Bioinformatische Analysen der cyanobakteriellen Komponente von Pflanzen und Algen: Auswirkungen und Implikationen für Genomevolution und Stoffwechsel

Inaugural-Dissertation

zur Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Oliver Deusch

aus Düsseldorf

Juli 2009

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. Martin Lercher

Tag der mündlichen Prüfung: 01.09.2009

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

Publikationen in Fachzeitschriften

Ma, Y., Jakowitscha, J., Deusch, O., Henze, K., Martin, W. und Löffelhardt, W. Transketolase from *Cyanophora paradoxa*: *in vitro* import into cyanelles and pea chloroplasts and a complex history of a gene often, but not always, transferred in the context of secondary endosymbiosis. *J Eukaryot Microbiol*, im Druck.

Atteia, A., Adrait, A., Brugiere, S., van Lis, R., Tardif, M., Deusch, O., Dagan, T., Kuhn, L., Gontero, B., Martin, W., Garin, J., Joyard, J. und Rollanda, N. A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the α-proteobacterial mitochondrial ancestor. *Mol Biol Evol*, 2009. 26:1533-48.

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. 25:748–61.

Basu, M.K., Rogozin, I.B., Deusch, O., Dagan, T., Martin, W. und Koonin, E.V. Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues. *Mol Biol Evol*, 2008. 25:111–9.

Kilian, B., Ozkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W. und Salamini, F. Independent wheat B and G genome origins in outcrossing aegilops progenitor haplotypes. *Mol Biol Evol*, 2007. 24:217–27.

Kilian, B., Ozkan, H., Kohl, J., von Haeseler, A., Barale, F., Deusch, O., Brandolini, A., Yucel, C., Martin, W. und Salamini, F. Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol Genet Genomics*, 2006. 276:230–41.

Tagungsbeitrge

Deusch, O. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Jahrestagung der SMBE. Barcelona, Spanien, 2008. Vortrag.

Deusch, O., Landan, G., Röttger, M., Grünheit, N., Dagan, T. und Martin, W. Ancient phylogenies: Alignments make a difference. Jahrestagung der SMBE. Halifax, Kanada, 2007. Posterpräsentation.

Deusch, O., Dagan, T. und Martin, W. Trees for all *Arabidopsis thaliana* proteins compared against 237 reference genomes reveals that 16% come from cyanobacteria (with NJ). Jahrestagung des SFB TR1 der DFG. München, Deutschland, 2006. Posterpräsentation.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Zusammenfassung

Die Entstehung der Plastiden in Algen und Pflanzen beruht auf der Etablierung einer Endosymbiose zwischen einem heterotrophen Eukaryoten (Wirt) sowie einem cyanobakterien-ähnlichen Prokaryoten (Symbiont) und liegt mindestens 1,2 Milliarden Jahre zurück. Im Laufe der Zeit gingen viele Gene des cyanobakteriellen Genoms verloren oder wurden in das Kerngenom des Wirts übertragen, was als endosymbiontischer Gentransfer bezeichnet wird. Wieviele Pflanzengene auf Cyanobakterien zurückgehen ist – wie die Art des Symbionten – weitgehend ungeklärt. Die vorliegende Arbeit stellt eine vergleichende Genomanalyse von 83.138 Genen aus den Gattungen *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* mit deren Homologen aus einer Referenzdatenbank von 851.607 Genen aus neun Cyanobakterien, 215 weiteren prokaryotischen und 15 eukaryotischen Genomen dar. Auf diese Weise sollte der Anteil cyanobakterieller Gene in diesen vier Genomen aus Pflanzen und Algen bestimmt werden. Die Analyse ergab 11.569 Stammbäume, die sowohl mit der Methode der maximalen Wahrscheinlichkeit als auch mit einer Distanzmethode abgeleitet wurden. Die Verlässlichkeiten der phylogenetischen Bäume sowie der zugrundeliegenden Alignments wurden mit der *heads-or-tails*-Methode quantifiziert. Für durchschnittlich 14 % der analysierten Pflanzengene wurde ein cyanobakterieller Ursprung abgeleitet, unabhängig von der Methode zur Baumrekonstruktion. Die Alignmentverlässlichkeit hatte einen großen Einfluss auf die Ableitung des Gentransfers. Für die verlässlicheren Alignments lag dieser Schätzer zwischen 17 % und 25 %. Die Identifizierung cyanobakterieller Gene ermöglichte es, jene Cyanobakterien zu ermitteln, die dem freilebenden Vorfahren der Plastiden am ähnlichsten waren. Unter den neun in dieser Arbeit verwendeten rezenten Cyanobakterien hatten *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 die größte Ähnlichkeit zu den Pflanzengenen cyanobakteriellen Ursprungs. Beide gehören in der Systematik von Stanier zur Abteilung IV, die eine Gruppe von filamentösen, heterozysten-bildenden Cyanobakterien

bezeichnet, welche in der Lage sind, Stickstoff zu fixieren. Diese Beobachtung und die Tatsache, dass Mitglieder der Abteilung IV häufig in modernen Symbiosen zu finden sind, stützen die Hypothese, dass die Stickstofffixierung bei der Etablierung der Symbiose der Chloroplasten eine treibende Kraft dargestellt haben könnte. Für 60 % der cyanobakteriellen Pflanzenproteine wurde eine Lokalisierung in Chloroplasten abgeleitet. Die cyanobakteriellen Gene wurden funktionell charakterisiert und die entsprechenden Enzyme in den Stoffwechselkarten der KEGG-Datenbank hervorgehoben. Analysen der Introns cyanobakterieller Gene (die zum Zeitpunkt des Transfers intronfrei waren) und alter eukaryotischer Gene zeigten, dass die Introndichte cyanobakterieller Gene geringer ist, aber der Prozess des Introngewinns anhält.

2 Abstract

Plastids in algae and plants arose by an endosymbiotic event, which dates back at least 1.2 billion years and involved a cyanobacterium-like prokaryote (symbiont) and a heterotrophic eukaryote (host). During the establishment of that endosymbiosis many prokaryotic genes were lost from the symbiont's genome or relocated to the host nucleus (endosymbiontic gene transfer). The issue of how many genes were transferred is unresolved, as well as the identity of the symbiont. This work represents a comparative analysis of 83,138 plant genes from *A. thaliana*, *O. sativa*, *C. reinhardtii* and *C. merolae* to a reference database of 851,607 genes from nine cyanobacteria, 215 other prokaryotes and 15 eukaryotic genomes to determine the fraction of cyanobacterial genes in those genomes. The study yielded 11,569 phylogenies which were inferred using maximum likelihood as well as neighbor-joining approaches. The question of alignment reliability was addressed using the *heads-or-tails* method. On average 14 % of the phylogenies indicated a cyanobacterial origin of the plant gene, no matter which method of tree inference was used. Alignment reliability had a big impact on the percentage of inferred gene transfer yielding estimates between 17 % and 25 % for the most reliable alignments. The identification of cyanobacterial genes allowed to search for the cyanobacterium most similar to the ancestor of plastids. Among the nine cyanobacteria sampled, *Anabaena variabilis* ATCC 29413 and *Nostoc sp.* PCC 7120 had the collections of genes most similar to the plant genes of cyanobacterial ancestry. Both belong to section IV of Stanier's cyanobacterial classification which describes a group of filamentous, heterocyst-forming and nitrogen-fixing cyanobacteria. This—as well as studies describing members of section IV as common partners in contemporary symbiotic associations—supports the hypothesis, that nitrogen fixation may have played an important role during the origin of plastids. 60 % of cyanobacterial plant proteins were predicted to be targeted to the chloroplast. The enzymatic functions of cyanobacterial genes were analysed and highlighted in the metabolic maps of

