik verwendet und keine globalen sondern lokale Alignments erstellt, ist es sehr schnell. Der Algorithmus, mit dem ähnliche Sequenzen entdeckt werden, besteht aus mehreren Schritten. Als erstes wird eine Tabelle mit den in der Suchsequenz vorkommenden Wörtern der Länge 3 und deren Nachbarwörtern erstellt. Zusätzlich dazu wird zu jedem Wort ein Punktwert berechnet, indem die Austausche zu dem Originalwert anhand einer Matrix (z.B. BLOSUM62) bewertet werden.

Je wahrscheinlicher der Austausch zwischen zwei Aminosäuren ist, desto höher ist die Wertung. Wird zum Beispiel das Wort HEP in der Suchsequenz gefunden, so hat HEP selber unter Verwendung der BLOSUM62 Matrix die Wertung 20 ($H \rightarrow H = 8$, $E \rightarrow E = 5$, $P \rightarrow P = 7$), das Nachbarwort HEG 11. In der Datenbank werden nur diejenigen Wörter gesucht, deren Wertung über einem bestimmten Schwellenwert T liegt. Der Standardwert für T ist 13.

Im nächsten Schritt wird versucht, die gefundenen Treffer in beide Richtungen so weit wie möglich zu verlängern. Hierbei wird zunächst nach Treffern gesucht, die dicht beieinanderliegen und in vielen Fällen zu einem Treffer vereint werden können. Durch diese später eingeführte "*Two-Hit*-Methode" wurde eine deutliche Zeitersparnis erzielt. Das Einfügen von Lücken (engl. *gaps*) ist erlaubt, dies wird aber mit negativen Wertungen bestraft. Kann ein Treffer nicht mehr verändert

13 http://www.mysql.de/

(verkürzt oder verlängert) werden ohne die Wertung zu verschlechtern, wird dieser Schritt beendet. Jeder Treffer mit einer Wertung über dem Schwellenwert S wird als sogenanntes "*high scoring pair*" (HSP) bezeichnet.

Für jedes HSP wird außerdem die Signifikanz (Erwartungswert E) berechnet. Diese gibt an, mit welcher Wahrscheinlichkeit der Treffer durch Zufall in einer Datenbank dieser Größe gefunden werden könnte. Eine Eingrenzung der Ausgabe kann dadurch erzielt werden, dass man für den Erwartungswert einen Schwellenwert angibt. Dann werden nur diejenigen HSPs ausgegeben, deren Erwartungswerte über diesem Schwellenwert liegen.

4.3.6. Alignments

Der erste Schritt bei einem Vergleich zweier Proteinsequenzen ist die Berechnung eines globalen Alignments. Anhand einer Bewertungsfunktion wird außerdem das optimale also das Alignment mit der höchsten Wertung ermittelt. Für diese Bewertungsfunktion werden verschiedene Substitutionsmatrizen verwendet, wie zum Beispiel die BLOSUM62 Matrix. Werden Lücken eingefügt, die Insertionen oder Deletionen darstellen, so wird dies mit Punktabzügen bestraft.

4.3.6.1. `needle` - Paarweise globale Alignments

Ein Programm, das ein Alignment von zwei Sequenzen berechnet ist `needle`, das Teil des Programmpakets EMBOSS (Rice et al., 2000) ist. Es berechnet ein optimales, paarweises, globales Alignment von zwei Sequenzen nach dem Needleman-Wunsch-Algorithmus (Needleman und Wunsch, 1970). Die Menge aller möglichen globalen Alignments zweier Sequenzen der Länge N ist sehr groß:

$$\frac{2^{2N}}{\sqrt{\pi N}}$$

Für zwei Sequenzen der Länge 100 sind dies $\sim 10^{58}$ Möglichkeiten. Deswegen wurde der Needleman-Wunsch-Algorithmus entwickelt, der diese Menge an Möglichkeiten reduziert, indem er dynamische Programmierung verwendet und so das Gesamtproblem in viele kleine, schnell zu lösende Probleme unterteilt. Hierfür wird eine zweidimensionale Matrix ausgefüllt, bei der die Spalten für die Buchstaben der einen Sequenz und die Zeilen für die Buchstaben der anderen Sequenz stehen. Jeweils eine Zeile und Spalte wird für Lücken hinzugefügt. Dann

wird die Matrix rekursiv nach folgender Formel ausgefüllt:

$$D(i,j) = max \begin{cases} D(i-1,j-1) + S(x_i,y_j) \\ D(i-1,j) + g \\ D(i,j-1) + g \end{cases} \tag{4.1}$$

Hierbei ist i die jeweilige Zeile, j die Spalte, $S(x_i,y_j)$ der Punktwert aus der Substitutionsmatrix, der die Wahrscheinlichkeit angibt, dass die Aminosäure x_i in y_j mutiert. Dieser Wert kann aus der verwendeten Substitutionsmatrix abgelesen werden. g bezeichnet die Strafpunkte, die für das Einfügen (engl. *gap creation penalty*) und das Erweitern (engl. *gap extension penalty*) einer Lücke verrechnet werden. Setzt man diese beiden Werte auf 10 beziehungsweise 1, erhält man im Initialisierungsschritt folgende Punktematrix, in der vermerkt wird, ob das Maximum in $D(i-1,j-1)$, $D(i-1,j)$ oder $D(i,j-1)$ stand.

		W	E	L	T
	0	-10	-11	-12	-13
W	-10				
E	-11				
R	-12				
T	-13				
E	-14				

D(1,1), der erste Wert der Tabelle wird auf 0 gesetzt, in der ersten Zeile und der ersten Spalte werden die Strafpunkte für Lücken addiert. Dabei fallen nur das erste Mal die 10 Strafpunkte an, dann wird die Lücke nur noch verlängert. Somit ist der erste Wert, der berechnet werden muss D(2,2). Hier ergibt sich:

$$D(2,2) = max \begin{cases} D(1,1) + S(W,W) = 0 + 11 = 11 \\ D(1,2) + g = -10 - 1 = -11 \\ D(2,1) + g = -10 - 1 = -11 \end{cases}$$

In diesem Fall wird der Wert aus $C(1,1)$ übernommen und der Wert für die Substitution von W nach W aus der BLOSUM62 Matrix addiert. An dieser Stelle eine Lücke einzufügen gäbe einen niedrigeren Wert. Auf diese Weise wird die komplette Matrix Zeile für Zeile ausgefüllt, was in folgender Matrix resultiert:

		W	E	L	T
	0	-10	-11	-12	-13
W	-10	**11**	1	-9	-14
E	-11	1	**16**	15	14
R	-12	-9	15	**14**	14
T	-13	-14	14	14	**19**
E	-14	-15	13	13	**18**

Ausgehend von dieser Tabelle wird nun das Alignment rückwärts aufgebaut. Als erstes wird also das Feld in der letzten Zeile und der letzten Spalte betrachtet. Das verwendete Maximum kam in diesem Fall von dem oberen Feld, das heißt es wird das E mit einer Lücke aligniert. Die restlichen Maxima wurden aus der Diagonalen erhalten, somit werden die Buchstaben immer untereinander geschrieben. Mit dieser Methode wird das Alignment

$$\begin{array}{l} WELT- \\ WERTE \end{array}$$

mit einer Wertung von 18 erstellt.

4.3.6.2. MUSCLE - multiple Alignments

Das Programm MUSCLE (Edgar, 2004*a*,*b*) berechnet aus multiplen Fasta-Dateien ein progressives, multiples Alignment. Hierfür wird in drei Schritten erst ein progressives Alignment berechnet, das dann horizontal verbessert wird.
Für die Erstellung des Alignments müssen zuerst die paarweisen Ähnlichkeiten – die evolutionären Distanzen – zwischen den verwendeten Sequenzen berechnet werden. Dies geschieht im ersten Schritt, indem die Anzahl der identischen k-Tupel in zwei Sequenzen gezählt werden:

$$F = \sum_{\tau} \frac{min(n_X(\tau), n_Y(\tau))}{min(L_X, L_Y) - k + 1}$$

Wobei τ ein bestimmtes k-Tupel, L_X und L_Y die Längen der Sequenzen X und Y sind. $n_X(\tau)$ und $n_Y(\tau)$ geben an, wie oft das k-Tupel in der jeweiligen Sequenz auftritt. Da hierfür kein paarweises Alignment berechnet werden muss, ist diese Methode schneller als die Berechnung der paarweisen Identität zwischen zwei

Sequenzen. Aus diesen Ähnlichkeiten wird dann mit

$$d_{kmer} = 1 - F$$

die Distanz berechnet und eine Distanzmatrix erstellt. Aus dieser Distanzmatrix wird mithilfe des *single-* oder *average-linkage*-Clusteringverfahrens (vgl. Abschnitt 3.2.2) ein Initialbaum berechnet. Mithilfe dieses Baums wird dann ein erstes progressives Alignment berechnet. In den folgenden Schritten des Algorithmus wird durch die Verbesserung des Initialbaumes versucht, dieses Alignment zu optimieren.

Als erstes wird die prozentuale paarweise Identität D zwischen allen Sequenzen aus dem erstellten multiplen Alignment berechnet. Da bei weiter entfernten Sequenzen auch Rückmutationen berücksichtigt werden müssen, wird hier die Kimura-Distanz (Kimura, 1983) verwendet:

$$d_{Kimura} = -log_e(1 - D - \frac{D^2}{5})$$

Aus diesen Distanzen wird wiederum eine Distanzmatrix und ein Baum berechnet. Als nächstes werden die beiden erstellten Bäume verglichen, indem diejenigen internen Knoten identifiziert werden, für die sich die Topologie geändert hat. Für diese Knoten werden neue partielle Alignments erstellt, während die partiellen Alignments der Knoten, die in identischen Teilen der Bäume vorkommen, erhalten bleiben. Dieser Schritt wid so lange wiederholt bis entweder die beiden Bäume identisch sind oder die Anzahl der unterschiedlichen Knoten in der nächsten Iteration nicht mehr verringert werden kann.

Auch der dritte Schritt wird iterativ angewendet, wobei eine maximale Anzahl an Iterationen beim Start des Programms angegeben werden kann. Hier wird versucht, das Alignment zu verbessern, indem der Baum partitioniert und das Alignment für jede Partition neu berechnet wird (Hirosawa et al., 1995). Ausgehend von den Kanten, die am weitesten von der Wurzel entfernt liegen, wird in jeder Iteration ein anderer Ast aus dem Baum entfernt, wodurch jedes Mal zwei Teilbäume entstehen. Für diese Teilbäume wird das entsprechende Profil aus dem multiplen Alignment extrahiert und Positionen, die nur Lücken enthalten, entfernt. Dann werden die beiden Profile miteinander aligniert, wobei Spalten innerhalb der Profile erhalten

bleiben. Lücken werden ebenfalls als komplette Spalten in ein Profil eingefügt. Das neue Alignment wird nun mit dem ursprünglichen verglichen, indem der *sum-of-pairs* (SP) Wert berechnet wird. Der SP-Wert ist die Summe der Substitutionswerte $S(i,j)$ für zwei Aminosäuren i und j über alle Sequenzen s und alle Spalten x des Alignments:

$$SP = \sum_{x} \sum_{s} \sum_{t>s} S(s[x], t[x])$$

Je ähnlicher sich die Aminosäuren in einer Spalte des Alignments sind, desto größer ist der SP-Wert. Zusätzlich dazu werden Strafpunkte für Lücken eingeführt. Ist der SP-Wert des neuen Alignments höher als der des Alten, so wird es durch das Neue ersetzt, anderenfalls wird das Alte behalten.

4.3.7. Die *heads-or-tails*-Analyse

Die *heads-or-tails*-Methode (HoT) wurde von Giddy Landan und Dan Graur (Landan und Graur, 2007) entwickelt, um die Qualität beziehungsweise die Akkuratheit eines Alignments aufgrund seiner Reproduzierbarkeit zu bestimmen. Bei dieser Methode wird das normale Alignment mit demjenigen verglichen, das aus den invertierten Sequenzen mit denselben Einstellungen erstellt wurde. In der Theorie sollten sich diese beiden Alignments, nachdem das Rückwärtsalignment wieder umgedreht wurde, nicht unterscheiden. Das Ausmaß der Unterschiede zwischen diesen Alignments kann dafür verwendet werden, die Verlässlichkeit des Alignments und des daraus berechneten Baumes zu bestimmen. Hierfür wurden zum Beispiel in Deusch et al. (2008) zwei Parameter bestimmt. Zum einen der *sum-of-pairs* Wert (SPS) und zum anderen der *column score* (CS).

Der SPS geht auf (Thompson et al., 1999a) zurück. Er basiert auf der Analyse der alignierten Paare in dem Vorwärts- und Rückwärtsalignment. Sind alle Aminosäurepaare, die in dem Vorwärtsalignment aligniert sind, ebenfalls in dem Rückwärtsalignment in einer Spalte zu finden, so ist der Wert des SPS 100 %, das heißt 100 % aller Aminosäurepaare sind in beiden Alignments identisch aligniert. Ein solches Alignment wird als verlässlich eingestuft. Formal ist der SPS folgendermaßen definiert:

Definition 4.1 ***Der SPS***

Sei N die Anzahl der Taxa in einem Alignment, M die Anzahl der Spalten und $A_{i1},\ldots,A_{iN}$ *die Aminosäuren in Spalte i. Dann gilt:*

$p_{ijk} = 1$, *falls* A_{ij} *und* A_{ik} *in beiden Alignments in einer Spalte auftreten, und* 0 *sonst.*

Diese Werte werden für jede Spalte in dem Alignment berechnet:

$$S_i = \sum_{j=1, j\neq k}^{N} \sum_{k=1}^{N} p_{ijk}$$

Der Gesamtwert für das Alignment ergibt sich nun aus der Summe dieser Einzelwerte, die zusätzlich normiert wird, indem durch die doppelte Anzahl aller möglichen Aminosäurepaare geteilt wird.

$$SPS = \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{M_r} S_{ri}}$$

Der zweite Parameter, der CS (Thompson et al., 1999a), beruht auf derselben Idee, betrachtet jedoch nicht die einzelnen Aminosäurepaare sondern die Spalten selber. Finden sich alle Spalten des Vorwärtsalignments ebenfalls in dem Rückwärtsalignment, so ist der *CS* gleich 1 oder 100 %:

Definition 4.2 ***Der CS***

$$CS = \frac{1}{M}\sum_{i=1}^{M} C_i \quad mit \quad \begin{cases} C_i = 1, & \textit{wenn die Spalte i des Vowärtsalignments} \\ & \textit{ebenfalls im Rückwärtsalignment zu finden ist,} \\ C_i = 0, & \textit{sonst.} \end{cases}$$

Diese Werte beschreiben jedoch die Alignments, in denen Lücken enthalten sind. Für die Rekonstruktion der phylogenetischen Bäume werden diese aus den Alignments entfernt, da sie die Distanz zwischen den Proteinen verfälschen können. Für Alignments ohne Lücken ist deswegen eher interessant, ob das Vorwärts- und das Rückwärtsalignment nach dem entfernen der Lücken dieselben Informationen enthalten. Dies ist der Fall, wenn dieselben Spalten in den Ausgangsalignments keine Lücken enthalten und somit für die phylogenetische Analyse nicht entfernt werden. Aus diesem Grund wurde der nCS (für *nogap column score*) entwickelt, der beschreibt, wieviele gleiche Spalten aus den beiden Alignments extrahiert wurden.

Definition 4.3 ***Der nCS***

$$nCS = \frac{e}{c_1} + \frac{e}{c_2}$$

e entspricht der Anzahl der Spalten, die sowohl in dem Vorwärtsalignment ohne Lücken als auch in dem Rückwärtsalignment ohne Lücken vorkommen und somit dieselben Informationen liefern. c_1 *ist die Anzahl der aus dem Vorwärtsalignment extrahierten Spalten,* c_2 *ist die Anzahl der aus dem Rückwärtsalignment extrahierten Spalten. Sind alle Spalten, die in dem einen Alignment vorhanden sind, ebenfalls in dem anderen vorhanden, so hat der nCS-Wert 2, beide Alignments ohne Lücken sind identisch und liefern denselben phylogenetischen Baum.*

4.3.8. `mcl` - Markov-Clustering

Wie bereits in Abschnitt 3.2.3 erläutert, versucht das Markov-Clustering Regionen innerhalb eines Graphen zu finden, die stark miteinander verbunden sind oder sogar Zusammenhangskomponenten bilden. Der genaue Algorithmus soll nun hier beschrieben werden.

Das erste Problem, das sich stellt, ist einen Graphen so zu kodieren, dass zum einen alle Informationen des Graphen enthalten sind und zum anderen Aussagen über benachbarte und stark zusammenhängende Knoten gemacht werden können. Hierfür wird im Allgemeinen die sogenannte Adjazenzmatrix (zwei Knoten heißen adjazent oder benachbart, wenn sie durch eine Kante verbunden sind) verwendet.

Definition 4.4 ***Die Adjazenzmatrix eines Graphen***
Ein gerichteter oder ungerichteter Graph mit n Knoten kann in einer zugehörigen $n \times n$*-Matrix so kodiert werden, dass ein Eintrag* N_{ij} *nur dann größer als 0 ist, wenn die beiden Knoten i und j benachbart, also durch eine Kante verbunden sind. Hat die Kante eine Gewichtung, so wird dieser Wert in die Matrix eingetragen, ansonsten sind die Einträge entweder gleich* 0*, wenn die Knoten nicht benachbart sind oder gleich* 1 *sonst.*

Um mit dieser Matrix rechnen zu können muss sie noch in eine **stochastische Matrix** transformiert werden, indem die Einträge der Matrix so normalisiert werden, dass sowohl die Zeilen- als auch die Spaltensummen gleich 1 und alle Einträge positiv sind. Ausgehend von dieser zu dem Graphen G gehörenden Matrix M_G kann eine Markovmatrix wie folgt definiert werden:

Definition 4.5 ***Die Markovmatrix eines Graphen***
Sei G ein Graph mit n Knoten und M_G die zugehörige stochastische Adjazenzmatrix. So kann die Markovmatrix T_G folgendermaßen aus M_G erstellt werden:

$$T_G = M_G\, d^{-1}$$

wobei d einer Diagonalmatrix entspricht, auf deren Hauptdiagonalen die Summe der ausgehenden Kanten eines Knotens steht:

$$d_{kk} = \sum_i M_{ik} \ \wedge \ d_{ij} = 0 \ f\ddot{u}r\ i \neq j$$

Ein Eintrag T_{ij} in einer Markovmatrix T_G gibt also an, wie stark zwei Vertices miteinander verbunden sind. Allerdings muss dies immer im Zusammenhang mit den anderen Werten in der Spalte gesehen werden. Je nachdem wieviele Nachbarn ein Knoten hat, kann der Wert von T_{ij} sehr groß sein (bei wenigen Nachbarn) oder klein (bei vielen Nachbarn). Sind in einer Spalte alle Werte gleich groß, so ist ein Knoten zu allen seinen Nachbarn gleich stark verbunden, bzw. die Wahrscheinlichkeit, von diesem Knoten aus zu einem seiner Nachbarn zu gehen, ist gleich für alle Nachbarn.

Wird eine stochastische Matrix mit sich selbst multipliziert, so ist das Ergebnis ebenfalls eine stochastische Matrix. Sie wird als zweite Potenz der Matrix (M^2) bezeichnet. Ist die $i - te$ Potenz einer stochastischen Matrix gleich der Matrix, also $M^i = M^{i-1}$ so heißt die Matrix M^i idempotent und es wurde ein Gleichgewichtszustand erreicht (M^∞).

Um jedoch ein Clustering aus einem Graphen erstellen zu können, muss ein Weg gefunden werden, um gut verbundene Nachbarschaften von anderen abzugrenzen. Das heißt, dass nicht nur die Markovmatrix so lange mit sich selber multipliziert werden kann bis ein Gleichgewichtsszustand erreicht ist (**Expansion**), sondern dass auch die Verbindungen zu "starken Nachbarn" – mit einer hohen Gewichtung an der zugehörigen Kante – verbessert und zu "schwachen Nachbarn" weiter verschlechtert werden müssen. Dies wird mit der sogenannten **Inflation** und dem Inflationsparameter Γ_r erreicht.

Definition 4.6 ***Der Inflationsparameter*** Γ_r
Sei $M \in \mathbb{R}^{k \times l}$ *eine Matrix mit* $M \geq 0$ *und* r *eine reelle nicht negative Zahl.* Γ_r *ist der Inflationsparameter mit dem Koeffizienten* r *mithilfe dessen jede Spalte von* M *folgendermaßen verändert wird:*

$$(\Gamma_r M)_{pq} = \frac{(M_{pq})^r}{\sum_{i=1}^{k} (M_{iq})^r}$$

Jeder Eintrag einer Spalte wird mit dem Faktor r *potenziert und dann die Spalte wieder so normalisiert, dass die Spaltensumme gleich* 1 *ist. Gilt* $0 < r < 1$ *so wird die Homogenität in der entsprechenden Spalte erhöht, ist* $r > 1$ *so wird die Spalte inhomogener.*

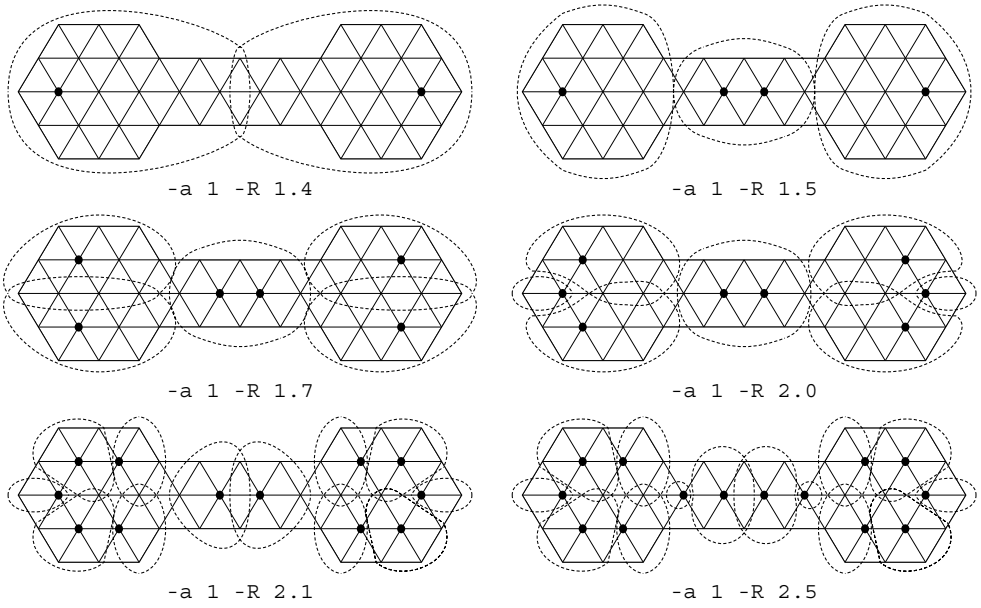

Abbildung 4.2.: Beispiel für die Auswirkungen des Inflationsparameters auf die Clustergranularität. Dargestellt ist ein Graph, der je nach Wert des Inflationsparameters (1,4-2,5) in unterschiedlich viele Cluster eingeteilt wird. Mit $R = 1{,}4$ entstehen zwei große Cluster, die bei den mittleren Vertices überlappen, schwarz markierte Vertices sind Attraktoren. Je größer der Inflationsparameter gewählt wird, desto kleiner und zahlreicher werden die Cluster. (Abbildung aus van Dongen (2000*c*))

Der Algorithmus des Programms `mcl` iteriert nun so lange über diese beiden Schritte – Expansion und Inflation – bis eine doppelt-idempotente Matrix M_{mcl}^{∞} erreicht wird, die sich weder durch das Potenzieren noch durch Inflation verändert. Aus dieser Matrix wird dann das Clustering der Vertices des Eingangsgraphen G erstellt. Angenommen, der zu clusternde Graph G verfügt über eine Cluster-

struktur, so gibt es in allen Clustern einen Vertex, der als Attraktor bezeichnet wird.

Definition 4.7 ***Der Attraktor***
Sei $G = (V,w)$ ein Graph mit N Knoten und einer zugehörigen idempotenten Matrix M. Der Vertex $a \in V$ wird Attraktor genannt, wenn die Anzahl der ausgehenden Kanten ungleich 0 ist, d.h. $M_{aa} \neq 0$. Die Menge der Nachbarn dieses Vertex wird ***Attraktorsystem*** *genannt.*

Von der Menge der Attraktoren eines Graphen ausgehend, kann nun ein überlappendes Clustering, das heißt ein Clustering, in dem ein Vertex zu mehreren Clustern gehören kann, definiert werden.

Definition 4.8 ***Das überlappende Clustering eines Graphen***
Sei $G = (V,w)$ ein Graph mit N Knoten, einer zugehörigen idempotenten Matrix M und einer Kantenrelation $\rightarrow$. V_x sei außerdem die Menge der Attraktoren in G und $E = E_1, \ldots, E_d$ die Menge der Äquivalenzklassen von $\rightarrow$ auf V_x. Dann ist eine Relation $\mathtt{v}$ auf $E \times V$ definiert mit $\mathtt{v}(E,a) = 1$ wenn $\exists b \in E$ mit $a \rightarrow b$ und $\mathtt{v}(E,a) = 0$ sonst. Ein überlappendes Clustering $CL_M = C_1, \ldots, C_d$ zu M und V hat d Elemente. Der $i-te$ Cluster C_i ist definiert als:

$$C_i = \{v \in V \mid \mathtt{v}(E_i, v) = 1\}$$

In diesem Cluster befinden sich also alle Vertices, die sich in der $i-ten$ Äquivalenzklasse befinden.

In der Praxis kommt es nur in Spezialfällen wie in Abbildung 4.2 vor, dass `mcl` ein überlappendes Clustering erstellt. Anstatt eines mathematischen Beweises soll hier nur ein Zitat des Autors von `mcl` angeführt werden: "*Mathematically speaking, this is a conjecture and not a theorem, but the present author will eat his shoe if it fails to be true (for marzipan values of shoe).*"

Zusätzlich zu dem Inflationsparameter können Parameter angegeben werden, um die Matrix nach jedem Iterationsschritt "auszudünnen". Wenn eine Spalte in der Matrix hinreichend heterogen ist, das heißt es gibt manche Vertices, die stark verbunden sind und andere für die ein vergleichsweise kleiner Wert in der Spalte steht, so können diese Werte, die nach verschiedenen Kriterien ausgewählt werden, auf 0 gesetzt werden. Die verschiedenen Parameter sind:

-p Der Schwellenwert: Alle Werte in einer Spalte, die kleiner als dieser Wert sind werden auf 0 gesetzt.

-P Die Genauigkeit: $P = \frac{1}{p}$ Es bleiben P Werte in der Spalte übrig (default: 4000).

-R/-pct Die Wiederherstellung / der Wiederherstellungsprozentsatz: Sind in einer Spalte weniger als P Werte übrig und die Summe aller Einträge ist kleiner als $pct/100$, so versucht mcl die größten R Werte wiederherzustellen (default: 600).

-S Die Selektion: Sind in einer Spalte mehr als P Werte übrig, so wird versucht die Anzahl der Einträge in der Spalte auf höchstens S Einträge zu reduzieren.

4.3.9. Evolutionäre Distanzen

Aus den berechneten multiplen Alignments sollen Phylogenien abgeleitet werden. Hierfür gibt es viele Algorithmen und Methoden (siehe Abschnitt 3.3.2). Aufgrund der großen Anzahl an Proteinfamilien in dieser Arbeit wurde der gesamte Datensatz mithilfe einer Distanzmethode analysiert. Hierbei wird aus den alignierten Distanzen eine Distanzmatrix berechnet, die alle paarweisen Distanzen zwischen den Sequenzen enthält. Das Programm `protdist` aus dem PHYLIP-Paket wurde für diese Berechnungen verwendet (Felsenstein, 2005).

4.3.9.1. `protdist`

Das Programm `protdist` benötigt als Eingabedatei ein Alignment im PHYLIP-Format. Da es Lücken in einer Sequenz, also das Zeichen "–" als eine zusätzliche Aminosäure zählt und sie somit in die Berechnung der Distanz eingehen, müssen alle Positionen, die in mindestens einer Sequenz eine Lücke aufweist, aus dem Alignment entfernt werden.

Der Algorithmus für die Berechnung der Distanz ist ein *Maximum-Likelihood*-Schätzer, das heißt es wird diejenige Distanz geschätzt, die die Wahrscheinlichkeit, gegeben ein Modell und die Daten, maximiert. Für Aminosäuren können mehrere Modelle gewählt werden: Die Kimura-Distanz (Kimura, 1980), die Dayhoff-Distanz (Schwartz und Dayhoff, 1979), die JTT-Distanz (Jones et al., 1992), das PMB-Modell (Smith et al., 2004) und eine Distanz basierend auf dem Kategorien-Modell. Da die

JTT Matrix auf einer aktuelleren und größeren Datenbank beruht, wird gemeinhin diese verwendet. Für die Berechnung wird folgende Formel herangezogen:

Definition 4.9 ***Die Maximum-Likelihood-Distanz***
Sei das evolutionäre Modell $P(x|y,t)$ gegeben, so ist die Maximum-Likelihood-Distanz (ML) zwischen zwei Sequenzen x_i und x_j der Länge N definiert als

$$d_{ij}^{ML} = max \prod_{u=1}^{N} P(x_i^u | x_j^u, t)$$

Hierbei ist $P(x_i^u|x_j^u,t)$ die Übergangswahrscheinlichkeit von der Aminosäure an der Position u in Spezies i zu der Aminosäure an der Position u in Spezies j. Diese Wahrscheinlichkeiten können aus der Matrix $P(T)$ des gewählten Modells ermittelt werden.

4.3.10. Phylogenetische Bäume

4.3.10.1. `neighbor` – *neighbor-joining*

Die *neighbor-joining*-Methode (Saitou und Nei, 1987) zur Erstellung von phylogenetischen Bäumen basiert auf dem in Abschnitt 3.2.2 beschriebenen hierarchischen *single-linkage*-Clusteringverfahren und ist in dem Programm `neighbor` des PHYLIP-Pakets implementiert. Hierbei werden so lange diejenigen Taxa beziehungsweise Cluster miteinander vereint, die die niedrigste Distanz aufweisen (also nächste Nachbarn sind), bis ein Cluster entstanden ist, in dem alle Taxa enthalten sind.

Für die Erstellung der Distanzmatrix, in der alle paarweisen Distanzen zwischen den Sequenzen gespeichert werden, können verschiedene Distanzen verwendet werden. Ausgehend von dieser Matrix wird für jedes Taxon i zu jedem anderen Taxon j die durchschnittliche Distanz nach folgender Formel berechnet:

$$r_i = \frac{1}{N-2} \sum_{k=1}^{N} d_{i,k}$$

Hierbei bezeichnet N die Anzahl der Taxa, die in der Distanzmatrix vertreten sind. Als nächstes wird aus diesen Durchschnittswerten und den ursprünglichen paarweisen Distanzen eine Zwischenmatrix M berechnet:

$$M_{i,j} = d_{i,j} - (r_i + r_j)$$

Diese Matrix wird als Grundlage für das Clustering verwendet. Da es sich um das *single-linkage*-Clusteringverfahren handelt, werden nun diejenigen Cluster zu einem größeren Cluster u(i,j) vereint, die die kleinste Distanz aufweisen. Die Kantenlängen v der Taxa zu dem neu erstellten inneren Knoten werden folgendermaßen berechnet:

$$v_{i,u} = \frac{d_{i,j} + r_i + r_j}{2} \quad und \quad v_{j,u} = d_{i,j} - v_{i,u}$$

Der innere Knoten wird nun ebenfalls in die Distanzmatrix eingetragen und die Distanzen von diesem zu allen anderen Knoten neu berechnet:

$$d_{u,k} = \frac{d_{i,k} + d_{j,k} - d_{ij}}{2}$$

Als nächstes muss sowohl die mittlere Distanz r_u neu erstellt werden als auch die Matrix M. Dieser Algorithmus wird so lange wiederholt, bis alle Taxa in einem Cluster vereint sind.

4.4. Arbeitsablauf

4.4.1. Bestimmung der bidirektionalen besten BLAST-Treffer

Zu Beginn der Analyse wurden die Datenbanken in ein einheitliches Format gebracht und für einen schnelleren Zugriff indiziert. Jedem Protein wurde eine Zahl zugewiesen, wodurch Arbeitsspeicher in den folgenden Bearbeitungsschritten gespart werden konnte. Für jeden Organismus wurde eine multiple Fasta-Datei erstellt, in der in einer Zeile das Erkennungszeichen “>“ und danach die eindeutige Nummer, die zu dem Protein gehört, und in der nächsten die Proteinsequenz selber steht (Abbildung 4.3).

```
>1
MNCQKLKGVTELRKGCSSESSLKAASRHFDLFLEEVILSGTIGGKEKKQLFPIRSVNNSTDSNNIDNSKR
>2
MSSNRFAILDDDDTAPAVKKDSKPAKAAVEASKPDDRRRPNQNDRNTKFGRGGRAPSRDGKRAYDRRSGT
```

Abbildung 4.3.: Beispiel einer multiplen Fasta-Datei.

Danach wurde eine Suche mit dem lokal installierten BLAST-Programm (Abschnitt 4.3.5) durchgeführt, bei der jedes der 710 Genome gegen jedes andere verglichen und für jeden Vergleich eine Datei erstellt wurde. Ein Genom mit sich

selber wurde nicht verglichen. Da aber für die Ermittlung des bidirektionalen besten BLAST-Treffers (BBH), sowohl Organismus 1 mit Organismus 2 als auch Organismus 2 mit Organismus 1 verglichen werden musste, wurden 503.390 Dateien erstellt.

Ein Erwartungswert-Schwellenwert wurde nicht gesetzt, da im Weiteren nur die bidirektionalen besten BLAST-Treffer verwendet wurden, das heißt wenn für ein Protein A aus Organismus 1 ein Protein B aus Organismus 2 als bester BLAST-Treffer gefunden wurde, so muss auch für Protein B das Protein A als bester Treffer gefunden werden, um die beiden Proteine als Homologe identifizieren zu können. Alle auf diese Weise ermittelten 161.708.991 Proteinpaare wurden in einer Datei zur weiteren Analyse gespeichert.

4.4.2. Bestimmung der globalen Identitäten

Da BLAST nur lokale Alignments erstellt, bei der sich teilweise nur sehr kleine Bereiche, sogenannte Domänen der Proteine ähneln, wurde anschließend für jedes BBH-Paar ein globales Alignment mit dem Programm `needle` (Abschnitt 4.3.6.1) und den voreingestellten Standardparametern erstellt und die paarweise globale Identität berechnet.

Da das Programm `needle` aus dem EMBOSS-Paket als Eingabe zwei Sequenzen einliest, die einzeln in jeweils einer Datei vorliegen müssen, und deswegen aufgrund der vielen Festplattenzugriffe sehr langsam ist, wurde eine von Mayo Röttger veränderte Version des Programms (`powerneedle`) verwendet. Dadurch konnte die gesamte Datenbank und die vorher erstellte Datei mit den BBHs in den Arbeitsspeicher geladen und die langsamen Festplattenzugriffe minimiert werden.

Für jedes in der Datei gespeicherte Sequenzpaar wurde sowohl die paarweise Identität als auch die paarweise Ähnlichkeit berechnet und in einer Datei gespeichert. Die ersten drei Spalten dieser Datei (Protein1 Protein2 Identität) können als Eingabe für das Clusteringprogramm `mcl` verwendet werden. Außerdem wurden die Werte in eine mysql-Tabelle importiert.

geneid1	geneid2	identity
158333236	75910091	20.5
158333238	75908625	22
158333241	75907911	25
158333242	75907910	90
158333244	75909735	89
158333247	75910072	87
158333255	75909870	11.5
158333256	75906668	45
158333261	75906758	46

Abbildung 4.4.: Beispiel einer Eingabedatei für `mcl`. In der ersten Spalte steht der Index für das erste Protein, in der zweiten der Index des zugehörigen bidirektionalen Treffers. Die dritte Spalte beinhaltet die paarweise Identität.

4.4.3. Clusteranalyse

Die Gesamtheit der berechneten Identitäten kann als Graph mit 161.708.991 Kanten und 2.825.466 Knoten interpretiert werden, wobei die Proteine die Knoten und die Identitäten die Kanten zwischen den Knoten darstellen. Diesen Graphen kann `mcl` (Abschnitt 4.3.8) verarbeiten und in Proteinfamilien clustern.

Es wurden folgende Optionen und Parameter verwendet: Die Option `--abc` wurde verwendet, damit `mcl` keinen weiteren Index von den Knotennamen erstellt sondern die Proteinnamen so übernimmt wie sie in der Eingabedatei stehen. Die Genauigkeit (P) wurde auf 20. 000 gesetzt. Das heißt 20.000 Werte bleiben pro Spalte nach dem Ausdünnen erhalten. Der Parameter S wurde auf 5.000 gesetzt und R auf 1.600. Außerdem wurde als Wiederherstellungsprozentsatz 90 % gewählt.

Als Ausgabe gibt `mcl` zum einen eine Datei mit den berechneten Qualitätsmerkmalen des Clusterings und der Anzahl der berechneten Cluster aus. Abbildung 4.5 zeigt die Ausgabe für das Clustering mit den Identitäten größer als 20 %. Je näher die Werte bei 100 liegen, desto besser ist das Clustering. Zum anderen erstellt `mcl` eine Datei, die alle berechneten Cluster enthält. Dabei repräsentiert jede Zeile einen Cluster derart, dass die Proteine, die zu diesem Cluster gehören, durch Tabstopps getrennt aufgelistet werden. Ein Cluster entspricht einer Proteinfamilie von der im weiteren Alignments und phylogenetische Bäume erstellt werden sollen.

```
[mcl] jury pruning marks (worst 100 cases): <92,91,96>, out of 100
[mcl] jury pruning synopsis: <92.2 or ripping> (cf -scheme, -do log)
[mcl] output is in 710_clustering_20
[mcl] 176758 clusters found
```

Abbildung 4.5.: Beispiel einer Ausgabe von `mcl` mit Qualitätsmerkmalen

4.4.4. Erstellung der multiplen Alignments

Für die Berechnung eines multiplen Alignments musste für jeden Cluster, der mehr als drei Proteine enthält (90.354), eine multiple Fasta-Datei erstellt werden. Hierfür wurde ein Perlskript geschrieben, dass die Liste mit den zu einem Cluster gehörenden Proteinen einliest, die entsprechenden Sequenzen aus der Sequenzdatenbank extrahiert und in eine Datei im Fasta-Format schreibt.

Diese Datei kann von dem Programm `muscle` (Abschnitt 4.3.6.2) gelesen werden. Die Berechnung des Alignments erfolgte unter Verwendung der Standardparameter, das heißt es wurden maximal 16 Iterationen durchgeführt und die Ausgabe des Alignments erfolgte im Fasta-Format. Anschließend wurde ein Perlskript verwendet, um Positionen, in denen in mindestens einer Sequenz Lücken auftreten, aus dem Alignment zu löschen. Aus den verbleibenden Positionen wurde eine Distanzmatrix erstellt. Außerdem wurden Alignmentmerkmale wie die durchschnittliche paarweise Sequenzidentität und die Länge des Alignments bestimmt und in einer mysql-Tabelle für spätere Analysen gespeichert.

4.4.4.1. Qualität der Alignments

Eine Methode, um die Qualität eines Alignments zu bestimmen ist die HoT-Methode (Abschnitt 4.3.7). Bei dieser Methode wird das Alignment mit dem Alignment verglichen, das aus den invertierten Sequenzen mit den gleichen Parametern erstellt wurde. Hierfür wurden mit einem Perlskript alle Sequenzen in allen multiplen Fasta-Dateien invertiert und neue Fasta-Dateien erstellt. Dann wurde von allen invertierten Fasta-Dateien mit den gleichen Einstellungen wie sie auch für die Vorwärtsalignments verwendet wurden, Alignments berechnet und mit einem von Mayo Röttger erstellten C-Programm die SPS- und CS-Werte und mithilfe eines Perlskripts die nCS-Werte der Alignments berechnet und in einer mysql-Tabelle gespeichert.

4.4.5. Erstellung der Bäume

Für die Berechnung der phylogenetischen Bäume wurde mit dem Programm `protdist` (Abschnitt 4.3.9.1) eine Distanzmethode verwendet. Hierfür wurden die Alignments verwendet, aus denen die Positionen, in denen sich Lücken befinden, entfernt wurden. Da manche Alignments keine Positionen ohne Lücken aufwiesen oder nicht berechnet werden konnten, konnte nur noch mit 90.354 Alignments weitergearbeitet werden. Für jedes Alignment wurde eine Distanzmatrix unter Verwendung der JTT-Matrix berechnet.

```
 4
25894     0.000000  2.018293  2.191488  2.074053
1095      2.018293  0.000000  1.831470  1.858253
98292     2.191488  1.831470  0.000000  1.532907
68386     2.074053  1.858253  1.532907  0.000000
```

Abbildung 4.6.: Beispiel einer Distanzmatrix, die von `protdist` aus einem Alignment mit 4 Taxa berechnet wurde.

Diese Distanzmatrix wurde dann als Eingabedatei für das Programm `neighbor` (Abschnitt 4.3.10.1) verwendet, das einerseits den berechneten phylogenetischen Baum im Newick-Format in einer Datei ausgibt und in einer anderen Datei eine graphische Darstellung des Baums und die zugehörigen Astlängen in einer Tabelle abbildet.

Für die Auswertung der Bäume konnte nicht weiter mit den indizierten Proteinen gearbeitet werden. Deswegen wurden in allen Bäumen die Indizes durch eine Kombination aus dem Phylum, dem Organismennamen und der *gi* (engl. *gene id*) ersetzt. Bei Organismen, für die eine detaillierte Analyse vorgesehen war wie zum Beispiel für die Protisten, wurde anstatt des Phylums eine Abkürzung des Organismennamens verwendet.

a)

```
Neighbor-Joining/UPGMA method version 3.68


   4 Populations

 Neighbor-joining method

 Negative branch lengths allowed


  +-------------------------1095
  !
  !      +----------------------68386
  2------1
  !      +------------------------98292
  !
  +-----------------------------------25894


remember: this is an unrooted tree!

Between        And            Length
-------        ---            ------
   2          1095            0.86519
   2             1            0.21322
   1          68386           0.74379
   1          98292           0.78912
   2          25894           1.15310
```

b)

```
(1095:0.86519,(68386:0.74379,98292:0.78912):0.21322,25894:1.15310);
```

Abbildung 4.7.: Beispiel der Ausgabedateien von `neighbor`. In Abbildung a) ist die graphische Darstellung zusammen mit der Tabelle der Astlängen abgebildet, Abbildung b) zeigt den Baum im Newick-Format

```
(crypto_Cryptosporidium_parvum_Iowa_II-66359392:0.86519,(plasmo_Plasmodium_falciparum_
3D7-124809530:0.74379,theil_Theileria_parva_strain_Muguga-71031176:0.78912):0.21322,
para_Paramecium_tetraurelia_strain_d4_2-145498148:1.15310);
```

Abbildung 4.8.: Beispiel eines Baums mit ersetzten Indizes im Newick-Format.

Insgesamt wurden 90.354 Bäume berechnet, die im weiteren auf die nächsten Nachbarn bestimmter Organismengruppen untersucht werden sollen.

4.4.5.1. Bestimmung des nächsten Nachbarn

Das Ziel der Analyse ist, die nächsten Nachbarn zum Beispiel aller Sequenzen der Archaeplastida und Chromalveolaten in den Bäumen zu untersuchen. Hierfür wurde das Bioperl Modul `Bio::Tree` verwendet. Dieses Modul stellt Funktionen

bereit, um einen Baum im Newick-Format einzulesen und auf die in dem Baum vorhandenen Taxa zuzugreifen. Ein Problem ist allerdings, dass die Bäume auf jeden Fall gewurzelt eingelesen werden. Ist ein Taxon Teil der zu untersuchenden Organismengruppe (fängt der Taxonname zum Beispiel mit "plant" an, wenn die Pflanzenproteine untersucht werden sollen), so werden erst alle Distanzen zu allen anderen Blättern berechnet und der Baum bei dem Taxon gewurzelt, das die größte Distanz zu dem Taxon aufweist.

Als nächstes wird derjenige interne Knoten bestimmt, der der direkte Vorgänger zu dem Taxon ist. Nun können alle Nachfahren dieses internen Knotens bestimmt werden. Sollen nur die nächsten Nachbarn dieser Spezies bestimmt werden, so können sowohl die Gruppen als auch die Gene in einer Datei gespeichert werden. Komplizierter wird es, wenn der nächste Nachbar einer Gruppe von Organismen gesucht werden soll. Dann wird, nachdem alle Nachfahren des Vorgängertaxons abgefragt wurden, überprüft, ob alle diese Taxa zu der speziellen Gruppe gehören.

Ist dies der Fall, so wird derjenige innere Knoten gesucht, der der Vorfahre des Vorfahren des zu untersuchenden Taxons ist. Sind alle Nachfahren dieses inneren Knotens wieder Mitglieder der Gruppe, so wird dieses Verfahren so lange rekursiv angewendet, bis mindestens ein Taxon unter den Nachfahren des Taxons ist, das nicht zu dieser Gruppe gehört. Sind nur Taxons in dem Baum, die auch zu dieser Gruppe gehören, wird die Gruppe als nächster Nachbar eingetragen.

In Abbildung 4.9 ist die Vorgehensweise für den phylogenetischen Baum des Proteins Cytochrom C6 von *Arabidopsis thaliana* dargestellt. Der Cluster enthält 36 Proteine, davon sind drei aus Pflanzen, eins aus einer Grünalge und 32 aus Cyanobakterien. In Abbildung 4.9 a) ist der ungewurzelte Baum abgebildet.

Das Protein mit der größten Distanz zu dem Protein aus *A. thaliana* ist das Protein aus *Synechococcus* sp RS9917. Als erstes wird der Baum dort gewurzelt, wodurch Abbildung 4.9 b) entsteht. Das erste HTU, das als Vorfahre des zu untersuchenden Proteins gelten kann, ist mit 1) markiert. Als einziges zusätzliches Protein, das unterhalb dieses HTUs zu finden ist, wird das Protein von *Populus trichocarpa* als nächster Nachbar identifiziert. Bei einer Analyse der **direkten** nächsten Nachbarn würde an dieser Stelle "plant" als Gruppe und Populus_trichocarpa-666852 als Protein gespeichert. Soll jedoch der nächste Nachbar eines Proteins **außerhalb der**

eigenen Gruppe identifiziert werden, so muss der Baum rekursiv bis zu dem mit 4) markierten HTU durchsucht werden. Erst an dieser Stelle sind auch Cyanobakterien in der Gruppe der Nachfahren des HTUs enthalten. Da kein einzelner nächster Nachbar identifiziert werden kann, sondern nur die Gruppe von Cyanobakterien, wird jedes dieser Proteine als nächster Nachbar angesehen und gespeichert. Als benachbarte Gruppe kann jedoch "cyano" angesehen werden. Im Folgenden wird in einem solchen Fall von einem cyanobakteriellen Signal des Proteins gesprochen.

Besteht eine Gruppe von nächsten Nachbarn aus mehreren verschiedenen Gruppen, so werden diese zu der nächsthöheren Gruppe zusammengefasst. So gehört zum Beispiel eine Mischung aus Cyano- und Archaebakterien zu der Gruppe der Bakterien. Ist eine Mischung aus Prokaryoten und Eukaryoten in dieser Gruppe, so wird der nächste Nachbar mit "universal" bezeichnet. Aufgrund der Tatsache, dass es sich bei dieser Analyse nicht um eine Sternsuche, sondern um durch Clustering identifizierte Proteinfamilien handelt, ist es möglich, dass in einer Proteinfamilie mehrere Proteine einer Spezies, sogenannte Paraloge, auftreten. Ist dies der Fall, so wird jedes Protein einzeln gezählt, da es vorkommen kann, dass in einer Spezies mehrere Proteine mit derselben Funktion, aber unterschiedlichen Ursprüngen (zum Beispiel ein Protein aus Mitochondrien und ein Protein aus den Chloroplasten) vorhanden sind.

4.4.6. Erstellung eines Supernetzwerks

Für die Erstellung des Supernetzwerks wurden alle Splits, die in allen Bäumen vorhanden sind, ermittelt und in einer Supermatrix gespeichert. Da für die Erstellung eines Supernetzwerks nur ein Protein pro Organismus in einem Baum vorhanden sein darf, wurden diejenigen Cluster, in denen Paraloge vorkommen, auf die Verbindungen zwischen den Proteinen überprüft. Die älteren Proteine sollten mehr Verbindungen zu allen anderen Proteinen aufweisen als später entstandene Kopien.

Das Protein einer Spezies, das die meisten Verbindungen also die größte Konnektivität zu anderen Proteinen in dem Cluster aufweist, wurde behalten, alle anderen Proteine dieser Spezies wurden aus dem Cluster entfernt. Dann wurden diese Cluster neu aligniert und die Distanzmatrizen und die Bäume neu berechnet. Aus den Bäumen wurde eine Supermatrix mithilfe des Programms `consense` und eines von Oliver Deusch erstellten Perl-Skripts erstellt.