the KEGG database. Comparative analyses of cyanobacterial genes (which were intronless at the time of transfer) and ancient eukaryotic genes showed that intron density in cyanobacterial genes is lower and intron gain still continues.

3 Einleitung

3.1 Die Entstehung des Pflanzenreichs durch Endosymbiose

Die Theorie, dass die Plastiden der Algen und Pflanzen auf eine Endosymbiose mit Cyanobakterien zurückgehen, ist über 100 Jahre alt und geht auf den russischen Biologen Constantin Sergejewitsch Mereschkowsky (1855–1921) zurück. Anfang des 20. Jahrhunderts beschrieb er, aufbauend auf den Arbeiten von Andreas Franz Wilhelm Schimper, erstmalig die Ähnlichkeit zwischen den Chloroplasten der Kieselalgen und freilebenden Cyanobakterien, die zu dieser Zeit noch „Blaualgen“ genannt wurden. In seinem Artikel „Über Natur und Ursprung der Chromatophoren im Pflanzenreiche“ postulierte er 1905 die Existenz von urtümlichen „Plasmaarten“, deren Kooperation und Integration zu höherentwickelten Lebensformen führe. Er beschrieb erstmalig den Mechanismus der Symbiogenese als wesentlichen Mechanismus der Evolution. Mereschkowskys Hypothese wurde zunächst wenig Beachtung geschenkt und sie geriet in Vergessenheit bis sie 1967 von Lynn Margulis wiederentdeckt wurde. In ihrem Artikel „On the Origin of Mitosing Cells“ postulierte sie, dass sich die Mitochondrien, die Chloroplasten und die Flagellen aus freilebenden Prokaryoten entwickelt haben (Sagan, 1967). Molekularbiologische Experimente und der Vergleich der Genome von Cyanobakterien und Plastiden sowie Proteobakterien und Mitochondrien lieferten eine Vielzahl von Indizien, die den endosymbiontischen Ursprung dieser Zellorganellen unterstützen. Heutzutage wird die Endosymbiontentheorie als Erklärung für den Ursprung der Chloroplasten (Archibald, 2006; Delwiche, 1999; Douglas, 1998; Matsuzaki et al., 2004; McFadden und van Dooren, 2004) und Mitochondrien (Embley und Martin, 2006; Embley et al., 2003; Martin und Muller, 1998) allgemein akzeptiert, während sie für die Flagellen abgelehnt wird (Cowan, 2000; Martin et al., 2001). Nach der Endosymbiontentheorie der Chloroplasten nahm vor mindestens 1,2 Milliarden Jahren (Butterfield, 2000; Yoon et al., 2004) ein heterotropher Eukaryot vermut-

lich über den Mechanismus der Phagocytose ein photoautotrophes Bakterium, welches den rezenten Cyanobakterien ähnlich war, auf. Der Eukaryot verdaute das Bakterium nicht, sondern behielt es als Endosymbionten. Der Wirt machte sich so die Fähigkeit des Symbionten zu eigen, Lichtenergie in chemische Energie umzuwandeln und wurde zum ersten photoautotrophen Eukaryoten. Aus diesem Organismus entwickelten sich im Laufe der Evolution alle Pflanzen und Algen mit primärer Plastide. Diese große taxonomische Gruppe wird in der Systematik von Adl et al. (2005) „Archeaplastida“genannt und umfasst Pflanzen und Grünalgen (Chloroplastida), Rotalgen (Rhodophyceae) und Glaucophyten (Glaucophyta). In dieser Systematik nehmen die Archeaplastida neben den Amoebozoa, Opisthokonta, Rhizaria, Chromalveolata und den Excavata den Rang einer Supergruppe der Domäne der Eukaryoten (Eukaryota) ein.

Neben den von zwei Biomembranen eingeschlossenen primären Plastiden existieren noch sekundäre Plastiden mit drei oder vier Membranen. Die einfachste Hypothese zur sekundären Endosymbiose postuliert drei unabhängige Ereignisse mit Rotalgen (Chromalveolata, Cavalier-Smith (2003)) sowie Grünalgen (Euglenophyta und Chlorarachniophyta (Archibald et al., 2003; Baldauf et al., 2000; Gilson et al., 2006; Leander, 2004; Rogers et al., 2007)) als Symbionten. Mit Glaucophyten als Symbionten sind bislang keine sekundären Endosymbiosen bekannt. In den Gattungen *Lepidodinium*, *Kryptoperidinium*, *Karlodinium* und *Dinophysis* führte eine tertiäre Endosymbiose dazu, dass die ursprüngliche sekundäre Plastide durch eine eukaryotische Alge mit sekundärer Plastide ausgetauscht wurde und so tertiäre Plastiden entstanden (Hackett et al., 2004).

3.2 Vom Endosymbionten zum Organell

Im Laufe der evolutionären Zeit wurde das Genom des cyanobakteriellen Endosymbionten immer weiter reduziert. Viele Gene gingen verloren oder wurden in das Kerngenom des eukaryotischen Wirts übertragen. Dieser Prozess wird als endosymbiontischer Gentransfer (EGT) bezeichnet (Martin et al., 1993). Die transferierten cyanobakteriellen Gene dienten dem Wirt als Quelle neuer und divergenter Gene. Die Produkte vieler dieser Gene wurden in den Chloroplasten importiert, wo sie ihre ursprüngliche Funktion ausüben konnten. Andere übernahmen neue Funktionen im Cytoplasma oder auch im Mitochondrium oder ersetzten vorhandene eukaryotische (oder prokaryotische) Proteine des Wirts (Allen, 2003; Bogorad, 2008; Martin und Herrmann, 1998; Martin und Schnarrenberger, 1997). Auch wenn

es Gegenbeispiele gibt (Abdallah et al., 2000; Martin und Schnarrenberger, 1997), wird für die Mehrheit der Proteine angenommen, dass Genursprung und Proteinlokalisierung übereinstimmen (Horiike et al., 2001). Martin et al. (2002) zeigten, dass die Hälfte der Pflanzenproteine cyanobakteriellen Ursprungs nicht in Chloroplasten lokalisiert sind, während Proteine anderen Ursprungs dort lokalisiert sein können.