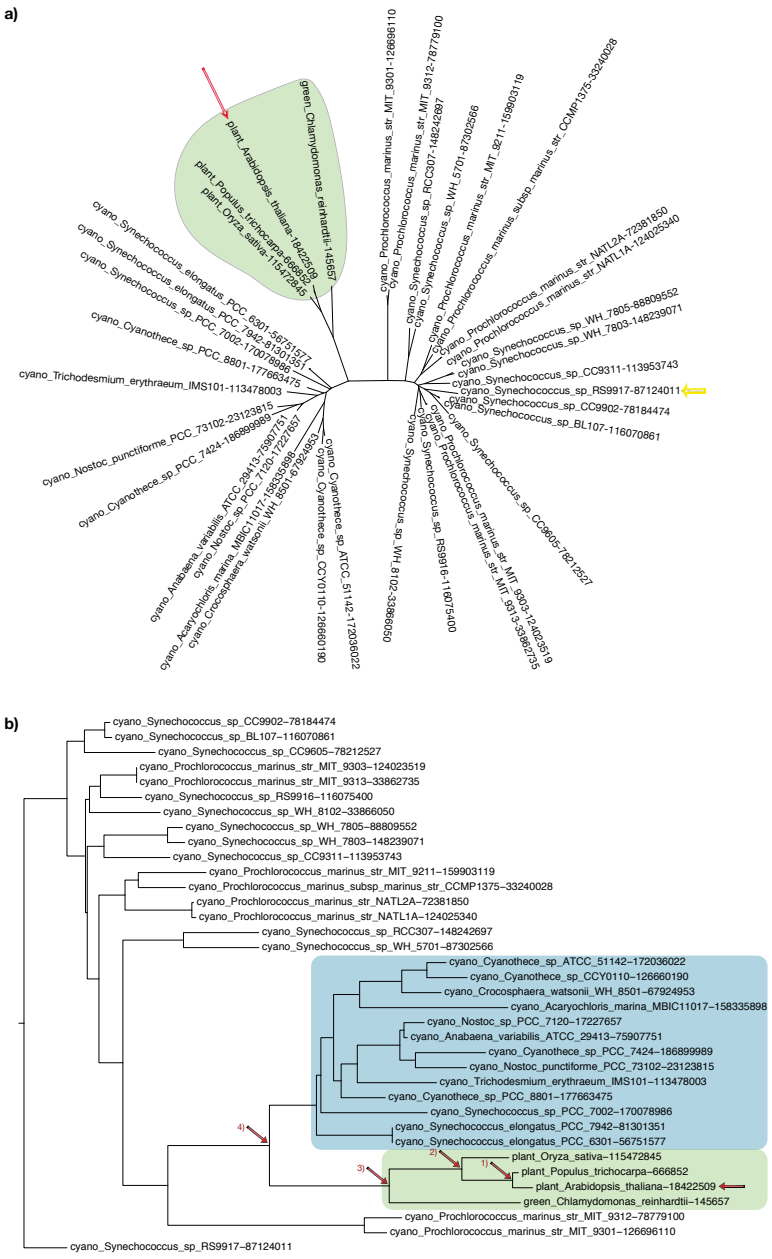

Abbildung 4.9.: Beispiel für die Identifizierung des nächsten Nachbarn eines Proteins (Cytochrom C6) von *Arabidopsis thaliana* oder einer Gruppe von Proteinen bestehend aus Pflanzen und einer Grünalge. In Abbildung a) ist der ungewurzelte Baum abgebildet, Die Position des zu untersuchenden Proteins (plant_ Arabidopsis_thaliana-18422509) ist mit einem roten Pfeil markiert. Die Position des Proteins, das die größte Distanz zu diesem aufweist (cyano_Synechococcus_sp_RS9917-87124011), ist ebenfalls mit einem gelben Pfeil markiert. Die Gruppe der Pflanzen und Grünalgen ist grün hinterlegt. In Abbildung b) ist der bei cyano_Synechococcus_sp_RS9917-87124011 gewurzelte Baum gezeigt. Die Gruppe der Pflanzen und Grünalgen ist grün hinterlegt und die Position des zu untersuchenden Proteins (plant_Arabidopsis_thaliana-18422509) mit einem roten Pfeil markiert. Mit roten Pfeilen sind die vier HTUs markiert, die man in dem Baum zurückgehen muss, um zusätzlich zu den Pflanzenproteinen andere Proteine in den Nachfahren des HTUs zu finden. Blau hinterlegt ist die Gruppe der nächsten Nachbarn zu dem Protein von *A. thaliana*.

Nach der Fertigstellung der Supermatrix wurde diese in das Nexus-Format, das als Eingabe für das Programm Splitstree (Huson und Bryant, 2006; Huson, 1998) verwendet wird, überführt. Splitstree berechnet aus dieser Charaktermatrix die paarweisen Hamming-Distanzen und die Splits zwischen den Organismen, die dann graphisch dargestellt werden können.

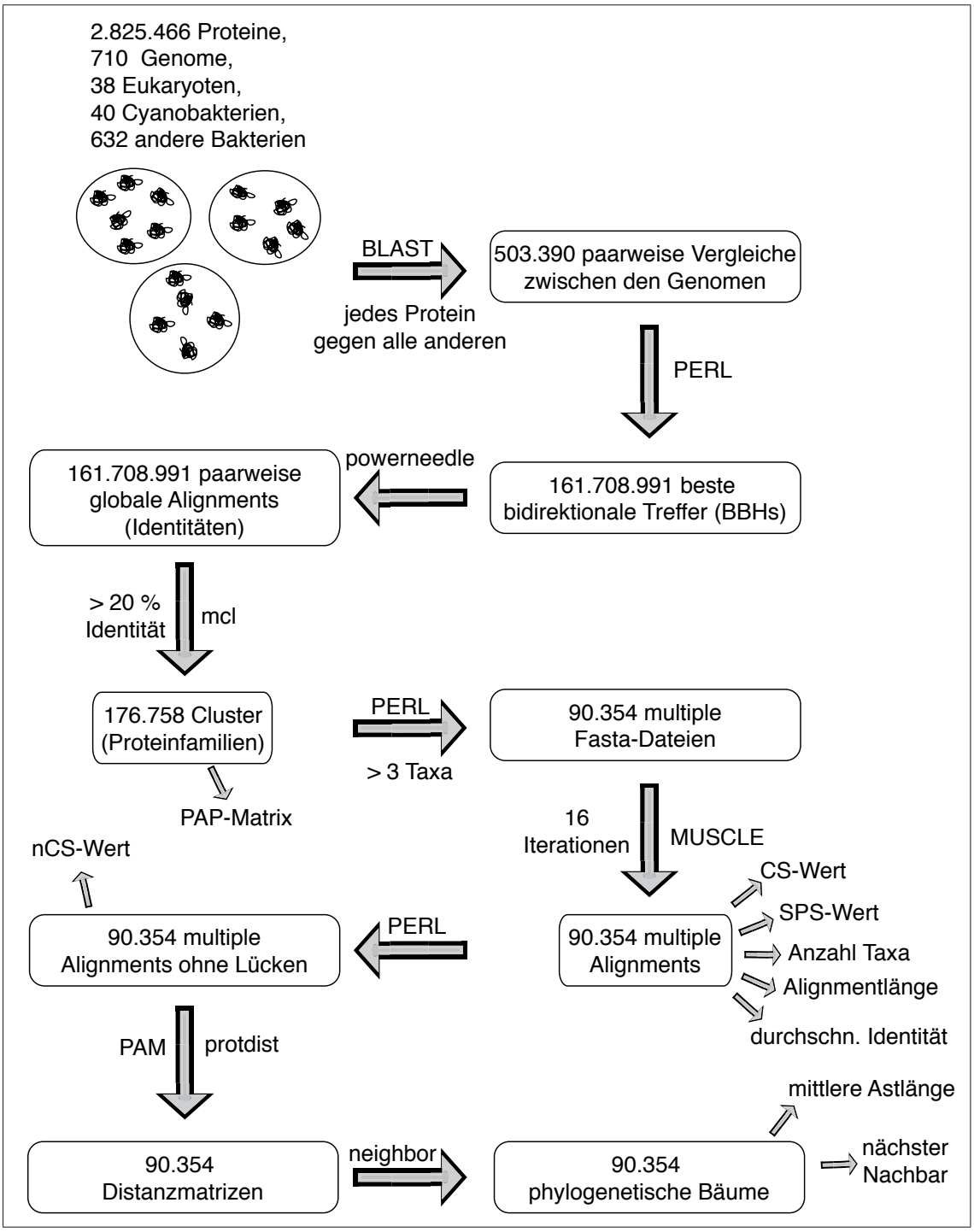

Abbildung 4.10.: Schematische Darstellung des Arbeitsablaufs der durchgeführten Analyse. Ausgehend von 2.825.466 Proteinen aus 710 vollständig sequenzierten Genomen wurde eine Homologiesuche mit BLAST durchgeführt. Die dann erhaltenen bidirektionalen besten BLAST-Treffer wurden nach der Erstellung eines globalen Alignments mit `mcl` in Cluster oder Proteinfamilien eingeteilt. Aus allen Clustern, die mehr als drei Sequenzen enthalten, wurden multiple Alignments und phylogenetische Bäume erstellt, um den nächsten Nachbarn der Proteine bestimmter Organismen zu ermitteln.

5 Ergebnisse

In dieser Arbeit sollen durch einen Vergleich der Genome der sieben in der Datenbank enthaltenen Algen und Pflanzen zu den Genomen der 40 vollständig sequenzierten Cyanobakterien der Anteil der cyanobakteriellen Gene in den Kerngenomen der Algen und Pflanzen bestimmt werden. Außerdem soll über diesen Vergleich der mögliche Vorfahre der primären Plastiden identifiziert werden. In einer weiteren Analyse sollen diejenigen Proteine in elf Protisten identifiziert werden, die ihren Ursprung in einer Grün- oder Rotalge haben. Auf diese Weise soll eine Aussage darüber getroffen werden, ob der Vorfahr der sekundären Plastide der Chromalveolaten eine Grün- oder Rotalge war.

Dazu wurden im ersten Schritt dieser Analyse das Programm BLAST (siehe Abschnitt 4.3.5 verwendet, um eine Homologiesuche von allen 2.825.466 Proteinen aus 710 vollständig sequenzierten Genomen durchzuführen. Aus dieser Analyse wurden 161.708.991 homologe Proteinpaare ermittelt, indem von allen besten Treffern pro Species nur diejenigen ausgewählt wurden, die selber auch das Protein als besten Treffer gefunden haben (siehe Abschnitt 4.4.1). 291.506 Proteine sind nicht in diesen bidirektionalen besten Treffern (BBH) enthalten. Von den gefundenen Proteinpaaren wurden globale paarweise Identitäten berechnet (siehe Abschnitt 4.4.2) und die ermittelten Werte als Eingangsgraph für `mcl` (siehe Abschnitt 4.3.8) verwendet.

5.1. Eigenschaften der Cluster

Insgesamt wurden acht Clusterings erstellt: Jeweils eins für verschiedene Identitätsschwellenwerte beginnend bei 0 % bis zu 70 %. Für jeden Schwellenwert wurden nur diejenigen BBHs verwendet, deren globale paarweise Identität größer als der jeweilige Schwellenwert ist. Für alle weiteren Analysen wurde das Clustering mit

dem Identitätsschwellenwert von 20 % verwendet, wodurch nur noch 147.401.786 Kanten in dem Eingangsgraphen vorhanden waren.

Tabelle 5.1.: Statistiken zu den verschiedenen Clusterings. Für jedes erstellte Clustering mit verschiedenen Identitätsschwellenwerten (0 % - 70 %) wurde die Anzahl der Cluster und die durchschnittliche Clustergröße berechnet. Außerdem wurde die Anzahl der in das Clustering eingehenden Proteine und die Anzahl der Cluster der Größe 1 (Waisen) ermittelt.

	Anzahl an Clustern	**Anzahl an Proteinen**	**durchschn. Größe der Cluster**	**Anzahl an Clustern der Größe 1**
0 %	214.589	2.825.466	11,81	1.978
10 %	202.190	2.384.119	12,39	1.619
20 %	176.758	2.196.887	12,69	516
30 %	195.265	2.063.623	10,69	168
40 %	230.236	1.903.249	8,33	44
50 %	260.758	1.722.414	6,64	16
60 %	279.067	1.525.124	5,46	4
70 %	282.060	1.312.593	4,68	4

In Tabelle 5.1 sind grundlegende Werte für jedes berechnete Clustering aufgeführt. Die Anzahl der erstellten Cluster betrug bei dem Clustering ohne Identitätsschwellenwert 214.589, fiel bei 10 % auf 202.190 und weiter bei 20 % auf 176.758. Ab 30 % stieg die Anzahl der Cluster wieder an bis sie 282.060 Cluster bei einem Schwellenwert von 70 % erreichte. Die Anzahl der Proteine, die in die Clusteranalyse eingingen also Teil eines BBH-Paares waren (2.533.960 ohne Schwellenwert), fiel bei jedem Clustering um 100.000 - 200.000 bis zu einem Wert von 1.312.593. Dies entspricht 46 % aller verwendeten Proteine. Die durchschnittliche Größe der Cluster betrug zwischen 4,68 Proteine und 12,69. Der Durchschnittswert war bei dem Clustering ohne Schwellenwert 11,81, stieg bis zum Clustering mit dem 20 %-Schwellenwert auf 12,69 an und fiel dann bis auf 4,68 Proteine pro Cluster ab. Die Anzahl der Waisen, also der Proteine, die zwar in mindestens einem BBH-Paar vorkamen, aber alleine in einem Cluster zu finden waren, lag bei 1.978 bei dem Clustering ohne Schwellenwert und fiel ab bis auf vier Proteine bei dem 60 %- und

70 %-Clustering. Das Clustering mit dem 20 %-Schwellenwert zeigte die wenigsten und größten Cluster.

Abbildung 5.1 zeigt für acht verschiedene Identitätsschwellenwerte die Anzahl der in das Clustering eingegangenen bidirektionalen besten BLAST-Treffer. In das Clustering ohne Schwellenwert gingen 161.708.991 Proteinpaare ein. Für das 20 %-Clustering wurden 147.401.786 Paare verwendet, für das 30 %-Clustering waren es 106.256.566 und für das 40 %-Clustering 60.446.254. Die kleinste Anzahl an BBHs (7.362.461) ging mit in das 70 %-Clustering ein.

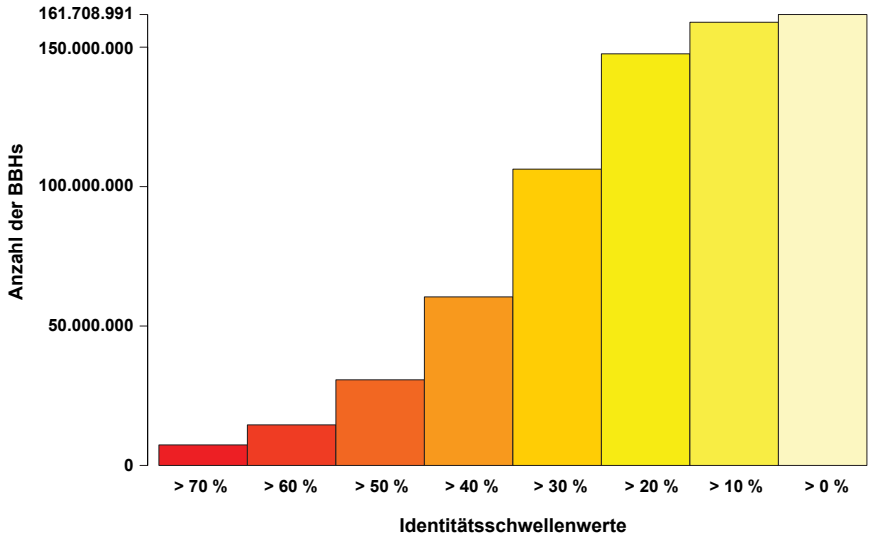

Abbildung 5.1.: Anzahl der bidirektionalen besten BLAST-Treffer (BBHs, siehe Abschnitt 4.4.1) bei verschiedenen paarweisen globalen Identitäten.

Aufgrund der Analysen der Clusteranzahl und -größe und aufgrund der Anzahl der eingehenden Proteine und BBHs wurde das 20 %-Clustering ausgewählt, um die weiteren Analyseschritte durchzuführen. Abbildung 5.2 zeigt ein Histogramm der Clustergrößen dieses Clusterings. 86.404 Cluster beinhalteten weniger als vier Proteine, weswegen sie für phylogenetische Analysen ungeeignet waren. In 59.279 Clustern waren mindestens vier aber höchstens zehn Proteine vorhanden. Dies entsprach circa 30 % aller Cluster aber nur 15 % (342.887) der Proteine. 3.394 Cluster beinhalteten mehr als 100 Proteine, der größte Cluster bestand aus 1013 Proteinen. Die durchschnittliche Clustergröße betrug, wie oben erwähnt, 12,69. Von jedem Cluster, der mehr als drei Sequenzen enthïelt (90.354), wurde jeweils eine multiple Fasta-Datei erstellt und Alignments berechnet (siehe Abschnitt 4.4.4). Außerdem wurden die Alignments der invertierten Sequenzen erstellt und als

ein Maß für die Qualität der Alignments die SPS-, CS- und nCS-Werte berechnet (siehe Abschnitt 4.3.7). Zusätzlich wurde die mittlere Identität der Sequenzen in einem Cluster ermittelt.

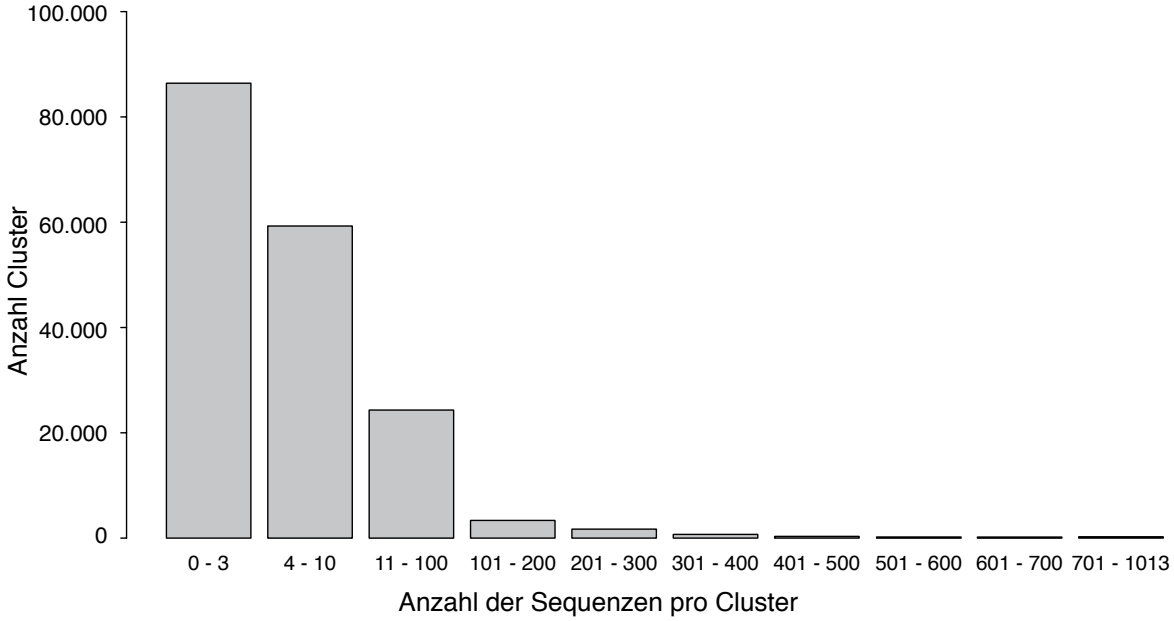

Abbildung 5.2.: Histogramm der Clustergrößen für das Clustering mit dem 20 %-Schwellenwert (siehe Abschnitt 4.4.3). Mehr als 30 % der Cluster bestehen aus 10 Proteinen oder weniger.

Für die weitere Analyse wurden alle Positionen aus den Alignments entfernt, in denen an mindestens einer Position eine Lücke auftrat. In einigen Alignments war nach diesem Schritt keine Position mehr vorhanden, so dass die phylogenetischen Distanzen von den verbleibenden 90.354 Alignments erstellt wurden. Von 90.354 Distanzmatrizen wurden phylogenetische Bäume abgeleitet. Von diesen Bäumen wurden die mittleren Astlängen berechnet (siehe Abschnitt 4.4.5). Abbildung 5.3 zeigt die Verteilung der berechneten Cluster-, Alignment- und Baumparameter.

Der Großteil der Alignments enthielt 100 bis 200 Positionen (26.407), sehr wenige waren länger als 1.000 (914) oder sogar länger als 1.500 (467) Positionen. Die mittlere Identität der Sequenzen lag bei den meisten Alignments (63.646, 80 %) zwischen 30 % und 70 %. In 3.851 Clustern waren die Sequenzen zwischen 95 % und 100 % identisch und in 59 Clustern lag die mittlere Sequenzidentität unter 20 %. Die mittlere Astlänge der Bäume variierte zwischen 0 und 1. 554 Bäume wiesen eine mittlere Astlänge auf, die größer als 1 war. In den meisten Bäumen

(70.137) war die mittlere Astlänge kleiner als 0,5. 58.191 Cluster bestanden aus maximal zehn Sequenzen, weitere 14.033 Cluster hatten zwischen 10 und 20 Sequenzen während 3.256 Cluster aus mehr als 100 Sequenzen bestanden. Fast alle (78.470, 88 %) berechneten Alignments hatten einen SPS-Wert zwischen 90 % und 100 %. Bei 7.136 Alignments lagen die Werte zwischen 80 % und 90 % und bei 2.131 Alignments zwischen 70 % und 80 %. Der CS-Wert für diese Alignments lag unter dem SPS-Wert. 52.734 (59 %) wiesen einen CS-Wert zwischen 90 % und 100 % auf, 13.120 zwischen 80 % und 90 % und 7.453 zwischen 70 % und 80 %. Für 210 Alignments wurde ein CS-Wert kleiner als 10 % berechnet.

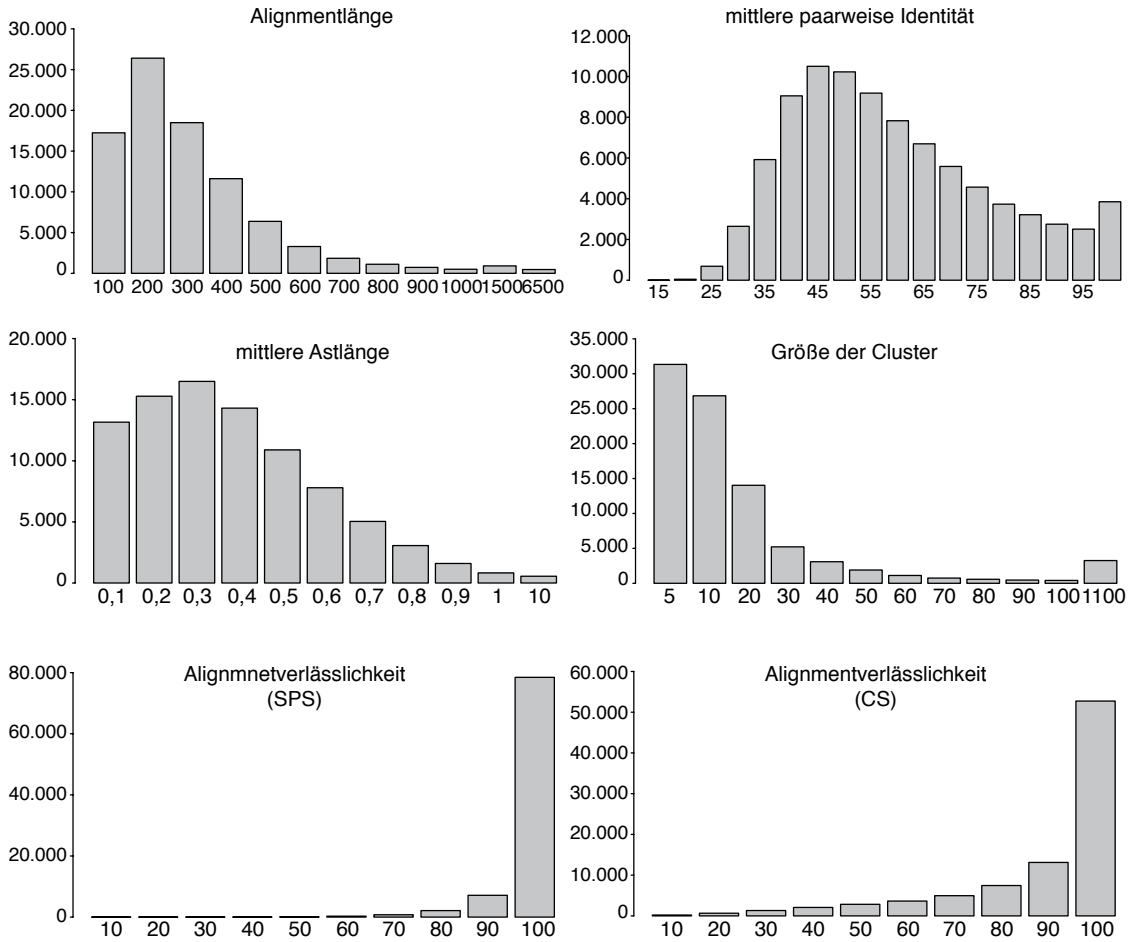

Abbildung 5.3.: Histogramme verschiedener Parameter des 20 % Clusterings. Dargestellt ist die Alignmentlänge der von den Sequenzen der Cluster berechneten Alignments (siehe Abschnitt 4.4.4), die mittlere paarweise Identität der Sequenzen in diesen Alignments, die mittlere Astlänge der aus diesen Alignments berechneten phylogenetischen Bäume (siehe Abschnitt 4.4.5), die Anzahl der in diesen Alignments vorkommenden Taxa, der SPS-Wert und der CS-Wert (siehe Abschnitt 4.3.7) der Alignments.

Abbildung 5.4 zeigt die paarweisen Korrelationen zwischen diesen Parametern. In der oberen Dreiecksmatrix sind die Korrelationen in Ellipsen abgebildet. Je mehr eine Ellipse gefüllt ist, desto stärker war die Korrelation. Ein roter Farbton bedeutet eine negative, ein blauer eine positive Korrelation. Eine intensive, dunkle Farbe deutet auf stärkere Korrelation hin, eine blasse, helle Farbe auf eine schwache Korrelation. In der unteren Dreiecksmatrix ist die Richtung der Korrelation ebenfalls durch die Schraffur in den Rechtecken deutlich gemacht. Eine Schraffur von links unten nach rechts oben bedeutet positive, eine Schraffur von links oben nach rechts unten bedeutet negative Korrelation.

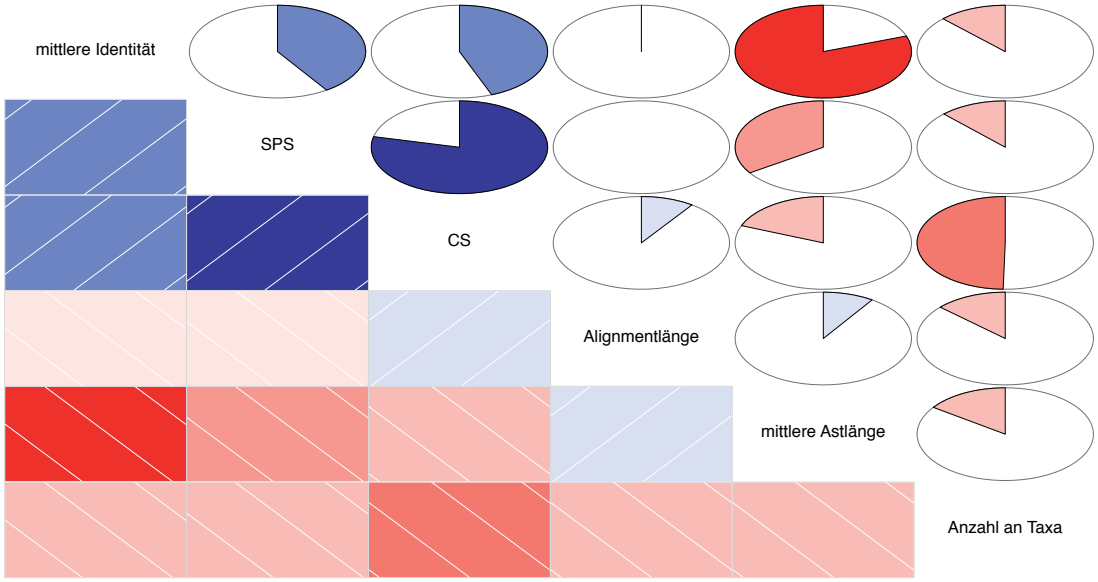

Abbildung 5.4.: Korrelogramm sechs verschiedener Parameter. Verglichen wurden die Alignmentlänge der Alignments ohne Lücken, die mittlere paarweise Identität der Sequenzen in diesen Alignments, die mittlere Astlänge der daraus berechneten Bäume, die Anzahl der in den Alignments enthaltenen Taxa, der SPS- und der CS-Wert. Die Stärke der Korrelation zwischen den Parametern ist durch die Farbkodierung abgebildet. Je intensiver die jeweilige Farbe ist, desto stärker ist die Korrelation. Ein roter Farbton bedeutet negative und ein blauer positive Korrelation. Die Richtung der Korrelation ist in der unteren Hälfte der Matrix ebenfalls durch die Schraffur in den Rechtecken in der unteren Dreiecksmatrix deutlich gemacht. Eine Schraffur von links unten nach rechts oben bedeutet positive, eine Schraffur von links oben nach rechts unten bedeutet negative Korrelation.

Der SPS- und der CS-Wert waren stark positiv korreliert ($r = 0{,}78$). Eine schwache positive Korrelation konnte zwischen diesen beiden Parametern und der mittleren paarweisen Identität beobachtet werden ($r = 0{,}41$ bzw. $r = 0{,}44$). Eine leichte negative Korrelation bestand zwischen dem CS-Wert und der Anzahl der

Taxa in einem Cluster ($r = -0{,}49$). Diese wurde jedoch nicht zu dem SPS-Wert beobachtet ($r = -0{,}12$), der eine leichte Korrelation zu der mittleren Astlänge der Bäume aufwies ($r = -0{,}34$). Zu allen anderen Parametern bestand keine Korrelation. Eine starke negative Korrelation wurde zwischen der mittleren paarweisen Identität der Alignments und der mittleren Astlänge der Bäume ermittelt ($r = -0{,}80$).

5.2. Cyanobakterielle Gene im Kerngenom der Pflanzen und Algen

Ein Ziel dieser Arbeit ist es die cyanobakteriellen Gene in den Kerngenomen der in der Analyse enthaltenen Pflanzen und Algen (*Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Cyanidioschyzon merolae*) zu identifizieren und zu quantifizieren. *A. thaliana*, *O. sativa* und *P. trichocarpa* sind Pflanzen und haben mit 31.480, 26.777 und 58.036 Proteinen die größten Genome. *Chlamydomonas reinhardtii*, *Ostreococcus tauri* und *Ostreococcus lucimarinus* sind einzellige Grünalgen. 14.598, 7.725 und 1.319 der Grünalgenproteine wurden in dieser Analyse verwendet.

Da die Anzahl der Proteine von *O. lucimarinus* ungewöhnlich klein war im Vergleich zu der sehr nah verwandten Grünalge *Ostreococcus tauri*, wurden diese Proteine von der weiteren Analyse ausgeschlossen. *Cyanidioschyzon merolae* ist eine Rotalge und 4.773 ihrer Proteine wurden verwendet. Abbildung 5.5 zeigt die Anzahl der Proteine und Cluster pro Organismus, die in das 20 %- und das Clustering ohne Identitätsschwellenwert (0 %) eingingen. Von den 31.480 Proteinen von *A. thaliana* waren 24.685 Proteine in 21.136 Clustern des 0 %-Clusterings zu finden. In dem 20 %-Clustering kamen 16.558 Proteine (53 % aller Proteine von *A. thaliana*) in 14.301 Clustern mit mehr als drei Sequenzen vor. 7.676 Proteine waren in kleineren Clustern zu finden.

Die Proteine und die zugehörigen Cluster der beiden anderen Pflanzen verhielten sich ähnlich. Für *O. sativa* waren 13.189 Proteine in 11.804 Clustern vertreten (21.004 Proteine in 18.529 Clustern für das 0 %-Clustering) und 5.532 Proteine befanden sich in kleinen Clustern. Von den 58.036 Proteinen von *P. trichocarpa* waren 20.087 Proteine in 16.868 Clustern des 20 %-Clusterings vorhanden (35.415

Proteine in 28.958 Clustern für das 0 %-Clustering) und 8.727 Proteine in kleinen Clustern. Von den Proteinen der beiden verwendeten Grünalgen *C. reinhardtii* und *O. tauri* gingen 5.522 (12.333) beziehungsweise 4.850 (7.416) Proteine in 5.107 (11.353) und 4.606 (7.048) Cluster ein. 661 und 895 Proteine waren nur in kleinen Clustern zu finden. 3.317 (4.715) Proteine der Rotalge *C. merolae* kamen in 3.134 (4.481) Clustern vor. 312 Proteine waren in Clustern mit weniger als vier Sequenzen enthalten.

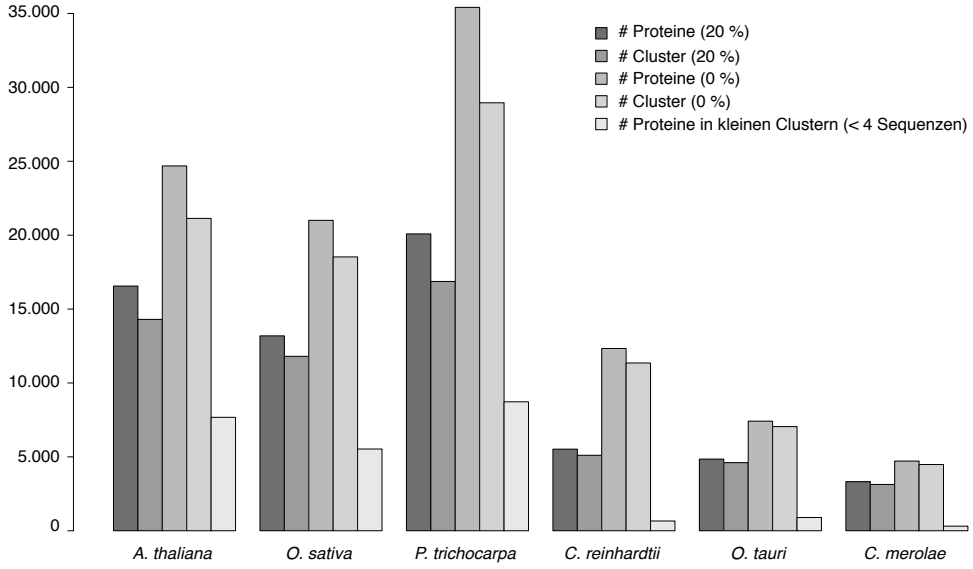

Abbildung 5.5.: Die Anzahl der verwendeten Pflanzen- und Algenproteine in verschiedenen Clusterings. Für jede Pflanze und Alge ist abgebildet, wie viele Proteine dieses Organismus in den Clustern des 20 %- und des Clusterings ohne Identitätsschwellenwert (0 %) vorkamen. Für diese Clusterings ist ebenfalls die Anzahl der Cluster, in denen ein Protein dieses Organismus vorhanden war, vermerkt. Der fünfte Balken pro Organismus zeigt die Anzahl der Proteine, die in Clustern des 20 %-Clusterings vorkamen und für die keine phylogenetischen Bäume erstellt wurden, da sie in Clustern mit weniger als vier Sequenzen enthalten waren.

Für die Analyse der Algen- und Pflanzenproteine in den Clustern mit mehr als drei Sequenzen wurden alle Cluster bestimmt, die mindestens ein Protein von mindestens einer Pflanze oder Alge enthielten (11.899). Zu jedem Algen- und Pflanzenprotein wurde dann nach der in Abschnitt 4.4.5.1 beschriebenen Methode der nächste Nachbar bestimmt. Da außer den Pflanzen und Algen noch zwei photosynthetische Protisten – *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* – in der Datenbank enthalten waren, wurden diese ebenfalls in die Gruppe der Pflanzen und Algen aufgenommen, das heißt es wurde automatisch der nächste Nachbar der Gruppe der photosynthetischen Eukaryoten bestimmt. Die beiden

Protisten besitzen eine sekundäre Plastide, weswegen sich ebenfalls cyanobakterielle Proteine in den Kerngenomen befinden können. Da Cluster, die weniger als vier Sequenzen beinhalteten nicht in diese Analyse mit eingingen und viele Bäume nicht berechnet werden konnten, wurde nur für einen Teil der Proteine ein nächster Nachbar bestimmt. Insgesamt wurden von *A. thaliana* 8.766 Proteine, von *O. sativa* 7.550, von *P. trichocarpa* 11.162, von *C. reinhardtii* 4.773, von *O. tauri* 3.895, von *C. merolae* 2.943 Proteine in Clustern mit mindestens vier Taxa betrachtet.

Alle Organismen wurden in insgesamt acht phylogenetische Gruppen eingeteilt: Archaebakterien, Cyanobakterien, sonstige Eubakterien, Pilze, Pflanzen und Algen, Proteobakterien, Protisten und Tiere. Waren Organismen aus verschiedenen Gruppen unter den nächsten Nachbarn (vergleiche Abschnitt 4.4.5.1), so wurden diese je nach Zusammensetzung in die Gruppe verschiedene Bakterien (Eubakterien und Proteobakterien), verschiedene Eukaryoten (Pflanzen, Algen, Tiere, Protisten und Pilze) und verschiedene Bakterien und Eukaryoten (mindestens ein Bakterium und ein Eukaryot) eingeteilt. Befanden sich Proteine von *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* in einer Gruppe mit Cyanobakterien, so wurde dies als cyanobakterielles Signal gewertet.

Abbildung 5.6 zeigt das Ergebnis dieser Analyse. Die wenigsten nächsten Nachbarn kamen für alle Pflanzen und Algen aus der Gruppe der Archaebakterien (0,5 % – 0,9 %), wohingegen die höchsten Werte in der Gruppe der Protisten zu finden waren (22,7 % in *P. trichocarpa* bis 30,9 % in *O. tauri*). Der Anteil der Cyanobakterien an allen Pflanzen- und Algenproteinen lag zwischen 6,2 % in *P. trichocarpa* und 9,1 % in *O. tauri*. Der Anteil der Pilze an den nächsten Nachbarn der Pflanzen und Grünalgen betrug zwischen 4,2 % und 5,2 %. In der Rotalge war er mit 7,8 % in *C. merolae* höher. Dies traf ebenfalls auf die Gruppe der gemischten Eukaryoten zu, in der die Pflanzen und Grünalgen Werte zwischen 11,5 % und 16,8 % aufwiesen, die Rotalge jedoch 19,2 %. In der Gruppe der Tiere zeigten die Pflanzen und Grünalgen mit Ausnahme von *O. tauri* einen höheren Anteil (12,1 % – 13,8 %) wohingegen die Rotalge und *O. tauri* einen Anteil zwischen 9,3 % und 10,4 % aufwiesen.

Eine starke Abstufung war in der Gruppe der Pflanzen zu finden, in die der nächste Nachbar eines Pflanzen- oder Algenproteins nur eingeordnet wurde, wenn der gesamte Cluster aus Pflanzen- und Algenproteinen bestand. Die Pflanzen

zeigten hier deutlich höhere Werte (14,4 % in *P. trichocarpa* bis 18,2 % in *O. sativa*). Die Grünalgen kamen dagegen in nur 9,1 % (*C. reinhardtii*) beziehungsweise 8,1 % (*O. tauri*) aller Fälle in reinen Pflanzen- und Algenclustern vor. Die kleinsten Anteile an diesen Clustern hatte die Rotalge mit 3,1 % (*C. merolae*). In der Gruppe der verschiedenen Bakterien (2,5 % – 3,2 %) und der verschiedenen Eukaryoten und Bakterien (5,2 % – 8,3 %) wiesen die Algen höhere Werte auf. Die Werte in der Gruppe der sonstigen Eubakterien lagen zwischen 4,0 % und 4,5 %. In der Gruppe der Proteobakterien hatte nur *P. trichocarpa* einen hohen Wert von 16,4 %, die Werte der anderen Pflanzen und Algen lagen zwischen 4,1 % und 7,5 %.

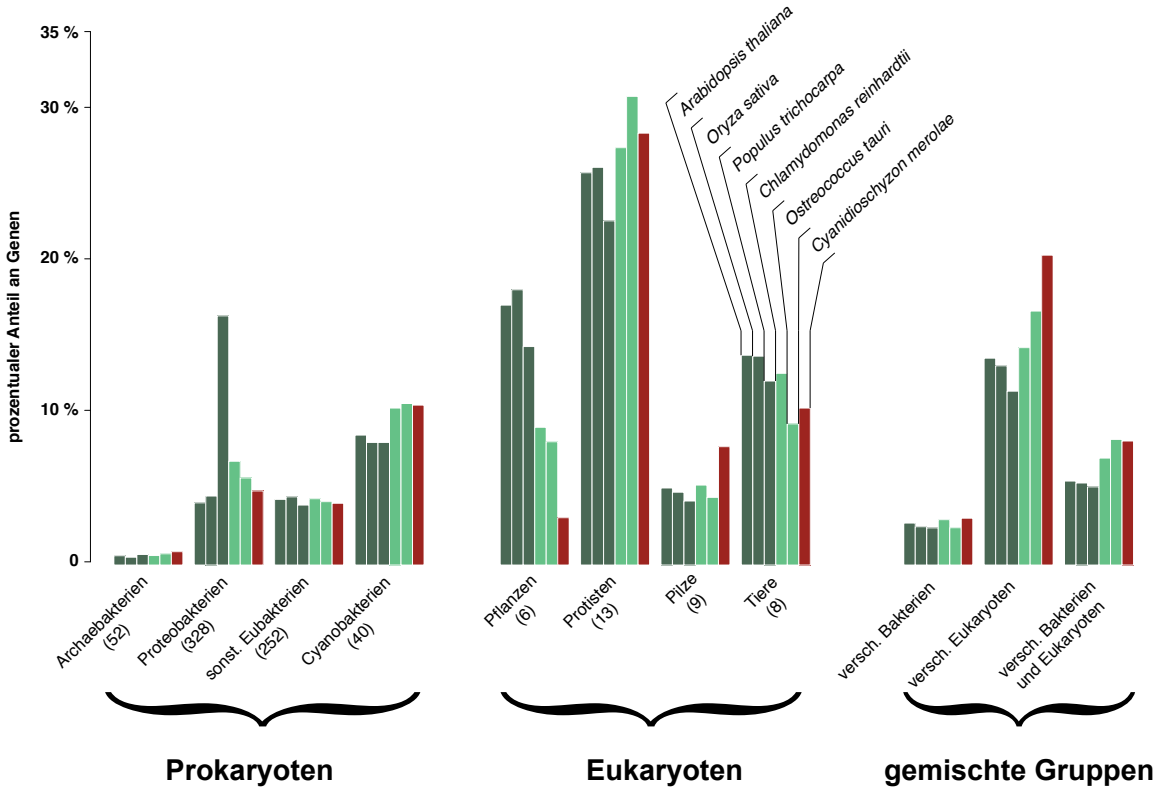

Abbildung 5.6.: Balkendiagramm der Anteile der nächsten Nachbarn für die Proteine von sechs verschiedenen Pflanzen und Algen. Insgesamt wurden von *A. thaliana* 8.766, von *O. sativa* 7.550, von *P. trichocarpa* 11.162, von *C. reinhardtii* 4.773, von *O. tauri* 3.895 und von *C. merolae* 2.943 Proteine in Clustern mit mindestens vier Taxa betrachtet.

Für jeden Cluster, in dem mindestens eine cyanobakterielle Sequenz und eine Sequenz von den Pflanzen und Algen vorkam (2.787), wurden im Weiteren diejenigen bestimmt, in denen nur Proteine aus Pflanzen, Algen und Cyanobakterien vorhanden waren.

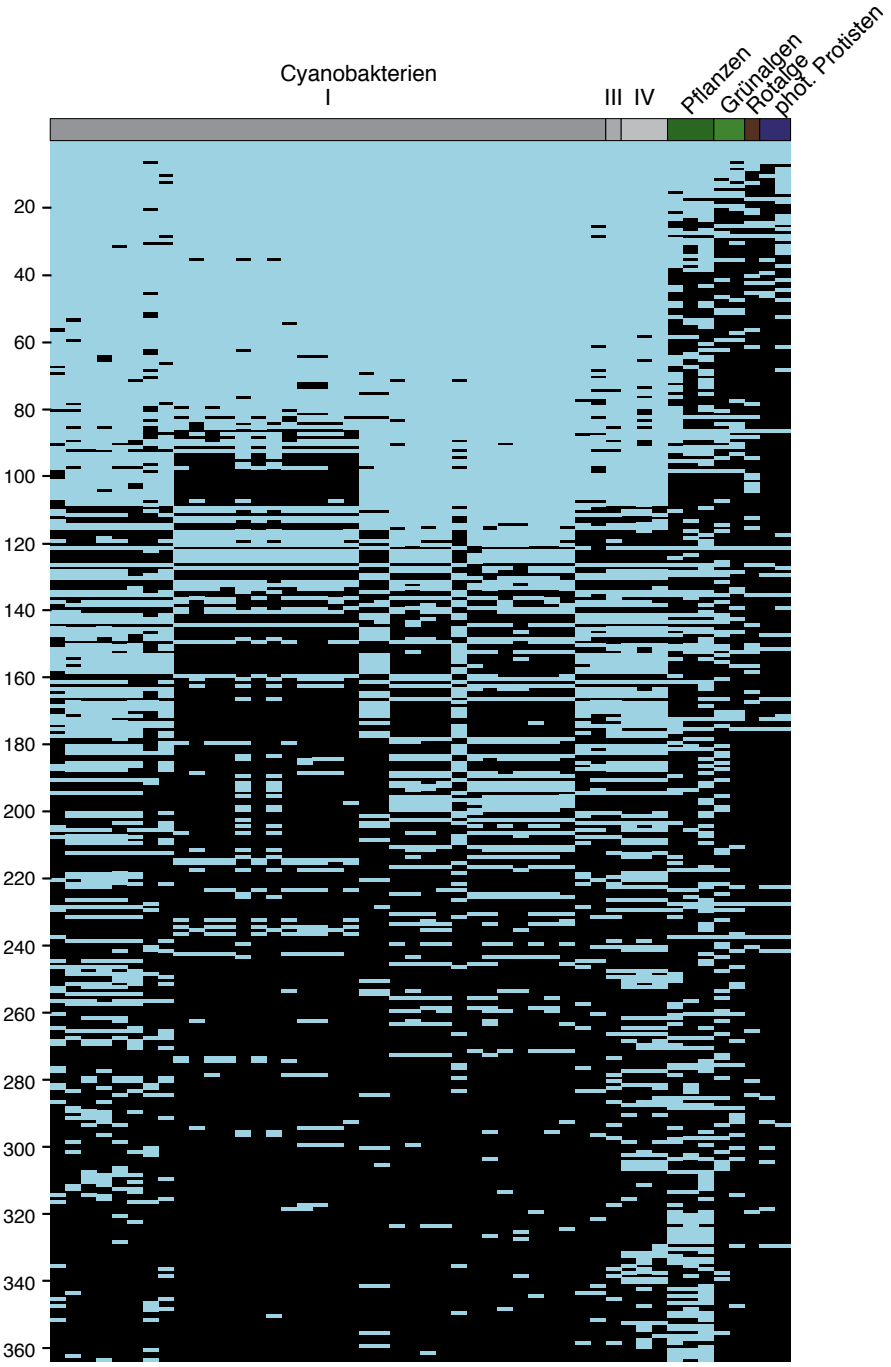

Abbildung 5.7.: Präsenz der Cyanobakterien und der photosynthetischen Eukaryoten in 364 cyanobakteriellen Clustern. Jede Zeile der Abbildung entspricht einem Cluster. Jede Spalte entspricht einem Organismus, wobei die Cyanobakterien nach ihren Gruppen (siehe Abschnitt 3.1.1) geordnet sind. Dann folgen die Pflanzen und Algen in folgender Reihenfolge: *A. thaliana*, *O. sativa*, *P. trichocarpa*, *C. reinhardtii*, *O. tauri*, *C. merolae*. Die letzten zwei Spalten symbolisieren die Präsenz der Proteine von *P. tricornutum* und *T. pseudonana*. Ein blauer Balken markiert die Präsenz eines Proteins des jeweiligen Organismus in einem Cluster.

Die beiden photosynthetischen Protisten wurden ebenfalls in diesen Clustern geduldet. Für diese Proteine war eine Auswertung der phylogenetischen Bäume unnötig, da auf jeden Fall ein cyanobakterielles Signal ermittelt worden wäre. Insgesamt waren dies 364 Cluster, wobei *A. thaliana* in 151 Clustern vorkam, *O. sativa* in 119, *P. trichocarpa* in 178, *C. reinhardtii* in 105, *O. tauri* in 70 und *C. merolae* in 52. Abbildung 5.7 zeigt die Präsenz der Cyanobakterien, Pflanzen, Algen und photosynthetischen Protisten in den gemeinsamen Clustern. Nur in sechs Clustern kamen alle 49 Organismen vor, ungefähr 90 Proteine kamen in fast allen Cyanobakterien vor. Wurde das gleichzeitige Auftreten der Cyanobakterien in einem Cluster betrachtet, so fiel auf, dass die Cyanobakterien der Gruppe IV, III und einige Cyanobakterien der Gruppe I (*Trichodesmium erythraeum*, *Cyanothece* sp. PCC 7424, *Cyanothece* sp. ATCC 51142, *Cyanothece* sp. CCY0110, *Microcystis aeruginosa* NIES 843, *Synechocystis* sp. PCC 6803, *Crocosphaera watsonii* WH 8501, *Synechococcus* sp. PCC 7002, *Synechococcus elongatus* PCC 6301, *Synechococcus elongatus* PCC 7942, *Cyanothece* sp. PCC 8801, *Thermosynechococcus elongatus* und *Gloeobacter violaceus*) in sehr vielen Clustern zusammen mit den Pflanzen und Algenproteinen auftraten.

Für alle Cluster, in denen mindestens ein Pflanzen- oder Algenprotein und ein Protein aus einem Cyanobakterium vorkamen (2.949) wurden diejenigen Cluster bestimmt, in denen in dem zugehörigen Baum, der aus dem Vorwärtsalignment berechnet wurde (im Folgenden Vorwärtsbaum genannt) oder in dem Baum, der aus dem Rückwärtsalignment berechnet wurde (im Folgenden Rückwärtsbaum genannt), mindestens ein Pflanzen- oder Algenprotein ein Cyanobakterium als nächsten Nachbar hatte.