Die Erfindung einer komplexen Proteinimportmaschinerie durch den Wirt ermöglichte es, die Genprodukte zuvor transferierter cyanobakterieller Gene in den Endosymbionten zu importieren. Dies bedeutete eine gravierende Veränderung für die Beziehung zwischen dem cyanobakteriellen Endosymbionten und dem Wirt. Zuvor diente das Genom des Endosymbionten dem Wirt als Quelle genetischen Materials. Das Genom des Endosymbionten stand jedoch weiterhin unter Selektionsdruck. Durch die Proteinimportmaschinerie konnten die biochemischen Eigenschaften des cyanobakteriellen Endosymbionten erhalten bleiben, wenn einzelne Gene durch Pseudogenisierung oder Verlust bzw. EGT nicht mehr zur Verfügung standen (Allen, 2003). Der Selektionsdruck auf die im cyanobakteriellen Chromosom kodierten Gene nahm ab und der Endosymbiont verlor im Laufe der evolutionären Zeit immer mehr Gene. Diese Vorgänge führten dazu, dass aus dem einst freilebenden Symbionten ein Zellorganell wurde, welches ohne den Wirt nicht länger lebensfähig war.

3.3 Endosymbiontischer Gentransfer in Zahlen

Als Konsequenz von Genverlust und endosymbiontischem Gentransfer enthalten die Plastidengenome der verschiedenen photosynthetischen Gruppen nur 60 bis 200 Gene, während Cyanobakteriengenome zwischen 1.884 und ca. 7.400 Genen besitzen (Timmis et al., 2004). Mittels bioinformatischer Analysen zur Lokalisierung von Proteinen in der Zelle für *Arabidopsis thaliana* und *Oryza sativa* wurde geschätzt, dass deren plastidäre Proteome aus 2.100 beziehungsweise 4.800 Proteinen bestehen (Richly und Leister, 2004). Somit sind plastidäre Proteome in etwa so groß wie die freilebender Cyanobakterien, während ihre Genome nur ca. fünf bis zehn Prozent der Gene von Cyanobakteriengenomen tragen.

Dieser Vergleich der Genomgröße rezenter Cyanobakterien mit Plastiden zeigt die Reduktion des Symbiontengenoms während des Übergangs vom Endosymbionten zum Organell. Analysen, die die Auswirkungen auf das Genom des Wirts untersuchen, sind komplexer und selten. Martin et al. (2002) folgern aus phylogene-

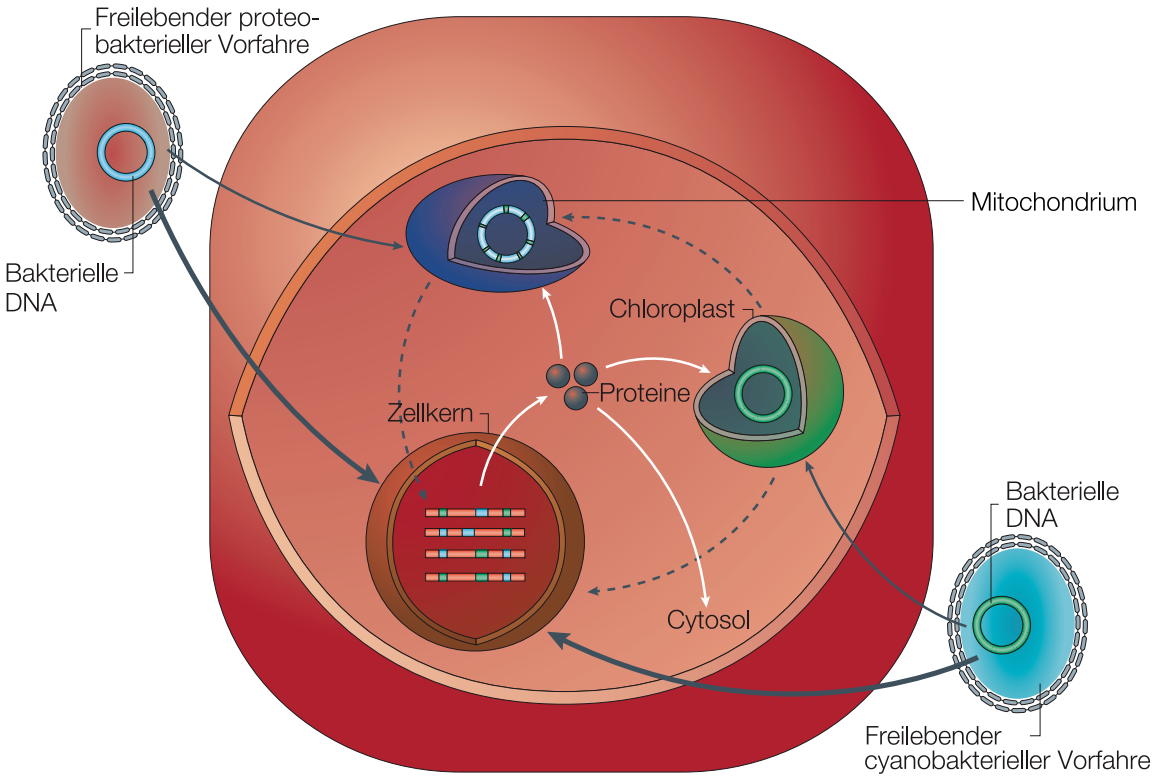

Abbildung 3.1: Endosymbiontischer Gentransfer während der Etablierung der Mitochondrien und Plastiden. Mitochondrien stammen von einem freilebenden Proteobakterium und Plastiden von einem freilebenden Cyanobakterium ab. Im Laufe der Evolution wurde der Großteil der prokaryotischen Gene in das Genom des eukaryotischen Wirts verlagert (dicke Pfeile). Organellen enthalten nur einen geringen Anteil der Gene der freilebenden Vorfahren (dünne Pfeile). Als Konsequenz werden mehr als 90 % des Proteoms der Organellen von kernkodierten Genen kodiert, deren Produkte aus dem Cytosol importiert werden (weiße Pfeile). Der Gentransfer von Organellen zum Kern hält bis heute an (gestrichelte Pfeile) und plastidäre DNA kann sogar in den Mitochondrien nachgewiesen werden. Verändert nach Timmis et al. (2004).

tischen Analysen aller kernkodierten Proteine von *Arabidopsis thaliana* sowie deren Homologe aus drei Cyanobakterien, 16 weiteren Prokaryoten und *Saccharomyces cerevisiae*, dass 18 % des Kerngenoms von *Arabidopsis thaliana* cyanobakteriellen Ursprungs ist. Eine andere Studie, die auf einer EST-Bibliothek des Glaucophyten *Cyanophora paradoxa* basiert, leitet eine cyanobakterielle Komponente von 11 % ab (Archibald, 2006; Reyes-Prieto et al., 2006). Für die Rotalgen – die dritte Gruppe der Archaeplastida – wurde in der Vorveröffentlichung der Ergebnisse dieser Arbeit (Deusch et al., 2008) erstmalig der cyanobakterielle Anteil von *Cyanidioschyzon merolae* mit mindestens 17,1 % abgeleitet.

3.4 Stoffwechselwege

Im Verlauf der Etablierung der Endosymbiose der Plastiden wurde aus einem heterotrophen Eukaryoten und einem photoautotrophen Prokaryoten der erste photoautotrophe Eukaryot. Aus metabolischer Sicht beinhaltete dieser Prozess das Verschmelzen zweier sehr verschiedener Kollektionen von Stoffwechselwegen zu einem neuen Organismus (Weeden, 1981). Die metabolischen Repertoires von Symbiont und Wirt unterschieden sich höchstwahrscheinlich nicht nur in Bezug auf die Fähigkeit zur Photosynthese und die damit verbundene Fähigkeit zur Fixierung energiereicher organischer Kohlenstoffverbindungen (Gould et al., 2008). Autotrophe Organismen synthetisieren alle Makromoleküle von Grund auf aus einfachen anorganischen Stoffen. Zwar gibt es einige Ausnahmen, in denen Vitamine benötigt werden, generell sind autotrophe Organismen jedoch weitgehend metabolisch unabhängig.

Heterotrophe Organismen haben über die Nahrung Zugang zu Makromolekülen, die sie nicht komplett verdauen. Stattdessen werden Zwischenprodukte des Katabolismus direkt für den Anabolismus verwendet. Als Konsequenz benötigen autotrophe Organismen viele Stoffwechselwege, die in heterotrophen Organismen fehlen. Im Zuge der Etablierung der Endosymbiose der Plastiden hat der Wirt vermutlich viele Stoffwechselwege – oder einzelne Enzyme daraus – vom Symbionten übernommen, die über die Fähigkeit zur Photosynthese hinausgehen. Ein gut untersuchtes Beispiel ist der Calvin-Zyklus (Abbildung 3.2) in photo- und auch chemoautotrophen Organismen. In einer zyklischen Folge lichtunabhängiger biochemischer Reaktionen erfolgt in diesem Stoffwechselweg die Assimilation von Kohlenstoff aus Kohlendioxid. Die Energie für diesen endergonen Prozess stammt aus Adenosintriphosphat (ATP), das bei photoautotrophen Organismen in der

Lichtreaktion gebildet wird. Bei chemoautotrophen Organismen kann das ATP aus verschiedenen exergonen Umsetzungen ihres Stoffwechsels stammen.

Martin und Schnarrenberger (1997) leiteten für acht der elf Enzyme des Calvin-Zyklus in Spinat einen endosymbiontischen Ursprung ab. Die Gene für sechs Enzyme stammen aus Cyanobakterien, zwei aus Proteobakterien und für drei Enzyme ist der Ursprung noch ungeklärt. Von dem Enzym Glycerinaldehyd-3-Phosphat Dehydrogenase (GAPDH) existieren in Spinat zwei Isoformen. Das Gen für das in Chloroplasten lokalisierte Nicotinamidadenindinukleotidphosphat (NADP)-abhängige Enzym stammt aus Cyanobakterien. Die cytosolische Isoform geht auf Proteobakterien zurück, benötigt Nicotinamidadenindinukleotid (NAD) als Kofaktor und ist ein Enzym der Glykolyse bzw. Glukoneogenese. Die Gene für die in den Chloroplasten lokalisierten Enzyme Triosephosphat Isomerase (TPI) und Fruktose-1,6-Bisphosphatase (FBP) stammen aus Proteobakterien und sind Gegenbeispiele für die Hypothese, dass bei Proteinen endosymbiontischen Ursprungs Donororganismus und Zielorganell übereinstimmen (Horiike et al., 2001). Diese beiden Enzyme üben Funktionen im Calvin-Zyklus und in der Glykolyse bzw. Glukoneogenese aus.

3.5 Introns

Spleißosomale Introns, welche proteinkodierende Gene unterbrechen, sowie das zugehörige Spleißosom sind charakteristische Merkmale der Eukaryoten (Deutsch und Long, 1999; Doolittle, 1978; Gilbert, 1978; Mattick, 1994). Bereits kurz nach der Entdeckung der Introns wurden zwei unterschiedliche Hypothesen entworfen, um den Ursprung und die Evolution der Introns zu erklären. Die *introns-early*-Hypothese (Darnell, 1978; Doolittle, 1978; Gilbert, 1978, 1987; Gilbert und Glynias, 1993; Gilbert et al., 1997) postuliert, dass Introns ein gemeinsames Merkmal proteinkodierender Gene waren, welches bereits im letzten gemeinsamen Vorfahren von Pro- und Eukaryoten vorhanden war und das Entstehen neuer Proteine durch Rekombination gefördert hat. Nach der Aufspaltung in die drei großen Domänen Eukaryoten, Eubakterien und Archaebakterien gingen die Introns in den Prokaryoten verloren.

Die *introns-late*-Hypothese (Logsdon, 1998; Logsdon und Palmer, 1994; Logsdon et al., 1995; Stoltzfus, 1994; Stoltzfus et al., 1994) postuliert, dass Prokaryoten nie Introns besessen haben und spleißosomale Introns und das Spleißosom Produkte der eukaryotischen Evolution sind. Bis heute konnten keine ursprünglichen In-