Dies war in 991 Clustern der Fall und die jeweiligen Signale sind in Abbildung 5.8 abgebildet. Jede Zeile in dieser Abbildung entspricht einem Cluster. Die linke Hälfte der Abbildung zeigt die Signale der Vorwärtsbäume und die rechte Hälfte der Abbildung entspricht den Signalen in den Rückwärtsbäumen. Jeweils eine Spalte zeigt die Signale für die verschiedenen Pflanzen und Algen. Die mittlere Spalte zeigt die Höhe des nCS-Werts - ein Maß für die Verlässlichkeit des Alignments (siehe Abschnitt 4.3.7) - für die jeweiligen Alignments.

Nach diesem Wert sind die Cluster geordnet, so dass Cluster mit einem höheren nCS-Wert weiter oben stehen. In jeder Zeile ist pro Spalte mit einem grünen Balken markiert, ob für diese Pflanze oder Alge in diesem Cluster ein Protein vorkam, das als nächsten Nachbarn ein Cyanobakterium aufwies.

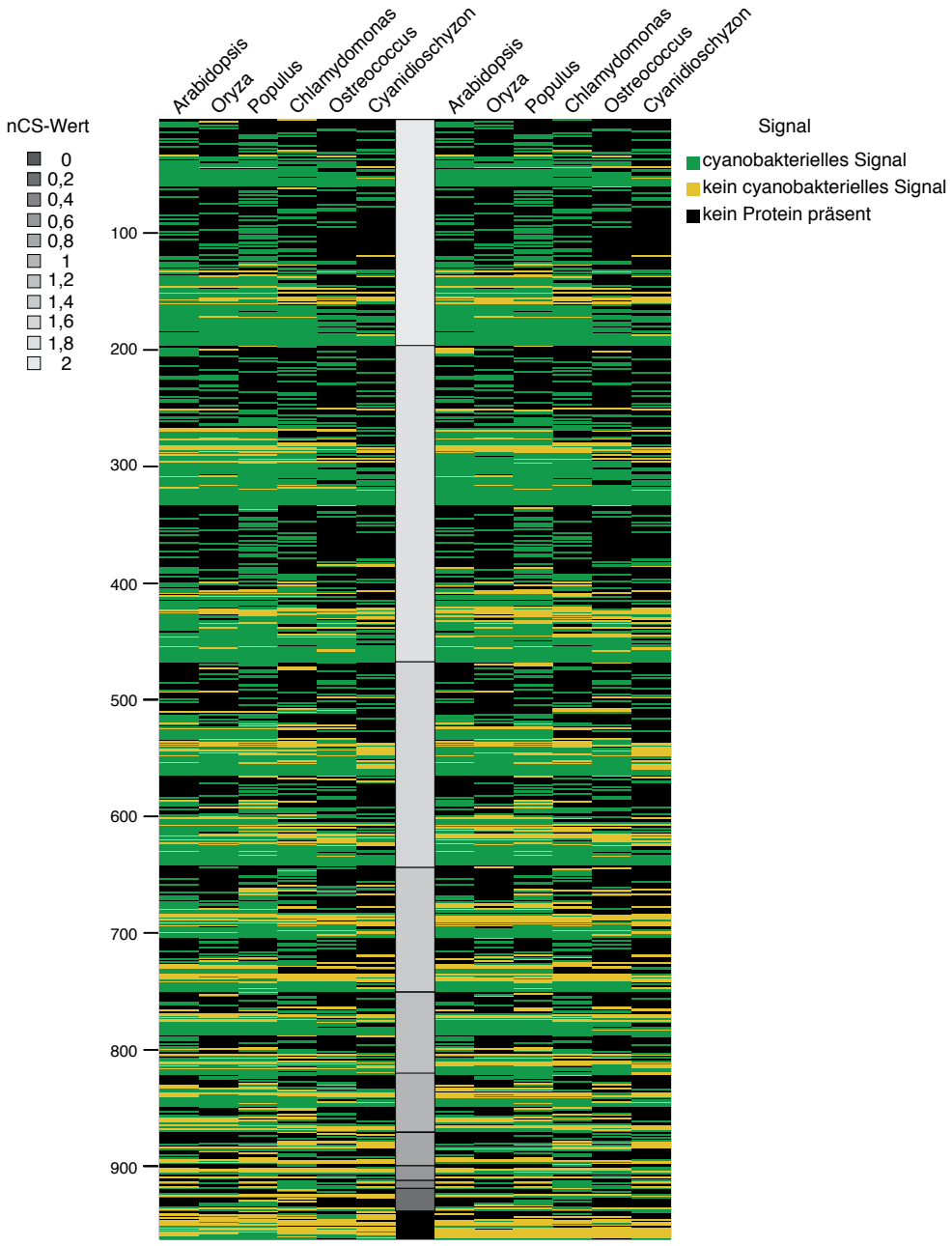

Abbildung 5.8.: Cyanobakterielle Signale aus 991 Clustern, in denen mindestens ein Pflanzen- oder Algenprotein ein Cyanobakterium als nächsten Nachbarn (siehe Abschnitt 4.4.5.1) in entweder dem Vorwärtsbaum (links) oder in dem Rückwärtsbaum hatte. Die Cluster sind nach dem nCS-Wert der Alignments – also nach ihrer Verlässlichkeit (siehe Abschnitt 4.3.7) – geordnet. In den Spalten sind die Signale der Proteine der Pflanzen und Algen abgebildet. Jede Zeile steht für einen Cluster, wobei ein grüner Balken ein cyanobakterielles Signal des Proteins des jeweiligen Organismus in dem Cluster markiert, ein gelber Balken ein nicht cyanobakterielles Signal. Die Farbkodierung in dem mittleren Balken zeigt die Intervallzugehörigkeit des nCS-Werts des jeweiligen Clusters. Die Pflanzen und Algen sind in folgender Reihenfolge sortiert: *A. thaliana*, *O. sativa*, *P. trichocarpa*, *C. reinhardtii*, *O. tauri*, *C. merolae*.

Ein schwarzer Balken weist darauf hin, dass für diese Pflanze oder Alge kein Protein in dem Cluster vorkam und ein gelber Balken bedeutet, dass zwar ein Protein dieser Pflanze oder Alge in dem Cluster vorhanden war, dieses aber kein Cyanobakterium als nächsten Nachbarn hatte. 59 Cluster hatten einen nCS-Wert von 2, somit zeigten beide Bäume dieselbe Phylogenie, da diejenigen Spalten ohne Lücken, die aus dem Vorwärts- und Rückwärtsalignment extrahiert wurden, identisch waren. Für 282 Cluster wurden vorwiegend identische Spalten extrahiert, weswegen sie in das nächste Intervall von 1,8 bis 2 eingeordnet wurden. Dann folgen 239 Cluster, deren nCS-Werte zwischen 1,6 und 1,8 lagen und 144 Cluster in dem darauffolgenden Intervall. 86 Cluster wiesen einen nCS-Wert zwischen 1,2 und 1,4 auf, während bei 64 Clustern nur noch circa die Hälfte der Spalten in den beiden Alignments identisch waren. Unter diesem Wert befanden sich 117 Cluster, von denen in 25 Fällen keine einzige Spalte zwischen dem Vorwärts- und Rückwärtsalignment identisch war.

Proteine von *A. thaliana* kamen in 557 Clustern in dieser Abbildung vor. 612 Proteine in 447 Clustern zeigten in den Vorwärtsbäumen ein cyanobakterielles Signal, 572 Proteine in 420 Clustern in den Rückwärtsbäumen. In 120 beziehungsweise 137 Clustern hatte kein Protein von *A. thaliana* ein Cyanobakterium als nächsten Nachbarn. 541 Proteine (88 %) in 398 Clustern zeigten sowohl im Vorwärts- als auch im Rückwärtsbaum ein cyanobakterielles Signal, wohingegen 62 Proteine in 43 Clustern dieses nur im Vorwärtsbaum und 31 Proteine in 22 Clustern nur im Rückwärtsbaum zeigten. In 94 Clustern konnte weder im Vorwärts- noch im Rückwärtsbaum ein Cyanobakterium als nächster Nachbar ermittelt werden. Für *O. sativa* waren es 503 Cluster insgesamt, davon zeigten 476 (446) Proteine in 389 (374) Clustern ein cyanobakterielles Signal. Für 441 Proteine in 346 Clustern war in beiden Bäumen ein cyanobakterielles Signal nachweisbar, für 55 Proteine in 36 Clustern nur im Vorwärtsbaum, für 35 Proteine in 28 Clustern nur im Rückwärtsbaum und in 93 Clustern in keinem der beiden Bäume. Proteine von *P. trichocarpa* waren in 624 Clustern in dieser Abbildung vertreten, von denen 701 (654) Proteine in 488 (457) Clustern als nächsten Nachbarn ein Cyanobakterium hatten.

Für 607 Proteine in 432 Clustern war dieses Signal in beiden Bäumen zu finden, für 80 Proteine in 47 Clustern beziehungsweise 47 Proteine in 25 Clustern nur in einem der Bäume und in 120 Clustern in keinem Baum. Die Proteine in 512 Clustern, in denen Sequenzen der Grünalge *C. reinhardtii* vorkamen, zeigten für 437 (420)

Proteine und 380 (375) Cluster eine Verwandtschaft zu Cyanobakterien, in 382 Proteinen verteilt auf 340 Cluster in beiden Bäumen, während in 47 (33 Cluster) und 38 (35 Cluster) Fällen die Verwandtschaft nur in einem der Bäume nachgewiesen wurde. In 104 Clustern war keine Verwandtschaft zu Cyanobakterien nachweisbar. 348 (328) Proteine von *O. tauri* hatten einen potenziellen cyanobakteriellen Ursprung. In 270 Clustern (298 Proteine) konnte dies in beiden Bäumen beobachtet werden, in 33 und 25 nur in einem Baum. In 92 Clustern wurde kein cyanobakterieller Ursprung abgeleitet. Das cyanobakterielle Signal wurde ebenfalls für 253 (234) Proteine in 230 (215) der 341 Cluster von *C. merolae* abgeleitet, wobei es für 219 Proteine in 201 Clustern in beiden Bäumen abgeleitet wurde, für 28 und 15 Proteine in 25 und 14 Clustern nur in einem der Bäume und in 101 Clustern in keinem.

Zusätzlich zu den Proteinen, für die durch die phylogenetische Analyse unter Verwendung der Cluster, die mindestens vier Sequenzen enthalten, ein cyanobakterieller Ursprung ermittelt werden konnte, gab es einige Proteine, die in den kleineren Clustern ein cyanobakterielles Signal lieferten. Dies ist der Fall, wenn in einem Cluster, in dem zwei oder drei Sequenzen enthalten waren, keine anderen Organismengruppen als Pflanzen, Algen und Cyanobakterien vorkamen. Insgesamt konnten für die betrachteten Algen und Pflanzen folgende Zahlen als sehr verlässlich angesehen werden: Aus *A. thaliana* stammten 541 Proteine, für die in der phylogenetischen Analyse sowohl im Vorwärts- als auch im Rückwärtsbaum der cyanobakterielle Ursprung hergeleitet wurde. Außerdem kamen 21 Proteine in kleineren Clustern ausschließlich mit anderen Pflanzen und Algen und Cyanobakterien vor, weswegen für insgesamt 561 Proteine ein cyanobakterieller Ursprung angenommen wurde.

Für 441 Proteine aus *O. sativa* wurde durch phylogenetische Analysen ein cyanobakterieller nächster Nachbar ermittelt, zusätzlich dazu gab es 29 Proteine in kleineren Clustern, wodurch eine Gesamtzahl von 470 Proteinen entstand. Die Proteine von *P. trichocarpa* zeigten in 607 Fällen einen cyanobakteriellen Ursprung in den phylogenetischen Bäumen und in 89 Fällen in kleineren Clustern. Im Kerngenom von *C. reinhardtii* gab es mindestens 382 und zusätzlich 32 also insgesamt 414 cyanobakterielle Gene. In *O. tauri* waren es 298 Proteine, die durch die phylogenetische Analyse bestätigt wurden und 29 aus kleinen Clustern also insgesamt 327 Proteine. 219 Proteine aus *C. merolae* zeigten in der phylogenetischen Analyse einen cyanobakteriellen Ursprung und zusätzliche acht in kleinen Clustern. Für

eine Zusammenfassung dieser Werte siehe Tabelle A.2, Tabelle A.3.

Für die Berechnung des cyanobakteriellen Anteils an den Kerngenomen der Pflanzen und Algen wurden nur diejenigen Proteine berücksichtigt, die sowohl im Vorwärts- als auch im Rückwärtsbaum dasselbe phylogenetische Signal zeigten. Insgesamt war das bei *A. thaliana* für 7.003 Proteine der Fall. Der cyanobakterielle Anteil betrug also 8 %. Wurden nur diejenigen Cluster betrachtet, in denen sowohl Proteine aus Cyanobakterien als auch Proteine von *A. thaliana* vorkamen (1.316), so gaben 932 Proteine in diesen Clustern in beiden Bäumen dasselbe Signal und der Anteil erhöhte sich auf 37,3 %.

Die phylogenetische Analyse der Proteine von *O. sativa* zeigte für 6.026 Proteine dasselbe Ergebnis. Davon kamen 1.198 Proteine, die in beiden Bäumen dasselbe Signal gaben, in 875 Clustern vor, in denen sowohl ein Protein von *O. sativa* als auch mindestens ein cyanobakterielles Protein vorhanden waren. Dies ergab einen Anteil von 7,7 % beziehungsweise 38,3 %, wenn die 29 kleinen Cluster mit eingerechnet wurden. Proteine von *P. trichocarpa* kamen in 1.363 Clustern zusammen mit einem Cyanobakterium vor. Insgesamt gaben 9.005 Proteine dasselbe Signal in beiden Bäumen, 2.090 dieser Proteine kamen in den geteilten Clustern vor. Der Anteil der cyanobakteriellen Proteine betrug also 7,7 % und 31,9 %.

Für 3.677 Proteine von *C. reinhardtii* wurde sowohl im Vorwärts- als auch im Rückwärtsbaum dasselbe Signal identifiziert. In 867 Clustern, in denen Proteine von *C. reinhardtii* und von einem Cyanobakterium vorkamen, zeigten 995 Proteine dasselbe Signal in beiden Bäumen. Somit ergaben sich Anteile von 10,3 % und 40,3 %. 2.914 Proteine von *O. tauri* zeigten in beiden Bäumen dasselbe Signal, in 672 Clustern waren sowohl Proteine von *O. tauri* als auch Proteine von Cyanobakterien vorhanden. In diesen Clustern zeigten 748 Proteine dasselbe Signal, wodurch Anteile von 11,1 % und 42,1 % berechnet werden konnten. Von 2.119 Proteinen von *C. merolae* kamen 623 Proteine in 565 geteilten Clustern vor. Für die Rotalge ergaben sich somit Anteile von 10,7 % und 36 %.

Abbildung 5.9 zeigt eine Analyse von denjenigen Proteine der Pflanzen und Algen, die sowohl im Vorwärts- als auch im Rückwärtsbaum ein cyanobakterielles Signal zeigten. Insgesamt wurden 785 Cluster mit mehr als drei Sequenzen betrachtet. Die Proteine für jede Pflanze und Alge, die in beiden Bäumen ein cyanobakterielles Signal zeigten, sind grün markiert, alle anderen schwarz.

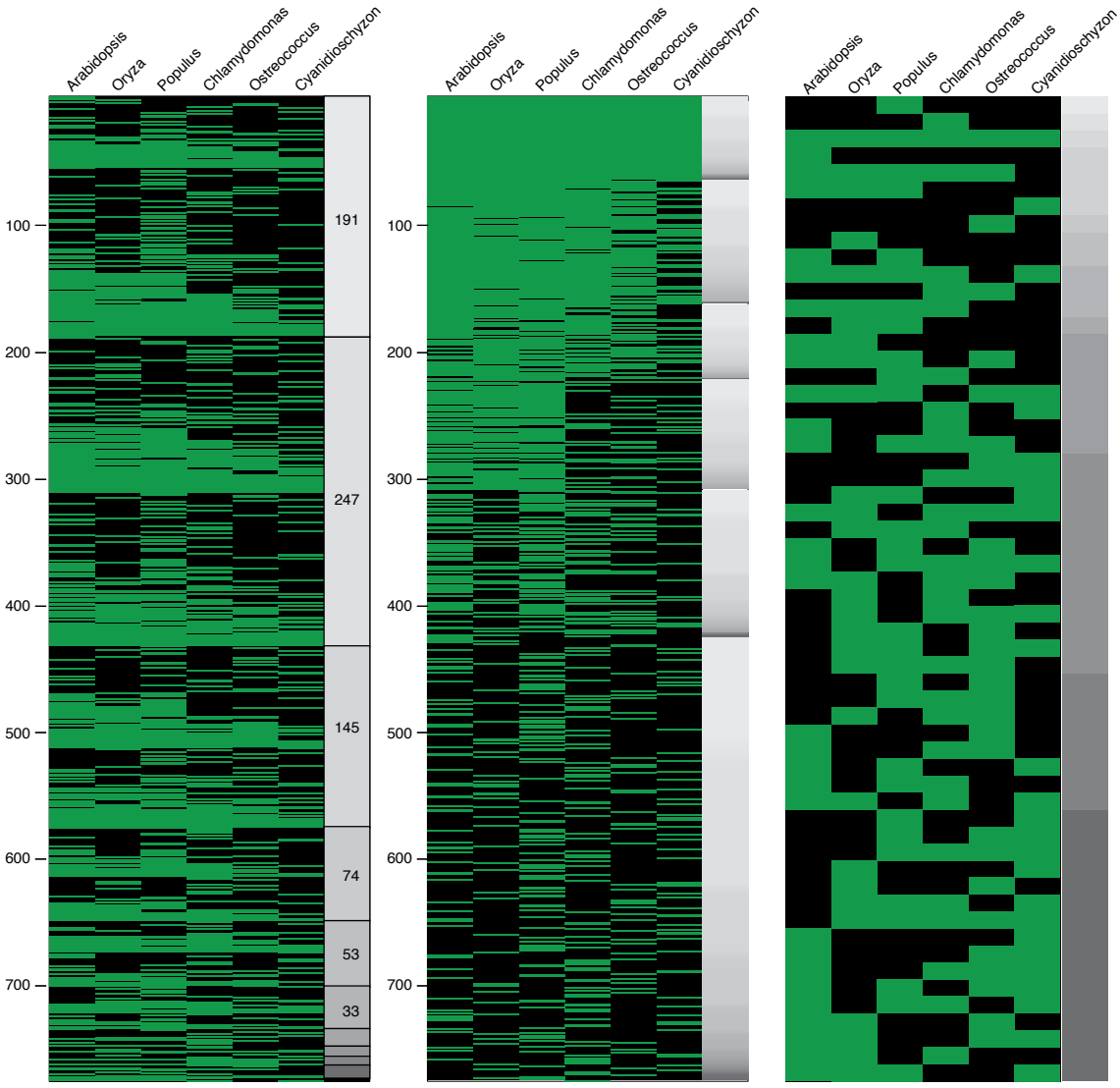

Abbildung 5.9.: Cyanobakterielle Signale für Pflanzen- und Algenproteine in 785 Clustern. Abgebildet sind 785 Cluster mit mehr als drei Sequenzen, in denen für ein beliebiges Pflanzen- oder Algenprotein sowohl im Vorwärts- als auch im Rückwärtsbaum ein Cyanobakterium als nächster Nachbar (siehe Abschnitt 4.4.5.1) gefunden wurde. Eine Zeile entspricht einem Cluster, eine Spalte einer Pflanze oder Alge. Die grünen Balken symbolisieren Proteine einer Pflanze oder Alge, die in dem Cluster ein cyanobakterielles Signal zeigten. Die linke Abbildung zeigt die Cluster sortiert nach dem nCS-Wert (der Verlässlichkeit der zugehörigen Alignments, siehe Abschnitt 4.3.7), wobei die Zahlen in dem grauen Balken angeben, wie viele Cluster einen nCS-Wert in dem jeweiligen Intervall hatten. In der mittleren Abbildung sind dieselben Cluster nach der Anzahl der Pflanzenproteine, die in diesem Cluster vorkamen, sortiert. Cluster, die mehr Pflanzenproteinen enthielten, stehen weiter oben. Der graue Balken zeigt den nCS-Wert der Alignments. Die rechte Abbildung zeigt wie oft ein Muster, das heißt ein gemeinsames Auftreten der Proteine verschiedener Pflanzen, vorkam. Die Pflanzen und Algen sind in jeder Abbildung in folgender Reihenfolge sortiert: *A. thaliana*, *O. sativa*, *P. trichocarpa* , *C. reinhardtii*, *O. tauri*, *C. merolae*

Das heißt in einem Cluster kann durchaus ein Protein dieser Pflanze oder Alge vorhanden gewesen sein, aber es zeigt entweder gar kein cyanobakterielles Signal oder nur in einem der Bäume. Die linke Abbildung zeigt die Cluster sortiert nach dem nCS-Wert der zugehörigen Alignments (siehe Abschnitt 4.3.7). 191 Cluster hatten einen nCS-Wert zwischen 1,8 und 2, 247 Werte lagen zwischen 1,6 und 1,8 und weitere 145 zwischen 1,4 und 1,6. Werte zwischen 1,2 und 1,4 kamen in 74 Clustern vor, während 53 Cluster Werte zwischen 1 und 1,2 aufwiesen. Für die restlichen 75 Cluster wurden Alignments berechnet, die einen nCS-Wert unter 1 hatten.

Die mittlere Abbildung zeigt dieselben Cluster. Allerdings sind sie nicht nach dem nCS-Wert geordnet, sondern nach der Anzahl der in den Clustern aufgetretenen Proteine, die in beiden Bäumen ein cyanobakterielles Signal zeigten. So stehen in den ersten Zeilen diejenigen 66 Cluster, in denen alle sechs Pflanzen und Algen mindestens ein Protein mit einem potenziellen cyanobakteriellen Ursprung aufwiesen. Der nCS-Wert ist farbkodiert neben der Abbildung zu finden. Je heller das grau ist, desto höher ist der nCS-Wert.

In der rechten Abbildung sind die auftretenden Muster nach der Häufigkeit ihres Auftretens sortiert zu finden. *A. thaliana* hatte in 60 Clustern als einzige Pflanze und Alge ein Protein mit cyanobakteriellem Ursprung. *O. sativa* und *O. tauri* waren in jeweils 40 und 43 Clustern einzeln zu finden und *P. trichocarpa* in 85 Clustern. Proteine von *C. reinhardtii* kamen in 75 einzeln vor, *C. merolae* in 51 Clustern. Alle Pflanzen und Algen waren in 66 Clustern vorhanden, die Pflanzen zusammen mit den Grünalgen in 59 Clustern und die Pflanzen alleine in 54 Clustern. Die Grünalgen waren in 22 Clustern einzeln vertreten. Alle Pflanzen und Algen ohne *O. tauri* oder *C. merolae* kamen in jeweils 22 Clustern vor. Insgesamt waren 58 verschiedene Muster zu beobachten, 16 kamen nur einmal vor, weitere acht zweimal und fünf dreimal.

5.3. Der Vorfahr der primären Plastiden

Aus den Signalen für die Proteine, für die ein cyanobakterieller Ursprung nachgewiesen werden konnte, können Rückschlüsse darauf gezogen werden, welches Cyanobakterium dem Vorläufer der Chloroplasten am Ähnlichsten ist. Abbildung 5.10 zeigt für jede Pflanze und Alge wie oft jedes Cyanobakterium zu den in Abbildung 5.9 als cyanobakteriell identifizierten Proteinen nächster Nachbar war. Da nicht immer nur ein Cyanobakterium nächster Nachbar zu einem Protein war sondern häufig eine Gruppe, übersteigt die Summe der nächsten Nachbarn in der Abbildung die Anzahl der cyanobakteriellen Proteine.

Von den 541 Proteinen aus *A. thaliana* (nur die Proteine in den Clustern mit mehr als drei Sequenzen) hatten 315 Proteine ein Protein aus *Acaryochloris marina* als nächsten Nachbarn. 301mal kam *Nostoc punctiforme* PCC 73102 als nächster Nachbar vor, 300mal *Nostoc* sp. PCC 7120 und 299mal *Anabaena variabilis* ATCC 29413. *Trichodesmium erythraeum*, *Cyanothece* sp. PCC 7424, *Cyanothece* sp. ATCC 51142, *Cyanothece* sp. CCY0110, *Microcystis aeruginosa* NIES 843, *Synechocystis* sp. PCC 6803, *Crocosphaera watsonii* WH 8501, *Synechococcus* sp. PCC 7002, *Synechococcus elongatus* PCC 6301, *Synechococcus elongatus* PCC 7942, *Cyanothece* sp. PCC 8801 und *Thermosynechococcus elongatus* BP-1 wiesen Werte zwischen 244 und 284 auf. Proteine von *Gloeobacter violaceus* waren 200mal nächster Nachbar zu einem Protein von *A. thaliana*.

Die 23 anderen Cyanobakterien zeigten Werte zwischen 154 und 170 und lagen somit deutlich unter dem Mittelwert von 208. Für die 441 cyanobakteriellen Proteine von *O. sativa* wurden 6.206 cyanobakterielle nächste Nachbarn gefunden. Die Werte für die einzelnen Cyanobakterien waren geringer als bei *A. thaliana*, trotzdem zeigte *A. marina* ebenfalls die meisten nächsten Nachbarn mit einem Wert von 242. *N. punctiforme* PCC 73102 (237), *Nostoc* sp. PCC 7120 (235), *Cyanothece* sp. PCC 7424 (292) und *A. variabilis* (232) zeigten die nächst höheren Werte. Die Werte derselben Cyanobakterien wie bei *A. thaliana* lagen über dem Mittelwert von 155. Dieselbe Verteilung konnte ebenfalls für die 607 Proteine von *P. trichocarpa* beobachtet werden. *A. marina* hatte mit 331 nächsten Nachbarn den höchsten Wert, dann folgten die oben erwähnten Cyanobakterien und dann mit einer deutlichen Differenz von 22 Proteinen die 23 *Prochlorococcus marinus* und *Synechococcus* sp. Stämme.

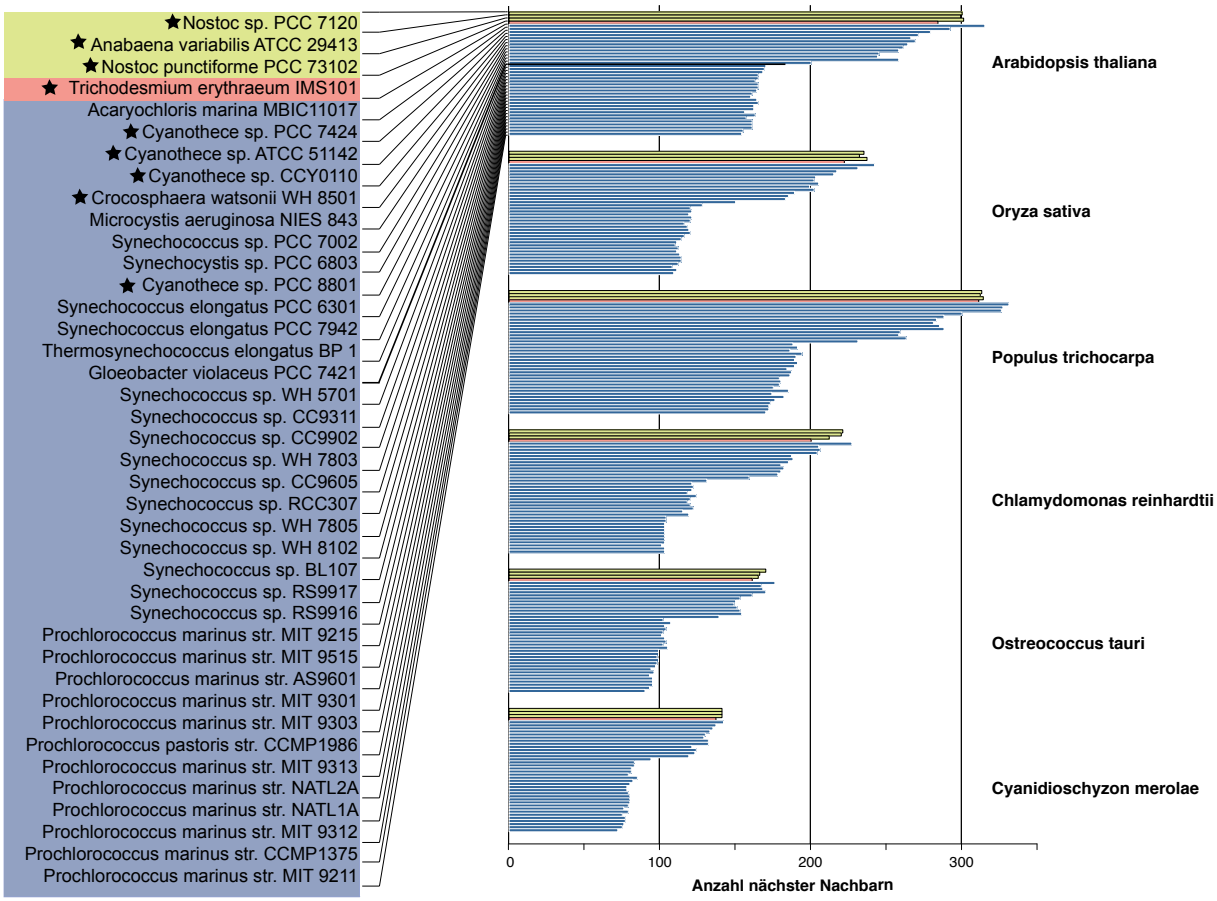

Abbildung 5.10.: Anteile der einzelnen Cyanobakterien an den nächsten Nachbarn der Pflanzen- und Algenproteine. Dargestellt sind Histogramme für jede Pflanze und Alge, die zeigen wie oft welches Cyanobakterium nächster Nachbar (siehe Abschnitt 4.4.5.1) zu einem Protein der entsprechenden Spezies war. Die Cyanobakterien sind nach den von Stanier eingeteilten Gruppen sortiert: 36 Cyanobakterien gehören zu Gruppe I, eins zu Gruppe III und drei zu Gruppe IV (siehe Abschnitt 3.1.1). Die Cyanobakterien, die mit Sternen markiert sind, besitzen eine Nitrogenase, mit der sie elementaren Stickstoff fixieren können. Das Histogramm links unten zeigt wie oft jedes Cyanobakterium insgesamt zu allen Pflanzen und Algenproteinen nächster Nachbarn war.

Auch bei den Grünalgen und *C. merolae* konnte diese Verteilung beobachtet werden. Dann folgten die Cyanobakterien der Gruppe IV, *T. erythraeum* aus der Gruppe III und einige Cyanobakterien der Gruppe I. Zwischen *Gloeobacter violaceus*, das die niedrigsten Werte dieser Gruppe aufwies und den restlichen 23 Cyanobakterien bestand eine deutliche Differenz.

Die Aufteilung in diese beiden Gruppen ist auch in Abbildung 5.11 erkennbar. Abbildung 5.10 legt nahe, dass in den meisten Fällen nicht ein Cyanobakterium der nächste Nachbar war, sondern eine Gruppe.

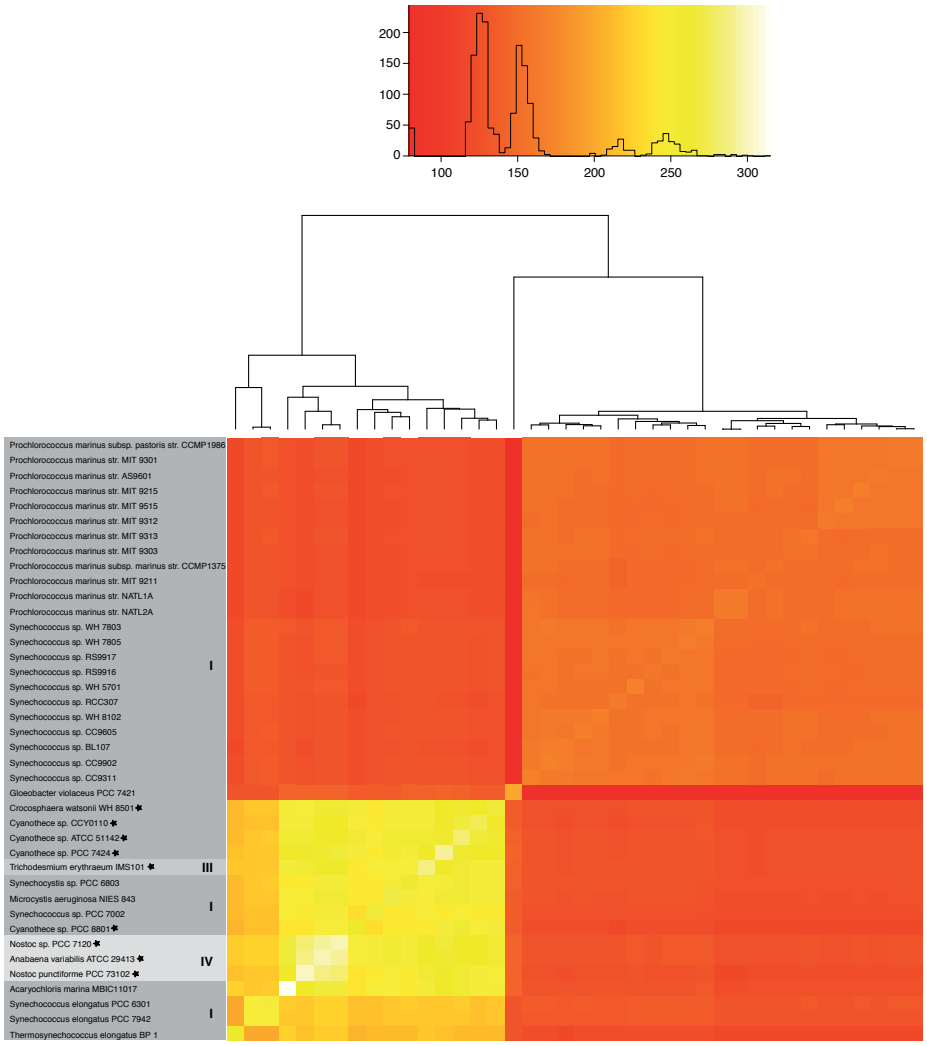

Abbildung 5.11.: Gleichzeitiges Auftreten der Cyanobakterien in der Gruppe der nächsten Nachbarn zu Proteinen von *A. thaliana*. Jede Zeile und Spalte dieser Abbildung repräsentiert ein Cyanobakterium. Wie oft die Cyanobakterien gemeinsam in der Gruppe der nächsten Nachbarn zu jedem Protein von *A. thaliana* vorkamen ist farbkodiert. Je öfter zwei Cyanobakterien gemeinsam in einer Gruppe vorkamen, desto heller ist das entsprechende Feld. Auf der Hauptdiagonalen sind die Werte von Abbildung 5.10 zu finden. Das Dendrogramm auf der linken Seite zeigt das hierarchische Clustering der Werte (siehe Abschnitt 3.2.2). Die farbliche Unterlegung der Namen der Cyanobakterien symbolisiert die Zugehörigkeit zu der Gruppe (siehe Abschnitt 3.1.1). Diejenigen Cyanobakterien, die eine Nitrogenase besitzen sind mit einem Stern markiert.

Deswegen wurde das gemeinsame Auftreten der Cyanobakterien als nächste Nachbarn zu Proteinen von *A. thaliana* untersucht. Die Hauptdiagonale der farbkodierten Matrix zeigt die Werte, die in Abbildung 5.10 zu sehen sind. Außerdem ist deutlich zu sehen, dass in fast allen 315 Gruppen, in denen *A. marina* vorhanden war, auch die Cyanobakterien der Gruppe IV vorkamen (255 - 265 geteilte Grup-

pen).

Hohe Werte (237 - 253) wurden ebenfalls zu den anderen Cyanobakterien, die eine Nitrogenase besitzen, beobachtet. Außerdem waren die *Synechococcus* sp. und *Thermosynechococcus elongatus* Stämme in dieser Gruppe zu finden (219 - 225). *Gloeobacter violaceus* war in deutlich weniger gemeinsamen Gruppen vorhanden (144). Eine deutliche Abstufung erfolgte zu den *Prochlorococcus marinus* und *Synechococcus* sp. Stämmen. Sie waren nur in 123 - 130 gemeinsamen Gruppen mit *A. marina* zu finden. Diese waren in 140 bis 170 gemeinsamen Gruppen, während zwischen den zwei Hauptgruppen höchstens 130 gemeinsame Gruppen zu finden war.

Auch das hierarchische Clustering dieser Werte, das durch das Dendrogramm in der Abbildung dargestellt ist, bestätigte die Aufteilung der Cyanobakterien in mindestens zwei Gruppen. Die Verteilung der Werte, die in der Legende zu sehen ist, zeigt, dass es viele Werte gab, die im roten Farbspektrum (circa 120 Gruppe) lagen. Dieses erste Maximum der Verteilung repräsentiert die Vergleiche zwischen den Gruppen. Das nächste Maximum zeigt die Werte innerhalb der zweiten Gruppe, die aus den *Prochlorococcus marinus* und *Synechococcus* sp. Stämmen bestand. Sie kamen zwar häufig in derselben Gruppe vor insgesamt jedoch nicht so häufig wie die Cyanobakterien aus der ersten Gruppe. Die anderen beiden kleineren Maxima zeigten die Vergleiche innerhalb der ersten Gruppe, wobei es eine Abstufung zwischen den *Synechococcus* sp. und *Thermosynechococcus elongatus* Stämmen und den anderen Cyanobakterien dieser Gruppe gab.

Aus allen 90.354 Bäumen und den darin enthaltenen 1.587.805 Splits wurde ein Supernetzwerk unter Verwendung der Hamming-Distanz erstellt (siehe Abschnitt 4.4.6). Mithilfe des Programms `splitstree` wurden für die graphische Darstellung nur die Pflanzen, Algen und Cyanobakterien ausgewählt. Abbildung 5.12 zeigt das Supernetzwerk für diese 47 Organismen. Die 40 Cyanobakterien trennten sich in zwei Gruppen auf: In der einen Gruppe befanden sich die *Prochlorococcus marinus* Stämme und bis auf *Synechococcus* sp. PCC 7002 und die zwei *Synechococcus elongatus* Stämme alle *Synechococcus* sp. Stämme. In der anderen Gruppe befanden sich die Spezies, die eine Nitrogenase für die Stickstofffixierung besitzen (markiert mit einem Stern).

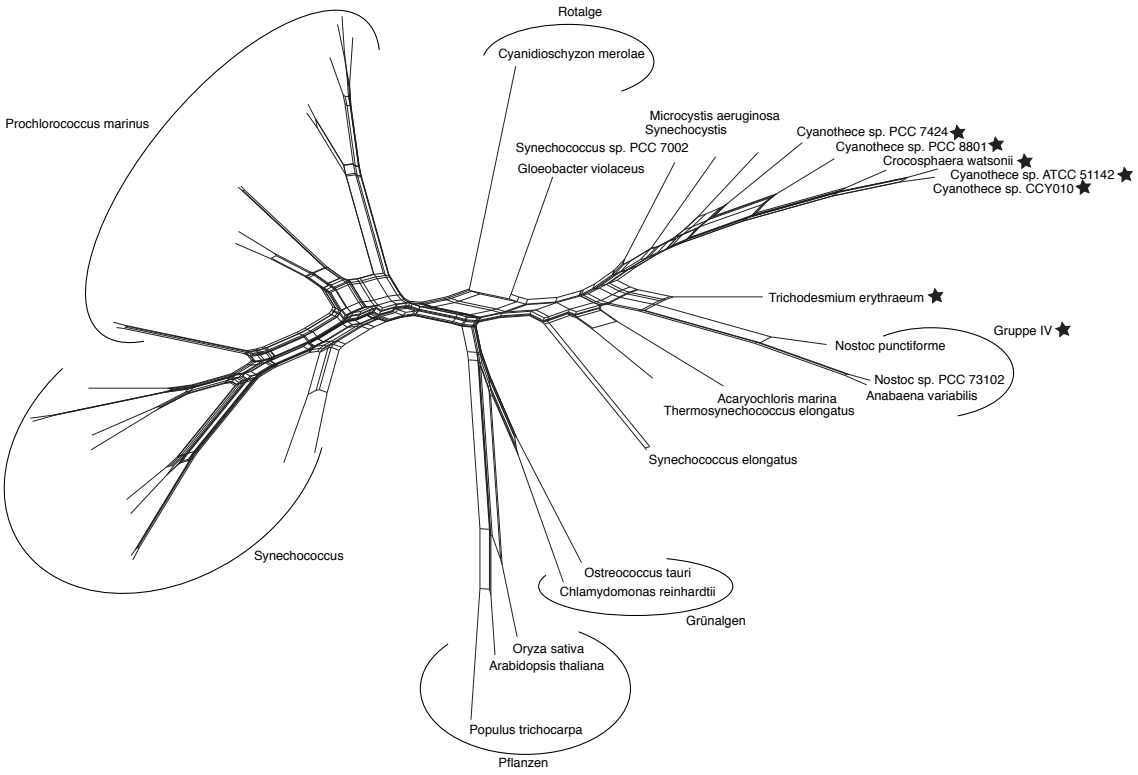

Abbildung 5.12.: Supernetzwerk der Pflanzen, Algen und Cyanobakterien erstellt aus 1.587.805 Splits von 90.354 berechneten Bäumen (siehe Abschnitt 4.4.6). Die Splits der Bäume wurden mit `consense` aus den Bäumen extrahiert und zu einer Supermatrix assembliert. Aus dieser Supermatrix wurden mithilfe des Programms `splitstree` Hamming-Distanzen zwischen den Organismen berechnet, die Pflanzen, Algen und Cyanobakterien ausgewählt und ein Netzwerk berechnet, das mithilfe eines C-Programms graphisch dargestellt wurde. Diejenigen Cyanobakterien, die eine Nitrogenase zur Stickstofffixierung enthalten, sind mit Sternen gekennzeichnet.

Zusätzlich dazu kamen alle Organismen, die in Abbildung 5.10 einen Wert zeigten, der höher war als der Mittelwert, in dieser Gruppe vor. *Gloeobacter violaceus* war basal zu dieser Gruppe platziert, allerdings durch einen deutlichen Split von allen anderen getrennt. Außerdem bildeten die Cyanobakterien der Gruppe IV und III (*Anabaena variabilis*, *Nostoc* und *Trichodesmium erythraeum*) eine Gruppe, die *Cyanothece* Stämme, *Crocosphaera watsonii*, *Microcystis aeruginosa*, *Synechocystis* und *Synechococcus* sp. PCC 7002 eine weitere. Die *Synechococcus elongatus* Stämme nahmen ähnlich wie *G. violaceus* eine basale Sonderstellung ein. *Thermosynechococcus elongatus* und *Acaryochloris marina* waren beide durch einen großen Split von den anderen Cyanobakterien getrennt.

Die Gruppe der Pflanzen und Algen trennte sich in zwei Gruppen auf. In der

einen befanden sich die Pflanzen und Grünalgen und in der anderen die Rotalge. Beide Gruppen entsprangen zwischen den beiden cyanobakteriellen Gruppen und waren durch deutliche Splits voneinander getrennt. Für die Pflanzen und Grünalgen gab es einen sehr großen Split, durch den sie mit den Stickstoff fixierenden Cyanobakterien gruppierten. Die Rotalge zeigte kleine Splits zu beiden cyanobakteriellen Gruppen.

Außerdem wurde eine Matrix erstellt, deren Zeilen jeweils einem Cluster entsprechen, in dem mindestens ein Protein einer Pflanze und Alge oder einem Cyanobakterium vorkamen. Die Spalten repäsentieren die 46 Organismen und die Einträge sind 1, wenn ein Organismus in diesem Cluster vorkam und 0 sonst. (engl. *presence absence matrix* (PAP)). Von dieser Matrix wurden Hamming-Distanzen zwischen den Organismen berechnet und eine hierarchische Clusteranalyse durchgeführt. Abbildung 5.13 zeigt das Ergebnis dieser Analyse.

In diesem Dendrogramm gruppierten die Grünalgen und die Rotalge, die Pflanzen bildeten eine eigene Gruppe. Fast alle Cyanobakterien, die eine Nitrogenase besitzen, und *M. aeruginosa* bildeten eine zweite Gruppe. *T. erythraeum* war das einzige Cyanobakterium mit einer Nitrogenase, das mit den *Synechococcus elongatus* Stämmen, *Synechococcus* sp. PCC 7002, *Synechocystis* und *Thermosynechococcus elongatus* gruppierte. Sowohl *G. violaceus* als auch *A. marina* gehörten zu keiner Gruppe. Die vierte Gruppe wurde von den *Prochlorococcus marinus* und *Synechococcus* sp. Stämmen gebildet.

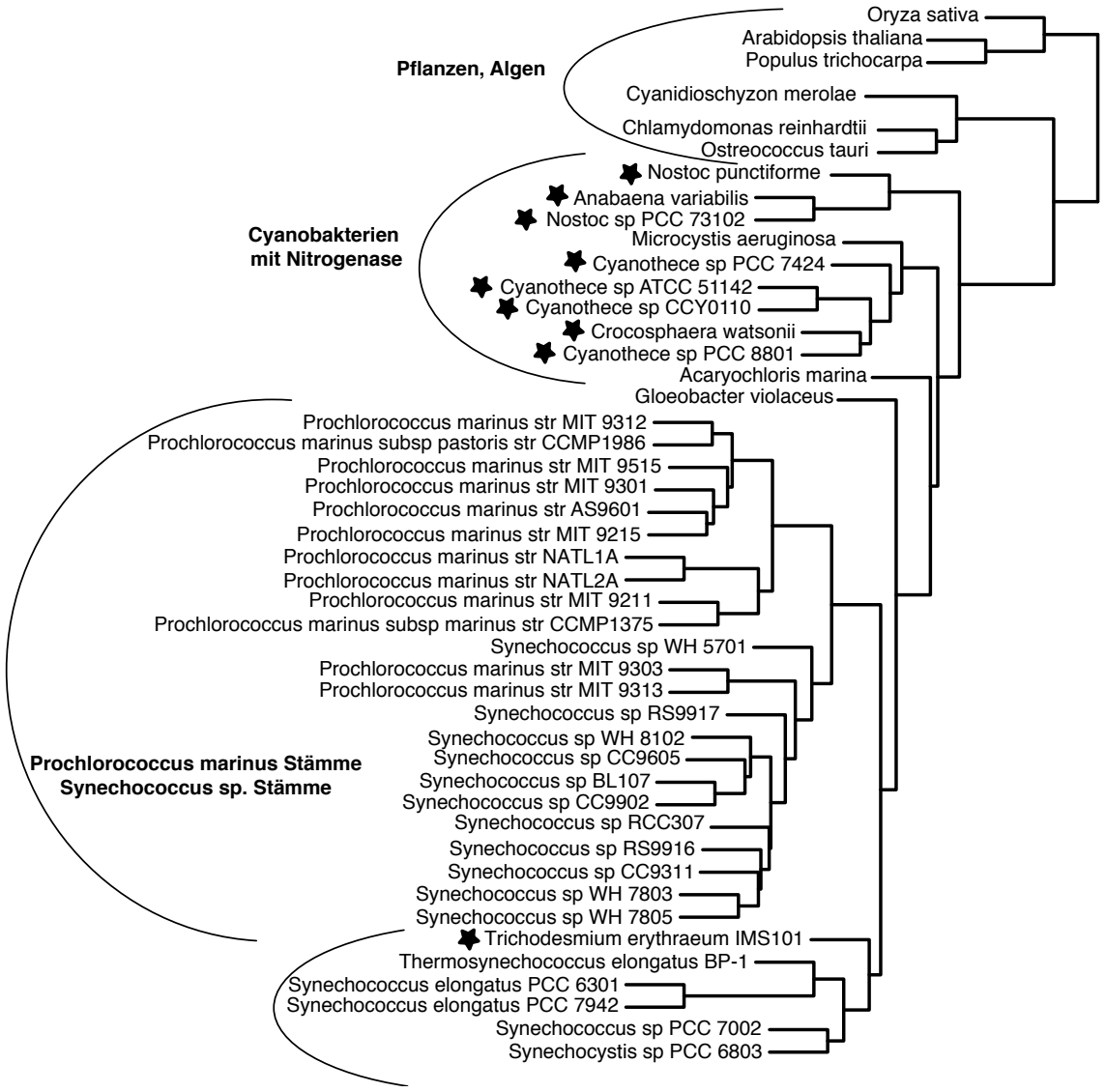

Abbildung 5.13.: Hierarchische Clusteranalyse der PAP-Matrix der Pflanzen, Algen und Cyanobakterien. Aus der PAP-Matrix für die Pflanzen, Algen und Cyanobakterien wurden Hamming-Distanzen zwischen diesen Organismen berechnet und eine hierarchische Clusteranalyse durchgeführt. Das Ergebnis dieser Analyse wurde in einem Dendrogramm graphisch dargestellt. Die Cyanobakterien, die eine Nitrogenase zur Stickstofffixierung besitzen, sind mit einem Stern markiert.

5.4. Grüne und rote Signale in Protisten

5.4.1. Die Bacillariophyta

Eine weitere Fragestellung, die in dieser Arbeit beantwortet werden sollte, ist die Frage nach dem Anteil der Proteine aus Grünalgen und Rotalgen in den Chromalveolaten und Excavata. Im Weiteren soll untersucht werden, wieviele Gene in den Kerngenomen von *T. pseudonana* und *P. tricornutum* eher den Proteinen der Grünalgen und wieviele den Proteinen der Rotalge stärker ähneln, um eine Aussage darüber treffen zu können, ob der Vorfahre des sekundären Endosymbionten eine Rotalge oder eine Grünalge war (vgl. Abschnitt 3.1.3).