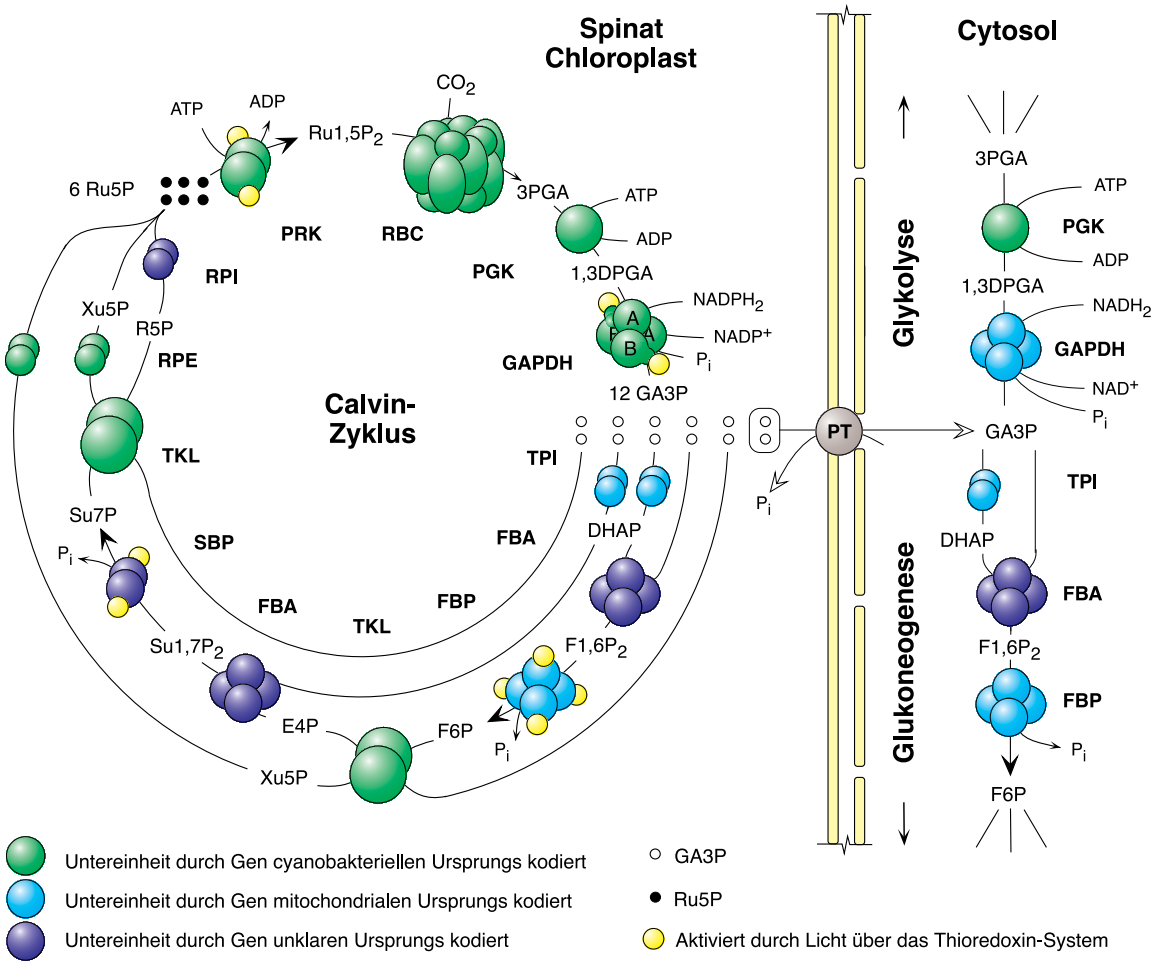

Abbildung 3.2: Calvin-Zyklus, Glykolyse und Glukoneogenese in Spinat. Der abgeleitete evolutionäre Ursprung der kernkodierten Gene ist farblich gekennzeichnet. Die Abkürzungen der Enzyme sind: FBA Fruktose-1,6-Bisphosphat Aldolase; FBP Fruktose-1,6-Bisphosphatase; GAPDH Glycerinaldehyd-3-Phosphat Dehydrogenase; PGK 3-Phosphoglycerat Kinase; PRI Ribose-5-Phosphat Isomerase; PRK Phosphoribulokinase; RBC Ribulose-1,5-Bisphosphat Carboxylase/Oxygenase; RPE Ribulose-5-Phosphat 3-Epimerase; SBP Sedoheptulose-1,7-Bisphosphatase; TKL Transketolase; TPI Triosephosphat Isomerase; PT Phosphat Translokator. Verändert nach Martin und Schnarrenberger (1997).

trons identifiziert werden, die die *introns-early*-Hypothese unterstützen würden. Da alle bekannten rezenten Eukaryoten Introns aufweisen, wird zudem davon ausgegangen, dass bereits der letzte gemeinsame Vorfahre aller Eukaryoten Introns besessen hat und diese die Errungenschaften einer sehr frühen eukaryotischen Evolution sind (Belshaw und Bensasson, 2006; Cho und Doolittle, 1997; Koonin, 2006; Logsdon et al., 1998; Stoltzfus et al., 1994).

Selbst-spleißende Introns der Gruppe II in Prokaryoten und Organellen weisen große Ähnlichkeiten mit spleißosomalen Introns sowie RNAs des Spleißosoms auf und werden als Vorläufer der eukaryotischen Introns angesehen (Lambowitz und Zimmerly, 2004; Toro et al., 2007; Zimmerly et al., 2001). Bis heute sind kaum Details darüber bekannt, wie sich spleißosomale Introns in eukaryotischen Genomen ausgebreitet haben (Koonin, 2006; Martin und Koonin, 2006). Ebensowenig ist über die Dynamik und die Mechanismen der Intronevolution bekannt. Die Häufigkeit von Introns wird oft mit der effektiven Populationsgröße und der Mutationsrate einer Spezies erklärt (Lynch und Conery, 2003; Lynch und Richardson, 2002). Neben diesen statistischen Erklärungen gibt es auch Indizien dafür, dass Introngewinn und -verlust von verschiedenen selektiven Kräften abhängig sind (Jeffares et al., 2006). Vergleichende Genomanalysen zeigen eine beeindruckende Konservierung von Intronpositionen in verschiedenen Tierarten (Raible et al., 2005). Auch zwischen orthologen Genen entfernt verwandter Eukaryoten wie den Tieren und den Pflanzen gibt es solche konservierten Positionen (Fedorov et al., 2002; Rogozin et al., 2003). Viele aktuelle Modelle zur Intronevolution sind widersprüchlich. Einige Modelle sehen Introngewinn (Qiu et al., 2004), andere Intronverlust (Roy und Gilbert, 2005*a*,*b*) als treibende Kraft der Evolution eukaryotischer Gene. Andere Szenarios beschreiben ein Zusammenspiel von Introngewinn und -verlust (Nguyen et al., 2005; Rogozin et al., 2003).

Bis heute konnten bei vergleichenden Genomanalysen noch keine Ereignisse des Introngewinns nachgewiesen werden. Coghlan und Wolfe (2004) beschreiben zwar den Gewinn von 122 Introns in Nematoden des Genus *Caenorhabditis* seit der Aufspaltung in *C. elegans* und *C. briggsae*. Eine erneute Analyse der Daten durch Roy und Penny (2006) konnte jedoch zeigen, dass die meisten Ereignisse des Introngewinns wahrscheinlich auf den Verlust alter Introns in der jeweils anderen Linie zurückgehen. Daher gibt es auch keine Indizien, die ein Modell zum Introngewinn unterstützen oder widerlegen würden, das die Entstehung neuer Introns durch Transposition existierender Introns erklärt (Fedorov et al., 2003). Als Mechanismus des Intronverlusts wird ein Rekombinationsereignis mit intronfreier cDNA vorgeschlagen (Fink, 1987; Mourier und Jeffares, 2003).

Ein Modell, das erklärt, warum Introngewinn so selten beobachtet wurde, geht davon aus, dass die Mehrzahl rezenter Introns zu einem sehr frühen Zeitpunkt der eukaryotischen Evolution entstanden sind und seitdem der Intronverlust die vorherrschende Kraft ist (Roy, 2006; Roy und Gilbert, 2005 *a*,*b*). Im Zuge des endosymbiontischen Gentransfers gelangten viele cyanobakterielle Gene in das Kerngenom des eukaryotischen Wirts (Abschnitt 3.2). Zum Zeitpunkt des Transfers waren diese Gene frei von Introns. Alle Introns in diesen Genen gehen somit auf den Prozess des Introngewinns zurück und können daher helfen, charakteristische Eigenschaften dieses Prozesses aufzudecken.

3.6 Der freilebende Vorfahre der Plastiden

Während über den Wirt nur wenig bekannt ist, wird angenommen, dass der Symbiont in biochemischer Hinsicht den rezenten Cyanobakterien sehr ähnlich war (Gould et al., 2008). Nach der Klassifizierung von Stanier (Rippka et al., 1979) werden rezente Cyanobakterien aufgrund morphologischer Eigenschaften in fünf Abteilungen (engl. *sections*) eingeteilt, die weitgehend botanischen Ordnungen entsprechen. Zu welcher dieser Abteilungen der freilebende Vorfahre der Plastiden gehört hat, ist noch nicht vollständig geklärt (Abschnitt 3.6.2).

3.6.1 Systematik der Cyanobakterien

Zur Abteilung I gehören Mitglieder der Ordnung Chroococcales. Dabei handelt es sich um einzellige Cyanobakterien, die sphärisch, zylindrisch oder oval sein können. Die meisten Vertreter dieser Abteilung vermehren sich durch binäre Teilung, manche durch asymmetrische binäre Teilung (Knospung). Einige Cyanobakterien dieser Abteilung bilden Kolonien, die durch Schleim oder eine Umhüllung zusammengehalten werden. Vertreter dieser Abteilung sind *Gloeobacter violaceus* sowie die Gattung *Synechococcus* (Abbildung 3.3 I).

Cyanobakterien in Abteilung II gehören zur Ordnung Pleurocapsales und zeichnen sich durch eine charakteristische Form der Fortpflanzung aus, die noch in keiner anderen prokaryotischen Gruppe beobachtet wurde. Bei der multiplen Spaltung finden ohne Wachstum schnell nacheinander Zellteilungen statt, bei denen mehr als 1.000 Tochterzellen entstehen können. Cyanobakterien aus der Abteilung II können sich durch multiple Spaltung oder eine Kombination aus dieser und binärer Teilung fortpflanzen. Diese Cyanobakterien können einzellig

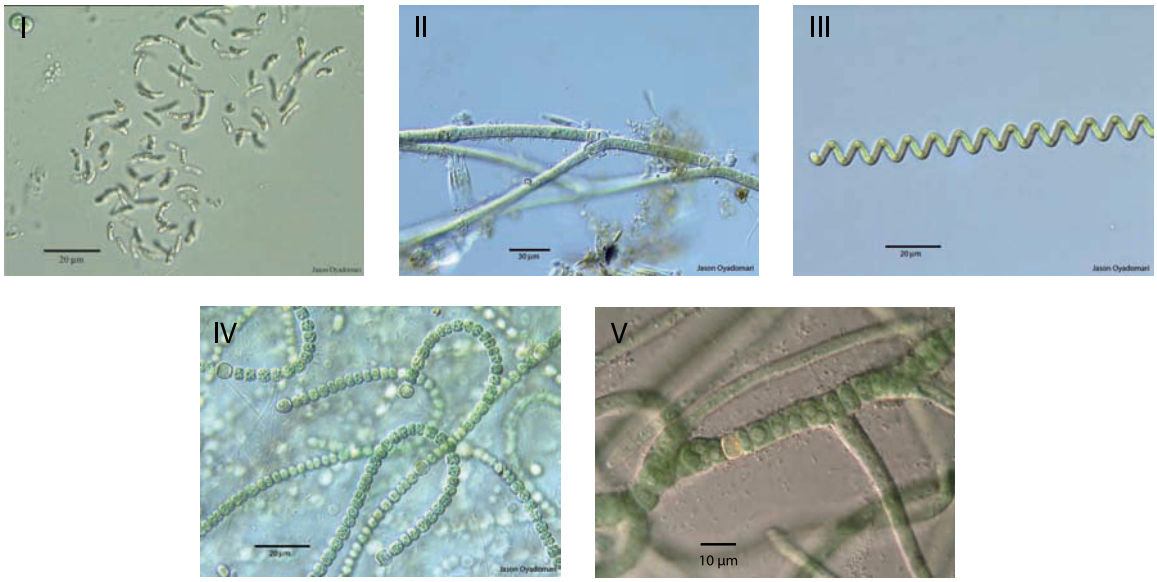

Abbildung 3.3: Lichtmikroskopische Aufnahmen von Repräsentanten der fünf cyanobakteriellen Abteilungen nach Rippka et al. (1979). I) *Synechococcus sp.*, – 20 μm, II) *Tolypothrix sp.*, – 30 μm, III) *Spirulina sp.*, – 20 μm, IV) *Nostoc sp.*, – 20 μm und V) *Fischerella sp.* – 10 μm. Quellen: I-IV) Jason K. Oyadomari[1], V) Culture Collection of Autotrophic Organisms[2].

sein oder Kolonien bilden, die durch eine faserige Hülle zusammengehalten werden. Zur Abteilung II gehören beispielsweise Vertreter der Gattung *Tolypothrix* (Abbildung 3.3 II).

Zur Abteilung III gehören fadenförmige Cyanobakterien der Ordnung Oscillatoriales, die durch Zellteilung in der selben Ebene ein Längenwachstum durchmachen. Cyanobakterien aus dieser Abteilung zeigen keine Zelldifferenzierung in Heterozysten und Aktineten wie die Cyanobakterien in den Abteilungen IV und V. Ebensowenig weisen die Filamente echte Verzweigungen auf. Zur Abteilung III gehören beispielsweise die zylindrischen Cyanobakterien der Gattung *Spirulina* (Abbildung 3.3 III).

Cyanobakterien in Abteilung IV gehören zur Ordnung Nostocales und bilden fadenförmige Gebilde wie bei der Gattung *Oscillatoriales*. Einige Zellen können sich zu Heterozysten differenzieren, in denen – abgeschottet vom aeroben Umgebungsmillieu – die Fixierung molekularen Stickstoffs in Form von Ammonium stattfindet. Darüberhinaus können Aktineten ausgebildet werden, die ein Überleben unter

1 http://www.keweenawalgae.mtu.edu/
2 http://www.butbn.cas.cz/ccala/index.php

extremen Umweltbedingungen wie Kälte und Trockenheit sicherstellen. Vertreter der Abteilung IV sind *Anabaena variabilis* und *Nostoc punctiforme* (Abbildung 3.3 IV).