Abbildung 5.14 zeigt die Analyse der nächsten Nachbarn von *T. pseudonana*. Insgesamt wurden 4.447 Proteine in 3.983 Clustern untersucht, in denen Proteine des Protisten vorkamen. In 2.619 Fällen war ein Protein von *P. tricornutum* der nächste Nachbar und in 332 Fällen andere Protisten. 147 Proteine hatten Proteine aus Cyanobakterien als nächsten Nachbarn, 122 aus Pflanzen, 155 aus Grünalgen, 85 aus der Rotalge. 37 Proteine hatten Proteine aus Pflanzen und Grünalgen als nächsten Nachbarn, neun eine Gruppe aus Pflanzen, Grünalgen und Rotalgen. Da die Pflanzen und Grünalgen ein weitaus größeres Genom besitzen als die Rotalge, wurden aus diesen Clustern diejenigen ausgewählt, in denen ein Protein aus *T. pseudonana*, mindestens ein Rotalgen- und mindestens ein Grünalgenprotein vorhanden waren.

Dies war für 2.310 Proteine der Fall, wobei wiederum die meisten nächsten Nachbarn aus *P. tricornutum* stammten, 21 aus Cyanobakterien, 28 aus Pflanzen, 53 aus Grünalgen, 16 aus Grünalgen und Pflanzen und 73 aus der Rotalge. Neun Proteine hatten eine Mischgruppe aus Pflanzen, Grünalgen und Rotalgen als nächsten Nachbarn, 114mal wurden andere Protisten gefunden. Betrachtet man von diesen Clustern nur diejenigen, in denen auch *P. tricornutum* vorkam, so bildeten die beiden Protisten in 1.636 Fällen eine Gruppe, während 25mal eine Rotalge der nächste Nachbar war, 27mal eine Grünalge und 10mal eine Pflanze.

Als nächstes wurde in diesen Clustern der nächste Nachbar zu der Gruppe dieser beiden Organismen gesucht. Für diese Proteine war 713mal ein anderer Protist nächster Nachbar, 172mal eine Rotalge, 35mal eine Pflanze, 122mal eine Grünalge und 50mal eine Gruppe aus Pflanzen und Grünalgen.

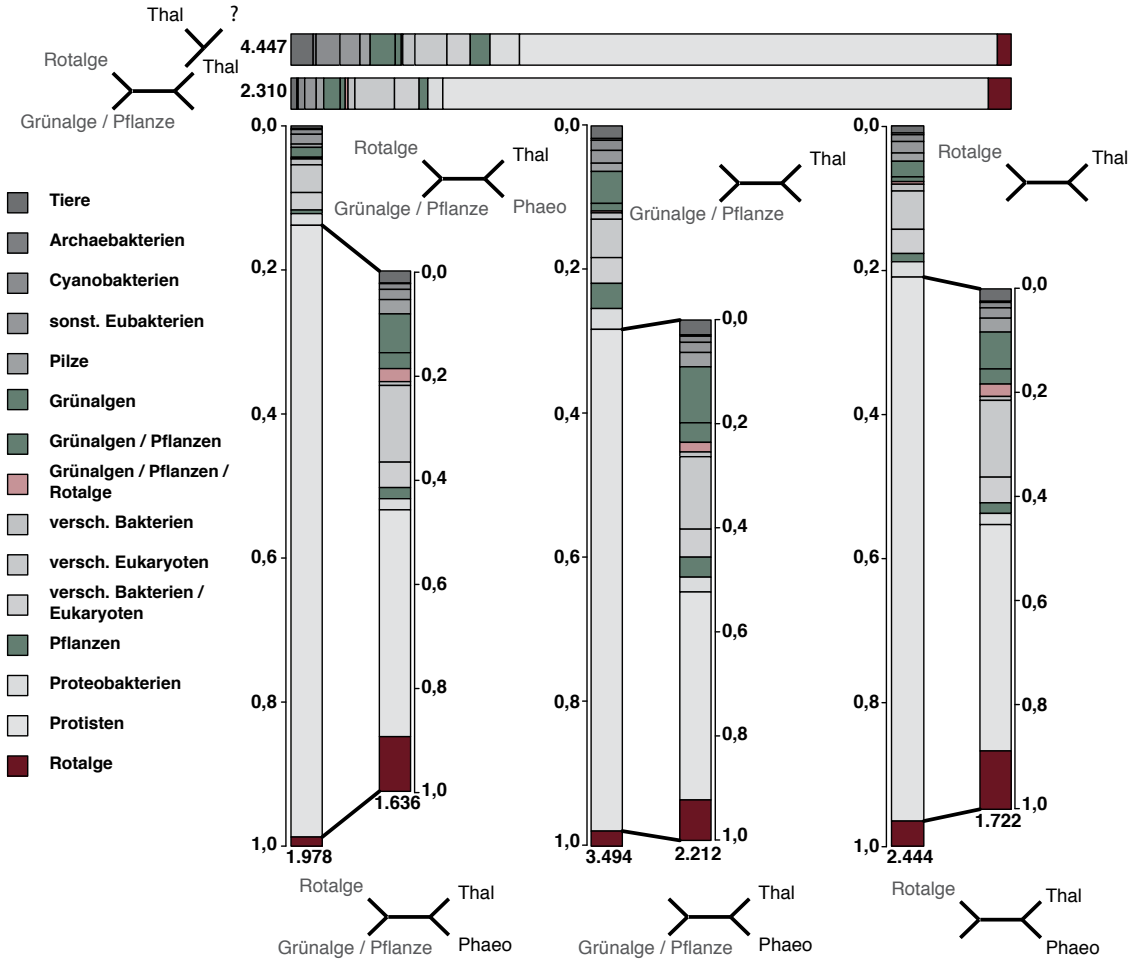

Abbildung 5.14.: Nächste Nachbarn von *Thalassiosira pseudonana* in verschiedenen Datensätzen. In jedem Stapeldiagramm sind die prozentualen Anteile der in der Legende dargestellten Gruppen an den nächsten Nachbarn (siehe Abschnitt 4.4.5.1) der Proteine von *T. pseudonana* dargestellt. Für das obere Diagramm wurden alle Bäume verwendet, in denen mindestens ein Protein von *T. pseudonana* vorkam (4.447). Für den nächsten Graphen wurden alle Bäume analysiert, in denen mindestens ein Protein von *T. pseudonana* ein Grünalgen- oder Pflanzenprotein und ein Rotalgenprotein vorkamen (2.310). Der untere linke Graph zeigt die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein von *T. pseudonana*, ein Protein von *Phaeodactylum tricornutum*, ein Grünalgen- oder Pflanzenprotein und ein Rotalgenprotein vorhanden waren (1.978). In 1.636 Bäumen war *P. tricornutum* nächster Nachbar. Für diese Fälle sind die nächsten Nachbarn beider Proteine dargestellt. Für den mittleren Graphen wurden alle Bäume verwendet, in denen mindestens ein Protein von *T. pseudonana* und ein Grünalgen- oder Pflanzenprotein vorkamen (3.494). In 2.212 Bäumen war *P. tricornutum* nächster Nachbar und die Anteile der nächsten Nachbarn beider Proteine sind dargestellt. Der rechte Graph zeigt die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein von *T. pseudonana* und ein Rotalgenprotein vorhanden waren (2.444). In 1.722 Bäumen war *P. tricornutum* nächster Nachbar.

41mal wurde eine Gruppe aus Grünalgen und Rotalgen gefunden, so dass insgesamt 197 nächste Nachbarn aus der Rotalge, 149 aus Grünalgen und 45 aus Pflanzen stammten. 53 Proteine wiesen eine Mischgruppe aus Pflanzen und Grünalgen als nächsten Nachbar auf, 43 eine Mischgruppe aus Grünalgen und Rotalgen. Für die Beantwortung der Frage, ob die Anzahl der gefundenen Grünalgen- und Pflanzenproteine in den Kerngenomen der photosynthetischen Protisten abhängig ist von der Genomgröße der Rotalge, wurden im Folgenden alle Cluster betrachtet, in denen mindestens ein Pflanzen- oder Grünalgenprotein und mindestens ein Protein von *T. pseudonana* vorhanden waren.

Zu dieser Auswahl gehörten 3.494 Proteine, von denen 155 Grünalgen als nächsten Nachbar hatten, 122 Pflanzen, 73 die Rotalge, 37 eine Mischgruppe aus Pflanzen und Grünalgen und 9 eine Mischgruppe aus Grünalgen, Pflanzen und Rotalgen. 2.434 dieser Proteine wiesen einen anderen Protisten als nächsten Nachbarn auf, in 2.212 Fällen war dies *P. tricornutum*. Wurde im Weiteren die Gruppe der beiden photosynthetischen Protisten betrachtet, so wurden 172 Rotalgenproteine, 238 Grünalgenproteine und 85 Pflanzenproteine identifiziert. Außerdem gab es 83 Proteine mit Mischgruppen aus Grünalgen und Pflanzen und 41 Proteine mit Mischgruppen aus Grünalgen und Rotalgen als nächsten Nachbarn. Zusätzliche 884 Proteine hatten nächste Nachbarn in der Gruppe der nicht-photosynthetischen Protisten.

Als weitere Analyse wurden diejenigen Cluster betrachtet, die mindestens ein Protein von *T. pseudonana* und mindestens ein Protein einer Rotalge enthielten (2.444). Von diesen Proteinen hatten 85 eine Rotalge als nächsten Nachabarn, 53 eine Grünalge und 28 eine Pflanze. *P. tricornutum* war in 1.722 Fällen der nächste Nachbar, in 16 Fällen war es eine Mischgruppe aus Grünalgen und Pflanzen, in 9 eine Mischgruppe aus Grünalgen und Rotalgen. Als nächste Nachbarn für die Gruppe der beiden Protisten wurden 749mal andere Protisten gefunden, 193mal die Rotalge, 122mal Grünalgen, 35mal Pflanzen, 50mal eine Mischgruppe aus Grünalgen und Pflanzen und 41mal eine Mischgruppe aus Grünalgen und Rotalgen. Ähnliche Zahlen wurden für die nächsten Nachbarn der Proteine von *P. tricornutum* ermittelt (siehe Abbildung A.1).

Als nächstes wurde die obige Analyse nochmal durchgeführt, allerdings wurden diesmal die nächsten Nachbarn der Gruppe der Chromalveolaten für alle 4.447

Proteine von *T. pseudonana* identifiziert. 94 Proteine befanden sich in Clustern, die nur aus Chromalveolaten bestanden, in 547 Fällen war eine Grünalge nächster Nachbar dieser Gruppe, in 412 Fällen eine Rotalge. Für 279 Proteine war eine Gruppe aus Grünalgen und Pflanzen nächster Nachbar, in 275 Fällen waren es Pflanzenproteine. 242 Proteine waren nächste Nachbarn zu Cyanobakterien und 112 waren nächste Nachbarn zu einer Gruppe aus Grünalgen und Rotalgen. In 199 Fällen waren die restlichen Protisten nächste Nachbarn der Chromalveolaten. Die größte Gruppe mit 668 Proteinen bestand aus einer Mischgruppe aus verschiedenen Eukaryoten. Diese Gruppe bildete mit 449 Proteinen ebenfalls die größte Gruppe der nächsten Nachbarn, wenn nur die 1.746 Proteine betrachtet wurden, die sich in Clustern befanden, in denen Proteine von Grünalgen oder Pflanzen, Rotalgen, von *P. trichocarpa* und von *T. pseudonana* zu finden waren.

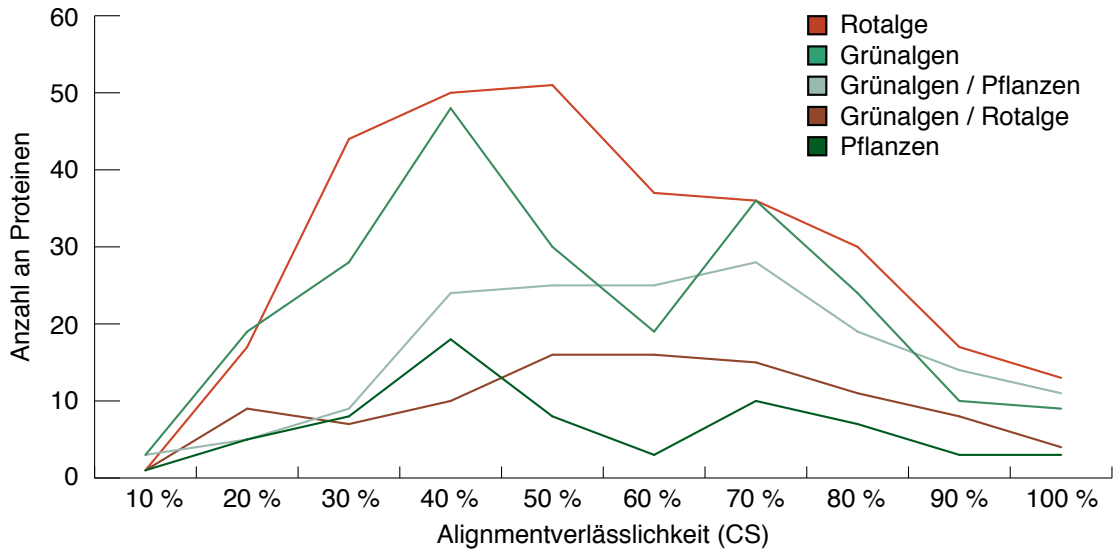

Abbildung 5.15.: Nächste Nachbarn in Pflanzen und Algen von *T. pseudonana* unter Verwendung von verschiedenen CS-Schwellenwerten, das heißt verschiedenen Alignmentverlässlichkeiten (siehe Abschnitt 4.3.7). Für jedes Intervall von CS-Werten wurden die Signale in den Pflanzen, Grünalgen, der Rotalge und den Mischgruppen aufgetragen.

Die nächstgrößere Gruppe bildete die Rotalge mit 296 und die Grünalgen mit 226 Proteinen. Dann folgten die gemischte Gruppe aus Grünalgen und Pflanzen, die 163mal den nächsten Nachbarn bildeten. Die Pflanzen alleine waren 66mal als nächste Nachbarn zu finden und die Mischgruppe aus Grünalgen und Rotalgen 97mal. In 102 Fällen waren die anderen Protisten nächster Nachbar und Cyanobakterien waren 34mal zu finden. Die nächsten Nachbarn von *P. tricornutum* zeigten

eine ähnliche Verteilung. Wurden die Signale für *T. pseudonana*, die aus Pflanzen und Algen kamen, nach dem CS-Wert der Alignments aufgeteilt, so ergab sich die in Abbildung 5.15 gezeigte Verteilung.

Außer in den beiden Intervallen mit den schlechtesten CS-Werten waren mehr nächste Nachbarn bei der Rotalge zu finden als bei den Grünalgen. Die wenigsten Signale in fast allen Intervallen stammten aus Pflanzen. Die Anzahl der nächsten Nachbarn, die aus einer Mischgruppe aus Algen und Pflanzen bestanden, lag in den Intervallen mit höheren CS-Werten über dem der Grünalgen. Die nächsten Nachbarn, die aus einer Mischgruppe aus Grünalgen und Rotalgen zusammengesetzt waren, bildeten in fast allen Intervallen die viertgrößte Gruppe.

5.4.2. Die Oomyceta

T. pseudonana und *P. tricornutum* gehören zusammen mit den Oomyceten *Phytophthora infestans*, *Phytophthora ramorum* und *Phytophthora sojae* zu dem Phylum der Stramenopilen. Die Oomyceten haben ihre sekundäre Plastide im Laufe der Evolution verloren. Gene, die vor diesem Ereignis in das Kerngenom transferiert wurden, sollen durch die nachfolgende Analyse identifiziert werden.

Abbildung 5.16 zeigt die Anteile der 15 definierten Gruppen an den nächsten Nachbarn der Gruppe der Oomyceten (siehe Abbildung 4.9). Die oberen drei Stapeldiagramme zeigen die Anteile der Gruppen an den nächsten Nachbarn der drei Oomyceten in allen Bäumen, in denen ein Protein des Organismus vorkam. Aus *P. infestans* wurden 5.552 Proteine untersucht, aus *P. ramorum* 5.013 und aus *P. sojae* 4.966. Für 1.924 Proteine wurde ein nächster Nachbar in der Gruppe der anderen Protisten gefunden, 645 in der Gruppe der Eukaryoten. 227 Proteine zeigten ein Signal in den Grünalgen, 103 in der Gruppe der Grünalgen und Pflanzen, 254 in der Gruppe der Pflanzen, 27 in der Gruppe der Grünalgen und Rotalgen, 164 in der Rotalge.

Viele nächste Nachbarn stammten außerdem aus der Gruppe der Proteobakterien (621) und den Tieren (481). Für die Proteine der anderen Oomyceten wurden sehr ähnliche Werte ermittelt. So zeigten 1.876 der 5.013 Proteine aus *P. ramorum* nächste Nachbarn in den anderen Protisten, 599 in den anderen Eukaryoten, 227

in den Grünalgen, 103 in der Gruppe der Grünalgen und Pflanzen, 220 in der Gruppe der Pflanzen, 25 in der Gruppe der Grünalgen und Rotalgen, 159 in der Rotalge. Außerdem hatten 418 Proteine nächste Nachbarn in der Gruppe der Proteobakterien und 422 in den Tieren.

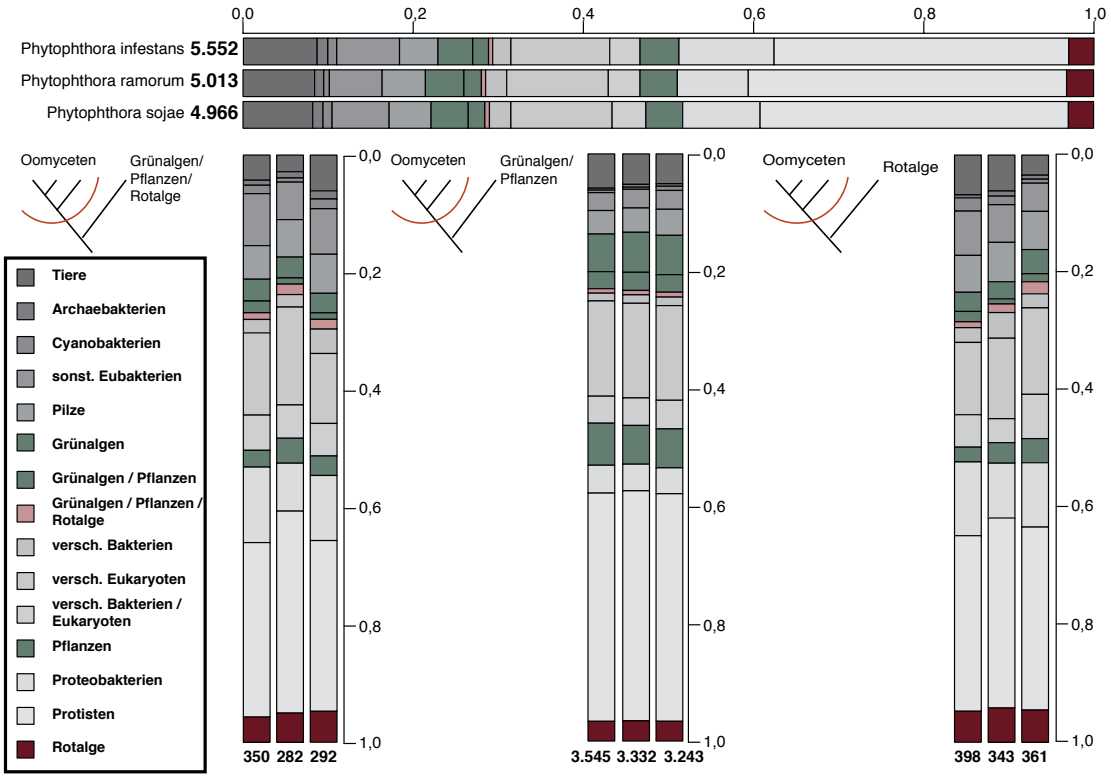

Abbildung 5.16.: Nächste Nachbarn von *Phytophthora infestans*, *Phytophthora ramorum* und *Phytophthora sojae* in verschiedenen Datensätzen. In jedem gestapelten Stapeldiagramm sind die prozentualen Anteile der in der Legende dargestellten Gruppen an den nächsten Nachbarn der Proteine der Oomyceten dargestellt (siehe Abschnitt 4.4.5.1). In den oberen drei Diagrammen wurden alle Bäume verwendet, in denen mindestens ein Protein des jeweiligen Oomyceten vorkam (5.552, 5.013, 4.966). Die unteren linken drei Diagramme zeigen die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein des jeweiligen Oomyceten, ein Grünalgen- oder Pflanzenprotein und ein Rotalgenprotein vorhanden waren (350, 282, 292). Für die mittleren drei Diagramme wurden alle Bäume verwendet, in denen mindestens ein Protein eines der Oomyceten und ein Grünalgen- oder Pflanzenprotein vorkamen (3.545, 3.332, 3.243). Die rechten drei Diagramme zeigen die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein eines der Oomyceten und ein Rotalgenprotein vorhanden waren (398, 343, 361).

Unter den 4.966 Proteinen von *P. sojae* wiesen 1.800 einen nächsten Nachbarn aus der Gruppe der anderen Protisten auf, 591 aus den anderen Eukaryoten. 217 Proteine hatten einen nächsten Nachbarn, der aus der Gruppe der Grünalgen stammte, für 97 Proteine konnte eine Gruppe aus Grünalgen und Pflanzen als

nächster Nachbar ermittelt werden. 215 nächste Nachbarn stammten aus den Pflanzen, 27 aus der Gruppe der Grünalgen und Rotalgen, 147 aus der Rotalge. In den drei linken Stapeldiagrammen sind die Ergebnisse der Analyse der Bäume, die mindestens ein Protein aus dem jeweiligen Oomyceten, aus mindestens einer Rotalge und aus mindestens einer Grünalge oder Pflanze enthielten, zu sehen. Dies entspricht 350, 282 und 292 Proteinen. Ungefähr ein Drittel (104, 97 und 91) dieser Proteine hatten nächste Nachbarn in den anderen Protisten. Für 13, zehn und zwölf Proteine wurde ein Grünalge als nächster Nachbar gefunden, in 15, 14 und 16 Fällen eine Rotalge, in zehn, zwölf und zwölf Fällen eine Pflanze.

Die mittleren drei Diagramme zeigen die nächsten Nachbarn der Proteine aus den Oomyceten, die in den Bäumen vorkamen, die mindestens ein Protein des jeweiligen Oomyceten enthielten und mindestens ein Protein aus einer Grünalge oder Pflanze. Für diesen Fall wurden 3.545, 3.332 und 3.243 Proteine analysiert. Mehr als ein Drittel der Proteine (1.378, 1.305, 1.257) zeigten einen nächsten Nachbarn in der Gruppe der Protisten, 227, 227 und 217 Proteine in den Grünalgen, 119, 115 und 109 in der Rotalge, 254, 220 und 215 in den Pflanzen. Weitere 103, 103 und 97 Proteine hatten einen nächsten Nachbarn in der Gruppe der Grünalgen und Pflanzen, 27, 25 und 27 in der Gruppe der Grünalgen und Rotalgen.

Die rechten Graphen zeigen die Signale in den Bäumen, in denen mindestens ein Protein des jeweiligen Oomyceten und mindestens ein Protein aus einer Rotalge enthalten waren. Es wurden 398, 343 und 361 Proteine untersucht, von denen 21, 20 und 19 ein Signal in der Rotalge aufwiesen, 13, 10 und 12 in den Grünalgen, 10, 12 und 12 in den Pflanzen.

Als nächstes wurden die Signale für diejenigen Proteine (350, 282, 292), die in den Bäumen enthalten sind, die mindestens ein Protein des jeweiligen Oomyceten, mindestens ein Protein aus einer Rotalge und mindestens ein Protein aus einer Grünalge oder Pflanze enthielten, nach unterschiedlichen CS-Schwellenwerten (siehe Abschnitt 4.3.7) unterteilt. Abbildung 5.17 zeigt für jeden Oomyceten und 10 verschiedene CS-Schwellenwerte – angefangen bei 0 % bis zu 90 % – die Anteile der 15 Gruppen an den nächsten Nachbarn dieser Proteine. Unter den Proteinen von *P. infestans* zeigten ohne Berücksichtigung eines CS-Schwellenwerts 13 Proteine ein Signal in den Grünalgen.

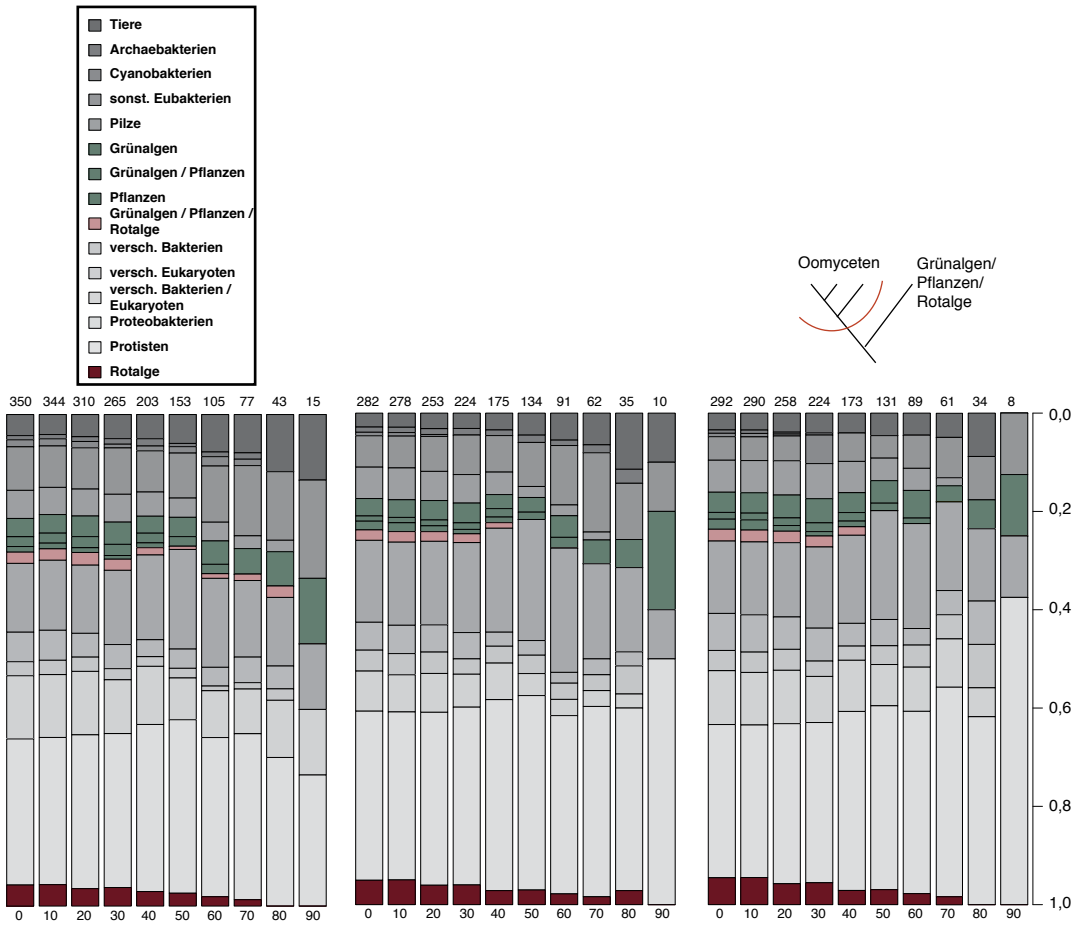

Abbildung 5.17.: Nächste Nachbarn von *Phytophthora infestans*, *Phytophthora ramorum* und *Phytophthora sojae* unter Verwendung von verschiedenen CS-Schwellenwerten. Die analysierten Bäume mussten mindestens ein Protein eines Oomyceten, mindestens einer Grünalge oder Pflanze und mindestens einer Rotalge enthalten. In jedem Stapeldiagramm sind die prozentualen Anteile der in der Legende dargestellten Gruppen an den nächsten Nachbarn der Proteine von *P. infestans* (links), *P. ramorum* (mitte) und *P. sojae* (rechts) dargestellt. Die Zahlen unter den Graphen geben an welcher CS-Schwellenwert (siehe Abschnitt 4.3.7) für den jeweiligen Datensatz verwendet wurde, über den Graphen ist die Gesamtzahl der betrachteten Bäume vermerkt.

Bei einem Schwellenwert von 30 % waren es noch zwölf, bei 60 % fünf und bei 90 % zwei. Da jedoch die Anzahl der Gesamtproteine mit jedem Schwellenwert sank, stieg der prozentuale Anteil dieser Proteine von 3,7 % auf 7 % bei einem Schwellenwert von 80 % und 13,3 % bei 90 %. Der Anteil der Signale in der Rotalge fiel dagegen stetig von 4,3 % auf 0 % bei einem Schwellenwert von 80 % und 90 %. Dies traf ebenfalls auf die Signale in den Mischgruppen zu, der prozentuale Anteil der Signale in den Pflanzen fiel bis zu einem Schwellenwert von 60 % von 2,9 % auf

0,9 % stark ab, stieg in den nächsten Intervallen bis 2,3 % lag aber ebenfalls bei 0 % im letzten Intervall. Das Signal in den Protisten blieb in jedem Intervall annähernd bei 30 %. Die Signale der Proteine der anderen beiden Oomyceten zeigten sehr ähnliche Verteilungen.

5.4.3. Die Chromalveolaten und die Excavata

Abbildung 5.18 zeigt die Signale der in der Datenbank enthaltenen Chromalveolaten, bestehend aus den Bacillariophyta *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* sowie den Oomyceten *Phytophthora infestans*, *Phytophthora sojae* und *Phytophthora ramorum*, den Ciliaten *Paramecium tetraurelia* und *Tetrahymena thermophila* und den Apicomplexa *Theileria parva*, *Plasmodium falciparum* und *Cryptosporidium parvum*. Außerdem wurde der Excavat *Trichomonas vaginalis* untersucht.

In Abbildung 5.18 sind diejenigen 3.791 Cluster abgebildet, in denen mindestens ein Protein der Protisten und mindestens ein Protein aus der Rotalge, den Grünalgen oder den Pflanzen vorkamen. Für diese Proteine wurde der nächste Nachbar der gesamten Gruppe der Chromalveolaten bestimmt (siehe Abbildung 4.9). Unter den nächsten Nachbarn der Bacillariophyta und der Oomyceten waren deutlich mehr Rotalgen- und Grünalgensignale vorhanden als unter den nächsten Nachbarn der anderen Chromalveolaten. So konnten in *T. pseudonana* 320 (299 in den Rückwärtsbäumen) Proteine mit einem Signal in der Rotalge identifiziert werden, in *P. tricornutum* 321 (297). Bei den Oomyceten waren es 219 (206), 197, (174) und 201 (179) Proteine, während in den Apicomplexa 114 (109), 107 (104) und 103 (95) Proteine gefunden wurden. Die Ciliaten zeigten mit 105 (104) und 99 (110) Proteinen ähnliche Werte. In *Trichomonas vaginalis* waren es 93 (90) Proteine. In Grünalgen hatten die Bacillariophyta 487 (484) und 437 (447) nächste Nachbarn, die Oomyceten 359 (371), 334 (340) und 345 (344).

P. tetraurelia und *T. thermophila* zeigten mit 161 (161) und 118 (125) Proteinen höhere Werte als die restlichen Chromalveolaten *T. parva*, *P. falciparum* und *C. parvum* mit 41 (39), 51 (57) und 45 (38) Proteinen. *T. vaginalis* hatte 122 (121) Proteine mit einem nächsten Nachbarn in Grünalgen. Den höchsten Wert in der Gruppe der Pflanzen zeigte *P. infestans* mit 328 (309) Proteinen, wohingegen in den Bacillariophyta 254 (248) und 245 (232) und den anderen Oomyceten 294 (290, 284) Signale in Pflanzen gefunden werden konnten.

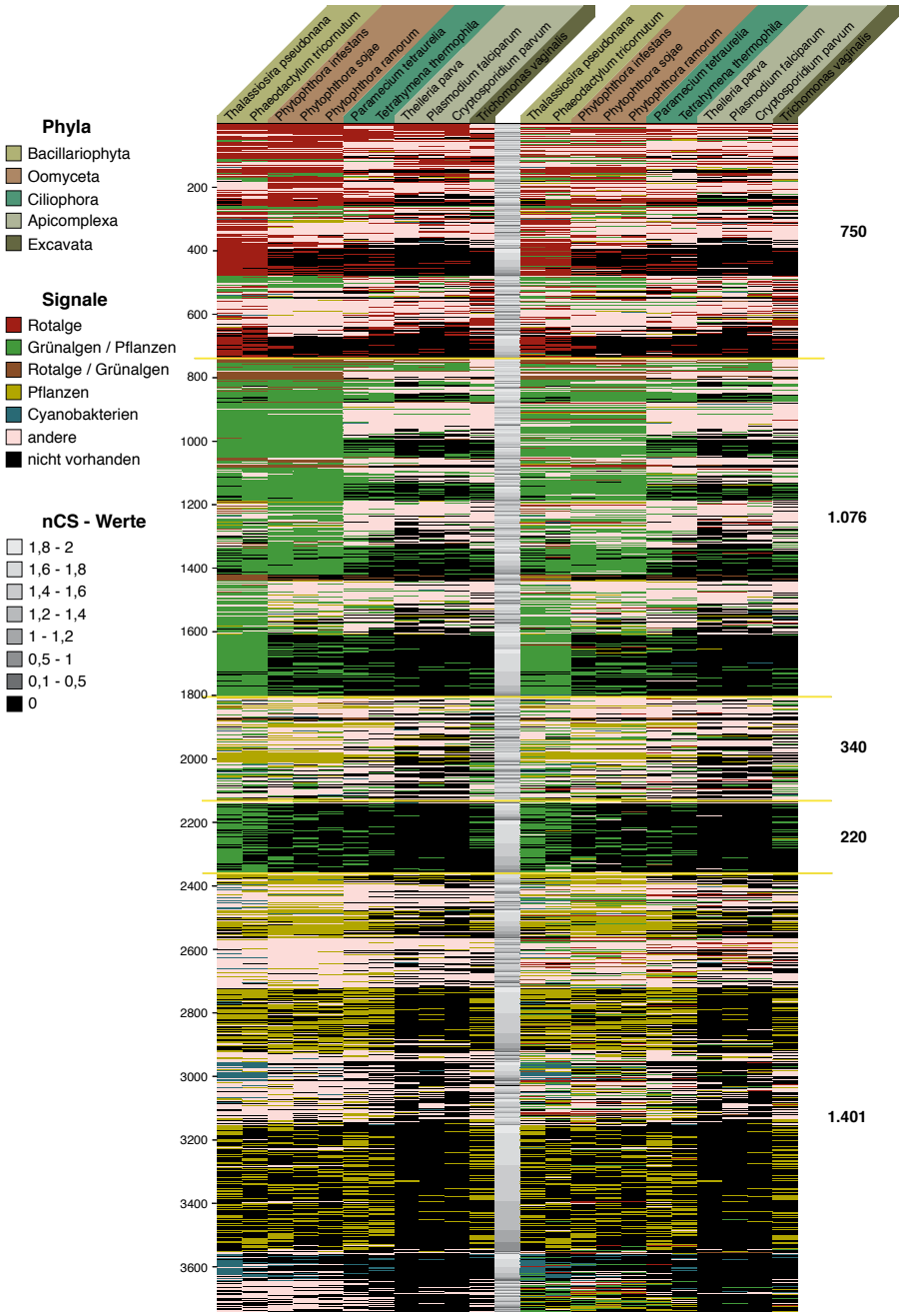

Abbildung 5.18.: Nächste Nachbarn der Chromalveolata und Excavata unter Verwendung von verschiedenen nCS-Werten – Werte, um die Verlässlichkeit der Alignments abzuschätzen – in 3.791 Clustern, in denen mindestens ein Protein der Protisten und mindestens ein Rotalgen-, Grünalgen- oder Pflanzenprotein vorkam. Abgebildet sind die jeweiligen Signale der Proteine der Protisten in Vorwärts- (links) und Rückwärtsbäumen (rechts). Zwischen den Datensätzen sind die nCS-Werte zu finden (siehe Abschnitt 4.3.7). Ein roter Balken markiert ein Protein eines Organismus dessen nächster Nachbar eine Rotalge ist, ein hellgrüner Balken markiert ein Protein mit einer Grünalge oder eine Mischgruppe aus Grünalgen und Pflanzen als nächsten Nachbarn und ein brauner Balken ein Protein mit einem Signal in einer Mischgruppe aus Grünalgen und Rotalgen. Gelb markiert sind die Proteine mit Signalen in Pflanzen, blau in Cyanobakterien, hellrot in einer anderen Gruppe. Ein schwarzer Balken zeigt, dass aus demjenigen Organismus kein Protein vorhanden war. Die Chromalveolaten und Excavata sind nach ihren Phyla sortiert.

Die Werte von *P. tetraurelia* und *T. thermophila* lagen bei 238 (243) und 169 (166), wohingegen die Apicomplexa geringere Werte von 41 (40), 51 (52) und 45 (41) Proteinen aufwiesen. *T. vaginalis* hatte 124 (126) Proteine mit einem nächsten Nachbarn in den Pflanzen. Die Gruppe der Proteine, deren nächster Nachbar eine Mischgruppe aus Grünalgen und Pflanzen war, wurde in Abbildung 5.18 zu den Grünalgensignalen hinzugezählt.

Dies waren in den Bacillariophyta 261 (268) und 258 (260) Proteine, bei den Oomyceten 249 (236), 237 (211) und 240 (222). Die Apicomplexa zeigten 74 (78), 71 (69) und 32 (40) dieser Proteine und die Ciliaten 38 (32) und 39 (38). In einigen Fällen konnte nicht entschieden werden, ob ein Protein ein rotes oder grünes Signal zeigte, da sowohl Grünalgen als auch die Rotalge in der Gruppe der nächsten Nachbarn auftraten. Dies waren bei den Bacillariophyta 99 (123) und 106 (117) Proteine, bei den Oomyceten 93 (106), 82 (97) und 88 (101), bei den Apicomplexa 41 (40), 29 (28) und 19 (22) und bei den Ciliaten 20 (23) und 29 (25) Proteine. Unter den Proteinen von *T. vaginalis* kamen 13 (elf) solcher Fälle vor. Außerdem hatten 166 (101) Proteine von *T. pseudonana* und 72 (65) von *P. tricornutum* nächste Nachbarn in Cyanobakterien.

In den Oomyceten war dies in 28 (28), 31 (31) und 21 (23) Fällen zu beobachten, in den Apicomplexa in zehn (vier), 15 (15) und fünf (fünf) Fällen. Die Ciliaten zeigten bei sieben (sieben) und zwei (fünf) Proteinen ein cyanobakterielles Signal, *T. vaginalis* in sechs (acht). Eine Korrelation zwischen dem nCS-Wert und den Signalen der Proteine konnte nicht beobachtet werden.

Die Cluster konnten in drei Gruppen aufgeteilt werden. In der ersten Gruppe waren 750 Cluster enthalten, in denen hauptsächlich rote Signale teilweise sogar in allen Chromalveolaten zu finden waren. Die Stramenopilen zeigten hier zwischen 77 und 93 rote Signale, die restlichen Protisten zwischen 30 und 53. Die zweite Gruppe enthielt 1.076 und 220 Cluster, in denen überwiegend grüne Signale zu finden waren. In diesen Clustern zeigten die Bacillariophyta und die Oomyceten höhere Werte als die anderen Protisten. Die dritte Gruppe, bestehend aus 340 und 1.401 Clustern zeigen hauptsächlich Signale in den Pflanzen und anderen Gruppen.

Um die auftretenden Muster in diesen Signalen zu untersuchen, wurden die aus den Vorwärtsbäumen gewonnen Werte von Abbildung 5.18 einer hierarchischen

Clusteranalyse unterzogen. Hierbei wurden sowohl die Zeilen und somit die Cluster als auch die Spalten, also die Organismen, nach ihrer Ähnlichkeit geordnet. Das Ergebnis der Analyse ist in Abbildung 5.19 zu sehen. Auch hier waren die drei Gruppen (Rotalgensignale (rot), Grünalgensignale (grün), Pflanzensignale (gelb)) gut zu erkennen. Zusätzlich fiel auf, dass diese Gruppen noch weiter unterteilt wurden: Die "rote" Gruppe bestand einerseits aus Clustern, in denen rote Signale vorwiegend in den Oomyceten gefunden wurden (151 Cluster). In weiteren 72 Clustern wurden auch rote Signale in den Bacillariophyta gefunden. Der zweite Teil dieser Gruppe bestand aus Clustern, in denen weder für die Oomyceten noch für die Bacillariophyta rote Signale abgeleitet wurden, dafür verteilt in den restlichen Protisten. In 49 Clustern war nur ein rotes Signal in *T. vaginalis* zu finden. In 72 Clustern zeigten beide Ciliaten ein rotes Signal, in 42 nur *P. tetraurelia* und nicht *T. thermophila* und in 34 nur *T. thermophila* und nicht *P. tetraurelia*. Bei den Apicomplexa waren ähnliche Werte zu finden. In 44 Clustern zeigten alle drei ein rotes Signal, in 33 nur *C. parvum*, in 21 nur *P. falciparum* und in 28 nur *T. parva*. In 46 Clustern wurden für Proteine von jeweils zwei Apicomplexa ein rotes Signal bestimmt.

Im unteren Teil der Abbildung sind diejenigen Cluster zu finden, in denen überwiegend die Bacillariophyta rote Signale zeigten. In 62 dieser Cluster war kein Protein der anderen Protisten vorhanden. Dies traf außerdem für 29 Proteine aus *T. pseudonana* und für 21 Proteine aus *P. tricornutum* einzeln zu. Die "grüne" Gruppe wurde in weitere vier Untergruppen unterteilt. In der ersten befanden sich Cluster, in denen die Stramenopilen gemeinsam ein grünes Signal zeigten, die anderen Protisten besaßen entweder das Protein nicht oder zeigten Signale in anderen Gruppen. So kamen in 182 Clustern Proteine aus allen Stramenopilen vor, die ein grünes Signal zeigten. In weiteren 59 Clustern fehlte das Protein in einem der Bacillariophyta, während in den anderen Organismen Proteine mit einem grünem Signal gefunden wurde und in 139 Clustern kamen Kombinationen der beiden Gruppen mit mindestens einem grünen Signal in den Bacillariophyta und in den Oomyceten vor.

Die zweite Untergruppe bestand aus Clustern, in denen nur die Proteine der Bacillariophyta einzeln oder gemeinsam ein grünes Signal zeigten. Besonders zu beachten ist hier eine Gruppe von 130 Proteinen, für die in beiden Bacillariophyta ein grünes Signal abgeleitet werden konnte, deren Homologe jedoch nicht in den anderen Protisten vorkamen.

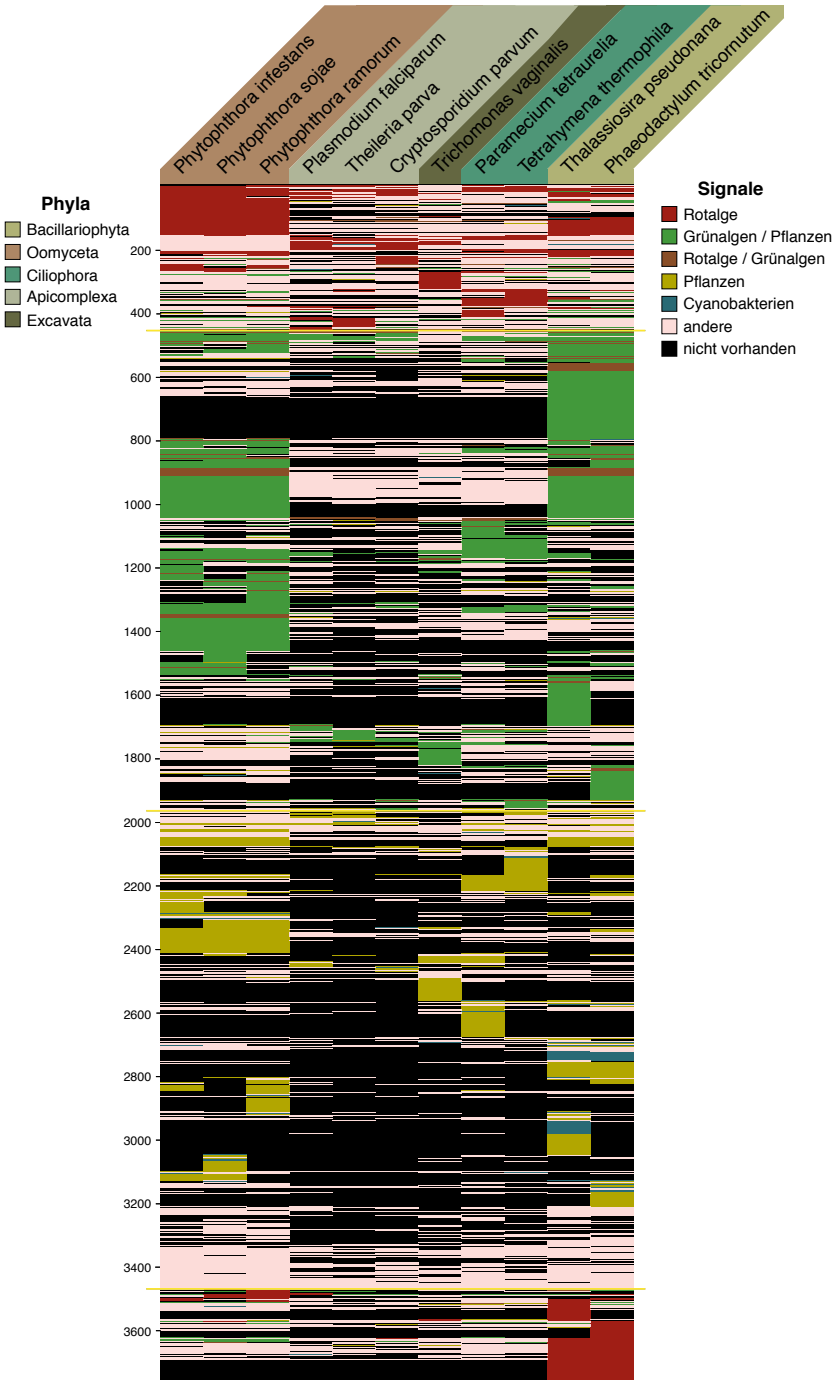

Abbildung 5.19.: Nächste Nachbarn der Chromalveolata und Excavata in 3.791 Clustern, in denen mindestens ein Protein der Protisten und mindestens ein Protein aus der Rotalge, den Grünalgen oder den Pflanzen vorkam. Abgebildet sind die jeweiligen Signale der Proteine aus Protisten in den Vorwärtsbäumen. Ein roter Balken markiert ein Protein dessen nächster Nachbar ein Rotalgenprotein ist, ein hellgrüner Balken ein Signal aus Grünalgen oder einer Mischgruppe aus Grünalgen und Pflanzen, ein brauner ein Signal aus einer Mischgruppe aus Grünalgen und Rotalgen. Gelb markiert sind die Proteine mit Signalen in Pflanzen, blau in Cyanobakterien, hellrot in einer anderen Gruppe. Ein schwarzer Balken zeigt, dass aus demjenigen Organismus kein Protein vorhanden war. Die Chromalveolaten und Excavata sind nach einem hierarchischen Clusteralgorithmus sortiert, ebenso die Zeilen, so dass der Graph nach dem Auftreten gleicher Muster geordnet ist.

Nochmal 83 Proteine kamen nur in *T. pseudonana* vor und zeigten ein grünes Signal, ebenso 51 Proteine, die nur in *P. tricornutum* zu finden waren. Die dritte Untergruppe zeigte Cluster, in denen vorwiegend die Proteine der Oomyceten unter Ausschluss der Bacillariophyta ein grünes Signal zeigten. Dies ist für alle drei Oomyceten in 31 Clustern der Fall, in 48 Clustern wies jeweils nur das Protein eines Oomyceten ein grünes Signal auf, während das Protein in den anderen Protisten nicht vorhanden war. Die letzte Gruppe beinhaltete Cluster, in denen die Proteine der Ciliaten und Alveolata ein grünes Signal zeigten. Die Proteine der Ciliaten zeigten in 142 Clustern zusammen ein grünes Signal, in 87 nur ein Protein von *P. tetraurelia* und in 39 Clustern ein Protein von *T. thermophila*.

Die Gruppe der Proteine, die einen nächsten Nachbarn in der Gruppe der Pflanzen aufwies, zeigte viele Proteine, die nur innerhalb der Phyla oder nur in einem Organismus vorhanden waren. So gab es kein Protein, dass in allen Protisten ein Signal in der Gruppe der Pflanzen hatte. In 18 Clustern zeigten die Stramenopilen ein Pflanzensignal, in 42 die Bacillariophyta, in 66 *T. pseudonana* und in 40 *P. tricornutum*. Die Proteine der drei Oomyceten hatten in 33 Fällen nächste Nachbarn in den Pflanzen, in 28 Clustern die Proteine beider Ciliaten, in 71 Clustern nur Proteine von *P. tetraurelia* und in 46 nur von *T. thermophila*. In sieben Clustern kamen Proteine aller drei Alveolaten vor, die ein grünes Signal zeigten. In 13, 28 und 19 Clustern waren die Proteine der Organismen jeweils einzeln und mit einem Pflanzensignal zu finden. Für Proteine von *T. vaginalis* konnte in 124 Clustern eine Pflanze als nächster Nachbar bestimmt werden.

Abbildung 5.20 zeigt für die Chromalveolata und Excavata die Anteile der nächsten Nachbarn in verschiedenen Gruppen. Die Gruppen der Rotalge, Grünalgen, Pflanzen, Cyanobakterien und die Mischgruppen aus Grünalgen und Rotalgen und aus Grünalgen und Pflanzen sind farblich gekennzeichnet. Die Verteilung der nächsten Nachbarn auf die verschiedenen Gruppen ist innerhalb der taxonomischen Untergruppen sehr ähnlich, zwischen diesen jedoch deutlich wenn auch meistens nicht signifikant unterschiedlich.

In den Bacillariophyta war das Signal der Grünalgen (12 %) leicht stärker als das der Rotalge (9 %). Außerdem waren 3 % bis 5 % der Proteine nächste Nachbarn zu Cyanobakterien. Dies entsprach 241 und 141 Proteinen, für die ein cyanobakterieller Ursprung abgeleitet wurde. Diese Werte waren um mindestens das Vierfache

höher als in allen anderen Gruppen. Die Oomyceta zeigten ebenfalls mehr nächste Nachbarn in den Grünalgen, Rotalgen und Pflanzen, als durchschnittlich zu erwarten wäre. Der prozentuale Anteil an Signalen in Grünalgen (7 % und 8 %) war auch in dieser Gruppe höher als in der Rotalge (5 % und 6 %).

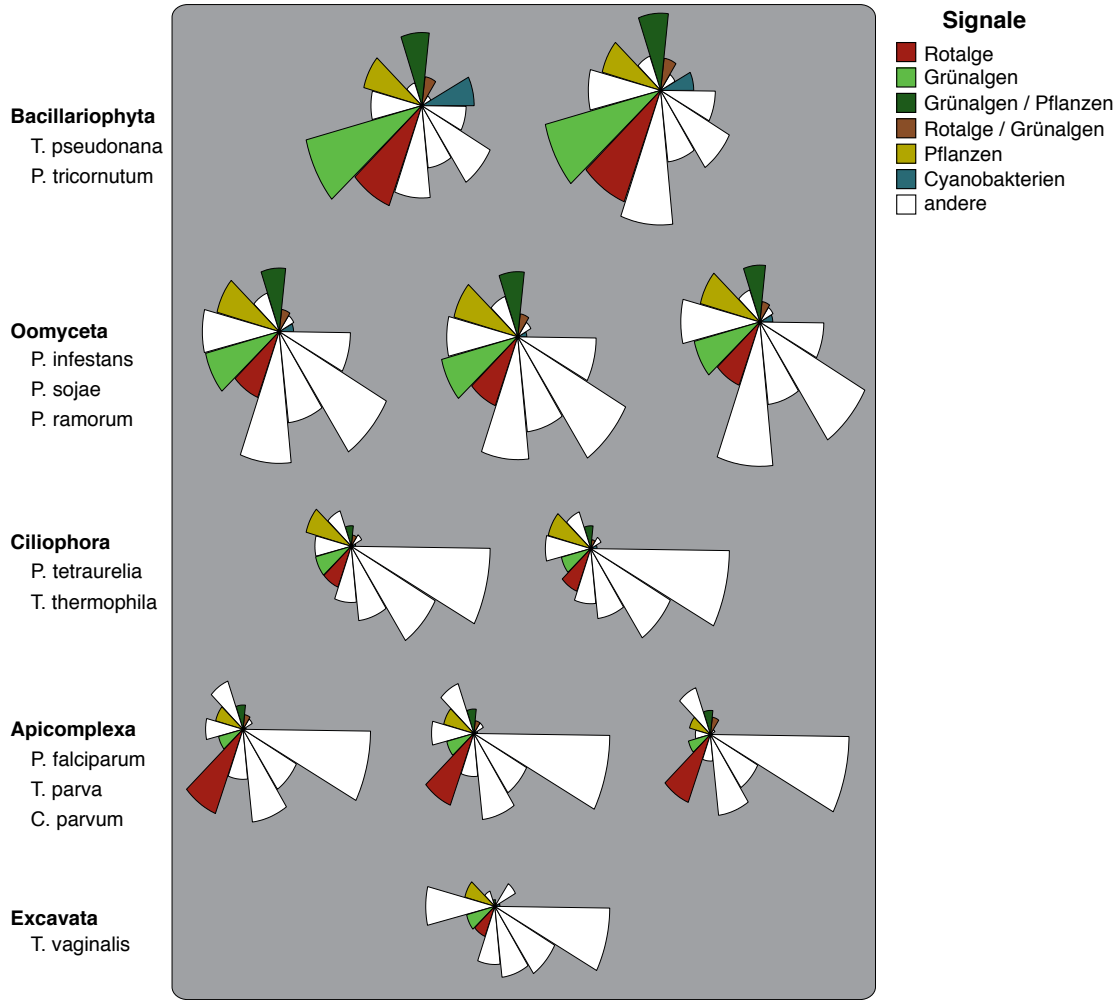

Abbildung 5.20.: Sterndiagramme der Signale der Chromalveolata und Excavata. Jedes einzelne Sterndiagramm zeigt die Anteile der nächsten Nachbarn der Proteine eines Organismus an den Gruppen der Tiere, Protisten, Cyanobakterien, Archaebakterien, Grünalgen und Rotalgen, Grünalgen und Pflanzen, Chromalveolaten, Pflanzen, Eubakterien, Grünalgen, Rotalge, Proteobakterien, Pilze. Die Gruppen der Rotalge, Grünalgen, Pflanzen, Cyanobakterien und die Mischgruppen aus Grünalgen und Rotalgen und aus Grünalgen und Pflanzen sind farblich gekennzeichnet. Je größer ein Segment eines Diagramms ist, desto höher ist der Anteil in der jeweiligen Gruppe.

Allerdings zeigten die Proteine der drei Oomyceten mehr nächste Nachbarn in der Gruppe der Proteobakterien (12 %, 11 %, 10 %), Pilze (7 %, 7 %, 8 %), Tiere (12 %) und Protisten (7 %, 7 %, 8 %). In den Ciliophora waren diese vier Gruppen deutlich

die häufigsten nächsten Nachbarn, während die Rotalge, Grünalgen und Pflanzen unterdurchschnittliche Werte aufwiesen. Die Proteine der Alveolata zeigten ähnlich hohe Werte in der Gruppe der Protisten, allerdings wiesen 8 % der Proteine nächste Nachbarn in der der Rotalge auf aber nur 3 % und 4 % in der Gruppe der Grünalgen. Somit war dies die einzige Gruppe unter den Chromalveolata, in der das Rotalgensignal stärker war als das Grünalgensignal. Der Excavat *T. vaginalis* zeigte nur geringe Signale in den Gruppen der Algen und Pflanzen. Die höchsten Werte waren in den Gruppen der Protisten (17 %), Eubakterien (10 %), Tiere (10 %) und Proteobakterien (7 %) zu finden.

Aus allen 90.354 berechneten Bäumen wurde ein Supernetzwerk berechnet, indem 1.587.805 Splits aus den Bäumen extrahiert und zu einer Supermatrix zusammengefügt wurden (siehe Abschnitt 4.4.6). Aus dieser Matrix wurde mithilfe des Programms `splitstree` ein Netzwerk erstellt, das die Verwandtschaft der Chromalveolaten, Excavata, Algen und Pflanzen darstellt (Abbildung 5.21). In diesem Netzwerk wurden die taxonomischen Gruppen der Bacillariophyta, Oomyceta, Ciliophora, Apicomplexa, Excavata, Rotalge, Grünalgen und Pflanzen farblich markiert.

Das Netzwerk zeigt die nahe Verwandtschaft der Apicomplexa und Ciliophora. *T. vaginalis* gruppierte ebenfalls mit diesen Organismen. Die Bacillariophyta und Oomyceta waren durch einen großen Split von den anderen Organismen getrennt, ebenso die Grünalgen und Pflanzen. Dieser Split separierte alle in dem Netzwerk enthaltenen Protisten und die Rotalge von den Grünalgen und Pflanzen. Außerdem abgebildet ist die Anzahl der nächsten Nachbarn der Proteine jedes Protisten in sechs verschiedenen Gruppen: Die Rotalge, Grünalgen, Pflanzen und Cyanobakterien sowie den Mischgruppen bestehend aus Grünalgen und Pflanzen und Grünalgen und Rotalgen.

Die Bacillariophyta wiesen mit insgesamt 1.874 und 1.718 Proteinen die meisten nächsten Nachbarn in diesen Gruppen auf. Dann folgten die Oomyceten mit 1.504, 1.378 und 1.402 Proteinen. Nur halb so viele Proteine wurden in den Ciliophora gefunden (863 und 581). 471 Proteine von *T. vaginalis* hatten in diesen Gruppen nächste Nachbarn, sowie 268, 295 und 282 Proteine der Apicomplexa. Die Verteilung auf die einzelnen Gruppen war innerhalb der taxonomischen Gruppen fast identisch zwischen diesen jedoch teilweise sehr unterschiedlich.

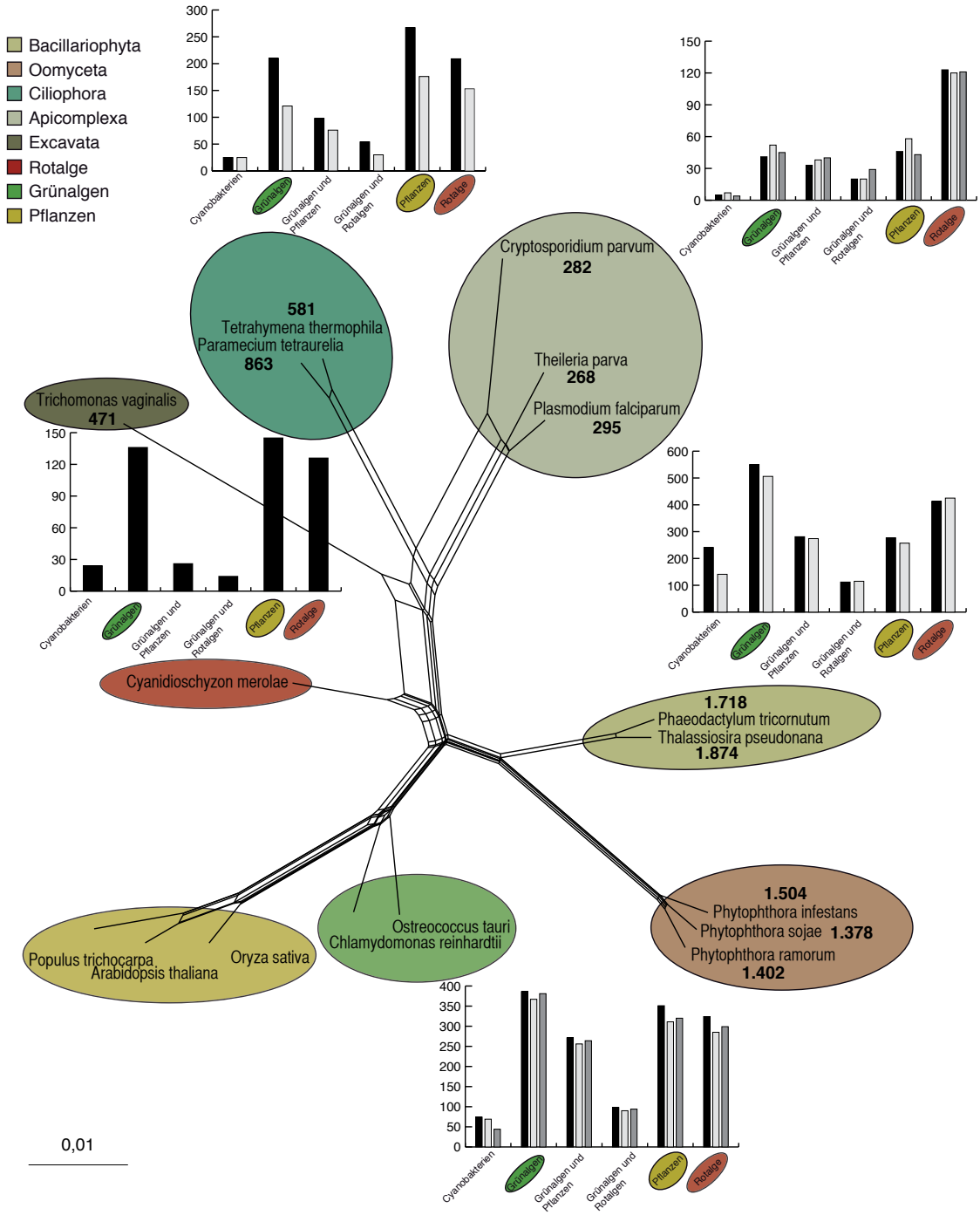

Abbildung 5.21.: Supernetzwerk der Chromalveolaten, Excavata, Rotalge, Grünalgen und Pflanzen erstellt aus 1.587.805 Splits von 90.354 berechneten Bäumen (siehe Abschnitt 4.4.6). Die Splits der Bäume wurden mit `consense` aus den Bäumen extrahiert und zu einer Supermatrix assembliert. Aus dieser Supermatrix wurden mithilfe des Programms `splitstree` Hamming-Distanzen zwischen den Organismen berechnet. Aus diesen wurden die Chromalveolaten, Excavata, Algen und Pflanzen und ausgewählt ein Netzwerk berechnet, das mithilfe eines C-Programms graphisch dargestellt wurde. Neben den Namen der Protisten ist die Anzahl der Proteine abgebildet, die nächste Nachbarn in den Pflanzen, Algen, Cyanobakterien oder Mischgruppen aus diesen zeigten. Die Anzahl der nächsten Nachbarn aus diesen Gruppen ist für jede taxonomische Untergruppe in einem Balkendiagramm dargestellt.

Die Proteine von *T. vaginalis* hatten in 136 Fällen eine Grünalge als nächsten Nachbarn, in weiteren 145 Fällen eine Pflanze und in 126 Fällen eine Rotalge. Die Werte der anderen Gruppen lagen zwischen 14 und 26. Bei den Proteinen der Ciliophora wurden 209 und 153 der nächsten Nachbarn in der Rotalge gefunden, 267 und 176 in den Pflanzen und 120 und 121 in den Grünalgen. 98 und 76 Proteine hatten eine Mischgruppe aus Pflanzen und Grünalgen als nächsten Nachbarn, 54 und 30 eine Mischgruppe aus Grünalgen und Rotalgen und 25 aus Cyanobakterien. Die Apicomplexa waren die einzige Gruppe, bei der mehr nächste Nachbarn in der Rotalge gefunden wurden (123, 120, 121) als in der Gruppe der Grünalgen (41, 52, 45). Die Anzahl der nächsten Nachbarn der Pflanzen lag bei 46, 58, 43. Die Proteine der Bacillariophyta hatten in 413, und 425 Fällen eine Rotalge als nächsten Nachbarn, in 277 und 257 Fällen eine Pflanze und in 550 und 506 Fällen eine Grünalge. Die Cyanobakterien zeigten mit 241 und 141 nächsten Nachbarn in dieser Gruppe das stärkste Signal. 281 und 274 der Proteine hatten eine Mischgruppe aus Pflanzen und Grünalgen als nächsten Nachbarn, 112 und 115 eine Mischgruppe aus Grünalgen und Rotalgen. Die Proteine der Oomyceten zeigten 386mal, 367mal und 381mal ein Signal in den Grünalgen, in 351mal, 311mal und 320mal in den Pflanzen und in 324mal, 285mal und 299mal in der Rotalge. 271, 256 und 264 der Proteine zeigten eine Mischgruppe aus Grünalgen und Pflanzen als nächsten Nachbarn, 98, 90, 94 eine Mischgruppe aus Grünalgen und Rotalgen.

5.4.4. Die Monophylie der Chromalveolaten

Für die Beantwortung der Frage, ob die Chromalveolaten eine monophyletische Gruppe bilden, wurde in einem ersten Schritt das gemeinsame Auftreten der vier verschiedenen Gruppen der Chromalveolaten in 19.094 Clustern, in denen mindestens ein Protein eines der Chromalveolaten vorkam, bestimmt. Das Ergebnis der Analyse ist in Abbildung 5.22 zu sehen. In 5.118 Clustern kamen aus den Chromalveolaten nur Proteine der Bacillariophyta vor, in 849 waren nur Proteine der Oomyceten vertreten, in 3.353 Clustern nur Ciliophora und in 4.208 nur Apicomplexa. In 1.175 Clustern kamen Proteine von Organismen aller vier Gruppen vor. Die Bacillariophyta und die Oomyceten traten in 561 Clustern gemeinsam auf, in 272 Clustern die Bacillariophyta und die Ciliophora, in 824 die Bacillariophyta und die Apicomplexa.

Die Oomyceten und die Ciliophora kamen in 123 Clustern gemeinsam vor,

die Oomyceten und die Apicomplexa in 127 Clustern. Die Ciliophora und die Apicomplexa waren in 1.379 Clustern gemeinsam zu finden, während Proteine der Bacillariophyta, Oomyceten und Ciliophora in 244 Clustern gemeinsam auftraten. Proteine der Bacillariophyta, Ciliophora und Apicomplexa waren in 599 Clustern gemeinsam vorhanden, Proteine der Oomyceten, Ciliophora und Apicomplexa waren in 244 Clustern zu finden. Diese Zahlen beinhalten die Cluster, in denen weniger als vier Sequenzen enthalten waren.

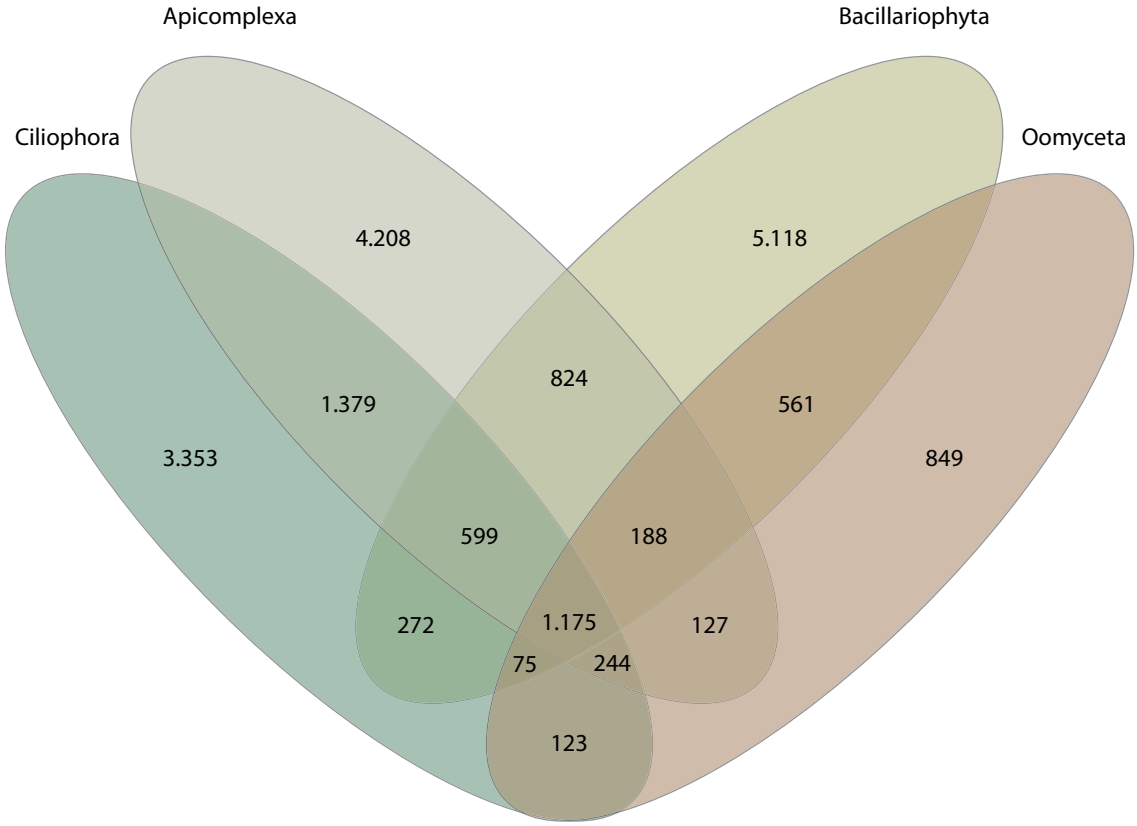

Abbildung 5.22.: Venn-Diagramm des gemeinsamen Auftretens der Bacillariophyta, Oomyceten, Ciliophora und Apicomplexa. Jedes Feld des Venn-Diagramms entspricht einer Kombination des gemeinsamen Auftretens der verschiedenen Taxa in 19.094 Clustern, in denen mindestens ein Protein eines der Chromalveolata vorkam.

Im Weiteren wurde ermittelt, wie oft die Chromalveolaten und die taxonomischen Untergruppen in den berechneten Bäumen monophyletisch waren (siehe Tabelle 5.2). In circa 50 % (582) der Fälle, in denen Chromalveolaten aus allen vier Gruppen in den Clustern vorhanden waren, bildeten die Taxa in den phylogenetischen Bäumen eine monophyletische Gruppe. In 2.435 Fällen bildeten die Bacillariophyta eine monophyletische Gruppe und in 3.343 Fällen die Oomyceten.

Die Ciliaten bildeten 1.627 monophyletische Gruppen, die Apicomplexa 901. Es gab 564 monophyletische Gruppen, in denen Proteine der Bacillariophyta und der Oomyceten vorkamen, 52 mit Proteinen der Bacillariophyta und der Ciliaten, 43 mit Proteinen der Bacillariophyta und der Apicomplexa. Außerdem bestanden 150 Gruppen aus Proteinen der Oomyceten und der Ciliophora und 25 aus den Proteinen der Oomyceten und Apicomplexa. 61 Gruppen wurden von Proteinen der Ciliophora und der Apicomplexa gebildet, während 248 aus Proteinen der Bacillariophyta, Oomyceten und Ciliophora gebildet wurden. Außerdem gab es 94 Gruppen, die Proteine der Bacillariophyta, Oomyceten und Apicomplexa enthielten, 34 enthielten Proteine der Oomyceten, Ciliophora und Apicomplexa.

Um die Werte zwischen den einzelnen taxonomischen Gruppen vergleichen zu können, wurde der Quotient aus der Anzahl der beobachteten monophyletischen Gruppen und der Anzahl der Bäume, in denen mindestens zwei Proteine der jeweiligen Gruppe vorkamen, gebildet.Den höchsten Wert zeigten die Gruppen, in denen Oomyceta und Ciliophora vorkamen. In allen der 150 ermittelten Bäume traten diese beiden Gruppen zusammen monophyletisch auf.

Die Bacillariophyta zeigten mit 90 % ebenfalls einen hohen Wert. Die Proteine der Oomyceten waren in 87 % aller Bäume, in denen mindestens zwei Proteine aus den Oomyceten vorkamen, monophyletisch, die Ciliophora in 78 % und die Apicomplexa in 70 %. Alle Chromalveolaten zusammen waren in 51 % aller Fälle monophyletisch. Das kleinste Verhältnis (24 %) war bei den Oomyceten, Ciliophora und Apicomplexa zu beobachten. Die Bacillariophyta und Ciliophora sind in 27 % der gemeinsamen Bäume monophyletisch, in jeweils 31 % haben Oomyceten und Apicomplexa und Ciliophora und Apicomplexa einen gemeinsamen Vorfahren, die Bacillariophyta sind in 94 von 228 Gruppen also 41 % monophyletisch. In 59 % der gemeinsamen Gruppen gehen Bacillariophyta und Oomyceten auf einen gemeinsamen Vorfahren zurück.

Um diese Werte einordnen zu können, wurden Vergleichswerte aus den Gruppen der Tiere, Pilze und Cyanobakterien berechnet. Außerdem wurde die Anzahl der monophyletischen Gruppen der Archaeplastida mit den vier Gruppen der Chromalveolata ermittelt (siehe Tabelle 5.3). Die einzelnen Gruppen der Tiere, Pilze, Archaeplastida und Cyanobakteria waren in mehr als 70 % aller gemeinsamer Bäume in monophyletischen Gruppen zu finden, wobei die Tiere mit 87 % den

höchsten Wert zeigten, die Archaeplastida mit 71 % den kleinsten. Die Opisthokonta (Tiere und Pilze) zeigten in 71 % der Gruppen Monophylie. Die Archaeplastida mit den Chromalveolata hatten in 47 % der Gruppen einen gemeinsamen Vorfahren, die Archaeplastida und die Opisthokonta in 71 %. Zusammen mit jeweils einer Gruppe der Chromalveolata erreichten die Archaeplastida Werte zwischen 60 % und 65 %, wobei sie zusammen mit den Bacillariophyta den höchsten Wert aufwiesen und zusammen mit den Ciliophora den kleinsten.

Tabelle 5.2.: Anzahl des gemeinsamen Auftretens und der monophyletischen Gruppen der Bacillariophyta, Oomyceta, Ciliophora, Apicomplexa und Kombinationen aus diesen Gruppen.

	gemeinsames Auftreten in einem Baum	**monophyletische Gruppen**	**Verhältnis**
Bacillariophyta	2.712	2.435	0,9
Oomyceta	3.863	3.343	0,87
Ciliophora	2.095	1.627	0,78
Apicomplexa	1.284	901	0,7
Bacillariophyta, Oomyceta	957	564	0,59
Bacillariophyta, Ciliophora	193	52	0,27
Bacillariophyta, Apicomplexa	91	43	0,47
Oomyceta, Ciliophora	150	150	1
Oomyceta, Apicomplexa	80	25	0,31
Ciliophora, Apicomplexa	195	61	0,31
Bacillariophyta, Oomyceta, Ciliophora	550	248	0,45
Bacillariophyta, Oomyceta, Apicomplexa	228	94	0,41
Oomyceta, Ciliophora, Apicomplexa	141	34	0,24
Bacillariophyta, Oomyceta, Ciliophora, Apicomplexa	1.144	582	0,51

Tabelle 5.3.: Anzahl des gemeinsamen Auftretens und der monophyletischen Gruppen der Archaeplastida, Tiere, Pilze, Cyanobakterien und Chromalveolaten.

	gemeinsames Auftreten in einem Baum	**monophyletische Gruppen**	**Verhältnis**
Archaeplastida	6.451	4.566	0,71
Tiere	6.300	5.496	0,87
Pilze	2.095	1.627	0,78
Cyanobakterien	6.323	4.684	0,74
Opisthokonta	7.283	5.206	0,71
Archaeplastida, Opisthokonta	8.849	5.427	0,61
Archaeplastida, Chromalveolata	9.141	4.331	0,47
Archaeplastida, Bacillariophyta	7.241	4.743	0,65
Archaeplastida, Oomyceten	8.271	5.185	0,63
Archaeplastida, Ciliophora	7.465	4.500	0,6
Archaeplastida, Apicomplexa	6.772	4.250	0,63

6 Diskussion

6.1. Die Cluster

In dieser Arbeit sollten die 2.825.466 Proteine von 710 vollständig sequenzierten Genomen miteinander verglichen werden, um die Gene, die durch endosymbiontischen Gentransfer aus den primären Plastiden in die Genome der Archaeplastida und aus den sekundären Plastiden in die Genome der Chromalveolata transferiert wurden. Außerdem sollten so Rückschlüsse auf die Natur der freilebenden Vorfahren der Plastiden gezogen werden.

Da Zufallstreffer zum einen durch die Ermittlung des bidirektionalen besten BLAST-Treffers (BBHs) reduziert wurden, und zum anderen für jedes Proteinpaar die globale Identität berechnet wurde, wurde die Homologiesuche mit BLAST (siehe Abschnitt 4.3.5) ohne Erwartungswert-Schwellenwert durchgeführt. In vielen Analysen werden BLAST-Treffer oder bidirektionale beste BLAST-Treffer als Ausgangsdatensatz für die Clusteranalyse verwendet (Remm et al., 2001; Tatusov et al., 1997*a*, 2000). Der Vorteil dieser Methode ist, dass man jedes Protein mit jedem anderen in annehmbarer Zeit vergleichen kann.

Die Berechnung der globalen Identitäten für jedes mögliche Proteinpaar in diesem Datensatz würde dagegen auch mit dem verbesserten Programm `powerneedle` (siehe Abschnitt 4.4.2) unter Verwendung von 32 Prozessoren circa sieben Jahre dauern. Durch die Verwendung von lokalen Proteinvergleichen kann es jedoch geschehen, dass Proteine mit mehreren funktionellen Domänen Proteine in einem Cluster vereint werden, die ansonsten keine Ähnlichkeit miteinander haben. Hat eine Proteinfamilie eine hohe Ähnlichkeit zu einer Domäne, eine andere zu der anderen Domäne, so ist dies bei der Betrachtung der Erwartungswerte nicht zu ermitteln. Werden jedoch die globalen Identitäten verglichen, so fällt auf, dass

es zwei Gruppen gibt, die innerhalb der Gruppen hohe, zwischen den Gruppen jedoch niedrige Identitäten aufweisen. Das Protein, in dem beide funktionelle Domänen enthalten sind, wird entweder in die Gruppe, zu der es die höhere Ähnlichkeit aufweist, oder in keine der beiden Gruppen eingeordnet. Aus diesem Grund wurde die mangelnde Information für viele Proteinpaare zugunsten besserer Informationen bei den Proteinpaaren, die als BBHs gefunden wurden, hingenommen.

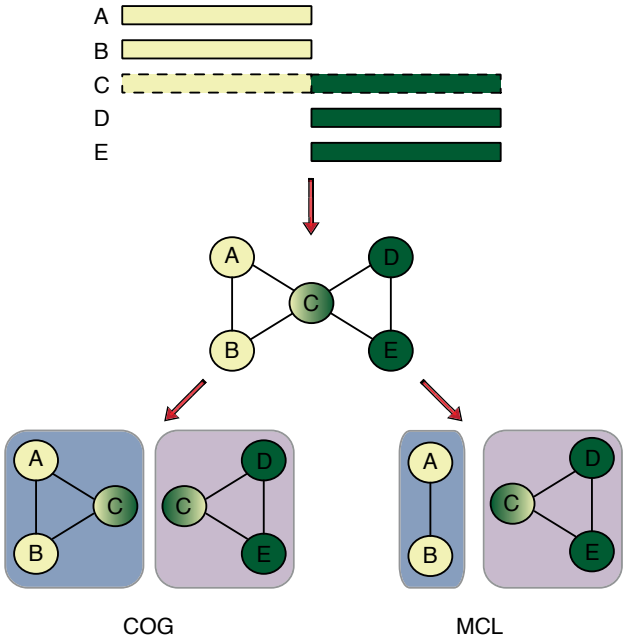

Abbildung 6.1.: Beispiel der Einordnung eines Multifunktionsproteins (C). Die Proteine A und B besitzen eine gemeinsame Domäne ebenso die Proteine D und E. Protein C weist beide Domänen auf und ist deswegen mit allen vier anderen Proteinen verbunden. In manchen Clusterings (z.B. COG, (Tatusov et al., 1997*b*)) werden zwei Cluster erstellt und Protein C ist in beiden vorhanden, `mcl` erstellt ebenfalls zwei Cluster, ordnet dieses Protein aber nur in einen Cluster ein.

Für die Durchführung des Clusterings musste ein Identitätsschwellenwert ausgewählt werden, bei dem möglichst viele Proteine des Ausgangsdatensatzes verwendet wurden, andererseits aber auch nur verlässliche Homologe in die Analyse mit eingingen. Gerade bei tiefen Phylogenien kann die globale Identität der Proteine sehr gering sein. Sequenzen, die eine globale Identität zwischen 20 % und 30 % aufweisen gehören zu einer Grauzone (engl. *twilight zone*), für die es zwar teilweise schwierig aber dennoch möglich ist, gute Alignments zu rekonstruieren (Doolittle, 1986). Andererseits ist die Rekonstruktion phylogenetischer Bäume von

Sequenzen, die weniger als 20 % globale Identität aufweisen sehr schwierig und nicht sehr verlässlich. Sie gehören zu der sogenannten "*midnight zone*" und die leichte Sequenzidentität ist eher auf konvergente Evolution als auf echte Homologie zurückzuführen (Rost, 1999).

Durch die Verwendung des globalen Identitätsschwellenwerts von 20 % wurden 14.307.205 Proteinpaare und 628.579 Proteine nicht verwendet, während es bei einem Schwellenwert von 30 % bereits 55.452.425 BBHs und 761.843 Proteine weniger waren (siehe Tabelle 5.1 und Abbildung 5.1). Aufgrund dieser Überlegungen wurde im Weiteren das Clustering mit dem 20 %-Identitätsschwellenwert verwendet. Zusätzlich zu dem Schwellenwert stellte sich die Frage nach der Wahl des Inflationsparameters I. Dieser bestimmt, wie ähnlich sich die Proteine in einem Cluster sein müssen (siehe 4.6). Wird der Inflationsparameter sehr klein gewählt (z.B. 1,4), so entstehen weniger und größere Cluster mit einer geringeren mittleren Identität der Proteine (van Dongen, 2000*b*). Wird dagegen ein sehr großer Wert wie 4,0 verwendet, so entstehen sehr viele kleine Cluster, die nur sehr nah verwandte Proteine beinhalten.

Wenn ein Cluster einer Proteinfamilie entspricht, so zerfällt diese bei einem Inflationsparameter von 4,0 in kleinere Untergruppen, in denen nur noch Proteine von nah verwandten Organismen vorkommen, und ist somit für eine phylogenetische Analyse, in der Proteine von mehreren Organismengruppen verglichen werden sollen, ungeeignet. Auf der anderen Seite sollte ein Cluster einer Proteinfamilie entsprechen und keiner Superfamilie, in der mehrere ähnliche Proteinfamilien vereinigt sind. Das Problem an der Wahl dieses Parameters ist also, dass keines der entstehenden Clusterings wirklich falsch ist, sondern nur die Granularität verändert wird, die bestimmt in welcher Auflösung die Proteinfamilien vorliegen.

Insbesondere da von den Proteinen, die sich in einem Cluster befinden, Alignments und phylogenetische Bäume berechnet werden sollen, muss darauf geachtet werden, dass die Cluster nicht Superfamilien entsprechen, da sowohl die Rechenzeit als auch die potentiellen Fehler in einem Alignment mit der Anzahl der Sequenzen steigen, die aligniert werden sollen (Landan und Graur, 2007, 2009). Wurde für die vorliegenden Daten kein Schwellenwert und ein Inflationsparameter von 2,0 verwendet, so enthielt der größte Cluster 1024 Proteine. Unter Verwendung des 20 %-Identitätsschwellenwerts waren nur noch 1013 Proteine in dem größten

Cluster, aber die mittlere Größe der Cluster war maximal (siehe Tabelle 5.1). Die Verwendung eines Inflationsparameters von 4,0 reduzierte die Anzahl der Sequenzen im größten Cluster auf 776, allerdings gab es in diesem Clustering auch circa 24.000 Cluster, die nur ein Protein enthielten, während es bei dem 20 %-Clustering mit einem Inflationsparameter von 2,0 nur 516 Cluster der Größe 1 (Waisen) gab.

Zusätzlich zu diesen Überlegungen wurde die Bewertung des Programms `mcl` herangezogen (engl. *jury pruning marks*, Abbildung 4.5). Hierbei wurde offensichtlich, dass das Clustering eher durch einen höheren Identitätsschwellenwert besser wurde als durch die Verwendung eines höheren Inflationsparameters. Der 20 %-Identitätsschwellenwert wurde also verwendet, um zu schwache Verbindungen zwischen den Proteinen herauszufiltern, und ein Inflationsparameter von 2,0, damit keine Superfamilien aber auch keine stark zerfallenen Proteinfamilien erstellt werden.

Obwohl die mittlere Clustergröße für dieses Clustering maximal war (siehe Tabelle 5.1), bestanden mehr als die Hälfte der Cluster aus weniger als vier Proteinen, wodurch sie für eine phylogenetische Analyse nicht geeignet waren, da in einem Baum, der drei Sequenzen enthält, jede Sequenz zu einer anderen der nächste Nachbar sein kann. Dies deutet darauf hin, dass sich viele kleinere Organismengruppen in dem Datensatz befanden, deren Proteine keine weiteren guten Homologe aufwiesen. Eine dieser Gruppen bestand zum Beispiel aus den drei Pflanzen, die 3.331 der Cluster der Größe drei und circa 5.000 Cluster der Größe zwei ausmachten.

Trotz des Identitätsschwellenwerts von 20 % gab es wie in Abbildung 5.3 zu sehen ist, auch Cluster, in denen die mittlere paarweise Identität unter 20 % lag. Dies ist dadurch zu erklären, dass nicht alle Proteine zu jedem anderen Protein in einem Cluster eine Identität über 20 % aufweisen mussten, um in den Cluster eingeordnet zu werden. Hatten manche Proteine in dem Cluster eine paarweise Identität von knapp über 20 %, so dass jedes Protein mindestens mit einem anderen verbunden war, die anderen paarweisen Identitäten aber unter 20 % lagen, so konnte die mittlere Identität aller Sequenzen ebenfalls unter 20 % liegen. Insgesamt handelte es sich um 47 Cluster, deren mittlere paarweise Identität unter dem Schwellenwert lag und 32 dieser Cluster enthielten weniger als zehn Sequenzen, die mittlere Astlänge lag bei all diesen Clustern über einer Substitution pro Position. Dies deutet

darauf hin, dass manche Proteine von `mcl` aufgrund nur sehr geringer Ähnlichkeit in kleine Cluster eingeordnet wurden, obwohl sie keine wirklichen Homologen in diesem Cluster hatten und erklärt auch die stark unterschiedliche Anzahl an Clustern der Größe 1 bei der Verwendung unterschiedlicher Clusterparameter.

Bei der Erstellung von Proteinfamilien aus vollständig sequenzierten Genomen ist jedoch zu erwarten, dass viele Proteine keine Homologen in anderen Organismen aufweisen. Diese fallen zu einem großen Teil bereits bei der BLAST-Suche aus dem Datensatz heraus. Diejenigen Proteine, die leichte Ähnlichkeiten zu anderen Proteinen aufweisen, sollten durch das Clustering identifiziert und vereinzelt werden. Da es sich jedoch nur um 47 von 90.354 Clustern handelte, und diese Cluster in den meisten Analysen nicht verwendet wurden, konnte dieses Verhalten des Programms `mcl` ignoriert werden.

Die Verteilung der mittleren paarweisen Identitäten hatte ihr Maximum in dem Intervall zwischen 40 % und 50 %. Für diese Werte sind gute und verlässliche Alignments rekonstruierbar. Die Verteilung der Alignmentlängen wies mit einem Maximum in dem Intervall zwischen 100 und 200 Aminosäuren ebenfalls eine typische Verteilung der Proteinlängen auf, und fast alle Bäume hatten eine mittlere Astlänge kleiner als eine Substitution pro Position, was auf konservierte Proteine hindeutet.

Die SPS- und CS-Werte, die die Alignmentverlässlichkeit bewerten sollten, wichen allerdings deutlich von den in anderen Analysen erhaltenen Werten ab. In Sternsuchen, in denen nur die Proteine eines Organismus verwendet werden, um homologe Proteinfamilien zu suchen, sind sehr viel mehr Alignments mit schlechten SPS- und CS-Werten zu beobachten (Deusch et al., 2008). In diesen Analysen konnten Korrelationen zwischen der Konserviertheit eines Proteins und den CS- und SPS-Werten ermittelt werden.

Da fast alle Cluster in dieser Analyse im Fall des SPS-Werts über 90 % und im Fall des CS-Werts über 80 % lagen, war eine solche Korrelation nicht zu beobachten. Dies liegt in erster Linie daran, dass so viele, sehr kleine Cluster erstellt wurden. Auch die leichte negative Korrelation zwischen der Anzahl der Sequenzen in einem Cluster und dem CS-Wert spiegelt sich hier wider (siehe Abbildung 5.4). Je weniger Sequenzen in einem Cluster sind, desto weniger Fehler geschehen

während der Berechnung des Alignments und dementsprechend besser sind die Werte der HoT-Analyse. Dies ist jedoch keine Aussage über die Konserviertheit der Proteine. Der SPS-Wert war aufgrund des kleinen Spektrums also nicht geeignet, um unverlässliche Alignments und Bäume zu identifizieren. Der CS-Wert und der nCS-Wert wurden trotzdem weiter verwendet, da es für empirische Daten im Moment keine anderen Möglichkeiten gibt, die Qualität eines Alignments zu bestimmen.

6.2. Die Entfernung der Lücken aus den Alignments

Die Entfernung der Lücken (engl. *gaps*) aus den Alignments ist umstritten. Allerdings wurde in Dwivedi und Gadagkar (2009) gezeigt, dass nur aus Alignments, in denen bis zu 20 % der Positionen Lücken aufweisen, zu 90 % bis 100 % richtige Bäume berechnet werden können. Positionen, in denen Lücken vorkommen, sind oftmals fehlerbehaftet, da viele Alignmentprogramme die Lücken nicht akkurat setzen (Kawakita et al., 2003).

In einigen Analysen wird ein Programm (`Gblocks`, Talavera und Castresana (2007)) verwendet, das je nachdem welche Parameter gewählt werden, nur Positionen aus den Alignments entfernt, in denen in mehr als 50 % der Sequenzen Lücken vorkommen. Zusätzlich zu diesen Positionen werden jedoch auch Positionen entfernt, die diese flankieren. Wenn man von Fehlern bei dem Einfügen der Lücken ausgeht, so finden sich diese Fehler ebenfalls in den flankierenden Positionen. Das Problem bei dieser Vorgehensweise ist jedoch, dass jede Sequenz auf eine andere Art und Weise verändert wird. In manchen Sequenzen bleiben viele Lücken in dem Alignment, in anderen werden die meisten entfernt, wodurch die Distanzen nicht systematisch verändert werden. Das Alignment aus dem dann der phylogenetische Baum berechnet wird, spiegelt im Zweifelsfall nicht mehr die Verwandtschaft der Organismen wider sondern die Anzahl der Positionen mit Lücken in den Sequenzen.

Werden jedoch alle Positionen mit Lücken aus dem Alignment entfernt, entstehen andere Probleme. So gibt es zum einen Alignments, die sehr kurz werden und somit für eine verlässliche phylogenetische Analyse ungeeignet sind. Zum anderen sind die Positionen, die keine Lücken aufweisen, zumeist sehr konservierte Bereiche in dem Protein, so dass für Proteine von nah verwandten Organismen keine

Unterschiede in den Sequenzen mehr gefunden werden können, und der phylogenetische Baum für diese Organismen nicht aufgelöst ist. Trotzdem erscheint dies im Vergleich zu den anderen verfügbaren Methoden als die bessere Art, um fehlende Informationen in Alignments zu behandeln, da teilweise nicht aufgelöste Bäume trotzdem besser sind als fehlerhafte (Castresana, 2000). Blouin et al. (2009) entwickelten kürzlich eine *support vector machine* (SVM), die von dem Benutzer lernt, wie er Alignments manuell korrigiert. Nachdem die Trainingsphase abgeschlossen ist, verbessert das Programm alle anderen Alignments auf dieselbe Art und Weise. Diese Vorgehensweise verhindert, dass bei einer manuellen Korrektur jedes Alignment anders verändert wird.

6.3. Die *heads-or-tails*-Analyse

Vor jeder phylogenetischen Analyse müssen mithilfe von Sequenzvergleichen Alignments erstellt werden, um homologe Positionen in den Sequenzen zu identifizieren. Dies bedeutet, dass das Resultat der phylogenetischen Analyse stark von der Erstellung des Alignments und dessen Genauigkeit abhängt (Hall, 2005; Kumar und Filipski, 2007; Ogden und Rosenberg, 2006; Rosenberg, 2005).

Bisher gibt es jedoch kaum Methoden, um die Genauigkeit eines Alignments objektiv bewerten zu können. In Landan und Graur (2007) wurde die *heads-or-tails*-Analyse (HoT-Analyse) vorgestellt, bei der das Alignment, das aus den Sequenzen in C-N-Orientierung erstellt wurde, mit dem verglichen wird, das aus den Sequenzen in N-C-Orientierung berechnet wurde, verglichen wird. Die zwei Werte, die dabei berechnet werden, vergleichen dabei insbesondere die Anzahl der identischen Spalten in den Alignments (CS-Wert, engl. *column score*) und die Anzahl der identischen Aminosäurepaare (SPS-Wert, engl. *sum-of-pairs score*).

In Deusch et al. (2008) wurde gezeigt, dass Alignments, für die höhere Werte in der HoT-Analyse berechnet wurden, erstens verlässlicher sind und zweitens aus den Sequenzen konservierter Proteine bestehen. Hall (2008) stellte in einer Analyse der Methode jedoch heraus, dass hiermit eigentlich die Reproduzierbarkeit des Alignments getestet wird und nicht wie akkurat es wirklich ist. Es werden nur die Alignments untereinander verglichen, da aber ein optimales Alignment für empirische Daten nicht existiert, kann nicht überprüft werden, wie stark diese von dem optimalen Alignment abweichen. Außerdem wurde deutlich, dass die

wirkliche Akkuratheit der Alignments durch die HoT-Werte überschätzt wird, beide aber trotzdem korreliert sind.

In dieser Arbeit konnten auch leichte Korrelationen zwischen den HoT-Werten und der mittleren Identität der Alignments und der mittleren Astlänge – beides Werte für die Konserviertheit der Proteine – beobachtet werden (siehe Abbildung 5.4), der CS-Wert war jedoch ebenfalls negativ korreliert mit der Anzahl der Taxa in den Alignments. Die Streuung der Werte war sehr klein und im Fall des SPS-Werts war ein Großteil der Werte größer als 80 % (siehe Abbildung 5.3). Dieser Wert wurde in Deusch et al. (2008) als ein Schwellenwert für die Unterscheidung zwischen verlässlichen und unverlässlichen Alignments verwendet. Die Streuung des CS-Werts war höher, die meisten Alignments hatten aber einen CS-Wert von über 90 %. Für eine Unterscheidung zwischen verlässlichen und unverlässlichen Aligments konnten diese Werte also nicht verwendet werden.

Der Hauptgrund für die guten HoT-Werte liegt einerseits in den vielen Alignments mit sehr wenig Taxa (siehe Abbildung 5.2). Mehr als 30.000 Alignments enthielten zwischen 5 und 10 Taxa. Geht man nun davon aus, dass die mittlere Identität der Sequenzen mindestens 20 % beträgt, ist es sehr unwahrscheinlich, dass bei der Berechnung dieser Alignments Fehler geschehen. Je mehr Sequenzen in den Alignments vorkommen, desto mehr Fehler können und werden von den Alignmentprogrammen gemacht (Landan und Graur, 2009). Die Alignments, die aus mehreren 100 Sequenzen berechnet wurden, haben also automatisch einen schlechteren CS-Wert als die Alignments, die nur nur wenige Sequenzen enthalten. Die Verteilung des CS-Werts spiegelt also nicht die erwartete Konserviertheit der Sequenzen beziehungsweise die Verlässlichkeit der Alignments wider, sondern die Anzahl der Sequenzen in den Alignments. Da bei der Berechnung des SPS-Werts ein Aminosäurepaar, das aus einer Aminosäure und einer Lücke besteht, als gleich angesehen wird, besteht keine Korrelation zwischen der Anzahl der Taxa in einem Alignment und dem SPS-Wert.

Der CS- und der SPS-Wert werden aus den Alignments mit Lücken berechnet. Das Alignment, aus dem dann der Baum berechnet wird, enthält jedoch keine Lücken mehr. Deswegen schien es sinnvoller, zu vergleichen, wieviele Informationen beziehungsweise Spalten in beiden Alignments nach der Entfernung der Spalten, die Lücken enthalten, zu finden sind. Auch dieser nCS-Wert (*nogap column*

score) scheint die Zuverlässigkeit der Alignments jedoch stark zu überschätzen und eine Korrelation zwischen dediesem Wert und den ermittelten Signalen in den Bäumen ist nicht erkennbar, weswegen darauf verzichtet wurde, einen Schwellenwert zu verwenden.

6.4. Die Analyse der Bäume

Die Ermittlung des nächsten Nachbarn eines Proteins kann bei 90.354 Bäumen nicht einzeln durchgeführt werden, da das Ergebnis der Analyse objektiv und reproduzierbar sein muss, und bei einer manuellen Inspektion nicht garantiert ist, dass für denselben Fall mehrfach die Entscheidung zugunsten des gleichen Ergebnis getroffen wird. Deswegen musste ein Weg gefunden werden, die Informationen, die in allen Bäumen enthalten sind, zusammenzufassen. Dies wurde mithilfe des Bioperl-Moduls `Bio::TreeIO` durchgeführt, in dem Algorithmen zum Einlesen von phylogenetischen Bäumen implementiert sind.

Ein Nachteil dieses Algorithmus ist die Tatsache, dass der Baum auf jeden Fall gewurzelt werden muss. Beim Einlesen geschieht dies automatisch, je nachdem welches Protein betrachtet werden soll, muss die Wurzel jedoch verändert werden. Ist das zu betrachtende Protein ein direkter Nachfolger der Wurzel, würden ansonsten alle anderen Sequenzen, die in dem Baum vorkommen als nächster Nachbar angesehen. Der Algorithmus zählt alle Sequenzen, die sich unterhalb des Knotens befinden, der der direkte Vorfahr des zu betrachtenden Blattes ist als nächste Nachbarn. Aus diesem Grund wurde der Baum an dem Knoten gewurzelt, der die größte Distanz zu dem zu betrachtenden Blatt hatte.

Diese Vorgehensweise führte nur bei sehr kleinen Bäumen zu Problemen, wenn zum Beispiel für die Bestimmung der grünen und roten Signale in den photosynthetischen Protisten *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* die Bäume analysiert werden sollten, in denen diese beiden Organismen, ein Protein aus einer Rotalge und ein Protein aus einer Grünalge zu finden waren (siehe Abschnitt 3.1.3). Die Wurzel wäre also entweder bei dem Grünalgenprotein oder bei dem Rotalgenprotein gesetzt worden, je nachdem welches die größere Distanz aufwies, und das andere Protein wäre automatisch der nächste Nachbar. Da es sich hierbei allerdings um Bäume handelte, die mithilfe einer Distanzmethode erstellt

wurden, die zum Beispiel mögliche Veränderungen in den Mutationsraten nicht berücksichtigen, ist ein solches Ergebnis nicht sehr verlässlich. Zusätzlich dazu entstanden durch diese Methode der Baumanalyse häufig Gruppen als nächste Nachbarn, die nicht weiter getrennt werden konnten. Die Distanz konnte aus oben genanntem Grund nicht verwendet werden, um aus dieser Gruppe einen einzigen nächsten Nachbarn mit der kleinsten Distanz zu dem zu betrachtenden Protein zu extrahieren, weswegen häufig der nächste Nachbar aus einer Mischgruppe aus Prokaryoten und Eukaryoten bestand (siehe zum Beispiel Abbildung 5.15), die zwar manuell hätten getrennt werden können, dies hätte jedoch wiederum die Objektivität der Analyse gefährdet.