Vertreter der Abteilung V gehören zur Ordnung Stigonematales und unterscheiden sich von den Vertretern aus Abteilung IV durch echte Verzweigungen in den Filamenten. Diese kommen durch eine Kombination von transversalen und longitudinalen Zellteilungen zustande. Vertreter der Abteilungen II, III und IV zeigen auch Verzweigungen, die jedoch auf Brüche im Filament zurückgehen und keine echten Verzweigungen darstellen. Zur Abteilung IV gehören beispielsweise Cyanobakterien der Gattung *Fischerella* (Abbildung 3.3 V).

3.6.2 Zuordnung

Phylogenetische Analysen ribosomaler RNA zeigen, dass Plastiden mit Cyanobakterien gruppieren, jedoch keine Affinität für eine bestimmte Abteilung aufweisen (Marin et al., 2005; Turner et al., 1999). Analysen konkatenierter Alignments plastidenkodierter Proteine kommen zu der selben Schlussfolgerung (Rodriguez-Ezpeleta et al., 2005). Im Gegensatz dazu konnte Sato (2006) schwache Hinweise darauf finden, dass der Vorfahre der Plastiden aus der *Anabaena-Synechocystis*-Linie stammen könnte. Das Konkatenieren von Daten ist jedoch problematisch. Bapteste et al. (2008) konnten zeigen, dass phylogenetische Signale dabei verloren gehen und nicht verstärkt werden. Besonders bei Prokaryoten ist dies ein Problem, da sie untereinander und mit anderen Prokaryoten Gene über den Prozess des lateralen Gentransfers austauschen (Raymond et al., 2002; Zhaxybayeva et al., 2006). Aus einem konkatenierten Alignment von Proteinfamilien mit vertikaler und horizontaler Abstammung werden daher Bäume abgeleitet, die weder die vertikale noch die horizontale Abstammung korrekt wiedergeben. Die Frage nach der Identität des freilebenden Vorfahrens der Plastiden von Pflanzen und Algen ist aus diesen Gründen immer noch offen.

3.7 Stammbäume und phylogenetische Rekonstruktion

Die Darstellung der Verwandtschaftsbeziehungen zwischen Organismen anhand eines Stammbaums geht auf Charles Robert Darwin (1809–1882) zurück. Unter der Überschrift „*I think*“ skizzierte er 1837 in seinen Notizen erstmalig einen Stammbaum, um zu veranschaulichen, wie sich die rezenten Arten aus gemeinsamen Vorfahren entwickelt haben könnten. Diese Darstellungsweise wird heute noch

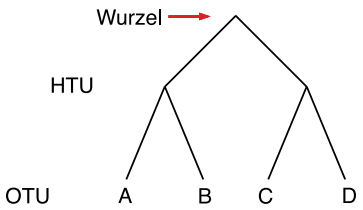

Abbildung 3.4: Gewurzelter Baum mit vier Taxa. Die Taxa A und B sowie die Taxa C und D (OTUs) haben sich aus einem gemeinsamen Vorläufer (HTU) entwickelt. Der gemeinsame Vorfahre aller Taxa ist die Wurzel.

verwendet, auch wenn heutzutage mit horizontalem und endosymbiontischem Gentransfer Vererbungsprozesse bekannt sind, die sich in einem Stammbaum nicht adäquat darstellen lassen (s.u.).

Ein Stammbaum oder phylogenetischer Baum ist ein gerichteter Graph, in dem Knoten (Taxa) über Kanten (Äste) verbunden sind (Abbildung 3.4). Die Wurzel repräsentiert den letzten gemeinsamen Vorfahren aller im Stammbaum dargestellten Arten. Die Blätter des Baums entsprechen den rezenten Arten (OTUs, engl. *operational taxonomic units*). Diese sind über gemeinsame Vorfahren (HTUs, engl. *hypothetical taxonomic units*) miteinander verbunden. Stammbäume sind in der Taxonomie bifurzierend, d. h. jeder interne Knoten hat den Eingangsgrad eins und den Ausgangsgrad zwei. Die Wurzel hat den Eingangsgrad null und den Ausgangsgrad zwei. Blätter haben den Eingangsgrad eins und den Ausgangsgrad null.

Die Basis einer jeden Rekonstruktion eines phylogenetischen Baums ist das multiple Alignment. Im multiplen Alignment werden die Aminosäuren – oder Nukleotide – der einzelnen biologischen Sequenzen derart ausgerichtet, dass Homologiemuster entstehen. Aus diesen werden anschließend phylogenetische Bäume abgeleitet. Sowohl das Erstellen von multiplen Alignments als auch das Ableiten von phylogenetischen Bäumen ist vom Rechenaufwand komplex: Die Berechnung eines optimalen multiplen Alignments hat eine Komplexität von $O(2^k n^k)$, wobei k die Anzahl der Sequenzen und n die längste zu alignierende Sequenz ist. Für n Taxa gibt es $\frac{(2n-5)!}{2^{n-3}(n-3)!}$ mögliche ungewurzelte Bäume. In der Praxis werden daher heuristische Methoden eingesetzt, die auf Kosten der Genauigkeit die Rechenzeit stark verkürzen. Als Konsequenz können abgeleitete Stammbäume fehlerbehaftet sein (Phillips et al., 2004).

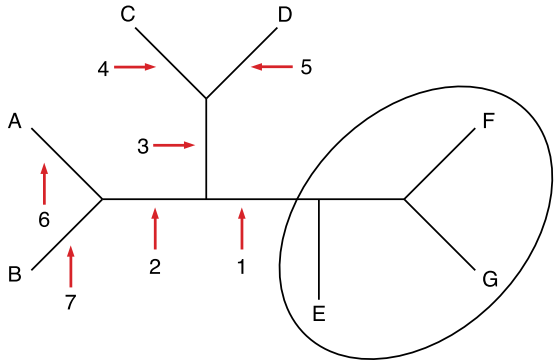

Abbildung 3.5: Ein ungewurzelter Baum mit dem „Klan“ der Taxa E, F und G. Die Taxa E, F und G sind „von den restlichen Taxa getrennt“. Würde eine Wurzel an den Positionen eins bis sieben platziert (mit Pfeilen markiert), so wäre der Klan auch eine Klade bzw. monophyletische Gruppe. Die drei „Nachbargruppen“ sind A und B (Wurzelposition 3, 4 oder 5), C und D (Wurzelposition 2, 6 oder 7) und A, B, C und D (Wurzelposition 1). Wenn der Klan eine Klade ist, so ist eine der Nachbargruppen die Schwestergruppe. Verändert nach Wilkinson et al. (2007).