6.5. Der Anteil der cyanobakteriellen Gene im Kerngenom von Pflanzen und Algen

Eine Zielsetzung dieser Arbeit bezieht sich auf die Ermittlung des Anteils der cyanobakteriellen Gene im Kerngenom der sieben inder Analyse enthaltenen Algen und Pflanzen. Für *Arabidopsis thaliana* wurde in früheren Studien ein Anteil von 12,7 % bis 18 % berechnet, für *Cyanidioschyzon merolae* ein Anteil von 17,1 %. Werte für *Oryza sativa* (13,6 %), *Chlamydomonas reinhardtii* (14,2 %) und *Cyanophora paradoxa* (11 %) wurden ebenfalls in anderen Analysen abgeleitet (Archibald, 2006; Deusch et al., 2008; Martin et al., 2002; Reyes-Prieto et al., 2006).

In dieser Arbeit konnten in den Kerngenomen der Pflanzen *Arabidopsis thaliana*, *Oryza sativa* und *Populus trichocarpa* 541, 441 und 607 Proteine identifiziert werden, für die ein cyanobakterieller Ursprung mittels endosymbiontischen Gentransfer angenommen werden können. Für die Grünalgen *Chlamydomonas reinhardtii* und *Ostreococcus tauri* waren es 382 und 298 und für die Rotalge *Cyanidioschyzon merolae* 219. Prozentual gesehen lag der Anteil der cyanobakteriellen Gene in den Algen (10 % - 11 %) höher als in den Pflanzen (8 %). Diese Zahlen sind niedriger, als die bisher abgeleiteten, was mehrere Ursachen hat. Die prozentualen Anteile verändern sich stark, je nachdem welche Untergruppe der Bäume oder Proteine man betrachtet. Werden die in dieser Arbeit ermittelten Werte nur auf diejenigen Bäume bezogen, in denen sowohl Pflanzen als auch Cyanobakterien vorkommen und außerdem sowohl im Vorwärts- als auch im Rückwärtsbaum dasselbe Signal beobachtet werden kann, so erhöhen sich diese Zahlen auf 32 % bis 44 % (siehe Tabelle A.2 und Tabelle A.3).

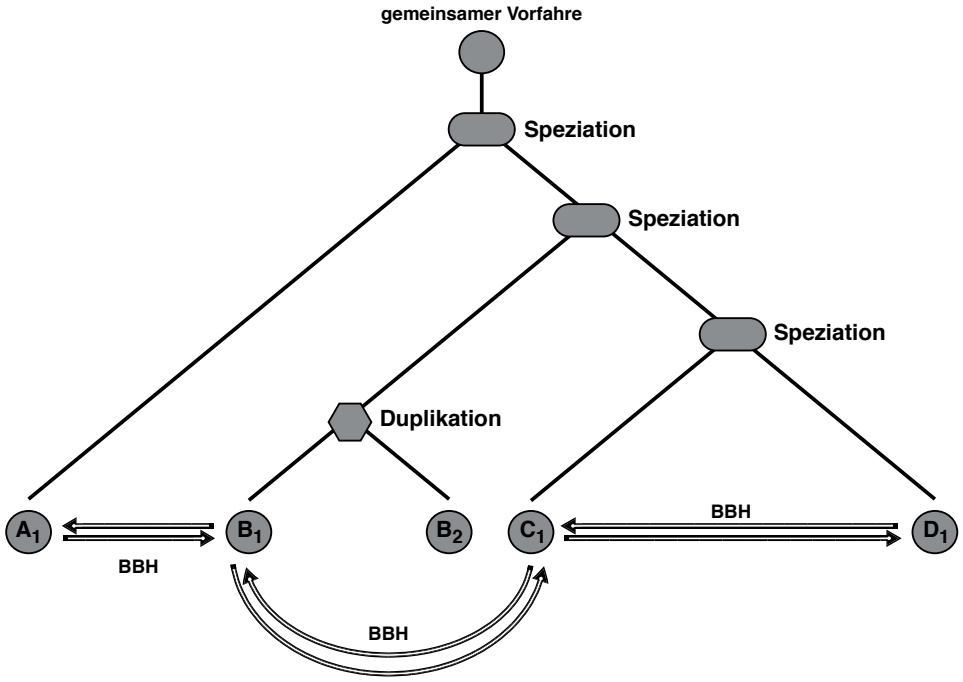

Abbildung 6.2.: Die Entstehung von Paralogen und Homologen durch Speziation und Duplikation

Während der Analyse des Anteils der cyanobakteriellen Gene im Kerngenom der Pflanzen und Algen fiel auf, dass viele Proteine gerade der Pflanzen nicht in die Analyse einbezogen wurden. Von der Pflanze *P. trichocarpa* waren 58.036 Proteine in der Datenbank enthalten, von denen fast die Hälfte schon durch die Verwendung der bidirektionalen BLAST-Suche aussortiert wurden.

Dies zeigt, dass sehr viele Duplikationen unter diesen Proteinen vorhanden waren. Bei der BLAST-Suche wurde nur der beste Treffer in jede Richtung pro Organismenvergleich als Homologie zwischen den Proteinen gewertet (siehe Abbildung 4.3). Da nur solche Proteinpaare in die Analyse eingingen, wurden die duplizierten Proteine nicht gefunden, wenn die Duplikation nach der Trennung eines Organismus von allen anderen Organismen in der Datenbank stattgefunden hat, also in einem Organismus mehrere Kopien vorlagen in dem anderen aber nur eine (siehe Abbildung 6.2).

Viele Proteine aus den Pflanzen waren außerdem in den kleinen Clustern zu finden, von denen keine Bäume berechnet wurden. Diese 3.331 Cluster beinhalte-

ten pflanzenspezifische Proteine, die entweder nur in Pflanzen zu finden waren, oder sich in diesem Phylum stark verändert hatten, so dass sie nicht in einen Cluster mit anderen eukaryotischen Proteinen mit derselben Funktion eingeordnet wurden. Auffällig ist, dass die Proteine der Grün- und Rotalgen keine derart hohe Anzahl an spezifischen Proteinen enthielten (siehe Abbildung 5.6).

Da ein Protein nur ein Signal in Pflanzen geben konnte, wenn alle Sequenzen eines Clusters aus Pflanzen oder Algen stammten, waren dies die pflanzen- und algenspezifischen Proteine, von denen wiederum die Pflanzen bis zu 20 % aufwiesen, wohingegen die einzelligen Grünalgen mit einem kleineren Genom deutlich weniger dieser Proteine zeigten. Die kleine, hochspezifische Rotalge kam nur in sehr wenigen dieser Cluster vor (14,4 % in *P. trichocarpa* bis 18,2 % in *O. sativa*, 9,1 % in *C. reinhardtii*, 8,1 % in *O. tauri*, und 3,1 % in *C. merolae*).

Das starke proteobakterielle Signal von *P. trichocarpa* ist wahrscheinlich auf eine Kontamination zurückzuführen. Entweder waren noch mitochondriale Proteine in dem Datensatz (siehe Abschnitt 3.1.2), der nur die Kernproteine enthalten sollte, oder es ist eine Kontamination während der Sequenzierung des Genoms aufgetreten. Eine andere Möglichkeit ist, dass diese Proteine keine Homologen in dem Datensatz hatten, da *P. trichocarpa* der einzige Baum ist, und sich somit deutlich von den anderen Pflanzen unterscheidet.

Aus Abbildung 5.9 wird ersichtlich, dass es unter den Clustern, in denen ein Protein ein cyanobakterielles Signal zeigte, zwei Gruppen gab: In der einen Gruppe befanden sich diejenigen Cluster, in denen fast nur Proteine von photosynthetischen Eukaryoten und Cyanobakterien vorkamen. In diesen Clustern zeigten alle diese Proteine einen cyanobakteriellen Ursprung. Auf der anderen Seite gab es diejenigen Cluster, in denen Proteine aus sehr vielen Organismengruppen vertreten waren. Es handelte sich also um sogenannte universelle Gene, die in fast jedem Organismus vorkommen. In diesen Fällen ist es eher unwahrscheinlich aber trotzdem möglich, dass das Protein aus dem Chloroplasten in das Kerngenom der Pflanze oder Alge gelangt ist, was durch die vielen nicht cyanobakteriellen Signale deutlich wird.

Deutete ein Protein in einem Cluster auf einen cyanobakteriellen Ursprung hin, alle anderen allerdings nicht, so ist es wahrscheinlich, dass das cyanobakteriel-

le Signal durch Zufall entstanden ist. Allerdings kann diese Beobachtung nicht verallgemeinert werden, da es auch in den großen Clustern Beispiele dafür gibt, dass alle Proteine sowohl in dem Vorwärts- als auch in dem Rückwärtsbaum ein cyanobakterielles Signal zeigten. Hier kann es sich also um ein universelles Protein des Wirts handeln, dass in den Pflanzen und Algen durch ein cyanobakterielles Protein ersetzt wurde (Martin et al., 1993).

In Deusch et al. (2008) wurden für *A. thaliana* 984 Proteine identifiziert, die aus dem Chloroplasten in das Kerngenom transferiert wurden. In dieser Analyse wurden 541 Proteine unter Verwendung der konservativsten Kriterien gefunden. Die niedrigere Anzahl an cyanobakteriellen Proteinen in dieser Analyse kann mehrere Ursachen haben: Zum einen enthielt die Datenbank von Deusch et al. (2008) keine Tiere und Protisten. Zum anderen wurde in der Analyse eine Sternsuche mit jeder Pflanze und Alge einzeln durchgeführt, während für diese Analyse Cluster verwendet wurden. Je nachdem wie ähnlich sich die Proteine der Pflanzen, Algen und Cyanobakterien sind, konnte es bei dem Clustering vorkommen, dass die Homologen in zwei verschiedene Cluster eingeordnet wurden, wodurch sie als reine Pflanzen- und Algencluster in das Ergebnis der Analyse eingingen. Um diese Proteine zu entdecken, müssten Clusterings betrachtet werden, die mit unterschiedlicher Granularität (verschiedenen Inflationsparametern) erstellt wurden.

Außerdem wurde für diese Arbeit eine sehr viel größere Datenbank verwendet, wodurch die einzelnen Cluster teilweise sehr groß wurden (siehe Abbildung 5.2) und eine verlässliche Erstellung der Alignments und Baumrekonstruktion für manche dieser Cluster infrage gestellt werden muss. Da die Proteobakterien einen sehr großen Anteil der Datenbank bildeten, ist es rein statistisch gerade in schlechteren Alignments möglich, dass sie durch Zufall als nächste Nachbarn zu den Pflanzen- und Algenproteine in den Bäumen gefunden wurden. Die in dieser Arbeit als cyanobakteriell identifizierten Proteine sind also eine sehr konservative Schätzung des Anteils der cyanobakteriellen Gene am Kerngenom der Pflanzen und Algen.

Abbildung 5.9 zeigt außerdem, dass der Prozess des endosymbiontischen Gentransfers in jeder Alge und Pflanze unterschiedlich abläuft. Die Gesamtzahl der transferierten Proteine scheint mit der Größe des Kerngenoms korrelliert zu sein. Die prozentualen Anteile innerhalb der taxonomischen Gruppen der Pflanzen, Grünalgen und Rotalgen waren dagegen sehr ähnlich (zwischen 8 % und 11 %

bzw. 32 % und 44 %). Es wurden nur 54 Cluster gefunden, in denen aus allen Pflanzen und Algen ein Protein vorhanden war und dieses einen cyanobakteriellen Ursprung aufwies. Dagegen wurden 85 cyanobakterielle Proteine nur in das Kerngenom von *P. trichocarpa* transferiert. Proteine, die in den Vorfahren der zwei Gruppen – 54 in den Pflanzen und 22 in Grünalgen – transferiert wurden, waren ebenfalls deutlich zu erkennen. Durch die sehr konservativen Kriterien, die die Proteine erfüllen mussten, konnten jedoch nicht alle Proteine gefunden werden und es ist zu erwarten, dass der Anteil der cyanobakteriellen Proteine in den Kerngenomen weitaus höher ist.

6.6. Der Vorfahr der Plastiden

Als nächstes sollte aus den ermittelten cyanobakteriellen Genen in den Kerngenomen der Pflanzen und Algen eine Aussage über den Vorfahren der primären Plastiden getroffen werden.

Zu allen untersuchten Proteinen von sechs der insgesamt sieben Pflanzen und Algen waren Proteine von *A. marina* am häufigsten (1.589mal) als nächster Nachbar zu finden (siehe Abbildung 5.10). Dann folgten die Cyanobakterien der Gruppen III und IV mit 1.476, 1545, 1.533 und 1.529 Gruppen und einige Cyanobakterien, von denen einige eine Nitrogenase – also die Fähigkeit zur Stickstofffixierung – besitzen (zwischen 1.258 und 1.505 Gruppen).

Gloeobacter violaceus (1.009 Gruppen) zeigte in allen Analysen eine Sonderstellung, was in erster Linie durch das Fehlen der Thylakoiden in diesem Organismus zu erklären ist (Sato, 2006; Tomitani et al., 2006). Durch Abbildung 5.11 wird jedoch klar, dass nicht ein einziges Cyanobakterium als Vorfahr identifiziert werden konnte, sondern lediglich eine Gruppe von Cyanobakterien. Dies liegt zum einen daran, dass die Proteine der Cyanobakterien im Vergleich zu denjenigen, die sich im Kerngenom der Pflanzen und Algen befinden, sehr stark ähneln.

Durch die Entfernung der Positionen aus den Alignments, in denen sich Lücken befinden, wurden zusätzlich Unterschiede zwischen den Proteinen der Cyanobakterien entfernt, wodurch in einer phylogenetischen Analyse innerhalb der Gruppe der Cyanobakterien keines identifiziert werden konnte, dass eine höhere Ähnlichkeit zu den Proteinen der Eukaryoten aufwies als eines der anderen. Somit

konnten diese Proteine zwar als cyanobakteriell klassifiziert werden, aus welchem Cyanobakterium das Protein jedoch stammt, ist nicht mehr zu ermitteln. Eine zusätzliche Schwierigkeit ist hier, dass Bakterien Gene über den Mechanismus des lateralen Gentransfers untereinander austauschen (Bapteste et al., 2009; Dagan et al., 2008; Koonin und Wolf, 2008).

Da *A. marina* als einziges Cyanobakterium Chlorophyll d, und somit beinahe Infrarotlicht, anstatt Chlorophyll a und b verwendet, kann es in direkter Nachbarschaft zu anderen Cyanobakterien leben ohne mit ihnen in Konkurrenz zu treten (Swingley et al., 2008). Auf diese Weise können viele Gene durch lateralen Gentransfer in das Genom von *A. marina* gelangt sein, was außerdem das große Genom dieses Cyanobakteriums erklären würde. Aufgrund dieses großen Genoms kommen Proteine von *A. marina* sehr oft in den Gruppen der nächsten Nachbarn vor. Dies deutet jedoch nicht unbedingt darauf hin, dass dies das Cyanobakterium ist, das dem Vorfahren der Plastiden am ähnlichsten ist. Es hat viel mehr den Anschein, als wäre die Eigenschaft Stickstoff fixieren zu können, ausschlaggebend dafür, wie oft Proteine eines Cyanobakteriums in der Gruppe der nächsten Nachbarn der Algen und Pflanzen vorkam.

Diejenigen Cyanobakterien, die die Fähigkeit zur Stickstofffixierung verloren haben (*Prochlorococcus marinus* und *Synechococcus* sp. Stämme), zeigten in allen Analysen einen geringen Anteil an den Gruppen der nächsten Nachbarn. In der Gruppe der Cyanobakterien, deren Genome nif-Gene enthalten, weisen diejenigen mehr Ähnlichkeit mit dem Vorfahren der Plastiden auf, die die größten Genome haben (siehe Abbildung 5.10). Somit gibt es zwei Möglichkeiten die Suche nach dem Vorfahren der Plastiden einzugrenzen: Auf der einen Seite kann die primäre Endosymbiose vor der Aufspaltung der Cyanobakterien geschehen sein. Dann wäre der Vorfahre der Plastiden ebenfalls der gemeinsame Vorfahre aller rezenten Cyanobakterien und die Identifikation eines Cyanobakteriums, das eine Endosymbiose mit einem heterotrophen Eukaryoten einging, fast unmöglich.

Auf der anderen Seite besteht die Möglichkeit, dass die Aufspaltung der Cyanobakterien in die in Abbildung 5.12 gezeigten Gruppen bereits vor der Endosymbiose stattfand. In dem Fall kann auf jeden Fall darauf geschlossen werden, dass der Vorfahr der Plastiden aus der Gruppe der Cyanobakterien stammt, die die Fähigkeit zur Stickstofffixierung besitzen. In dieser Gruppe konnte jedoch

kein Cyanobakterium eindeutig bestimmt werden, dessen Proteine eine höhere Ähnlichkeit zu den cyanobakteriellen Proteinen in den Kerngenomen der Pflanzen und Algen aufwiesen.

In anderen Analysen wurden die Cyanobakterien der Gruppe IV (*Nostocales*) als diejenigen herausgestellt, die dem Vorfahren der Plastiden am ähnlichsten sind (Deusch et al., 2008; Tomitani et al., 2006). Dies kann zwar in dieser Arbeit bestätigt werden, jedoch scheint nicht die Gruppenzugehörigkeit diese Ähnlichkeit hervorzurufen. Da die früheren Analysen sehr viel weniger Genome von Cyanobakterien enthielten, wurde dies erst durch die hier beschriebene Analyse deutlich. So können weder die *Cyanothece* Stämme noch *A. marina*, die alle zu der Gruppe I gehören, als Vorfahr der Plastiden ausgeschlossen werden.

6.7. "Grüne" und "rote" Gene im Kerngenom der Protisten

Die dritte Fragestellung, die in dieser Arbeit untersucht werden sollte, ist die Frage nach dem Verhältnis der "grünen" zu den "roten" Genen in den Genomen der Chromalveolata, um eine Aussage darüber treffen zu können, ob der Vorfahr der sekundären Plastide eine Grün- oder Rotalge war, beziehungsweise ob es Spuren einer seriellen Endosymbiose mit Vertretern beider Gruppen gibt.

6.7.1. Die Bacillariophyta

Die Bacillariophyta *Thalassiosira pseudonana* und *Phaeodactylum tricornutum* sind marine einzellige Algen, die ihre Plastide über eine sekundäre Endosymbiose erworben haben. Sowohl die Struktur als auch phylogenetische Analysen der Proteine der Plastide deuten darauf hin, dass der Vorfahre der sekundären Plastide eine Rotalge war, deren Zellkern im Laufe der Evolution verloren ging (Fast et al., 2001; Harper und Keeling, 2003; Patron et al., 2004). Damit die Protisten die photosynthetischen Eigenschaften der Plastide nutzen konnten, mussten diejenigen Proteine, die aus dem Genom der primären Plastide in das Kerngenom der Rotalge transferiert worden waren, nun in das Kerngenom des Protisten überführt werden. Diese Proteine können durch phylogenetische Analysen identifiziert werden.

Unter den Proteinen von *T. pseudonana* und *P. tricornutum* wurden 241 und 141 Proteine mit potenziellem cyanobakteriellen Ursprung gefunden. Obwohl die

sekundäre Plastide eindeutig der einer Rotalge ähnelt (Cavalier-Smith, 2003), da sie Chlorophyll a und c, das sich aus dem Phycobilin der Rotalgenplastiden entwickelt hat, anstatt Chlorophyll a und b als Antennenproteine verwendet, wurden in mehreren unabhängigen Analysen sowie in dieser Arbeit sehr viele Proteine identifiziert, die eine Grünalge oder Pflanze als nächsten Nachbarn aufweisen (Armbrust et al., 2004; Bowler et al., 2008; Frommolt et al., 2008; Moustafa et al., 2009; Stiller et al., 2009).

In dem Artikel zu der Erstveröffentlichung des Genoms von *T. pseudonana* (Armbrust et al., 2004) wurden durch eine Homologiesuche 2.026 Proteine gefunden, die einem Protein von *A. thaliana* am ähnlichsten waren. 3.783 Proteine wurden in *C. merolae* und *A. thaliana* gefunden und 922 in *Nostoc* sp. PCC 7120, *C. merolae* und *A. thaliana*. Moustafa et al. (2009) identifizierten ca. 1.700 Proteine in den Bacillariophyta, die einen nächsten Nachbarn in der Gruppe der Grünalgen und Pflanzen zeigen. In einer früheren Analyse des Genoms von *T. pseudonana* derselben Autoren zeigen diese sogar, dass die monophyletische Verwandtschaft der Archaeplastida und Bacillariophyta nicht zufällig ist, da fast 70 % der berechneten Bäume eine Monophylie dieser beiden Gruppen zeigten (Moustafa et al., 2008). In Moustafa et al. (2009) postulieren sie nun aufgrund dieser Beobachtung eine zweite sekundäre Endosymbiose des Vorfahren der Bacillariophyta mit einer Grünalge. Diese "grüne" Plastide soll später duch eine "rote" Plastide ersetzt worden sein. Die Proteine, die vor der Endosymbiose mit der Rotalge in das Kerngenom des Protisten aus dem Genom der Grünalge transferiert worden wären, zeigten in der Analyse ein Grünalgensignal, diejenigen Proteine, die nach der Ersetzung transferiert worden wären, ein Signal in der Rotalge. Die Proteine, die einen nächsten Nachbarn in den Pflanzen zeigten, sollen aus dem gemeinsamen Vorfahren der Pflanzen und Bacillariophyta stammen.

Eine andere Hypothese ist, dass die Pflanzen eine Schwestergruppe zu den Chromalveolaten darstellt, und somit wäre ein hoher Anteil an nächsten Nachbarn in der Gruppe der Pflanzen und Grünalgen sogar zu erwarten. In vielen Analysen sind weder die Archaeplastida noch die Chromalveolata monophyletisch, sondern vermischen sich, so dass Burki et al. (2008) diese Gruppe als die "Chromalveolaten-Archaeplastida Megagruppe" bezeichnen (Burki et al., 2007, 2009; Hampl et al., 2009). Auch in dieser Arbeit wurde eine Vermischung der Gruppen festgestellt (siehe Tabelle 5.2). Für die Proteine von *T. pseudonana* wurde zum Beispiel in 281

Fällen nur eine Gruppe aus Grünalgen und Pflanzen und in 112 Fällen eine Gruppe aus Rotalgen und Grünalgen als nächster Nachbar bestimmt. Werden alle Proteine mit nächsten Nachbarn in den Archaeplastida und der Rotalge zusammengezählt (1.633) so übersteigt diese Zahl die Anzahl der nächsten Nachbarn in den Tieren um das Vierfache.

Die Ergebnisse zu der Analyse der cyanobakteriellen Gene in dem Kerngenom der Pflanzen und Algen mit primärer Plastide lassen auch folgende Interpretation zu: Wie in Abbildung 5.9 gezeigt wurde, gab es nur 54 Proteine, die in allen untersuchten Pflanzen und Algen in das Kerngenom transferiert wurden, 22 nur in beiden Grünalgen und 28 nur in beiden Rotalgen. Die häufigsten Muster, die auftraten waren solche, in denen ein Gen nur in einer Pflanze oder Alge transferiert wurde.

Da der Vorfahre der sekundären Plastide nicht bekannt ist, kann nicht ermittelt werden, welche cyanobakteriellen Gene in das Kerngenom dieser Alge transferiert wurden, bevor sie die Endosymbiose einging. Wurde ein Gen in dieser Alge in den Kern transferiert in der Rotalge dieser Analyse jedoch nicht, dafür in einer Grünalge oder Pflanze, so wären letztere als nächste Nachbarn ermittelt worden. Dies ist jedoch kein Indiz für eine zweite Endosymbiose mit einer Grünalge, sondern zeigt, dass Analysen dieser Art stark von den verfügbaren Daten abhängen.

Die Rotalge in dieser Analyse ist hochspezialisiert und besitzt ein stark reduziertes Genom (Matsuzaki et al., 2004). Auch wenn es sich bei den Proteinen, für die ein nächster Nachbar aus den Grünalgen identifiziert wurde, nicht um cyanobakterielle Proteine handelt, so kann ohne die Analyse eines größeren Rotalgengenoms nicht mit Sicherheit behauptet werden, dass diese Proteine wirklich aus einer Grünalge stammen. Die gleiche Argumentation kann ebenfalls auf die hohe Anzahl an Pflanzenproteinen als nächste Nachbarn angewendet werden.

Derzeit sind Grünalgengenome aus nur zwei Ordnungen vollständig sequenziert: *Chlamydomonas reinhardtii* aus den Volvocales und *Ostreococcus tauri* aus den Mamiellales. Auch hier sind also nur wenige Daten verfügbar und die Anzahl der Pflanzenproteine übersteigt die der Algenproteine um das fünffache (siehe Abbildung 4.1). Alleine durch Zufall sollten hier also mehr nächste Nachbarn in der Gruppe der Pflanzen gefunden werden, insbesondere wenn die Chromal-

veolata enger mit den Archaeplastida verwandt sind als mit den Tieren oder Pilzen.

Jiroutová et al. (2007) argumentieren ebenfalls, dass ein Gen in den Kerngenomen der analysierten Rotalgen nach der sekundären Endosymbiose verlorengegangen oder ersetzt worden sein kann, weswegen diese Gene nicht als Proteine aus Rotalgen erkannt werden können. Außerdem ist nichts über die Natur des Endosymbionten bekannt, der sich deutlich von der thermophilen Rotalge in dieser Analyse unterschieden haben könnte. Wenn die Endosymbiose nur kurze Zeit nach der Trennung der Rotalgen und Grünalgen (vor 550 bis 1500 Millionen Jahren (Cavalier-Smith, 2003; Javaux, 2007; Martin et al., 2003; Yoon et al., 2004)) erfolgte, dann kann es sogar sein, dass der Endosymbiont sehr viel Ähnlichkeit zu beiden Linien aufgewiesen hat, wodurch in einer Analyse, die die Genome der heutigen Rotalgen und Grünalgen untersucht, nicht zwischen diesen beiden Linien unterschieden werden kann.

Eine andere Erklärung für die höheren Signale in Grünalgen könnte eine erhöhte Mutationsrate der transferierten Gene oder der Rotalgengene sein. Da die Rotalge in heißen Quellen vorkommt, könnte das Kerngenom unter erhöhtem Selektionsdruck stehen, wodurch die Proteine in einer phylogenetischen Analyse eine höhere Distanz zu den Proteinen aus den Bacillariophyta berechnet würde, und die Grünalgen- oder Pflanzenproteine als nächste Nachbarn erkannt würden.

Eine Vermutung, die in dieser Arbeit nicht bestätigt werden konnte, ist dass das überdurchschnittlich hohe grüne Signal aus der Verwendung von unzuverlässigen Alignments resultiert (Dagan und Martin, 2009). In den Ergebnissen der phylogenetischen Analysen zeigte sich keine Abhängigkeit zwischen den HoT-Werten, die als Maß für die Alignmentverlässlichkeit gelten (Landan und Graur, 2007), und der Gruppe, aus der der nächste Nachbar stammte. In Deusch et al. (2008) konnte nachgewiesen werden, dass die Alignments konservierterer Proteine höhere HoT-Werte aufweisen, wodurch die aus den Alignments berechneten phylogenetischen Bäume verlässlicher waren.

Wäre das grüne Signal lediglich ein systematischer Fehler in der Analyse, so wäre zu erwarten gewesen, dass die Alignments, aus denen phylogenetische Bäume mit einem nächsten Nachbarn aus der Gruppe der Grünalgen und Pflanzen berechnet wurden, signifikant schlechtere HoT-Werte aufweisen, als diejenigen,

aus denen nächsten Nachbarn zu Proteinen der Rotalge berechnet wurden. Abbildung 5.15 zeigt jedoch keine Unterschiede in den Verteilungen der verschiedenen Signale in verschiedenen Intervallen der CS-Werte der Alignments.

Zusammenfassend kann gesagt werden, dass ein starkes Signal in den Grünalgen und Pflanzen in allen bisherigen Analysen der Proteine der Bacillariophyta zu erkennen ist. Die Theorie, dass dieses Signal auf eine weitere sekundäre Endosymbiose mit einer Grünalge zurückzuführen ist, kann jedoch nicht bestätigt werden. Vielmehr unterstützen die Ergebnisse dieser Analyse die Theorie, dass die Bacillariophyta aufgrund ihrer Fähigkeit zur Photosynthese als Schwestergruppe der Archaeplastida angesehen werden kann und somit das starke grüne Signal für diese Gruppe zu erwarten ist, anstatt "You are what you eat!" (Doolittle, 1998) könnte man also sagen "You are what you do".

Da beide Hypothesen aufgrund der heutigen Datenlage gleichwahrscheinlich sind, sollte sich hier nach "Ockhams Rasiermesser" für die einfachere Hypothese entschieden werden. Hätte es zwei serielle Endosymbiosen gegeben, so hätten wahrscheinlich schon transferierte für die Etablierung der Endosymbiose essentielle Gene aus der Grünalge genutzt werden können. Allerdings stellt sich die Frage, wieso zum Beispiel das Gen ftsZ, das für die Teilung der Plastide zuständig ist und eindeutig aus einer Rotalge stammt (Jiroutová et al. (2007), diese Analyse), nochmal transferiert wurde. Die weitaus einfachere Erklärung ist, dass die photosynthetischen Eukaryoten sehr viele gemeinsame Gene aufweisen, die direkt oder indirekt mit der Photosynthese verbunden sind.

6.7.2. Die Oomyceta

Die Oomyceten gehören wie die Bacillariophyta zu den Stramenopilen und somit zu den Chromalveolaten, deren gemeinsamer Vorfahre eine sekundäre Plastide durch die Endosymbiose mit einer Rotalge erworben haben soll (Cavalier-Smith, 1999; Fry, 2008). Im Gegensatz zu den photosynthetisch aktiven Bacillariophyta haben die Oomyceten diese Plastide jedoch verloren, als sie sich zu Pfanzenparasiten entwickelt haben. Dass der gemeinsame Vorfahre der Oomyceten und Bacillariophyta einen sekundären Endosymbionten, der auf eine Endosymbiose mit einer Rotalge zurückgeht, besessen hat, gilt mittlerweise als gesichert (Cavalier-Smith, 1999; Lane und Archibald, 2008). In einigen Analysen wurden bis zu 855 cyano-

bakterielle als auch Proteine, die eine enge Verwandtschaft zu Rotalgenproteinen zeigen, gefunden (Maruyama et al., 2009; Tyler et al., 2006).

Auch in dieser Arbeit wurden unter den Proteinen der Oomyceten 74, 69 und 44 cyanobakterielle Proteine und 324, 285 und 299 Proteine mit einem nächsten Nachbarn zu den Proteinen der Rotalge identifiziert. Wie auch in den Ergebnissen der Analyse zu den Proteinen der Bacillariophyta zu sehen war, überstieg auch hier der Anteil an denjenigen Proteinen, die einen nächsten Nachbarn in Grünalgen zeigen, den Anteil der aus der Rotalge stammenden Proteinen (siehe Abbildung 5.20). 72 Proteine hatten in allen fünf Stramenopilen eine Rotalge als nächsten Nachbarn. Demgegenüber stehen 76 Proteine, die in beiden Bacillariophyta transferiert wurden, nicht jedoch in den Oomyceten und 22 Proteine, die nur in den Oomyceten transferiert wurden. 10 Proteine zeigen in den Oomyceten ein rotes in den Bacillariophyta jedoch ein grünes Signal, 7 Proteine zeigen in den Bacillariophyta ein rotes und in den Oomyceten ein grünes Signal. Für 182 Proteine aller fünf Organismen kann ein nächster Nachbar in der Gruppe der Grünalgen ermittelt werden, in 173 Fällen nur für Proteine der Bacillariophyta und in 94 Fällen für die Proteine der Oomyceten (siehe Abbildung 5.19).

Diese Zahlen zeigen eindeutig, dass die Oomyceten von einem Organismus abstammen, der eine sekundäre "rote" Plastide besessen hat. In dem gemeinsamen Vorfahren der Bacillariophyta und Oomyceten wurden mindestens 72 Proteine aus dem Rotalgengenom in den Kern transferiert. Der gemeinsame Vorfahre der drei untersuchten Oomyceten verlor die "rote" Plastide, wodurch in den Bacillariophyta über längere Zeit endosymbiontischer Gentransfer stattgefunden hat.

Dies erklärt die höheren Zahlen der cyanobakteriellen und "roten" Gene in den Bacillariophyta. Nach beziehungsweise während der Aufspaltung der Oomyceten in Pflanzenparasiten, die jeweils eine spezifische Pflanzenart befallen und somit unterschiedliche Pathogene benötigen, gingen viele Gene aus dem Kerngenom dieser Organismen verloren und neue Gene, die für die parasitische Lebensweise gebraucht wurden, kamen hinzu (Martens et al., 2008). Dies erklärt, wieso manche der transferierten Gene der Rotalge nur in einzelnen Phytophthoraarten vorkommen. In Obornik und Green (2005) wird außerdem die Hypothese formuliert, dass die sekundäre Plastide in der Gruppe der Oomyceten verloren ging, bevor die Endosymbiose vollständig etabliert war, was auf eine sehr schnelle Diversifikation

der Protisten nach der sekundären Endosymbiose hindeutet, was aber aufgrund der Ergebnisse dieser Arbeit abgelehnt werden kann.

Das starke grüne Signal, das auch in der Gruppe der Oomyceten zu finden war (siehe auch Abbildung 5.20) kann, genau wie das grüne Signal in den Bacillariophyta, durch zwei unterschiedliche Hypothesen erklärt werden. Einerseits kann es eine sekundäre Endosymbiose des Vorfahren der Chromalveolaten oder zumindest der Stramenopilen mit einer Grünalge gegeben haben, die vor der Endosymbiose der Rotalge stattfand. Diese Hypothese kann durch die Ergebnisse dieser Arbeit weder wiederlegt noch bestätigt werden, erscheint aber aufgrund des hohen Verwandtschaftsgrades der Stramenopilen und der Pflanzen eher unwahrscheinlich.

Zum anderen kann das grüne Signal und insbesondere das Pflanzensignal auf die wenigen verfügbaren Daten in den Gruppen der Rot- und Grünalgen zurückgeführt werden. Wenn ein Protein, das aus der Rotalge stammt, die die Endosymbiose mit dem Vorfahren der Chromalveolaten eingegangen ist, nicht in der sequenzierten Rotalgen vorkommt, so ist es sehr wahrscheinlich, dass eine Grünalge als nächster Nachbar gefunden wird. Hat das Protein auch keine Homologen in der Grünalge, so ist der nächste Nachbar mit großer Wahrscheinlichkeit eine Pflanze.

Zusätzlich dazu sind die Rotalge, die Grünalgen, die Pflanzen und die Bacillariophyta die einzigen photosynthetischen Eukaryoten in diesem Datensatz. Die Daten der Oomyceten zeigen zumindest eine photosynthetische Vergangenheit dieser Gruppe. Somit ist es auch sehr wahrscheinlich, dass nur die Organismen dieser Gruppen viele Gene besitzen, die direkt oder auch indirekt mit der Photosynthese zu tun haben.

Diese stärkere Verwandtschaft der Proteine dieser Gruppen im Vergleich zu den anderen Chromalveolata wird ebenfalls in Abbildung 5.21 deutlich. In diesem Netzwerk sind die Stramenopilen die einzige Gruppe der Chromalveolaten, die sich innerhalb der Gruppe der Grünalgen, Rotalgen und Pflanzen befindet. Die anderen Protisten sind durch einen großen Split von diesen Organismen getrennt. Außerdem ist durch einen ebenfalls sehr großen Split gut zu erkennen, wie stark sich die Oomyceten nach ihrer Abspaltung von den Bacillariophyta verändert haben.

6.7.3. Die Ciliata

Die Ciliaten sind eine Gruppe von Protisten, die Algen und Bakterien als Nahrungsquelle verwenden. Die meisten Ciliaten zeigen weder Hinweise darauf, dass ihr Vorfahre eine sekundäre Plastide besessen hat, noch sind bisher überzeugende Spuren von endosymbiontischem Gentransfer berichtet worden (Archibald, 2008; Eisen et al., 2006). Allerdings gibt es eine Gruppe von Ciliaten, zu denen zum Beispiel *Laboea strobila* gehört, die die Chloroplasten ihrer Beute – photosynthetische Cryptophyceen – nicht direkt verdauen, sondern sie im eigenen Cytosol behalten, wo sie weiter Photosynthese betreiben und so auch die Ciliaten für eine Weile ernähren (Agatha et al., 2004; Stoecker et al., 1988). Die Chloroplasten der Cryptophyceen können jedoch nur über eine bestimmte Zeit in den Ciliaten überleben, dann werden sie verdaut und die Chloroplasten anderer Cryptophyceen nehmen ihren Platz ein. Zusätzlich zu dieser – fast autotrophen Ernährungsweise – leben diese Ciliaten besonders in den Wintermonaten auch von anderen Organismen als Beute. Diese Lebensweise wurde besonders in Seen der Antarktis beobachtet, in denen den Organismen im Winter nur sehr wenig Sonnenlicht zur Verfügung steht, um Photosynthese betreiben zu können (Laybourn-Parry, 2002).

Ein anderer mixotroph lebender Ciliat ist *Myrionecta rubra*. In Taylor et al. (1969) wurde berichtet, dass das Cytoplasma von *M. rubra* nur die Chloroplasten einer Cryptophycee aufweist. In einer späteren Analyse wurden jedoch auch die Zellkerne der Cryptophyceen gefunden (Oakley und Taylor, 1978), die jedoch von den Chloroplasten separiert vorlagen. Dies deutet darauf hin, dass auch in diesem Fall noch keine Endosymbiose etabliert wurde, die Kerngenome der Cryptophycee also noch für die Photosynthese gebraucht werden.

Die in dieser Analyse enthaltenen Ciliaten *Paramecium tetraurelia* und *Tetrahymena thermophila* zeigen keine Spuren einer Plastide in ihren Zellen. Da keine Anzeichen für eine sekundäre Endosymbiose in dem Cytosol der Ciliaten zu finden sind, ist es umstritten, ob diese wirklich zu den Chromalveolata gehören. Es gibt bisher nur eine vergleichende Genomanalyse der Proteine von *P. tetraurelia* und *T. thermophila*, in der 16 mögliche Kandidaten für einen endosymbiontischen Gentransfer identifiziert wurden (Reyes-Prieto et al., 2008).

Eine alternative Erklärung für die Existenz dieser Gene in den Kerngenomen

der Ciliaten liefert Archibald (2008), indem er auf die Möglichkeit des horizontalen Gentransfers von der Beute der Ciliaten (Algen) zu den Ciliaten hinweist, wie es bereits für Ciliaten, die in Rinderpansen zu finden sind und für die Protisten *Entamoeba histolytica* und *Trichomonas vaginalis*, berichtet wurde (Alsmark et al., 2009; Doolittle, 1998; Ricard et al., 2006).

In der vorliegenden Arbeit wurden jeweils 25 Proteine gefunden, die nächste Nachbarn in den Cyanobakterien zeigten, 210 und 121 mit nächsten Nachbarn in Grünalgen, 209 und 153 in der Rotalge und 267 und 176 in Pflanzen. Übereinstimmungen mit den Proteinen aus anderen Chromalveolata, die nächste Nachbarn in der Rotalge und den Cyanobakterien zeigten, sind nur sehr spärlich vorhanden (siehe Abbildung 5.19). Obwohl in beiden Ciliaten 25 Proteine mit potentiellem cyanobakteriellen Ursprung gefunden wurden, waren nur fünf dieser Proteine in beiden Ciliaten vorhanden. Nach manueller Inspektion der Bäume wurden 15 Proteine aus *P. tetraurelia* bestimmt, für die die Phylogenie sehr überzeugend einen cyanobakteriellen Ursprung zeigte, obwohl in den meisten Fällen nur wenige Cyanobakterien überhaupt in den Clustern vorhanden waren. *A. marina* war häufig das einzige Cyanobakterium, das ebenfalls in dem Cluster vorkam. Vier diese Proteine sind in *Arabidopsis thaliana* plastidär kodiert.

Die größten Anteile der nächsten Nachbarn an den Proteinen der beiden Ciliaten war in den Gruppen der Protisten, Tiere und Pilze zu finden. Dies deutet darauf hin, dass die Ciliaten – unter anderem durch die Gene für die Cilien – näher mit den nicht-photosynthetischen Eukaryoten verwandt sind, genau wie die Stramenopilen durch die Gene für die Photosynthese öfter mit den photosynthetischen Eukaryoten gruppieren. Die Ergebnisse der Analyse zeigen, dass durchaus cyanobakterielle Gene in den Kerngenomen der Ciliaten gefunden werden können, allerdings waren dies auch in dieser Analyse nur sehr wenige.

Dies deutet entweder darauf hin, dass sie durch horizontalen Gentransfer aus der Beute des Ciliaten in das Kerngenom gelangt sind und sich die Ciliaten schon vor der sekundären Endosymbiose von den anderen Chromalveolaten abgespalten haben, oder dass die Ciliaten sehr früh nach der sekundären Endosymbiose die Plastide verloren haben, beziehungsweise in den Ciliaten die Endosymbiose nicht etabliert wurde.

6.7.4. Die Apicomplexa

Die Apicomplexa sind eine heterotrophe Gruppe der Chromalveolaten, die aus Parasiten der Tiere besteht. Obwohl sie die Fähigkeit verloren haben, Photosynthese zu betreiben, haben einige – wie der Malariaerreger *Plasmodium falciparum* – trotzdem die "rote" Plastide behalten. Das Genom der Plastide besitzt jedoch keine Gene für die Photosynthese, stattdessen sind hier Gene für den Isoprenoid- und Lipidstoffwechsel kodiert (Wilson et al., 1996).

Es gibt sowohl Studien, die zeigen, dass der Apicoplast auf eine Endosymbiose mit einer Rotalge zurückgeht (Cavalier-Smith, 1999; Fast et al., 2001; Williamson et al., 1994) als auch Studien, die eine Grünalge als Vorfahren postulieren (Funes et al., 2002; Köhler et al., 1997). Letzteres würde gegen die Chromalveolatenhypothese sprechen und kann mit den Ergebnissen dieser Arbeit nicht unterstützt werden. In *Cryptosporidium parvum* konnte kein Apicoplast gefunden werden (Barta und Thompson, 2006; Zhu et al., 2000).

Erst in den letzten Jahren wurde ein Organismus entdeckt, der die Eigenschaften des Vorfahren der Apicomplexa und ihrer Schwestergruppe den Dinoflagellaten aufwies (Moore et al., 2008). *Chromera velia* ist ein photosynthetisch aktiver Alveolat, der in Symbiose mit der Koralle *Plesiastrea versipora* lebt. Die nichtphotosynthetischen Gene der sekundären Plastide zeigen eine große Ähnlichkeit zu den Genen der Apicoplasten, während die photosynthetischen Gene starke Ähnlichkeit zu denen der Dinoflagellaten haben. Basierend auf dieser Beobachtung könnte der photosynthetische Vorfahre der Apicomplexa ebenfalls in einer Symbiose mit einem marinen Invertebraten gelebt haben, die sich im Laufe der Evolution zu einem parasitären Lebensstil entwickelt und auf Vertebraten ausgeweitet hat.

Die Vermutung, dass der Apicoplast sich aus der Plastide einer Rotalge entwickelt hat, kann mit den Ergebnissen dieser Arbeit unterstützt werden. Als einzige Gruppe der Chromalveolaten wiesen die Kerngenome der Apicomplexa *P. falciparum*, *T. parva* und *C. parvum* mehr Gene auf, die durch endosymbiontischen Gentransfer aus der Rotalge (120, 123, 121) aquiriert worden sein könnten, als aus einer Grünalge (52, 41, 45) oder Pflanze (58, 46, 43). In der Verteilung der nächsten Nachbarn zu den Proteinen der Apicomplexa wurden keine Unterschiede zwischen den Organismen, die einen Apicomplasten besitzen und denen, die diesen

verloren haben, beobachtet werden, was darauf hindeutet, dass *C. parvum* den Apicoplasten erst vor kurzer Zeit verloren hat. Die meisten nächsten Nachbarn der Proteine von den Apicomplexa stammen aus den Gruppen der Protisten, Pilze oder Tiere oder aus Mischgruppen der Eukaryoten. Es konnten nur sieben, fünf und vier cyanobakterielle Gene identifiziert werden.

Diese Daten zeigen eindeutig sowohl die "rote Natur" des Apicoplasten als auch die Korrelation des grünen Signals mit der Fähigkeit der Photosynthese. Aufgrund ihres parasitären Lebensstils haben die Apicomplexa alle Gene, deren Funktion mit der Photosynthese verbunden ist, verloren und haben neue parasitische Gene erworben. In diesen Kerngenomen wurde dementsprechend kein Hinweis auf eine potentielle Endosymbiose des Vorfahren der Chromalveolata mit einer Grünalge beobachtet.

6.7.5. Die Excavata

Trichomonas vaginalis ist ein Parasit des Menschen der weder jetzt noch früher eine sekundäre Plastide besessen hat (Carlton et al., 2007). Die Proteine dieses Excavaten können also als "Negativkontrolle" für die Analyse der Proteine der Chromalveolaten verwendet werden. Wie Abbildung 5.20 zeigt, wurden in keiner Pflanzen oder Algengruppe viele nächste Nachbarn gefunden.

Die meisten nächsten Nachbarn stammen aus den Protisten, dann folgen die Eubakterien, die Tiere, Pilze und Proteobakterien. Es wurden außerdem 24 Proteine identifiziert, die mit Cyanobakterien gruppierten. Dies zeigt, dass auch in dieser Analyse ein gewisses "Hintergrundrauschen" vorhanden ist. Somit sind manche Proteine durch Zufall in monophyletischen Gruppen mit Proteinen aus Organismen zu finden, die ansonsten keine nahe Verwandtschaft zu den anderen Organismen aufweisen. Dies ist nicht in jedem Fall ein Hinweis auf endosymbiontischen oder horizontalen Gentransfer.

6.7.6. Die Chromalveolata

Die Chromalveolatentheorie von Cavalier-Smith (1999) besagt, dass die sekundären Plastiden in der Gruppe der Chromalveolaten auf nur eine sekundäre Endosymbiose eines heterotrophen Eukaryoten mit einer Rotalge zurückgehen. Dies impliziert außerdem, dass dieser heterotrophe Eukaryot der Vorfahr aller

Chromalveolaten war. Seitdem die Sequenzen der plastidären und Kerngenome der Chromalveolaten verfügbar sind, wird versucht diese Theorie zu beweisen oder zu widerlegen.

Die stärksten Argumente für eine Monophylie der Chromalveolaten wurden in Keeling (2009) dargelegt: Sowohl die Entwicklung der Fructose-6-Phosphat-Aldolase (FPA) als auch der Glycerinaldehyd-3-Phosphat-Dehydrogenase (GAPDH), die in der Glykolyse und dem Calvin-Zyklus benötigt werden, weisen auf eine einzige sekundäre Endosymbiose in der Gruppe der Chromalveolaten hin (Keeling, 2009). In Pflanzen und Algen gibt es sowohl eine plastidäre als auch eine cytosolische Kopie dieser Gene. In Chromalveolaten wurde das cytosolische Gen dupliziert und eine Kopie wurde mit einer plastidären Transitsequenz ausgestattet (Fast et al., 2001; Harper und Keeling, 2003). Derselbe Mechanismus wurde für die FBA der Rot- und Grünalgen beobachtet (Gross et al., 1999), alle Chromalveolaten haben jedoch eine eigene FBA entwickelt und diese dupliziert (Patron et al., 2004). Wären die Chromalveolaten nicht monophyletisch, müsste die Präsenz dieser sehr ähnlichen Gene in den Untergruppen der Chromalveolaten durch lateralen Gentransfer zwischen diesen Organismen oder durch konvergenete Evolution erklärt werden. Letztere Hypothese kann jedoch aufgrund der sehr hohen Ähnlichkeit zwischen den Organismen abgelehnt werden.

Andere Studien, die teilweise die vollständigen Plastidengenome oder die Kerngenome der Chromalveolaten verglichen haben, konnten die Monophylie in viele Fällen nicht bestätigen (Daugbjerg und Andersen, 1997; Ishida et al., 2000; Yoon et al., 2002). Als mögliche Ursachen hierfür wurde die Verwendung einiger schnell evolvierender Gene sowie die lückenhafte Verfügbarkeit der Genomdaten verschiedener Chromalveolaten genannt (Hagopian et al., 2004; Khan et al., 2007). Außerdem haben die Genome gerade die Ciliaten einen sehr hohen AT-Gehalt (72 % in *P. tetraurelia*, citepAury:2006sw), wodurch Homologiesuchen und phylogenetische Analysen auf der DNA-Level erschwert werden.

In der vorliegenden Arbeit wurden 1.175 Cluster identifiziert, in denen mindestens ein Protein eines Protisten der vier Gruppen der Bacillariophyta, Oomyceten, Ciliophora und Apicomplexa vorkamen (siehe Abbildung 5.22). In etwas über der Hälfte (51 %) der erstellten Bäume sind die Chromalveolata monophyletisch zum Ausschluss aller anderen Organismen in den Bäumen (siehe Tabelle 5.2). Dieser

Wert war im Vergleich zu den Werten, die für andere taxonomische Gruppen wie zum Beispiel die Archaeplastida (71 %), Tiere (87 %), Pilze (78 %) und Cyanobakterien (74 %) sehr gering, obwohl die vier einzelnen betrachteten Gruppen der Chromalveolaten ebenfalls höhere prozentuale Anteile zeigten (70 % bis 90 %). Die von Burki et al. (2008) postulierte "Chromalveolaten-Archaeplastida Megagruppe", die auch in Nozaki (2005) postuliert wird, wurde nur in 47 % der gemeinsamen Bäume beobachtet. Verschiedene Kombinationen aus den Gruppen der Chromalveolaten zeigten Werte zwischen 24 % und 31 %.

Aufgrund dieser Beobachtungen müsste die Chromalveolatentheorie abgelehnt werden. Wurde die Zusammensetzung der Organismen, die die Monophylie in den 562 Gruppen störten, betrachtet so fiel auf, dass diese Gruppe in den meisten Fällen (397) nur aus anderen Eukaryoten bestanden. Das heißt, dass entweder diese Proteine in den Eukaryoten so stark konserviert sind, dass die Gruppen nicht mehr aufgelöst werden konnten, oder dass manche Eukaryoten, wie zum Beispiel der Pilz *Encephalitozoon cuniculi*, dazu tendieren, zusammen mit den Proteinen der Protisten in den Bäumen zu gruppieren.

Die Gruppierung mancher Proteine aus den Oomyceten mit den Pilzen läßt sich zum Beispiel durch das Chitin, das in ihren Zellwänden wie auch in den Zellwänden der Pilze vorkommt, erklären. In 412 Bäumen störten die Archaeplastida die Monophylie der Chromalveolata, in 471 Bäumen die Amöbe *Dictyostelium discoideum*, in 491 Gruppen die Pilze. Die Gruppen der Bakterien störten dagegen weniger häufig die monophyletischen Gruppen.

Die Ergebnisse deuten darauf hin, dass die Frage nach der Monophylie der Chromalveolaten, Archaeplastida und der potentiellen "Chromalveolaten-Archaeplastida Megagruppe" in einer vergleichenden Genomanalyse nicht beantwortet werden kann. Die Protisten zeigen aufgrund ihrer stark unterschiedlichen Lebensweisen in Teilen ihrer Gene Ähnlichkeiten zu allen eukaryotischen Gruppen. So tendieren die photosynthetischen Chromalveolaten stärker dazu mit den Archaeplastida zu gruppieren als zum Beispiel die Ciliophora, aus denen viele Proteine eher mit den Tieren und den Pilzen gruppieren. Viele Proteine der Oomyceten zeigen ebenfalls starke Ähnlichkeit zu denen der Pilze, weswegen sie früher als solche bezeichnet wurden.

In den Genomen der Chromalveolata zeigt sich eindrucksvoll, wie stark sich sogar in den Genomen dieser vermeintlich einfachen eigentlich aber hochkomplexen Eukaryoten ihre Lebensweise widerspiegeln und eine verlässliche Rekonstruktion der Phylogenie und somit eine Einordnung dieser Organismen basierend auf den vollständigen Genomen nahezu unmöglich machen.

6.8. Schlussfolgerung und Ausblick

In dieser Arbeit wurden in den Kerngenomen der Pflanzen *Arabidopsis thaliana*, *Oryza sativa* und *Populus trichocarpa* 541, 441 und 607 Proteine identifiziert, für die ein cyanobakterieller Ursprung mittels endosymbiontischen Gentransfers angenommen werden können. Für die Grünalgen *Chlamydomonas reinhardtii* und *Ostreococcus tauri* waren es 382 und 298 und für die Rotalge *Cyanidioschyzon merolae* 219. Prozentual gesehen lag der Anteil der cyanobakteriellen Gene in den Algen (10 % - 11 %) höher als in den Pflanzen (8 %).

Die *heads-or-tails*-Methode zur Ermittlung der Alignmentverlässlichkeit spiegelte bei den aus den Clustern berechneten Alignments eher die Anzahl der Taxa und die mittlere globale Identität der Sequenzen wider als die Konserviertheit, weswegen nur diejenigen Proteine ausgewählt wurden, die sowohl in dem Alignment, das aus den Sequenzen in N-C-Orientierung erstellt wurde, als auch in dem Baum, der aus dem reversen Alignment berechnet wurde, dasselbe Signal zeigten. Aus diesem Grund kann das Ergebnis dieser Analyse als eine sehr konservative Schätzung angesehen werden.

In den Gruppen, die die nächsten Nachbarn der Pflanzen- und Algenproteine in den Bäumen bildeten, kamen diejenigen Cyanobakterien, die die Fähigkeit besitzen, Stickstoff zu fixieren, am häufigsten vor. Insbesondere waren *Acaryochloris marina* und die Nostocales sehr oft in der Gruppe der nächsten Nachbarn zu finden. Die Hypothese, dass der Vorfahr der primären Plastiden die Fähigkeit zur Stickstofffixierung besessen hat, kann unterstützt werden. Allerdings muss aufgrund des starken lateralen Gentransfers infrage gestellt werden, ob es auch mit zusätzlichen Genomen aus den Pflanzen, Algen und Cyanobakterien möglich sein wird, ein konkretes Cyanobakterium als Vorläufer der Plastiden zu identifizieren.

Die Vermutung, dass der Vorläufer der Plastiden aus der Gruppe IV stammt,

kann nicht unterstützt werden, vielmehr zeigt sich eine Verwandtschaft zu allen Cyanobakterien, die Stickstoff fixieren können. Die ersten Genome von Cyanobakterien der Gruppe V wurden kürzlich sequenziert und werden eventuell Aufschluss darüber geben können, ob der Vorfahr der Plastiden wirklich fähig gewesen sein muss, Heterozysten auszubilden, oder ob die Cyanobakterien, die dazu in der Lage sind, nur aufgrund ihres größeren Genoms häufiger als nächste Nachbarn der Pflanzen- und Algenproteine gefunden werden.

Der hohe Anteil an "grünen" Genen für die photosynthetischen Bacillariophyta (550, 506) und die Oomyceten (386, 367, 381) wurde in dieser Arbeit bestätigt. Die Ciliophora, die keine sekundäre Plastide besitzen, zeigten um die Hälfte weniger (210, 121) und die parasitischen Apicomplexa sehr wenige (41, 52, 45) "grüne" Gene. Demgegenüber stehen 425 und 324 "rote" Gene in den Bacillariophyta, 285, 299 und 209 Gene in den Oomyceten, 153 und 123 in den Ciliophora und 123, 120, 121 in den Apicomplexa. Diese Gruppe war die einzige, die einen höheren Anteil an "roten" als an "grünen" Genen aufwies.

Zwar ist es möglich, dass die Apicomplexa als Parasiten alle Gene, die aus einer Grünalge stammten, verloren haben, allerdings stellt sich dann die Frage, wieso nicht ebenfalls die aus einer Rotalge transferierten Gene verloren gegangen sind. Aufschluss über diese Frage könnte das Genom des vermeintlich letzten photosynthetisch aktiven Vorfahren der Apicomplexa – *Chromera velia* – geben. Außerdem könnte die Sequenzierung eines der Ciliaten, die mixotroph in den Seen der Antarktis überleben und dabei Cryptophyceen für einige Zeit in ihrem Cytosol behalten, bevor sie verdaut werden, Aufschluss darüber geben, ob die "roten Gene", die in den Genomen der Ciliaten identifiziert wurden, durch lateralen Gentransfer zu erklären sind, oder ob eventuell der letzte gemeinsame Vorfahre der Chromalveolaten unter den Ciliaten zu suchen ist.

In dieser taxonomischen Gruppe kommen drei verschiedene Stadien auf dem Weg zu einer Endosymbiose vor: Diejenigen Ciliaten, die heterotroph leben, die mixotrophen Ciliaten und *Myrionecta rubra* als bisher einziger phototropher Ciliat. *M. rubra* hat zwar eine tertiäre anstatt ein er sekundären Endosymbiose etbaliert, allerdings ist es nicht schwer sich vorzustellen, dass ein anderer Ciliat eine Endosymbiose mit einer Rotalge anstatt mit einer Cryptophycee eingegangen ist, woraus die Gruppe der Chromalveolaten entstanden sind, andere Ciliaten

lebten jedoch weiter heterotroph oder haben die Plastide vor der vollständigen Etablierung der Endosymbiose wieder verloren.

Wertvolle Informationen für die Ermittlung der aus Rot- oder Grünalgen stammenden Gene in den Chromalveolaten könnte ebenfalls ein größeres Genom einer "normalen" Rotalge wie zum Beispiel von *Porphyra umbilicalis* liefern. Auch wenn das stärkere grüne Signal ebenso in den Bäumen beobachtet wurde, in denen sowohl Proteine der Rotalgen als auch der Grünalgen vorkamen, könnten die Proteine der hochspezialisierten Rotalge in dieser Analyse sich stark von den Proteinen des Vorfahren der Plastide unterscheiden, selbst wenn es sich dabei um eine andere Rotalge handeln sollte.

Die Monophylie der Chromalveolaten, die immer noch heftig diskutiert wird, kann nicht ohne Einwände unterstützt aber auch nicht widerlegt werden. So zeigten die photosynthetisch-aktiven Chromalveolaten viele Proteine, die sich in einer monophyletischen Gruppe mit den Proteinen der Archaeplastida befinden. Die Gruppen der nicht-photosynthetischen Chromalveolaten wurden oftmals durch Proteine der Pilze, Tiere und Archaeplastida gestört. Dies legt die Vermutung nahe, dass sich die Lebensweise der Protisten stark in deren Genomen und somit auch in den Ergebnisse einer vergleichenden Genomanalyse widerspiegelt. So zeigt die Gesamtheit der Proteine der Chromalveolata oftmals ein stark unterschiedliches Bild dieser Organismen, so dass die Herausforderung an eine Genomanalyse in der Suche nach den Gemeinsamkeiten besteht.

Die Sequenzierung vollständiger Genome ist in der letzten Zeit aufgrund der stark verbesserten Technik sehr viel einfacher und vor allen Dingen billiger geworden. Deswegen ist zu erwarten, dass in den nächsten Jahren sehr viele Genome sequenziert werden, die helfen könnten, die noch immer offenen Fragen zur Evolution der Endosymbiose in den Archaeplastida und Chromalveolata zu beantworten. Hierbei wäre in erster Linie die Entdeckung des photosynthetischen Vorfahren aller Chromalveolata, wenn es ihn noch gibt, zu nennen.

A Anhang

Tabelle A.1.: Aufstellung der verwendeten Organismen mit der Anzahl der Proteine pro Organismus und des jeweiligen Phylums

Organismus	# Proteine	Phylum
Acidothermus cellulolyticus 11B	2157	Actinobakteria
Arthrobacter aurescens TC1	4587	Actinobakteria
Arthrobacter sp. FB24	4506	Actinobakteria
Bifidobacterium adolescentis ATCC 15703	1631	Actinobakteria
Bifidobacterium longum	1729	Actinobakteria
Clavibacter michiganensis subsp. michiganensis NCPPB 382	3079	Actinobakteria
Clavibacter michiganensis subsp. sepedonicus	2941	Actinobakteria
Corynebacterium diphtheriae	2272	Actinobakteria
Corynebacterium efficiens YS-314	2950	Actinobakteria
Corynebacterium glutamicum ATCC 13032 Bielefeld	3057	Actinobakteria
Corynebacterium glutamicum ATCC 13032 Kitasato	2993	Actinobakteria
Corynebacterium glutamicum R	3080	Actinobakteria
Corynebacterium jeikeium K411	2120	Actinobakteria
Frankia alni ACN14a	6711	Actinobakteria
Frankia sp. CcI3	4499	Actinobakteria
Frankia sp. EAN1pec	7191	Actinobakteria
Kineococcus radiotolerans SRS30216	4681	Actinobakteria
Leifsonia xyli subsp. xyli str. CTCB07	2030	Actinobakteria
Mycobacterium abscessus	4941	Actinobakteria
Mycobacterium avium 104	5120	Actinobakteria
Mycobacterium avium paratuberculosis	4350	Actinobakteria

Organismus	# Proteine	Phylum
Mycobacterium bovis	3920	Actinobakteria
Mycobacterium bovis BCG Pasteur 1173P2	3952	Actinobakteria
Mycobacterium gilvum PYR-GCK	5579	Actinobakteria
Mycobacterium leprae	1605	Actinobakteria
Mycobacterium smegmatis MC2 155	6716	Actinobakteria
Mycobacterium sp. JLS	5739	Actinobakteria
Mycobacterium sp. KMS	5975	Actinobakteria
Mycobacterium sp. MCS	5615	Actinobakteria
Mycobacterium tuberculosis CDC1551	4189	Actinobakteria
Mycobacterium tuberculosis F11	3941	Actinobakteria
Mycobacterium tuberculosis H37Ra	4034	Actinobakteria
Mycobacterium tuberculosis H37Rv	3989	Actinobakteria
Mycobacterium ulcerans Agy99	4160	Actinobakteria
Mycobacterium vanbaalenii PYR-1	5979	Actinobakteria
Nocardia farcinica IFM 10152	5936	Actinobakteria
Nocardioides sp. JS614	4909	Actinobakteria
Propionibacterium acnes KPA171202	2297	Actinobakteria
Renibacterium salmoninarum ATCC 33209	3507	Actinobakteria
Rhodococcus sp. RHA1	9145	Actinobakteria
Rubrobacter xylanophilus DSM 9941	3140	Actinobakteria
Saccharopolyspora erythraea NRRL 2338	7197	Actinobakteria
Salinispora arenicola CNS-205	4917	Actinobakteria
Salinispora tropica CNB-440	4536	Actinobakteria
Streptomyces avermitilis	7676	Actinobakteria
Streptomyces coelicolor	8154	Actinobakteria
Thermobifida fusca YX	3110	Actinobakteria
Tropheryma whipplei TW08/27	783	Actinobakteria
Tropheryma whipplei Twist	808	Actinobakteria
Bos taurus	25255	Animalia
Danio rerio	35676	Animalia
Drosophila melanogaster	20722	Animalia
Gallus gallus	4840	Animalia
Homo sapiens	37868	Animalia
Mus musculus	35156	Animalia
Rattus norvegicus	31705	Animalia
Xenopus tropicalis	7083	Animalia

Organismus	# Proteine	Phylum
Aeropyrum pernix	1700	Archaea
Archaeoglobus fulgidus	2420	Archaea
Caldivirga maquilingensis IC-167	1963	Archaea
Candidatus Korarchaeum cryptofilum OPF8	1602	Archaea
Candidatus Methanoregula boonei 6A8	2450	Archaea
Haloarcula marismortui ATCC 43049	4240	Archaea
Halobacterium salinarum R1	2749	Archaea
Halobacterium sp.	2622	Archaea
Haloquadratum walsbyi	2646	Archaea
Hyperthermus butylicus	1602	Archaea
Ignicoccus hospitalis KIN4/I	1434	Archaea
Metallosphaera sedula DSM 5348	2256	Archaea
Methanobacterium thermoautotrophicum	1873	Archaea
Methanobrevibacter smithii ATCC 35061	1793	Archaea
Methanococcoides burtonii DSM 6242	2273	Archaea
Methanococcus aeolicus Nankai-3	1490	Archaea
Methanococcus jannaschii	1786	Archaea
Methanococcus maripaludis C5	1822	Archaea
Methanococcus maripaludis C6	1826	Archaea
Methanococcus maripaludis C7	1788	Archaea
Methanococcus maripaludis S2	1722	Archaea
Methanococcus vannielii SB	1678	Archaea
Methanocorpusculum labreanum Z	1739	Archaea
Methanoculleus marisnigri JR1	2489	Archaea
Methanopyrus kandleri	1687	Archaea
Methanosaeta thermophila PT	1696	Archaea
Methanosarcina acetivorans	4540	Archaea
Methanosarcina barkeri fusaro	3624	Archaea
Methanosarcina mazei	3370	Archaea
Methanosphaera stadtmanae	1534	Archaea
Methanospirillum hungatei JF-1	3139	Archaea
Nanoarchaeum equitans	536	Archaea
Natronomonas pharaonis	2822	Archaea
Nitrosopumilus maritimus SCM1	1795	Archaea
Picrophilus torridus DSM 9790	1535	Archaea
Pyrobaculum aerophilum	2605	Archaea

Organismus	# Proteine	Phylum
Pyrobaculum arsenaticum DSM 13514	2299	Archaea
Pyrobaculum calidifontis JCM 11548	2149	Archaea
Pyrobaculum islandicum DSM 4184	1978	Archaea
Pyrococcus abyssi	1898	Archaea
Pyrococcus furiosus	2125	Archaea
Pyrococcus horikoshii	1955	Archaea
Staphylothermus marinus F1	1570	Archaea
Sulfolobus acidocaldarius DSM 639	2223	Archaea
Sulfolobus solfataricus	2977	Archaea
Sulfolobus tokodaii	2825	Archaea
Thermococcus kodakaraensis KOD1	2306	Archaea
Thermofilum pendens Hrk 5	1876	Archaea
Thermoplasma acidophilum	1482	Archaea
Thermoplasma volcanium	1499	Archaea
Thermoproteus neutrophilus V24Sta	1966	Archaea
uncultured methanogenic archaeon RC-I	3085	Archaea
Chlamydia muridarum	911	Chlamydiae
Chlamydia trachomatis	895	Chlamydiae
Chlamydia trachomatis 434/Bu	874	Chlamydiae
Chlamydia trachomatis A/HAR-13	919	Chlamydiae
Chlamydia trachomatis L2b/UCH-1/proctitis	874	Chlamydiae
Chlamydophila abortus S26/3	932	Chlamydiae
Chlamydophila caviae	1005	Chlamydiae
Chlamydophila felis Fe/C-56	1013	Chlamydiae
Chlamydophila pneumoniae AR39	1112	Chlamydiae
Chlamydophila pneumoniae CWL029	1052	Chlamydiae
Chlamydophila pneumoniae J138	1069	Chlamydiae
Chlamydophila pneumoniae TW-183	1113	Chlamydiae
Parachlamydia sp. UWE25	2031	Chlamydiae
Chloroflexus aurantiacus J-10-fl	3853	Chloroflexi
Dehalococcoides ethenogenes 195	1580	Chloroflexi
Dehalococcoides sp. BAV1	1371	Chloroflexi
Dehalococcoides sp. CBDB1	1458	Chloroflexi
Herpetosiphon aurantiacus ATCC 23779	5278	Chloroflexi
Roseiflexus castenholzii DSM 13941	4330	Chloroflexi
Roseiflexus sp. RS-1	4517	Chloroflexi

Organismus	# Proteine	Phylum
Acaryochloris marina MBIC11017	8383	Cyanobakteria
Anabaena variabilis ATCC 29413	5661	Cyanobakteria
Crocosphaera watsonii WH 8501	5958	Cyanobakteria
Cyanothece sp. ATCC 51142	5304	Cyanobakteria
Cyanothece sp. CCY0110	6475	Cyanobakteria
Cyanothece sp. PCC 7424	6059	Cyanobakteria
Cyanothece sp. PCC 8801	4390	Cyanobakteria
Gloeobacter violaceus PCC 7421	4430	Cyanobakteria
Microcystis aeruginosa NIES-843	6312	Cyanobakteria
Nostoc punctiforme PCC 73102	7672	Cyanobakteria
Nostoc sp. PCC 7120	6130	Cyanobakteria
Prochlorococcus marinus str. AS9601	1921	Cyanobakteria
Prochlorococcus marinus str. MIT 9211	1855	Cyanobakteria
Prochlorococcus marinus str. MIT 9215	1983	Cyanobakteria
Prochlorococcus marinus str. MIT 9301	1907	Cyanobakteria
Prochlorococcus marinus str. MIT 9303	2997	Cyanobakteria
Prochlorococcus marinus str. MIT 9312	1810	Cyanobakteria
Prochlorococcus marinus str. MIT 9313	2269	Cyanobakteria
Prochlorococcus marinus str. MIT 9515	1906	Cyanobakteria
Prochlorococcus marinus str. NATL1A	2193	Cyanobakteria
Prochlorococcus marinus str. NATL2A	2163	Cyanobakteria
Prochlorococcus marinus subsp. marinus str. CCMP1375	1883	Cyanobakteria
Prochlorococcus marinus subsp. pastoris str. CCMP1986	1717	Cyanobakteria
Synechococcus elongatus PCC 6301	2527	Cyanobakteria
Synechococcus elongatus PCC 7942	2662	Cyanobakteria
Synechococcus sp. BL107	2507	Cyanobakteria
Synechococcus sp. CC9311	2892	Cyanobakteria
Synechococcus sp. CC9605	2645	Cyanobakteria
Synechococcus sp. CC9902	2307	Cyanobakteria
Synechococcus sp. PCC 7002	3186	Cyanobakteria
Synechococcus sp. RCC307	2535	Cyanobakteria
Synechococcus sp. RS9916	2961	Cyanobakteria
Synechococcus sp. RS9917	2770	Cyanobakteria
Synechococcus sp. WH 5701	3346	Cyanobakteria

Organismus	# Proteine	Phylum
Synechococcus sp. WH 7803	2533	Cyanobakteria
Synechococcus sp. WH 7805	2883	Cyanobakteria
Synechococcus sp. WH 8102	2519	Cyanobakteria
Synechocystis sp. PCC 6803	3569	Cyanobakteria
Thermosynechococcus elongatus BP-1	2476	Cyanobakteria
Trichodesmium erythraeum IMS101	4451	Cyanobakteria
Acholeplasma laidlawii PG-8A	1380	Eubakteria
Acidobacteria bacterium Ellin345	4777	Eubakteria
Aquifex aeolicus	1560	Eubakteria
Aster yellows witches'-broom phytoplasma AYWB	693	Eubakteria
Bacteroides fragilis NCTC 9343	4231	Eubakteria
Bacteroides fragilis YCH46	4625	Eubakteria
Bacteroides thetaiotaomicron VPI-5482	4816	Eubakteria
Bacteroides vulgatus ATCC 8482	4065	Eubakteria
Borrelia afzelii PKo	1214	Eubakteria
Borrelia burgdorferi	1640	Eubakteria
Borrelia garinii PBi	932	Eubakteria
Candidatus Sulcia muelleri GWSS	227	Eubakteria
Chlorobium chlorochromatii CaD3	2002	Eubakteria
Chlorobium phaeobacteroides DSM 266	2650	Eubakteria
Chlorobium tepidum TLS	2252	Eubakteria
Cytophaga hutchinsonii ATCC 33406	3785	Eubakteria
Deinococcus geothermalis DSM 11300	3062	Eubakteria
Deinococcus radiodurans	3181	Eubakteria
Fervidobacterium nodosum Rt17-B1	1750	Eubakteria
Flavobacterium johnsoniae UW101	5017	Eubakteria
Flavobacterium psychrophilum JIP02/86	2412	Eubakteria
Fusobacterium nucleatum	2067	Eubakteria
Gramella forsetii KT0803	3584	Eubakteria
Leptospira borgpetersenii serovar Hardjo-bovis JB197	2880	Eubakteria
Leptospira borgpetersenii serovar Hardjo-bovis L550	2945	Eubakteria
Leptospira interrogans serovar Copenhageni	3658	Eubakteria
Leptospira interrogans serovar Lai	4727	Eubakteria

Organismus	# Proteine	Phylum
Mesoplasma florum L1	682	Eubakteria
Mycoplasma agalactiae PG2	742	Eubakteria
Mycoplasma capricolum subsp. capricolum ATCC 27343	812	Eubakteria
Mycoplasma gallisepticum	726	Eubakteria
Mycoplasma genitalium	477	Eubakteria
Mycoplasma hyopneumoniae 232	691	Eubakteria
Mycoplasma hyopneumoniae 7448	663	Eubakteria
Mycoplasma hyopneumoniae J	665	Eubakteria
Mycoplasma mobile 163K	633	Eubakteria
Mycoplasma mycoides	1016	Eubakteria
Mycoplasma penetrans	1037	Eubakteria
Mycoplasma pneumoniae	689	Eubakteria
Mycoplasma pulmonis	782	Eubakteria
Mycoplasma synoviae 53	672	Eubakteria
Onion yellows phytoplasma str. 'onion yellows'	754	Eubakteria
Parabacteroides distasonis ATCC 8503	3850	Eubakteria
Pelodictyon luteolum DSM 273	2083	Eubakteria
Petrotoga mobilis SJ95	1898	Eubakteria
Pirellula sp.	7325	Eubakteria
Porphyromonas gingivalis W83	1909	Eubakteria
Prosthecochloris vibrioformis DSM 265	1753	Eubakteria
Salinibacter ruber DSM 13855	2833	Eubakteria
Solibacter usitatus Ellin6076	7826	Eubakteria
Thermosipho melanesiensis BI429	1879	Eubakteria
Thermotoga lettingae TMO	2040	Eubakteria
Thermotoga maritima	1858	Eubakteria
Thermotoga petrophila RKU-1	1785	Eubakteria
Thermus thermophilus HB27	2210	Eubakteria
Thermus thermophilus HB8	2238	Eubakteria
Treponema denticola ATCC 35405	2767	Eubakteria
Treponema pallidum	1036	Eubakteria
Ureaplasma parvum serovar 3 str. ATCC 27815	609	Eubakteria
Ureaplasma urealyticum	614	Eubakteria
Alkaliphilus metalliredigens QYMF	4625	Firmicute

Organismus	# Proteine	Phylum
Alkaliphilus oremlandii OhILAs	2836	Firmicute
Bacillus amyloliquefaciens FZB42	3693	Firmicute
Bacillus anthracis Ames	5311	Firmicute
Bacillus anthracis str. Ames 0581	5617	Firmicute
Bacillus anthracis str. Sterne	5287	Firmicute
Bacillus cereus ATCC 10987	5844	Firmicute
Bacillus cereus ATCC 14579	5255	Firmicute
Bacillus cereus subsp. cytotoxis NVH 391-98	3844	Firmicute
Bacillus cereus ZK	5641	Firmicute
Bacillus clausii KSM-K16	4096	Firmicute
Bacillus halodurans	4066	Firmicute
Bacillus licheniformis ATCC 14580	4178	Firmicute
Bacillus licheniformis DSM 13	4196	Firmicute
Bacillus pumilus SAFR-032	3681	Firmicute
Bacillus subtilis	4105	Firmicute
Bacillus thuringiensis serovar konkukian str. 97-27	5197	Firmicute
Bacillus thuringiensis str. Al Hakam	4798	Firmicute
Bacillus weihenstephanensis KBAB4	5653	Firmicute
Caldicellulosiruptor saccharolyticus DSM 8903	2679	Firmicute
Candidatus Desulforudis audaxviator MP104C	2157	Firmicute
Carboxydothermus hydrogenoformans Z-2901	2620	Firmicute
Clostridium acetobutylicum	3848	Firmicute
Clostridium beijerinckii NCIMB 8052	5020	Firmicute
Clostridium botulinum A	3590	Firmicute
Clostridium botulinum A ATCC 19397	3552	Firmicute
Clostridium botulinum A *str.* Hall	3404	Firmicute
Clostridium botulinum A3 *str.* Loch Maree	3984	Firmicute
Clostridium botulinum B1 *str.* Okra	3852	Firmicute
Clostridium botulinum F *str.* Langeland	3659	Firmicute
Clostridium difficile 630	3753	Firmicute
Clostridium kluyveri DSM 555	3913	Firmicute
Clostridium novyi NT	2315	Firmicute
Clostridium perfringens	2723	Firmicute
Clostridium perfringens ATCC 13124	2876	Firmicute
Clostridium perfringens SM101	2631	Firmicute
Clostridium phytofermentans ISDg	3902	Firmicute

Organismus	# Proteine	Phylum
Clostridium tetani E88	2432	Firmicute
Clostridium thermocellum ATCC 27405	3189	Firmicute
Desulfitobacterium hafniense Y51	5060	Firmicute
Desulfotomaculum reducens MI-1	3276	Firmicute
Enterococcus faecalis V583	3265	Firmicute
Finegoldia magna ATCC 29328	1813	Firmicute
Geobacillus kaustophilus HTA426	3540	Firmicute
Geobacillus thermodenitrificans NG80-2	3445	Firmicute
Heliobacterium modesticaldum Ice1	3000	Firmicute
Lactobacillus acidophilus NCFM	1862	Firmicute
Lactobacillus brevis ATCC 367	2218	Firmicute
Lactobacillus casei ATCC 334	2771	Firmicute
Lactobacillus delbrueckii bulgaricus	1562	Firmicute
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	1721	Firmicute
Lactobacillus gasseri ATCC 33323	1755	Firmicute
Lactobacillus helveticus DPC 4571	1610	Firmicute
Lactobacillus johnsonii NCC 533	1821	Firmicute
Lactobacillus plantaearum	3057	Firmicute
Lactobacillus reuteri F275	1900	Firmicute
Lactobacillus sakei subsp. sakei 23K	1879	Firmicute
Lactobacillus salivarius UCC118	2017	Firmicute
Lactococcus lactis	2321	Firmicute
Lactococcus lactis subsp. cremoris MG1363	2434	Firmicute
Lactococcus lactis subsp. cremoris SK11	2504	Firmicute
Leuconostoc citreum KM20	1820	Firmicute
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	2005	Firmicute
Listeria innocua	3043	Firmicute
Listeria monocytogenes	2846	Firmicute
Listeria monocytogenes 4b F2365	2821	Firmicute
Listeria welshimeri serovar 6b SLCC5334	2774	Firmicute
Lysinibacillus sphaericus C3-41	4771	Firmicute
Moorella thermoacetica ATCC 39073	2465	Firmicute
Oceanobacillus iheyensis	3500	Firmicute
Oenococcus oeni PSU-1	1691	Firmicute

Organismus	# Proteine	Phylum
Pediococcus pentosaceus ATCC 25745	1755	Firmicute
Pelotomaculum thermopropionicum SI	2920	Firmicute
Staphylococcus aureus COL	2618	Firmicute
Staphylococcus aureus NCTC 8325	2892	Firmicute
Staphylococcus aureus RF122	2509	Firmicute
Staphylococcus aureus subsp. aureus JH1	2780	Firmicute
Staphylococcus aureus subsp. aureus JH9	2726	Firmicute
Staphylococcus aureus subsp. aureus MRSA252	2656	Firmicute
Staphylococcus aureus subsp. aureus MSSA476	2598	Firmicute
Staphylococcus aureus subsp. aureus Mu3	2698	Firmicute
Staphylococcus aureus subsp. aureus Mu50	2731	Firmicute
Staphylococcus aureus subsp. aureus MW2	2632	Firmicute
Staphylococcus aureus subsp. aureus N315	2619	Firmicute
Staphylococcus aureus subsp. aureus str. Newman	2614	Firmicute
Staphylococcus aureus subsp. aureus USA300	2604	Firmicute
Staphylococcus aureus subsp. aureus USA300_TCH1516	2683	Firmicute
Staphylococcus epidermidis ATCC 12228	2485	Firmicute
Staphylococcus epidermidis RP62A	2526	Firmicute
Staphylococcus haemolyticus	2676	Firmicute
Staphylococcus saprophyticus	2514	Firmicute
Streptococcus agalactiae 2603V/R	2124	Firmicute
Streptococcus agalactiae A909	1996	Firmicute
Streptococcus agalactiae NEM316	2094	Firmicute
Streptococcus gordonii str. Challis substr. CH1	2051	Firmicute
Streptococcus mutans	1960	Firmicute
Streptococcus pneumoniae D39	1914	Firmicute
Streptococcus pneumoniae Hungary19A-6	2155	Firmicute
Streptococcus pneumoniae R6	2043	Firmicute
Streptococcus pneumoniae TIGR4	2105	Firmicute
Streptococcus pyogenes M1 GAS	1697	Firmicute
Streptococcus pyogenes Manfredo	1745	Firmicute
Streptococcus pyogenes MGAS10270	1986	Firmicute
Streptococcus pyogenes MGAS10394	1886	Firmicute
Streptococcus pyogenes MGAS10750	1979	Firmicute

Organismus	# Proteine	Phylum
Streptococcus pyogenes MGAS2096	1898	Firmicute
Streptococcus pyogenes MGAS315	1865	Firmicute
Streptococcus pyogenes MGAS5005	1865	Firmicute
Streptococcus pyogenes MGAS6180	1894	Firmicute
Streptococcus pyogenes MGAS8232	1839	Firmicute
Streptococcus pyogenes MGAS9429	1877	Firmicute
Streptococcus pyogenes SSI-1	1861	Firmicute
Streptococcus sanguinis SK36	2270	Firmicute
Streptococcus suis 05ZYH33	2186	Firmicute
Streptococcus suis 98HAH33	2185	Firmicute
Streptococcus thermophilus CNRZ1066	1915	Firmicute
Streptococcus thermophilus LMD-9	1716	Firmicute
Streptococcus thermophilus LMG 18311	1889	Firmicute
Symbiobacterium thermophilum IAM14863	3338	Firmicute
Syntrophomonas wolfei Goettingen	2504	Firmicute
Thermoanaerobacter pseudethanolicus ATCC 33223	2243	Firmicute
Thermoanaerobacter sp. X514	2349	Firmicute
Thermoanaerobacter tengcongensis	2588	Firmicute
Ashbya gossypii ATCC 10895	4725	Fungi
Candida glabrata CBS138	1269	Fungi
Cryptococcus neoformans JEC21	6594	Fungi
Debaryomyces hansenii	6317	Fungi
Encephalitozoon cuniculi GB-M1	1996	Fungi
Kluyveromyces lactis	5357	Fungi
Saccharomyces cerevisiae	5865	Fungi
Schizosaccharomyces pombe	5166	Fungi
Yarrowia lipolytica	6520	Fungi
Arabidopsis thaliana	31480	Plantae
Oryza sativa	26777	Plantae
Populus trichocarpa	58036	Plantae
Ostreococcus lucimarinus	1319	Grünalge
Ostreococcus tauri	7725	Grünalge
Chlamydomonas reinhardtti	14598	Grünalge
Cyanidioschyzon merolae	4773	Rotalge
Acidiphilium cryptum JF-5	3559	Proteobakteria

Organismus	# Proteine	Phylum
Acidovorax avenae subsp. citrulli AAC00-1	4709	Proteobakteria
Acidovorax sp. JS42	4155	Proteobakteria
Acinetobacter baumannii	3712	Proteobakteria
Acinetobacter baumannii ATCC 17978	3368	Proteobakteria
Acinetobacter baumannii SDF	2975	Proteobakteria
Acinetobacter sp. ADP1	3325	Proteobakteria
Actinobacillus pleuropneumoniae L20	2012	Proteobakteria
Actinobacillus pleuropneumoniae serovar 3 *str.* JL03	2036	Proteobakteria
Actinobacillus succinogenes 130Z	2079	Proteobakteria
Aeromonas salmonicida subsp. salmonicida A449	4437	Proteobakteria
Aeromonas sp. ATCC 7966	4122	Proteobakteria
Agrobacterium tumefaciens str. C58	5355	Proteobakteria
Alcanivorax borkumensis SK2	2755	Proteobakteria
Alkalilimnicola ehrlichei MLHE-1	2865	Proteobakteria
Anaeromyxobacter dehalogenans 2CP-C	4346	Proteobakteria
Anaeromyxobacter sp. Fw109-5	4466	Proteobakteria
Anaplasma marginale str. St. Maries	949	Proteobakteria
Anaplasma phagocytophilum HZ	1264	Proteobakteria
Arcobacter butzleri RM4018	2259	Proteobakteria
Azoarcus sp. BH72	3989	Proteobakteria
Azoarcus sp. EbN1	4599	Proteobakteria
Azorhizobium caulinodans ORS 571	4717	Proteobakteria
Bartonella bacilliformis KC583	1283	Proteobakteria
Bartonella henselae str. Houston-1	1488	Proteobakteria
Bartonella quintana str. Toulouse	1142	Proteobakteria
Bartonella tribocorum CIP 105476	2092	Proteobakteria
Baumannia cicadellinicola str. Hc (*Homalodisca coagulata*)	595	Proteobakteria
Bdellovibrio bacteriovorus	3587	Proteobakteria
Bordetella bronchiseptica	4994	Proteobakteria
Bordetella parapertussis	4185	Proteobakteria
Bordetella pertussis	3436	Proteobakteria
Bordetella petrii	5027	Proteobakteria
Bradyrhizobium japonicum	8317	Proteobakteria
Bradyrhizobium sp. BTAi1	7622	Proteobakteria

Organismus	# Proteine	Phylum
Bradyrhizobium sp. ORS278	6717	Proteobakteria
Brucella abortus bv. 1 *str.* 9-941	3085	Proteobakteria
Brucella canis ATCC 23365	3251	Proteobakteria
Brucella melitensis	3198	Proteobakteria
Brucella melitensis biovar Abortus	3034	Proteobakteria
Brucella ovis	2890	Proteobakteria
Brucella suis 1330	3271	Proteobakteria
Brucella suis ATCC 23445	3241	Proteobakteria
Buchnera aphidicola	507	Proteobakteria
Buchnera aphidicola str. Cc (*Cinara cedri*)	357	Proteobakteria
Buchnera aphidicola str. Sg (*Schizaphis graminum*)	546	Proteobakteria
Buchnera sp.	574	Proteobakteria
Burkholderia cenocepacia AU 1054	6477	Proteobakteria
Burkholderia cenocepacia HI2424	6919	Proteobakteria
Burkholderia cenocepacia MC0-3	7008	Proteobakteria
Burkholderia cepacia AMMD	6617	Proteobakteria
Burkholderia mallei ATCC 23344	5024	Proteobakteria
Burkholderia mallei NCTC 10229	5510	Proteobakteria
Burkholderia mallei NCTC 10247	5852	Proteobakteria
Burkholderia mallei SAVP1	5189	Proteobakteria
Burkholderia multivorans ATCC 17616	6259	Proteobakteria
Burkholderia pseudomallei 1106a	7183	Proteobakteria
Burkholderia pseudomallei 1710b	6347	Proteobakteria
Burkholderia pseudomallei 668	7230	Proteobakteria
Burkholderia pseudomallei K96243	5728	Proteobakteria
Burkholderia sp. 383	7717	Proteobakteria
Burkholderia thailandensis E264	5634	Proteobakteria
Burkholderia vietnamiensis G4	7617	Proteobakteria
Burkholderia xenovorans LB400	8702	Proteobakteria
Calyptogena magnifica	976	Proteobakteria
Campylobacter concisus 13826	1985	Proteobakteria
Campylobacter curvus 525.92	1931	Proteobakteria
Campylobacter fetus subsp. fetus 82-40	1719	Proteobakteria
Campylobacter hominis ATCC BAA-381	1687	Proteobakteria
Campylobacter jejuni	1634	Proteobakteria

Organismus	# Proteine	Phylum
Campylobacter jejuni RM1221	1838	Proteobakteria
Campylobacter jejuni subsp. doylei 269.97	1731	Proteobakteria
Campylobacter jejuni subsp. jejuni 81-176	1758	Proteobakteria
Campylobacter jejuni subsp. jejuni 81116	1626	Proteobakteria
Candidatus Blochmannia floridanus	583	Proteobakteria
Candidatus Blochmannia pennsylvanicus BPEN	610	Proteobakteria
Candidatus Carsonella ruddii PV	182	Proteobakteria
Candidatus Desulfococcus oleovorans Hxd3	3265	Proteobakteria
Candidatus Pelagibacter ubique HTCC1062	1354	Proteobakteria
Candidatus Vesicomyosocius okutanii HA	937	Proteobakteria
Caulobacter crescentus	3737	Proteobakteria
Caulobacter sp. K31	5438	Proteobakteria
Chromobacterium violaceum	4407	Proteobakteria
Chromohalobacter salexigens DSM 3043	3298	Proteobakteria
Citrobacter koseri ATCC BAA-895	5026	Proteobakteria
Colwellia psychrerythraea 34H	4910	Proteobakteria
Coxiella burnetii	2052	Proteobakteria
Coxiella burnetii Dugway 7E9-12	2125	Proteobakteria
Coxiella burnetii RSA 331	1975	Proteobakteria
Dechloromonas aromatica RCB	4171	Proteobakteria
Delftia acidovorans SPH-1	6040	Proteobakteria
Desulfotalea psychrophila LSv54	3234	Proteobakteria
Desulfovibrio desulfuricans G20	3775	Proteobakteria
Desulfovibrio vulgaris Hildenborough	3531	Proteobakteria
Desulfovibrio vulgaris subsp. vulgaris DP4	3091	Proteobakteria
Dichelobacter nodosus VCS1703A	1280	Proteobakteria
Dinoroseobacter shibae DFL 12	4187	Proteobakteria
Ehrlichia canis Jake	925	Proteobakteria
Ehrlichia chaffeensis Arkansas	1105	Proteobakteria
Ehrlichia ruminantium str. Gardel	950	Proteobakteria
Ehrlichia ruminantium str. Welgevonden	888	Proteobakteria
Enterobacter sakazakii ATCC BAA-894	4434	Proteobakteria
Enterobacter sp. 638	4240	Proteobakteria
Erwinia carotovora subsp. atroseptica SCRI1043	4472	Proteobakteria
Erythrobacter litoralis HTCC2594	3011	Proteobakteria
Escherichia coli 536	4629	Proteobakteria

Organismus	# Proteine	Phylum
Escherichia coli APEC O1	4880	Proteobakteria
Escherichia coli CFT073	5379	Proteobakteria
Escherichia coli DH10B	4126	Proteobakteria
Escherichia coli E24377A	4997	Proteobakteria
Escherichia coli HS	4384	Proteobakteria
Escherichia coli K12	4133	Proteobakteria
Escherichia coli O157:H7	5341	Proteobakteria
Escherichia coli O157:H7 EDL933	5423	Proteobakteria
Escherichia coli SMS-3-5	4913	Proteobakteria
Escherichia coli UTI89	5189	Proteobakteria
Escherichia coli W3110	4226	Proteobakteria
Francisella philomiragia subsp. philomiragia ATCC 25017	1915	Proteobakteria
Francisella tularensis subsp. holarctica	1754	Proteobakteria
Francisella tularensis subsp. holarctica FTA	1580	Proteobakteria
Francisella tularensis subsp. holarctica OSU18	1555	Proteobakteria
Francisella tularensis subsp. novicida U112	1719	Proteobakteria
Francisella tularensis subsp. tularensis	1603	Proteobakteria
Francisella tularensis subsp. tularensis FSC198	1605	Proteobakteria
Francisella tularensis subsp. tularensis WY96-3418	1634	Proteobakteria
Geobacter metallireducens GS-15	3532	Proteobakteria
Geobacter sulfurreducens	3446	Proteobakteria
Geobacter uraniumreducens Rf4	4357	Proteobakteria
Gluconacetobacter diazotrophicus PAl 5	3852	Proteobakteria
Gluconobacter oxydans 621H	2664	Proteobakteria
Granulobacter bethesdensis CGDNIH1	2437	Proteobakteria
Haemophilus ducreyi 35000HP	1717	Proteobakteria
Haemophilus influenzae	1657	Proteobakteria
Haemophilus influenzae 86-028NP	1792	Proteobakteria
Haemophilus influenzae PittEE	1619	Proteobakteria
Haemophilus influenzae PittGG	1667	Proteobakteria
Haemophilus somnus 129PT	1798	Proteobakteria
Haemophilus somnus 2336	1980	Proteobakteria
Hahella chejuensis KCTC 2396	6778	Proteobakteria
Halorhodospira halophila SL1	2407	Proteobakteria

Organismus	# Proteine	Phylum
Helicobacter acinonychis Sheeba	1618	Proteobakteria
Helicobacter hepaticus	1875	Proteobakteria
Helicobacter pylori 26695	1576	Proteobakteria
Helicobacter pylori HPAG1	1544	Proteobakteria
Helicobacter pylori J99	1489	Proteobakteria
Herminiimonas arsenicoxydans	3325	Proteobakteria
Hyphomonas neptunium ATCC 15444	3505	Proteobakteria
Idiomarina loihiensis L2TR	2628	Proteobakteria
Jannaschia sp. CCS1	4283	Proteobakteria
Janthinobacterium sp. Marseille	3697	Proteobakteria
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	5187	Proteobakteria
Lawsonia intracellularis PHE/MN1-00	1337	Proteobakteria
Legionella pneumophila Lens	2934	Proteobakteria
Legionella pneumophila Paris	3166	Proteobakteria
Legionella pneumophila str. Corby	3206	Proteobakteria
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	2942	Proteobakteria
Leptothrix cholodnii SP-6	4363	Proteobakteria
Magnetococcus sp. MC-1	3716	Proteobakteria
Magnetospirillum magneticum AMB-1	4559	Proteobakteria
Mannheimia succiniciproducens MBEL55E	2380	Proteobakteria
Maricaulis maris MCS10	3063	Proteobakteria
Marinobacter aquaeolei VT8	4272	Proteobakteria
Marinomonas sp. MWYL1	4439	Proteobakteria
Mesorhizobium loti	7272	Proteobakteria
Mesorhizobium sp. BNC1	4543	Proteobakteria
Methylibium petroleiphilum PM1	4449	Proteobakteria
Methylobacillus flagellatus KT	2753	Proteobakteria
Methylobacterium extorquens PA1	4829	Proteobakteria
Methylobacterium radiotolerans JCM 2831	6431	Proteobakteria
Methylobacterium sp. 4-46	6692	Proteobakteria
Methylococcus capsulatus Bath	2956	Proteobakteria
Myxococcus xanthus DK 1622	7331	Proteobakteria
Neisseria gonorrhoeae FA 1090	2002	Proteobakteria
Neisseria meningitidis 053442	2020	Proteobakteria

Organismus	# Proteine	Phylum
Neisseria meningitidis FAM18	1917	Proteobakteria
Neisseria meningitidis MC58	2063	Proteobakteria
Neisseria meningitidis Z2491	2049	Proteobakteria
Neorickettsia sennetsu Miyayama	932	Proteobakteria
Nitratiruptor sp. SB155-2	1843	Proteobakteria
Nitrobacter hamburgensis X14	4326	Proteobakteria
Nitrobacter winogradskyi Nb-255	3122	Proteobakteria
Nitrosococcus oceani ATCC 19707	3017	Proteobakteria
Nitrosomonas europaea	2461	Proteobakteria
Nitrosomonas eutropha C71	2551	Proteobakteria
Nitrosospira multiformis ATCC 25196	2805	Proteobakteria
Novosphingobium aromaticivorans DSM 12444	3937	Proteobakteria
Ochrobactrum anthropi ATCC 49188	4799	Proteobakteria
Orientia tsutsugamushi Boryong	1182	Proteobakteria
Paracoccus denitrificans PD1222	5077	Proteobakteria
Parvibaculum lavamentivorans DS-1	3636	Proteobakteria
Pasteurella multocida	2015	Proteobakteria
Pelobacter carbinolicus	3352	Proteobakteria
Pelobacter propionicus DSM 2379	3804	Proteobakteria
Photobacterium profundum SS9	5489	Proteobakteria
Photorhabdus luminescens	4683	Proteobakteria
Polaromonas naphthalenivorans CJ2	4929	Proteobakteria
Polaromonas sp. JS666	5453	Proteobakteria
Polynucleobacter necessarius STIR1	1508	Proteobakteria
Polynucleobacter sp. QLW-P1DMWA-1	2077	Proteobakteria
Pseudoalteromonas atlantica T6c	4281	Proteobakteria
Pseudoalteromonas haloplanktis TAC125	3486	Proteobakteria
Pseudomonas aeruginosa	5568	Proteobakteria
Pseudomonas aeruginosa PA7	6286	Proteobakteria
Pseudomonas aeruginosa UCBPP-PA14	5892	Proteobakteria
Pseudomonas entomophila L48	5134	Proteobakteria
Pseudomonas fluorescens Pf-5	6138	Proteobakteria
Pseudomonas fluorescens PfO-1	5736	Proteobakteria
Pseudomonas mendocina ymp	4594	Proteobakteria
Pseudomonas putida F1	5252	Proteobakteria
Pseudomonas putida GB-1	5409	Proteobakteria

Organismus	# Proteine	Phylum
Pseudomonas putida KT2440	5350	Proteobakteria
Pseudomonas putida W619	5182	Proteobakteria
Pseudomonas stutzeri A1501	4128	Proteobakteria
Pseudomonas syringae pv. phaseolicola 1448A	5171	Proteobakteria
Pseudomonas syringae pv. syringae B728a	5089	Proteobakteria
Pseudomonas syringae pv. tomato str. DC3000	5613	Proteobakteria
Psychrobacter arcticum 273-4	2120	Proteobakteria
Psychrobacter cryohalolentis K5	2511	Proteobakteria
Psychrobacter sp. PRwf-1	2385	Proteobakteria
Psychromonas ingrahamii 37	3545	Proteobakteria
Ralstonia eutropha H16	6626	Proteobakteria
Ralstonia eutropha JMP134	6446	Proteobakteria
Ralstonia metallidurans CH34	6319	Proteobakteria
Ralstonia solanacearum	5116	Proteobakteria
Rhizobium etli CFN 42	5963	Proteobakteria
Rhizobium leguminosarum bv. viciae 3841	7143	Proteobakteria
Rhodobacter sphaeroides ATCC 17023	4242	Proteobakteria
Rhodobacter sphaeroides ATCC 17025	4333	Proteobakteria
Rhodobacter sphaeroides ATCC 17029	4132	Proteobakteria
Rhodoferax ferrireducens T118	4418	Proteobakteria
Rhodopseudomonas palustris BisA53	4878	Proteobakteria
Rhodopseudomonas palustris BisB18	4886	Proteobakteria
Rhodopseudomonas palustris BisB5	4397	Proteobakteria
Rhodopseudomonas palustris CGA009	4820	Proteobakteria
Rhodopseudomonas palustris HaA2	4683	Proteobakteria
Rhodospirillum rubrum ATCC 11170	3841	Proteobakteria
Rickettsia akari str. Hartford	1259	Proteobakteria
Rickettsia bellii OSU 85-389	1476	Proteobakteria
Rickettsia bellii RML369-C	1429	Proteobakteria
Rickettsia canadensis str. McKiel	1093	Proteobakteria
Rickettsia conorii	1374	Proteobakteria
Rickettsia felis URRWXCal2	1512	Proteobakteria
Rickettsia massiliae MTU5	980	Proteobakteria
Rickettsia prowazekii	835	Proteobakteria
Rickettsia rickettsii Iowa	1384	Proteobakteria
Rickettsia rickettsii Sheila Smith	1345	Proteobakteria

Organismus	# Proteine	Phylum
Rickettsia typhi str. wilmington	838	Proteobakteria
Roseobacter denitrificans OCh 114	4129	Proteobakteria
Saccharophagus degradans 2-40	4008	Proteobakteria
Salmonella enterica subsp. arizonae serovar 62:z4,z23:–	4510	Proteobakteria
Salmonella enterica subsp. enterica serovar Choleraesuis	4648	Proteobakteria
Salmonella enterica subsp. enterica serovar Paratyphi A *str.* ATCC 9150	4093	Proteobakteria
Salmonella enterica subsp. enterica serovar Paratyphi B *str.* SPB7	5601	Proteobakteria
Salmonella typhi	4758	Proteobakteria
Salmonella typhi Ty2	4318	Proteobakteria
Salmonella typhimurium LT2	4527	Proteobakteria
Serratia proteamaculans 568	4942	Proteobakteria
Shewanella amazonensis SB2B	3645	Proteobakteria
Shewanella baltica OS155	4489	Proteobakteria
Shewanella baltica OS185	4394	Proteobakteria
Shewanella baltica OS195	4688	Proteobakteria
Shewanella denitrificans OS217	3754	Proteobakteria
Shewanella frigidimarina NCIMB 400	4029	Proteobakteria
Shewanella halifaxensis HAW-EB4	4278	Proteobakteria
Shewanella loihica PV-4	3859	Proteobakteria
Shewanella oneidensis	4467	Proteobakteria
Shewanella pealeana ATCC 700345	4241	Proteobakteria
Shewanella putrefaciens CN-32	3972	Proteobakteria
Shewanella sediminis HAW-EB3	4497	Proteobakteria
Shewanella sp. ANA-3	4360	Proteobakteria
Shewanella sp. MR-4	3924	Proteobakteria
Shewanella sp. MR-7	4014	Proteobakteria
Shewanella sp. W3-18-1	4044	Proteobakteria
Shewanella woodyi ATCC 51908	4880	Proteobakteria
Shigella boydii Sb227	4285	Proteobakteria
Shigella dysenteriae	4506	Proteobakteria
Shigella flexneri 2a	4445	Proteobakteria
Shigella flexneri 2a *str.* 2457T	4068	Proteobakteria

Organismus	# Proteine	Phylum
Shigella flexneri 5 *str.* 8401	4116	Proteobakteria
Shigella sonnei Ss046	4475	Proteobakteria
Silicibacter pomeroyi DSS-3	4252	Proteobakteria
Silicibacter sp. TM1040	3864	Proteobakteria
Sinorhizobium medicae WSM419	6213	Proteobakteria
Sinorhizobium meliloti	6205	Proteobakteria
Sodalis glossinidius	2516	Proteobakteria
Sorangium cellulosum So ce 56	9384	Proteobakteria
Sphingomonas wittichii RW1	5345	Proteobakteria
Sphingopyxis alaskensis RB2256	3195	Proteobakteria
Sulfurovum sp. NBC37-1	2438	Proteobakteria
Syntrophobacter fumaroxidans MPOB	4064	Proteobakteria
Syntrophus aciditrophicus SB	3168	Proteobakteria
Thiobacillus denitrificans ATCC 25259	2827	Proteobakteria
Thiomicrospira crunogena XCL-2	2196	Proteobakteria
Thiomicrospira denitrificans ATCC 33889	2096	Proteobakteria
Verminephrobacter eiseniae EF01-2	4947	Proteobakteria
Vibrio cholerae	3835	Proteobakteria
Vibrio cholerae O395	3875	Proteobakteria
Vibrio fischeri ES114	3802	Proteobakteria
Vibrio harveyi ATCC BAA-1116	6055	Proteobakteria
Vibrio parahaemolyticus	4832	Proteobakteria
Vibrio vulnificus CMCP6	4484	Proteobakteria
Vibrio vulnificus YJ016	5024	Proteobakteria
Wigglesworthia brevipalpis	617	Proteobakteria
Wolbachia endosymbiont of *Drosophila melanogaster*	1195	Proteobakteria
Wolbachia endosymbiont strain TRS of *Brugia malayi*	805	Proteobakteria
Wolinella succinogenes	2042	Proteobakteria
Xanthobacter autotrophicus Py2	5035	Proteobakteria
Xanthomonas campestris	4181	Proteobakteria
Xanthomonas campestris pv. campestris str. 8004	4273	Proteobakteria
Xanthomonas campestris pv. vesicatoria str. 85-10	4726	Proteobakteria
Xanthomonas citri	4427	Proteobakteria
Xanthomonas oryzae pv. oryzae KACC10331	4145	Proteobakteria

Organismus	# Proteine	Phylum
Xanthomonas oryzae pv. oryzae MAFF 311018	4372	Proteobakteria
Xylella fastidiosa	2832	Proteobakteria
Xylella fastidiosa M12	2104	Proteobakteria
Xylella fastidiosa Temecula1	2036	Proteobakteria
Yersinia enterocolitica subsp. enterocolitica 8081	4051	Proteobakteria
Yersinia pestis Angola	4045	Proteobakteria
Yersinia pestis Antiqua	4364	Proteobakteria
Yersinia pestis biovar Mediaevalis str. K1973002	4142	Proteobakteria
Yersinia pestis CO92	4066	Proteobakteria
Yersinia pestis KIM	4202	Proteobakteria
Yersinia pestis Nepal516	4094	Proteobakteria
Yersinia pestis Pestoides F	4069	Proteobakteria
Yersinia pseudotuberculosis IP 31758	4324	Proteobakteria
Yersinia pseudotuberculosis IP32953	4038	Proteobakteria
Yersinia pseudotuberculosis YPIII	4192	Proteobakteria
Zymomonas mobilis subsp. mobilis ZM4	1998	Proteobakteria
Cryptosporidium parvum Iowa II	3805	Protist
Dictyostelium discoideum	13408	Protist
Entamoeba histolytica	8163	Protist
Paramecium tetraurelia	40043	Protist
Phytophthora infestans	22658	Protist
Phaeodactylum tricornutum	10402	Protist
Phytophthora sojae	19027	Protist
Phytophthora ramorum	15743	Protist
Plasmodium falciparum	5262	Protist
Tetrahymena thermophila	26007	Protist
Thalassiosira pseudonana	11397	Protist
Theileria parva strain Muguga	4059	Protist
Trichomonas vaginalis	59679	Protist

Tabelle A.2.: Anzahl der Proteine aus *Arabidopsis thaliana*, *Oryza sativa* und *Populus trichocarpa*, für die ein cyanobakterieller Ursprung durch phylogenetische Analyse und gemeinsames Auftreten in Clustern mit weniger als vier Sequenzen hergeleitet werden kann.

	A. thaliana		**O. sativa**		**P. trichocarpa**	
	Proteine	Cluster	Proteine	Cluster	Proteine	Cluster
# Signale in V.bäumen	612	447	476	389	701	488
# Signale in R.bäumen	572	420	446	374	654	457
# Signale in V.- und R.bäumen	541	398	441	346	607	432
# Signale nur in V.bäumen	62	43	55	36	80	47
# Signale nur in R.bäumen	31	23	35	28	47	25
# Signale in kleinen Clustern	21	21	29	29	89	89
insgesamt	562	419	470	375	696	521
# gleiche Signale (alle)	7.003	5.461	6.026	5.140	9.005	6.837
# gleiche Signale (geteilt)	1.485	932	1.198	875	2.090	1.363
cyanobakterieller Anteil (alle)	8 %	8,1 %	7,7 %	7,6 %	7,7 %	7,9 %
cyanobakterieller Anteil (geteilt)	37,3 %	45 %	38,3 %	43 %	31,9 %	37 %

Tabelle A.3.: Anzahl der Proteine aus *Chlamydomonas reinhardtii*, *Ostreococcus tauri* und *Cyanidioschyzon merolae*, für die ein cyanobakterieller Ursprung durch phylogenetische Analyse und gemeinsames Auftreten in Clustern mit weniger als vier Sequenzen hergeleitet werden kann.

	C. reinhardtii		**O. tauri**		**C. merolae**	
	Proteine	Cluster	Proteine	Cluster	Proteine	Cluster
# Signale in V.bäumen	437	380	348	325	253	230
# Signale in R.bäumen	420	375	328	305	234	215
# Signale in V.- und R.bäumen	382	340	298	270	219	201
# Signale nur in V.bäumen	47	33	36	33	28	25
# Signale nur in R.bäumen	38	35	29	25	15	14
# Signale in kleinen Clustern	32	32	29	29	8	8
insgesamt	414	372	327	299	227	209
# gleiche Signale (alle)	3.677	3.464	2.914	2.802	2.119	2.034
# gleiche Signale (geteilt)	995	867	748	672	623	565
cyanobakterieller Anteil (alle)	10,3 %	11 %	11,1 %	10,9 %	10,7 %	10,8 %
cyanobakterieller Anteil (geteilt)	40,3 %	43 %	42,1 %	44,5 %	36 %	37 %

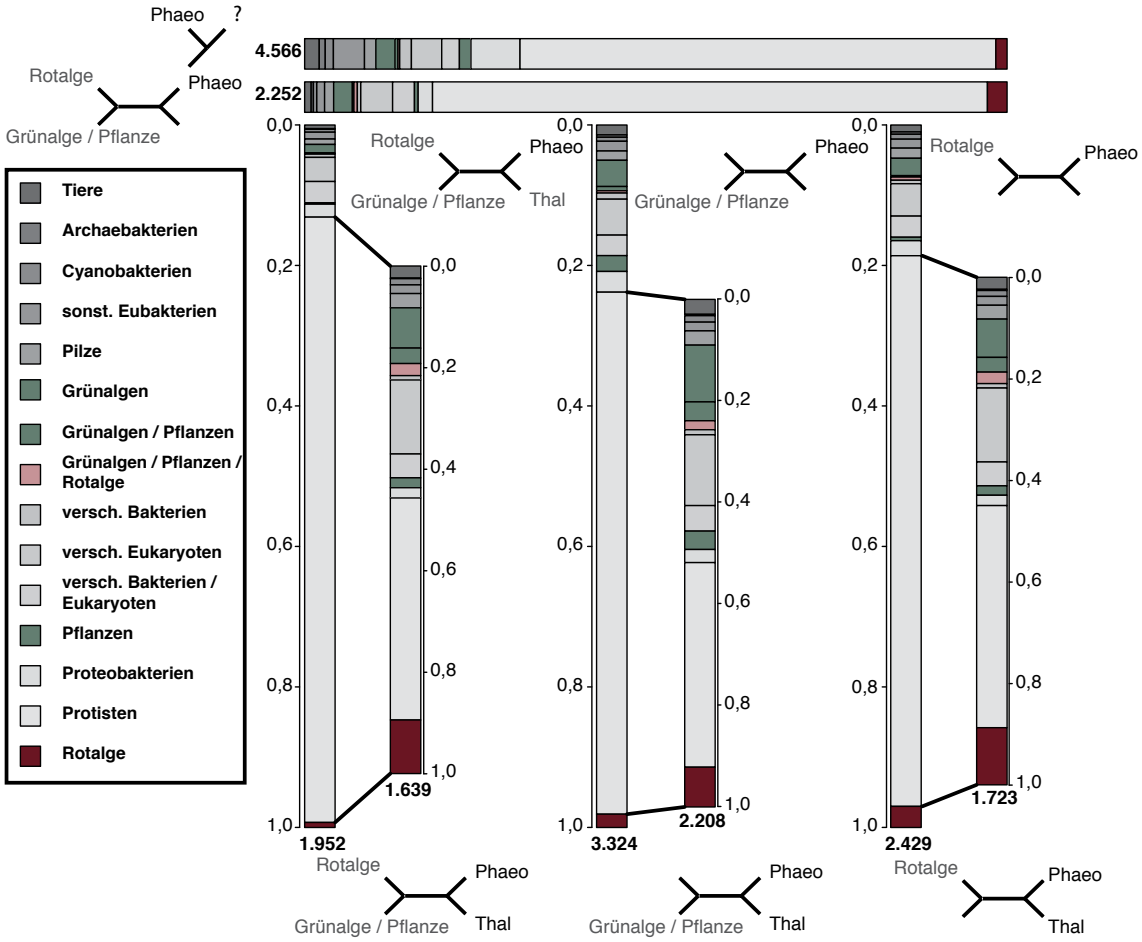

Abbildung A.1.: Nächste Nachbarn von *Phaeodactylum tricornutum* in verschiedenen Datensätzen. In jedem gestapelten Bargraphen sind die prozentualen Anteile der in der Legende dargestellten Gruppen an den nächsten Nachbarn (siehe Abschnitt 4.4.5.1) der Proteine von *P. tricornutum* dargestellt. In dem obersten Graphen wurden alle Bäume verwendet, in denen mindestens ein Protein von *P. tricornutum* vorkommt (4.566). Für den nächsten Graphen wurden alle Bäume verwendet, in denen mindestens ein Protein von *P. tricornutum* ein Grünalgen- oder Pflanzenprotein und ein Rotalgenprotein vorkommen (2.252). Der untere linke Graph zeigt die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein von *P. tricornutum*, ein Protein von *Thalassiosira pseudonana*, ein Grünalgen- oder Pflanzenprotein und ein Rotalgenprotein vorkommen (1.952). In 1.639 Bäumen ist *T. pseudonana* nächster Nachbar. Für diese Fälle sind die nächsten Nachbarn beider Proteine dargestellt. Für den mittleren Graphen wurden alle Bäume verwendet, in denen mindestens ein Protein von *P. tricornutum* und ein Grünalgen- oder Pflanzenprotein präsent sind (3.324). In 2.208 Bäumen ist *T. pseudonana* nächster Nachbar und die Anteile der nächsten Nachbarn beider Proteine sind dargestellt. Der rechte Graph zeigt die nächsten Nachbarn in allen Bäumen, in denen mindestens ein Protein von *P. tricornutum* und ein Rotalgenprotein vorkommen (2.429). In 1.723 Bäumen ist *T. pseudonana* nächster Nachbar.

Tabelle A.4.: Anzahl der nächsten Nachbarn der Proteine der Bacillariophyta in verschiedenen taxonomischen Gruppen

taxonomische Gruppe	*T. pseudonana*	*P. tricornutum*
Tiere	358	333
Archaebakterien	46	69
Chromalveolaten	94	130
Cyanobakterien	241	141
Eubakterien	235	308
Pilze	241	256
Grünalgen	550	506
Grünalgen/Pflanzen	281	274
Grünalgen/Rotalge	112	115
versch. Bakterien	136	158
versch. Eukaryoten	677	677
versch. Bakterien und Eukaryoten	232	211
Pflanzen	277	257
Proteobakterien	355	478
Protisten	204	233
Rotalge	413	425

Tabelle A.5.: Anzahl der nächsten Nachbarn der Proteine der Oomyceten in verschiedenen taxonomischen Gruppen

taxonomische Gruppe	*P. infestans*	*P. sojae*	*P. ramorum*
Tiere	683	589	596
Archaebakterien	85	79	72
Chromalveolaten	163	166	175
Cyanobakterien	74	69	44
Eubakterien	452	372	345
Pilze	396	370	387
Grünalgen	386	367	381
Grünalgen/Pflanzen	271	256	264
Grünalgen/Rotalge	98	90	94
versch. Bakterien	153	165	161
versch. Eukaryoten	840	763	790
versch. Bakterien und Eukaryoten	231	214	214
Pflanzen	351	311	320
Proteobakterien	688	531	497
Protisten	366	344	381
Rotalge	324	285	299

Tabelle A.6.: Anzahl der nächsten Nachbarn der Proteine der Ciliophora in verschiedenen taxonomischen Gruppen

taxonomische Gruppe	*P. tetraurelia*	*T. thermophila*
Tiere	546	362
Archaebakterien	67	46
Chromalveolaten	181	131
Cyanobakterien	25	25
Eubakterien	208	182
Pilze	357	235
Grünalgen	210	121
Grünalgen/Pflanzen	98	76
Grünalgen/Rotalge	54	30
versch. Bakterien	69	65
versch. Eukaryoten	1023	668
versch. Bakterien und Eukaryoten	156	104
Pflanzen	267	176
Proteobakterien	268	184
Protisten	795	555
Rotalge	209	153

Tabelle A.7.: Anzahl der nächsten Nachbarn der Proteine der Apicomplexa in verschiedenen taxonomischen Gruppen

taxonomische Gruppe	*T. parva*	*P. falciparum*	*C. parvum*
Tiere	99	112	107
Archaebakterien	17	20	10
Chromalveolaten	70	83	83
Cyanobakterien	5	7	4
Eubakterien	61	79	30
Pilze	127	135	135
Grünalgen	41	52	45
Grünalgen/Pflanzen	33	38	40
Grünalgen/Rotalge	20	20	29
versch. Bakterien	9	16	13
versch. Eukaryoten	474	487	493
versch. Bakterien und Eukaryoten	53	64	48
Pflanzen	46	58	43
Proteobakterien	68	67	44
Protisten	208	254	276
Rotalge	123	120	121

Tabelle A.8.: Anzahl der nächsten Nachbarn der Proteine von *T. vaginalis* in verschiedenen taxonomischen Gruppen

taxonomische Gruppe	*T. vaginalis*
Tiere	319
Archaebakterien	109
Chromalveolaten	64
Cyanobakterien	24
Eubakterien	324
Pilze	279
Grünalgen	136
Grünalgen/Pflanzen	26
Grünalgen/Rotalge	14
versch. Bakterien	86
versch. Eukaryoten	593
versch. Bakterien und Eukaryoten	105
Pflanzen	145
Proteobakterien	227
Protisten	539
Rotalge	126

Literaturverzeichnis

Adams III, E. N. Consensus techniques and the comparison of taxonomic trees. *Syst Zool*, 1972. **21**:390–397.

Agatha, S., Strüder-Kypke, M. C. und Beran, A. Morphologic and genetic variability in the marine planktonic ciliate *Laboea strobila* Lohmann, 1908 (ciliophora, oligotrichia), with notes on its ontogenesis. *J Eukaryot Microbiol*, 2004. **51**:267–281.

Allen, J. The function of genomes in bioenergetic organelles. *Philos Trans R Soc London, Ser B*, 2003. **358**:19–37.

Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P. und Burch, I. W. Stromatolite reef from the early archaean era of australia. *Nature*, 2006. **441**:714–718.

Alsmark, U. C., Sicheritz-Ponten, T., Foster, P. G., Hirt, R. P. und Embley, T. M. Horizontal gene transfer in eukaryotic parasites: a case study of *Entamoeba histolytica* and *Trichomonas vaginalis*. *Meth Mol Biol*, 2009. **532**:489–500.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. und Lipman, D. J. Basic local alignment search tool. *J Mol Biol*, 1990. **215**:403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. und Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997. **25**:3389–3402.

Anbar, A. D., Duan, Y., Lyons, T. W., Arnold, G. L., Kendall, B., Creaser, R. A., Kaufman, A. J., Gordon, G. W., Scott, C., Garvin, J. und Buick, R. A whiff of oxygen before the great oxidation event? *Science*, 2007. **317**:1903–1906.

Archibald, J. M. Algal genomics: exploring the imprint of endosymbiosis. *Curr Biol*, 2006. **16**:R1033–R1035.

Archibald, J. M. Nucleomorph genomes: structure, function, origin and evolution. *Bioessays*, 2007. **29**:392–402.

Archibald, J. M. Plastid evolution: remnant algal genes in ciliates. *Curr Biol*, 2008. **18**:R663–R665.

Archibald, J. M. und Keeling, P. J. Recycled plastids: a 'green movement' in eukaryotic evolution. *Trends Genet*, 2002. **18**:577–584.

Arisue, N., Hashimoto, T., Yoshikawa, H., Nakamura, Y., Nakamura, G., Nakamura, F., Yano, T.-A. und Hasegawa, M. Phylogenetic position of *Blastocystis hominis* and of stramenopiles inferred from multiple molecular sequence data. *J Eukaryot Microbiol*, 2002. **49**:42–53.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. W. Y., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P. und Rokhsar, D. S. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, 2004. **306**:79–86.

Bachvaroff, T. R., Sanchez Puerta, M. V. und Delwiche, C. F. Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages. *Mol Biol Evol*, 2005. **22**:1772–1782.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. und Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 2000. **290**:972–977.

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F.-J., Dupré, J., Dagan, T., Boucher, Y. und Martin, W. Prokaryotic evolution and the tree of life are two different things. *Biol Direct*, 2009. **4**:34.

Barta, J. R. und Thompson, R. C. A. What is cryptosporidium? Reappraising its biology and phylogenetic affinities. *Trends Parasitol*, 2006. **22**:463–468.

Bininda-Emonds, O. R., Gittleman, J. L. und Purvis, A. Building large trees by combining phylogenetic information: a complete phylogeny of the extant carnivora (mammalia). *Biol Rev Camb Philos Soc*, 1999. **74**:143–175.

Blouin, C., Perry, S., Lavell, A., Susko, E. und Roger, A. J. Reproducing the manual annotation of multiple sequence alignments using a svm classifier. *Bioinformatics*, 2009. **25**:3093–8.

Bodyl, A. Do plastid-related characters support the chromalveolate hypothesis? *J Phycol*, 2005. **41**:712–719.

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C. J., Coesel, S., De Martino, A., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P. G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Oudot-Le Secq, M.-P., Napoli, C., Obornik, M., Parker, M. S., Petit, J.-L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Van de Peer, Y. und Grigoriev, I. V. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 2008. **456**:239–244.

Buick, R. When did oxygenic photosynthesis evolve? *Philos Trans R Soc Lond B Biol Sci*, 2008. **363**:2731–2743.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S. I., Jakobsen, K. S. und Pawlowski, J. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One*, 2007. **2**:e790.

Burki, F., Shalchian-Tabrizi, K. und Pawlowski, J. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol Lett*, 2008. **4**:366–369.

Burki, F., Inagaki, Y., te, J. B., Archibald, J. M., Keeling, P. J., Cavalier-Smith, T., Sakaguchi, M., Hashimoto, T., Horak, A., Kumar, S., Klaveness, D., Jakobsen, K. S., Pawlowski, J. und Shalchian-Tabrizi, K. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biol Evol*, 2009. **1**:231–238.

Campbell, L., Nolla, H. A. und Vaulot, D. The importance of Prochlorococcus to community structure in the central north pacific-ocean. *Limnol Oceanogr*, 1994. **39**:954–961.

Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., Wortman, J. R., Bidwell, S. L., Alsmark, U. C. M., Besteiro, S., Sicheritz-Ponten, T., Noel, C. J., Dacks, J. B., Foster, P. G., Simillion, C., Van de Peer, Y., Miranda-Saavedra, D., Barton, G. J., Westrop, G. D., Müller, S., Dessi, D., Fiori, P. L., Ren, Q., Paulsen, I., Zhang, H., Bastida-Corcuera, F. D., Simoes-Barbosa, A., Brown, M. T., Hayes, R. D., Mukherjee, M., Okumura, C. Y., Schneider, R., Smith, A. J., Vanacova, S., Villalvazo, M., Haas, B. J., Pertea, M., Feldblyum, T. V., Utterback, T. R., Shu, C.-L., Osoegawa, K., de Jong, P. J., Hrdy, I., Horvathova, L., Zubacova, Z., Dolezal, P., Malik, S.-B., Logsdon, J. M., Jr, Henze, K., Gupta, A., Wang, C. C., Dunne, R. L., Upcroft, J. A., Upcroft, P., White, O., Salzberg, S. L., Tang, P., Chiu, C.-H., Lee, Y.-S., Embley, T. M., Coombs, G. H., Mottram, J. C., Tachezy, J., Fraser-Liggett, C. M. und Johnson, P. J. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science*, 2007. **315**:207–212.

Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, 2000. **17**:540–552.

Cavalier-Smith, T. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol*, 1999. **46**:347–366.

Cavalier-Smith, T. Genomic reduction and evolution of novel genetic membranes

and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Philos Trans R Soc Lond B Biol Sci*, 2003. **358**:109–133.

Cavalli-Sforza, L. L., Barrai, I. und Edwards, A. W. Analysis of human evolution under random genetic drift. *Cold Spring Harb Symp Quant Biol*, 1964. **29**:9–20.

Chaal, B. K. und Green, B. R. Protein import pathways in „complex" chloroplasts derived from secondary endosymbiosis involving a red algal ancestor. *Plant Mol Biol*, 2005. **57**:333–342.

Codd, G., Bell, S., Kaya, K., Ward, C., Beattie, K. und Metcalf, J. Cyanobacterial toxins, exposure routes and human health. *Eur J Phyco*, 1999. **34**:405–415.

Creevey, C. J. und McInerney, J. O. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, 2005. **21**:390–392.

Dagan, T. und Martin, W. Seeing green and red in diatom genomes. *Science*, 2009. **324**:1651–1652.

Dagan, T., Artzy-Randrup, Y. und Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA*, 2008. **105**:10039–10044.

Daugbjerg, N. und Andersen, R. A. Phylogenetic analyses of the rbcl sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. *Mol Biol Evol*, 1997. **14**:1242–1251.

Dayhoff, M. O., Eck, R. V. und Park, C. M. A model of evolutionary change in proteins. In **Dayhoff, M. O.** (Herausgeber) *Atlas of protein sequence and structure*, Band 5. Natl Biomed Res Found, Washington DC., 1972. 89–99.

Dayhoff, M. O., Hunt, L. T., Barker, W. C., Schwartz, R. M., Orcutt, B. C. und Young, C. L. A model of evolutionary change in proteins. In **Dayhoff, M. O.** (Herausgeber) *Atlas of Protein Sequence and Structure*, Band 5 suppl. 3. Natl Biomed Res Found, Washington DC., 1978. 345–352.

Des Marais, D. J. Earth's early biosphere. *Gravit Space Biol Bull*, 1998. **11**:23–30.

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W. und Dagan, T. Genes of cyanobacterial origin in plant nuclear

genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol*, 2008. **25**:748–761.

van Dongen, S. A cluster algorithm for graphs. Technischer Bericht INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, 2000*a*.

van Dongen, S. *Graph Clustering by Flow Simulation*. Dissertation, University of Utrecht, 2000*b*.

van Dongen, S. Performance criteria for graph clustering and markov cluster experiments. Technischer Bericht INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, 2000*c*.

Doolittle, R. F. *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences*. University Science Books, 1986.

Doolittle, W. F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*, 1998. **14**:307–311.

Dwivedi, B. und Gadagkar, S. R. Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol Biol*, 2009. **9**:211.

Edgar, R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 2004*a*. **5**:1–19.

Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 2004*b*. **32**:1792–1797.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Jr, Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P. und Orias, E. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol*, 2006. **4**:e286.

Elias, M. und Archibald, J. M. Sizing up the genomic footprint of endosymbiosis. *BioEssays*, 2009. **31**:1273–1279.

Falkowski, P. G. und Godfrey, L. V. Electrons, life and the evolution of Earth's oxygen cycle. *Philos Trans R Soc Lond B Biol Sci*, 2008. **363**:2705–2716.

Fast, N. M., Kissinger, J. C., Roos, D. S. und Keeling, P. J. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol*, 2001. **18**:418–426.

Fay, P. Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol Rev*, 1992. **56**:340–373.

Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 1981. **17**:368–376.

Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*, 2005.

Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H. und Zubrzycki, S. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Math*, 1951. **2**:282–285.

Foth, B. J. und McFadden, G. I. The apicoplast: a plastid in *Plasmodium falciparum* and other apicomplexan parasites. *Int Rev Cytol*, 2003. **224**:57–110.

Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, S., Maier, U. G., Grossman, A. R., Bhattacharya, D. und Lohr, M. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol*, 2008. **25**:2653–2667.

Fry, W. Phytophthora infestans: the plant (and r gene) destroyer. *Mol Plant Pathol*, 2008. **9**:385–402.

Funes, S., Davidson, E., Reyes-Prieto, A., Magallón, S., Herion, P., King, M. P. und González-Halphen, D. A green algal apicoplast ancestor. *Science*, 2002. **298**:2155.

Gallon, J. R., LaRue, T. A. und Kurz, W. G. W. Photosynthesis and nitrogenase activity in blue-green alga gloeocapsa. *Can J Microbiol*, 1974. **20**:1633–1637.

Gibbs, S. P. The chloroplasts of Euglena may have evolved from symbiotic green algae. *Can J Bot*, 1978. **56**:2883–2889.

Gould, S. B., Waller, R. R. und McFadden, G. I. Plastid evolution. *Annu Rev Plant Biol*, 2008. **59**:491–517.

Gross, W., Lenze, D., Nowitzki, U., Weiske, J. und Schnarrenberger, C. Characterization, cloning, and evolutionary history of the chloroplast and cytosolic class I aldolases of the red alga *Galdieria sulphuraria*. *Gene*, 1999. **230**:7–14.

Hackett, J. D., Yoon, H. S., Li, S., Reyes-Prieto, A., Rümmele, S. E. und Bhattacharya, D. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Mol Biol Evol*, 2007. **24**:1702–1713.

Hagopian, J. C., Reis, M., Kitajima, J. a. P., Bhattacharya, D. und de Oliveira, M. C. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J Mol Evol*, 2004. **59**:464–477.

Hall, B. G. Comparison of the accuracies of several phylogenetic methods using protein and dna sequences. *Mol Biol Evol*, 2005. **22**(3):792–802.

Hall, B. G. How well does the hot score reflect sequence alignment accuracy? *Mol Biol Evol*, 2008. **25**(8):1576–80.

Hamming, R. W. Error detecting and error correcting codes. *Bell System Technical Journal*, 1950. **29**:147–160.

Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B. und Roger, A. J. Phylogenomic analyses support the monophyly of excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A*, 2009. **106**:3859–3864.

Harper, J. T. und Keeling, P. J. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. *Mol Biol Evol*, 2003. **20**:1730–1735.

Harper, J. T., Waanders, E. und Keeling, P. J. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int J Syst Evol Microbiol*, 2005. **55**:487–496.

Hempel, F., Bullmann, L., Lau, J., Zauner, S. und Maier, U. G. ERAD-derived pre-protein transport across the 2nd outermost plastid membrane of diatoms. *Mol Biol Evol*, 2009. **26**:1781–1790.

Hirosawa, M., Totoki, Y., Hoshida, M. und Ishikawa, M. Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput Appl Biosci*, 1995. **11**:13–18.

Holland, B., Conner, G., Huber, K. und Moulton, V. Imputing supertrees and supernetworks from quartets. *Syst Biol*, 2007. **56**:57–67.

Holland, B. R., Benthin, S., Lockhart, P. J., Moulton, V. und Huber, K. T. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol Biol*, 2008. **8**:202.

Holland, H. D. The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B Biol Sci*, 2006. **361**:903–915.

Hughes, R., Robertson, M., Ellington, A. und Levy, M. The importance of prebiotic chemistry in the RNA world. *Curr Opin Chem Biol*, 2004. **8**:629–633.

Huson, D. und Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 2006. **23**:254–267.

Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 1998. **14**:68–73.

Iida, K., Takishita, K., Ohshima, K. und Inagaki, Y. Assessing the monophyly of chlorophyll-c containing plastids by multi-gene phylogenies under the unlinked model conditions. *Mol Phylogenet Evol*, 2007. **45**:227–238.

Imanian, B. und Keeling, P. J. The dinoflagellates durinskia baltica and kryptoperidinium foliaceum retain functionally overlapping mitochondria from two evolutionarily distinct lineages. *BMC Evol Biol*, 2007. **7**:172.

Ishida, K., Cavalier-Smith, T. und Green, B. Endomembrane structure and the chloroplast protein targeting pathway in *Heterosigma akashiwo* (raphidophyceae, chromista). *J Phycol*, 2000. **36**:1135–1144.

Jain, A. K. und Dubes, R. C. *Algorithms for clustering data*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

Javaux, E. J. The early eukaryotic fossil record. *Adv Exp Med Biol*, 2007. **607**:1–19.

Jiroutová, K., Horák, A., Bowler, C. und Oborník, M. Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. *J Mol Evol*, 2007. **65**:496–511.

Jones, D. T., Taylor, W. R. und Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 1992. **8**:275–282.

Jukes, T. H. und Cantor, C. R. *Evolution of Protein Molecules*. Academic Press, 1969.

Kasting, J. F. und Howard, M. T. Atmospheric composition and climate on the early earth. *Philos Trans R Soc Lond B Biol Sci*, 2006. **361**:1733–1741.

Kaufman, A., Johnston, D., Farquhar, J., Masterson, A., Lyons, T., Bates, S., Anbar, A., Arnold, G., Garvin, J. und Buick, R. Late archean biospheric oxygenation and atmospheric evolution. *Science*, 2007. **10**:1900–1903.

Kawakita, A., Sota, T., Ascher, J. S., Ito, M., Tanaka, H. und Kato, M. Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (bombus). *Mol Biol Evol*, 2003. **20**:87–92.

Keeling, P. J. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol*, 2009. **56**:1–8.

Kerr, R. Earth science - the story of O_2. *Science*, 2005. **308**:1730–1732.

Khan, H., Parks, N., Kozera, C., Curtis, B. A., Parsons, B. J., Bowman, S. und Archibald, J. M. Plastid genome sequence of the cryptophyte alga *Rhodomonas salina* CCMP1319: lateral transfer of putative DNA replication machinery and a test of chromist plastid phylogeny. *Mol Biol Evol*, 2007. **24**:1832–1842.

Kim, Y.-O., Chae, J., Hong, J.-S. und Jang, P.-G. Comparing the distribution of ciliate plankton in inner and outer areas of a harbor divided by an artificial breakwater. *Mar Environ Res*, 2007. **64**:38–53.

Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 1980. **16**:111–120.

Kimura, M. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

Köhler, S., Delwiche, C. F., Denny, P. W., Tilney, L. G., Webster, P., Wilson, R. J., Palmer, J. D. und Roos, D. S. A plastid of probable green algal origin in apicomplexan parasites. *Science*, 1997. **275**:1485–1489.

Koonin, E. V. und Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res*, 2008. **36**:6688–6719.

Kumar, S. und Filipski, A. Multiple sequence alignment: in pursuit of homologous dna positions. *Genome Res*, 2007. **17**(2):127–35.

Landan, G. und Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, 2007. **24**:1380–1383.

Landan, G. und Graur, D. Characterization of pairwise and multiple sequence alignment errors. *Gene*, 2009. **441**:141–147.

Lane, C. E. und Archibald, J. M. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol*, 2008. **23**:268–75.

Laybourn-Parry, J. Survival mechanisms in antarctic lakes. *Philos Trans R Soc Lond B Biol Sci*, 2002. **357**:863–869.

Li, S., Nosenko, T., Hackett, J. D. und Bhattacharya, D. Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. *Mol Biol Evol*, 2006. **23**:663–674.

Lió, P. und Goldman, N. Models of molecular evolution and phylogeny. *Genome Res*, 1998. **8**:1233–1244.

Margush, T. und McMorris, F. R. Consensus n-trees. *Bull Math Biol*, 1981. **43**:239–244.

Martens, C., Vandepoele, K. und Van de Peer, Y. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc Natl Acad Sci USA*, 2008. **105**:3427–3432.

Martin und Herrmann. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol*, 1998. **118**:9–17.

Martin, W. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proc R Soc Lond Ser B Biol Sci*, 1999. **266**:1387–1395.

Martin, W. und Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature*, 1998. **392**:37–41.

Martin, W. und Russell, M. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos Trans R Soc Lond B Biol Sci*, 2003. **358**:59–83.

Martin, W. und Schnarrenberger, C. The evolution of the calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr Genet*, 1997. **32**:1–18.

Martin, W., Brinkmann, H., Savonna, C. und Cerff, R. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc Natl Acad Sci USA*, 1993. **90**:8692–8696.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. und Penny, D. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA*, 2002. **99**:12246–12251.

Martin, W., Rotte, C., Hoffmeister, M., Theissen, U., Gelius-Dietrich, G., Ahr, S. und Henze, K. Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. *IUBMB Life*, 2003. **55**:193–204.

Maruyama, S., Matsuzaki, M., Misawa, K. und Nozaki, H. Cyanobacterial contribution to the genomes of the plastid-lacking protists. *BMC Evol Biol*, 2009. **9**:197.

Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S.-Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., Yoshida, Y., Nishimura, Y., Nakao, S., Kobayashi, T., Momoyama, Y., Higashiyama, T., Minoda, A., Sano, M., Nomoto, H., Oishi, K., Hayashi, H., Ohta, F., Nishizaka, S., Haga, S., Miura, S., Morishita, T., Kabeya, Y., Terasawa, K., Suzuki, Y., Ishii, Y., Asakawa, S., Takano, H., Ohta, N., Kuroiwa, H., Tanaka, K., Shimizu, N., Sugano, S., Sato, N., Nozaki, H., Ogasawara, N., Kohara, Y. und Kuroiwa, T. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 2004. **428**:653–657.

McFadden, G. I. und Waller, R. F. Plastids in parasites of humans. *BioEssays*, 1997. **19**:1033–1040.

Mereschkowsky, C. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol Centralblatt*, 1905. **25**:593–604.

Moore, R. B., Oborník, M., Janouskovec, J., Chrudimský, T., Vancová, M., Green, D. H., Wright, S. W., Davies, N. W., Bolch, C. J. S., Heimann, K., Slapeta, J., Hoegh-Guldberg, O., Logsdon, J. M. und Carter, D. A. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature*, 2008. **451**:959–963.

Moustafa, A., Chan, C. X., Danforth, M., Zear, D., Ahmed, H., Jadhav, N., Savage, T. und Bhattacharya, D. A phylogenomic approach for studying plastid endosymbiosis. *Genome Inform*, 2008. **21**:165–76.

Moustafa, A., Beszteri, B., Maier, U. G., Bowler, C., Valentin, K. und Bhattacharya, D. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*, 2009. **324**:1724–1726.

Mullineaux, P. M., R, G. J. und Chaplin, A. E. Acetylene reduction (nitrogen fixation) by cyanobacteria grown under alternating light-dark cycles. *FEMS Microbiol Lett*, 1981. **10**:245–247.

Needleman, S. B. und Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 1970. **48**:443–453.

Nosenko, T., Lidie, K. L., Van Dolah, F. M., Lindquist, E., Cheng, J.-F. und Bhattacharya, D. Chimeric plastid proteome in the florida "red tide" dinoflagellate *Karenia brevis*. *Mol Biol Evol*, 2006. **23**:2026–2038.

Notredame, C., Higgins, D. G. und Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 2000. **302**:205–217.

Nozaki, H. A new scenario of plastid evolution: plastid primary endosymbiosis before the divergence of the "plantae", emended. *J Plant Res*, 2005. **118**:247–55.

Nozaki, H., Iseki, M., Hasegawa, M., Misawa, K., Nakada, T., Sasaki, N. und Watanabe, M. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. *Mol Biol Evol*, 2007. **24**:1592–1595.

Nozaki, H., Maruyama, S., Matsuzaki, M., Nakada, T., Kato, S. und Misawa, K. Phylogenetic positions of glaucophyta, green plants (archaeplastida) and haptophyta (chromalveolata) as deduced from slowly evolving nuclear genes. *Mol Phylogenet Evol*, 2009. **53**:872–880.

Oakley, B. R. und Taylor, F. J. Evidence for a new type of endosymbiotic organization in a population of the ciliate *Mesodinium rubrum* from British Columbia. *Biosystems*, 1978. **10**(4):361–369.

Obornik, M. und Green, B. R. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol Biol Evol*, 2005. **22**:2343–2353.

Ogden, T. H. und Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol*, 2006. **55**:314–28.

Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E. V., Bowler, C. und Green, B. R. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. *Mol Genet Genomics*, 2007. **277**:427–439.

Patron, N. J., Rogers, M. B. und Keeling, P. J. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot Cell*, 2004. **3**:1169–1175.

Patron, N. J., Inagaki, Y. und Keeling, P. J. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. *Curr Biol*, 2007. **17**:887–891.

Remm, M., Storm, C. E. und Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 2001. **314**:1041–1052.

Reyes-Prieto, A., Hackett, J. D., Soares, M. B., Bonaldo, M. F. und Bhattacharya, D. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol*, 2006. **16**:2320–2325.

Reyes-Prieto, A., Moustafa, A. und Bhattacharya, D. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Curr Biol*, 2008. **18**:956–962.

Ricard, G., McEwan, N. R., Dutilh, B. E., Jouany, J.-P., Macheboeuf, D., Mitsumori, M., McIntosh, F. M., Michalowski, T., Nagamine, T., Nelson, N., Newbold, C. J., Nsabimana, E., Takenaka, A., Thomas, N. A., Ushida, K., Hackstein, J. H. P. und Huynen, M. A. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*, 2006. **7**:22.

Rice, P., Longden, I. und Bleasby, A. EMBOSS: the european molecular biology open software suite. *Trends Genet*, 2000. **16**:276–277.

Rippka, R. und Waterbury, J. B. The synthesis of nitrogenase by non-heterocystous cyanobacteria. *FEMS Microbiol Lett*, 1977. **2**:83–86.

Rippka, R., Deruelles, J., Waterbury, J., Herdman, M. und Stanier, R. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol*, 1979. **111**:1–61.

Rosenberg, M. S. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics*, 2005. **6**:278.

Rost, B. Twilight zone of protein sequence alignments. *Protein Eng*, 1999. **12**:85–94.

Sagan, L. On the origin of mitosing cells. *J Theor Biol*, 1967. **14**:255–274.

Saitou, N. und Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 1987. **4**:406–425.

Sanchez-Puerta, M. V., Bachvaroff, T. R. und Delwiche, C. F. Sorting wheat from chaff in multi-gene analyses of chlorophyll c-containing plastids. *Mol Phylogenet Evol*, 2007. **44**:885–897.

Sato, N. Origin and evolution of plastids: Genomic view on the unification and diversity of plastids. In **Springer Netherlands, D.** (Herausgeber) *The structure and function of plastids*. Wise, R R and Hoober, J K, 2006. 75–102.

Schopf, J. W. Microfossils of the early archean apex chert: new evidence of the antiquity of life. *Science*, 1993. **260**:640–646.

Schwartz, R. M. und Dayhoff, M. O. Protein and nucleic acid sequence data and phylogeny. *Science*, 1979. **205**:1038–1039.

Semple, C. und Steel, M. A supertree method for rooted trees. *Discr Appl Math*, 2000. **105**:147–158.

Simpson, A. G. B., Inagaki, Y. und Roger, A. J. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Mol Biol Evol*, 2006. **23**:615–625.

Smith, A. D., Lui, T. W. und Tillier, E. R. Empirical models for substitution in ribosomal RNA. *Mol Biol Evol*, 2004. **21**:419–427.

Sokal, R. R. und Michener, C. D. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 1958. **28**:1409–1438.

Sorensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skifter*, 1948. **5**:1–34.

Stanier, R. Y., Sistrom, W. R., Hansen, T. A., Whitton, B. A., Castenholz, R. W., Pfennig, N., Gorlenko, V. N., Kondratieva, E. N., Eimhjellen, K. E., Whittenbury, R., Gherna, R. L. und Truper, H. G. Proposal to place the nomenclature of the cyanobacteria (blue-green algae) under the rules of the international code of nomenclature of bacteria. *Int J Syst Bacteriol*, 1978. **28**:335–336.

Stewart, W. D. P. Some aspects of structure and function in N_2-fixing cyanobacteria. *Annu Rev Microbiol*, 1980. **34**:497–536.

Stiller, J. W., Huang, J., Ding, Q., Tian, J. und Goodwillie, C. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics*, 2009. **10**:484.

Stoecker, D. K., Silver, M. W., Michaels, A. E. und Davis, L. H. Obligate mixotrophy in *Laboea strobila*, a ciliate which retains chloroplasts. *Mar Biol*, 1988. **99**:415–423.

Swingley, W. D., Chen, M., Cheung, P. C., Conrad, A. L., Dejesa, L. C., Hao, J., Honchak, B. M., Karbach, L. E., Kurdoglu, A., Lahiri, S., Mastrian, S. D., Miyashita, H., Page, L., Ramakrishna, P., Satoh, S., Sattley, W. M., Shimada, Y., Taylor, H. L., Tomo, T., Tsuchiya, T., Wang, Z. T., Raymond, J., Mimuro, M., Blankenship, R. E. und Touchman, J. W. Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci U S A*, 2008. **105**:2005–2010.

Talavera, G. und Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 2007. **56**:564–577.

Tatusov, R. L., Koonin, E. V. und Lipman, D. J. A genomic perspective on protein families. *Science*, 1997*a*. **278**:631–637.

Tatusov, R. L., Koonin, E. V. und Lipman, D. J. A genomic perspective on protein families. *Science*, 1997*b*. **278**:631–637.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. und Koonin, E. V. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 2000. **28**:33–36.

Taylor, F. J. R. Implications and extensions of the serial endosymbiosis theory of the origin of eukaryotes. *Taxon*, 1974. **23**:229–258.

Taylor, F. J. R., Blackbourn, D. J. und Blackbourn, J. Ultrastructure of the chloroplasts and associated structures within the marine ciliate *Mesodinium rubrum* (lohmann). *Nature*, 1969. **224**:819 – 821.

Thompson, J. D., Higgins, D. G. und Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 1994. **22**:4673–4680.

Thompson, J. D., Plewniak, F. und Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 1999a. **27**:2682–2690.

Ting, C. S., Rocap, G., King, J. und Chisholm, S. W. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends Microbiol*, 2002. **10**:134–142.

Tomitani, A., Knoll, A. H., Cavanaugh, C. M. und Ohno, T. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A*, 2006. **103**:5442–5447.

Tyler, B. M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R. H. Y., Aerts, A., Arredondo, F. D., Baxter, L., Bensasson, D., Beynon, J. L., Chapman, J., Damasceno, C. M. B., Dorrance, A. E., Dou, D., Dickerman, A. W., Dubchak, I. L., Garbelotto, M., Gijzen, M., Gordon, S. G., Govers, F., Grunwald, N. J., Huang, W., Ivors, K. L., Jones, R. W., Kamoun, S., Krampis, K., Lamour, K. H., Lee, M.-K., McDonald, W. H., Medina, M., Meijer, H. J. G., Nordberg, E. K., Maclean, D. J., Ospina-Giraldo, M. D., Morris, P. F., Phuntumart, V., Putnam, N. H., Rash, S., Rose, J. K. C., Sakihama, Y., Salamov, A. A., Savidor, A., Scheuring, C. F., Smith, B. M., Sobral, B. W. S., Terry, A., Torto-Alalibo, T. A., Win, J., Xu, Z., Zhang, H., Grigoriev, I. V., Rokhsar, D. S. und Boore, J. L. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, 2006. **313**:1261–1266.

Vesteg, M., Vacula, R. und Krajcovic, J. On the origin of chloroplasts, import mechanisms of chloroplast-targeted proteins, and loss of photosynthetic ability - review. *Folia Microbiol (Praha)*, 2009. **54**:303–321.

Whitfield, J. B., Cameron, S. A., Huson, D. H. und Steel, M. A. Filtered z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. *Syst Biol*, 2008. **57**:939–947.

Williamson, D. H., Gardner, M. J., Preiser, P., Moore, D. J., Rangachari, K. und Wilson, R. J. The evolutionary origin of the 35 kb circular DNA of *Plasmodium*

falciparum: new evidence supports a possible rhodophyte ancestry. *Mol Gen Genet*, 1994. **243**:249–252.

Wilson, R. J., Denny, P. W., Preiser, P. R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D. J., Moore, P. W. und Williamson, D. H. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *J Mol Biol*, 1996. **261**:155–172.

Yoon, H. S., Hackett, J. D., Pinto, G. und Bhattacharya, D. The single, ancient origin of chromist plastids. *Proc Natl Acad Sci USA*, 2002. **99**:15507–15512.

Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. und Bhattacharya, D. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol*, 2004. **21**:809–818.

Zhu, G., J., M. M. und S., K. J. *Cryptosporidium parvum* appears to lack a plastid genome. *Microbiology*, 2000. **146**:315–321.

Danke

Ich danke Herrn Prof. Dr. William Martin dafür, dass er mir das interessante Thema überlassen hat. Insbesondere möchte ich mich für seine Geduld und Motivation während der Fertigstellung dieser Arbeit bedanken. Diverse Diskussionen, Teilnahmen an Konferenzen und der Auslandsaufenthalt in Neuseeland haben sowohl diese Arbeit bereichert als auch Perspektiven für die Zukunft aufgezeigt. Ebenso möchte ich mich für seine Bereitschaft bedanken die Gutachten für das Stipendium und für die vorliegende Arbeit in sehr kurzer Zeit anzufertigen. Außerdem bedanke ich mich für die Bereitstellung der Infrastruktur, ohne die die Fertigstellung dieser Arbeit nicht möglich gewesen wäre.

Ich danke Herrn Prof. Dr. Gerhard Steger für das meiner Arbeit entgegengebrachte Interesse, seine Bereitschaft das Korreferat zu übernehmen und das Gutachten für die vorliegende Arbeit in sehr kurzer Zeit anzufertigen.

Dr. Tal Dagan danke ich für die Einführung in das Programm `mcl` und für ihre Diskussionsbereitschaft.

Außerdem danke ich allen anderen Mitarbeitern des Instituts für Botanik III, die direkt oder indirekt bei dem Gelingen dieser Arbeit geholfen haben.

Dr. Gabriel Gelius-Dietrich danke ich für seine Hilfe bei diversen Programmier-, LaTeX- und Computerproblemen.

Mayo Röttger und Christian Eßer danke ich für die Programmierung des Programms `powerneedle`, ohne das ich wahrscheinlich immer noch paarweise Alignments berechnen würde. Auch für die zahlreichen Diskussionen möchte ich mich bedanken.

Dr. Nahal Ahmadinejad und Verena Zimorski danke ich für die langjährige gute Atmosphäre im "Mädels-Office" und die Erkenntnis, dass auch Frauen ohne Zickenterror zusammen arbeiten können. Die Möglichkeit einfach mal die Bürotür

schließen zu können, hat mir in einigen Fällen sehr geholfen. Verena Zimorski danke ich insbesondere für ihre Hilfe bei allen organisatorischen Problemen.

Ganz besonders bedanken möchte ich mich bei Dr. Oliver Deusch, der durch das Korrekturlesen der vorliegenden Arbeit diese erheblich verbessert hat. Durch diverse Diskussionen fielen mir oftmals andere und bessere Lösungswege für viele Probleme auf. Insbesondere bedanken möchte ich mich für die Geduld und die Motivation während der Fertigstellung dieser Arbeit bedanken. Danke!

Zu guter Letzt möchte ich mich bei meiner Familie – besonders bei meinen Großeltern Ursula und Martin Volkmer und Selma und Walter Grünheit – für ihre Unterstützung und ihr Verständnis bedanken. Bei meinen Eltern Angelika und Wolfgang Grünheit bedanke ich mich für die Gewissheit, im “Notfall“ immer nach Hause kommen zu können. Außerdem danke ich meiner Mutter für das Korrekturlesen der vorliegenden Arbeit. Meinem Bruder Thorsten Grünheit danke ich dafür, dass er immer für mich da war. Danke!

Die vorliegende Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde weder in der vorgelegten noch in ähnlicher Form bei einer anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 18. Januar 2010

.....

(Nicole Grünheit)