Neben methodischen Unzulänglichkeiten gibt es noch weitere Gründe dafür, dass ein abgeleiteter Genbaum nicht mit dem Speziesbaum übereinstimmt. Der horizontale Gentransfer, der hauptsächlich zwischen Prokaryoten vorkommt, sowie der endosymbiontische Gentransfer (Abschnitt 3.2) sind nicht-vertikale Vererbungsprozesse. Gene werden nicht vertikal von den Eltern auf die Nachkommen sondern horizontal über die Artgrenze hinweg übertragen. Diese Prozesse können in einem bifurzierenden Baum nicht adäquat dargestellt werden. Manche Autoren beschreiben Netzwerke (Dagan et al., 2008; Huson, 1998) oder Ringe (Martin und Embley, 2004; Rivera und Lake, 2004) als besser geeignete Visualisierungen der Verwandtschaftsbeziehungen.

Mit phylogenetischen Methoden abgeleitete Bäume sind in der Regel ungewurzelt, können jedoch nachträglich durch biologisches Wissen oder bioinformatische Methoden gewurzelt werden. Weil bei diesem Schritt von falschen Annahmen ausgegangen werden kann, wird er bei phylogenetischen Analysen oft ausgelassen. Ungewurzelte Bäume sind ungerichtete Graphen, in denen die Information über den gemeinsamen Vorläufer fehlen. Auch solche Bäume enthalten Informationen darüber, wie die darin enthaltenen Taxa miteinander verwandt sind. Bezeichnungen wie Klade, monophyletische Gruppe und Schwestergruppe (Tabelle 3.1) werden häufig auch für ungewurzelte Bäume benutzt, um die Beziehungen zwischen Taxa und taxonomischen Gruppen zu beschreiben (Ben Ali et al., 2001; Bowe et al., 2000; Brochier et al., 2004). Diese Bezeichnungen stammen aus der Kladi-

Tabelle 3.1: Terminologie, um die Beziehungen zwischen Taxa in gewurzelten und ungewurzelten Bäumen zu beschreiben. Die Bezeichnungen für ungewurzelte Bäume wurden von Wilkinson et al. (2007) vorgeschlagen.

Gewurzelter Baum	Ungewurzelter Baum
Klade, monophyletische Gruppe	Klan
Schwestergruppe	Nachbargruppe
„Manche Taxa sind mit diesen Taxa näher verwandt als mit jenen.“	“Einige Taxa sind von anderen Taxa abgespalten.“

stik, in der gewurzelte Bäume vorliegen und die Vorläufertaxa bekannt sind. Die Verwendung dieser Begriffe in ungewurzelten Bäumen ist problematisch. Die in Abbildung 3.5 markierte Gruppe der Taxa E, F und G wäre eine monophyletische Gruppe, wenn die Wurzel an den Positionen eins bis sieben platziert würde. Eine andere Platzierung der Wurzel würde die Gruppierung aufbrechen. Wilkinson et al. (2007) schlagen daher eine eigene Terminologie vor, um die Beziehungen in ungewurzelten Bäumen zu beschreiben (Tabelle 3.1 und Abbildung 3.5). Die Gruppe der Taxa E, F und G bildet danach einen „Klan“.

3.8 Zielsetzung

Die primären Plastiden in Pflanzen und Algen gehen auf die Etablierung einer Endosymbiose eines Eukaryoten mit einem Cyanobakterium zurück. Im Laufe der Evolution wurden viele Gene des Cyanobakteriums in das Kerngenom des Wirts transferiert oder gingen verloren (endosymbiontischer Gentransfer). Bisherige Studien, die auf limitierten Datensätzen beruhen, beziffern die cyanobakterielle Komponente der Genome von Algen und Pflanzen auf 18 % (Martin et al., 2002) beziehungsweise 11 % (Archibald, 2006; Reyes-Prieto et al., 2006). In dieser Arbeit sollen anhand einer größeren Auswahl an Referenzgenomen mittels bioinformatischer Methoden die verfügbaren vollständig sequenzierten Genome von Algen und Pflanzen auf Gene cyanobakteriellen Ursprungs untersucht werden. Die Identifizierung dieser Gene wird eine quantitative Aussage über die cyanobakterielle Komponente dieser Genome erlauben.

Darüberhinaus kann eine Ähnlichkeitssuche mit diesen Genen in den Genomen rezenter Cyanobakterien unter ihnen das Cyanobakterium identifizieren, das dem freilebenden Vorfahren der Chloroplasten am ähnlichsten ist. Dadurch können Rückschlüsse über die Biologie des Vorfahren gezogen werden, auch wenn er in seiner ursprünglichen Form nicht mehr existiert. Ferner können Pflanzengene cyanobakteriellen Ursprungs helfen, aus Cyanobakterien stammende Stoffwechselwege – oder einzelne Enzyme – in Pflanzen und Algen zu identifizieren. Weiterhin kann über Vorhersagen der Zielkompartimente verschiedener Pflanzenproteine der Zusammenhang von Proteinlokalisierung und Genursprung untersucht werden. Zudem wird eine vergleichende Analyse der Eigenschaften der Introns in cyanobakteriellen und alten eukaryotischen Genen Rückschlüsse über die Evolution von Introns zulassen.

Die Ergebnisse dieser Arbeit sollen dazu beitragen, das Verständnis über die Endsymbiose der primären Plastiden zu vertiefen.

4 Material und Methoden

4.1 Sequenzdaten

Sämtliche Analysen dieser Arbeit wurden anhand von Proteinsequenzen vollständig sequenzierter Organismen durchgeführt, die aus öffentlich zugänglichen Datenbanken heruntergeladen wurden. Dabei handelt es sich um die Translationen aller offenen Leseraster der jeweiligen Genome. Bei Eukaryoten wurde ausschließlich mit kernkodierten Sequenzen gearbeitet. Eine Ausnahme stellen die Plastidengenome dar, die verwendet wurden, um Plastiden-DNA in den Kerngenomen der Pflanzen und Algen nachzuweisen (Abschnitt 4.5.7). Ein Großteil der Sequenzen stammt aus der RefSeq-Datenbank (Pruitt et al., 2005) des NCBI [1] (National Center for Biotechnology Information) auf dem Stand von Januar 2006.

4.1.1 Pflanzensequenzen

Sequenzen von *Arabidopsis thaliana* ((Arabidopsis Genome Initiative, 2000) Stand Januar 2006) und *Oryza sativa* ((International Rice Genome Sequencing Project, 2005) Stand Mai 2006) wurden aus der RefSeq-Datenbank heruntergeladen. Sequenzen von *Chlamydomonas reinhardtii* (Merchant et al., 2007) wurden vom JGI [2] (Joint Genome Institute des U.S. Departement of Energy) bezogen (Version 2.0). Sequenzen für *Cyanidioschyzon merolae* (Matsuzaki et al., 2004) stammen von den Internetseiten des *Cyanidioschyzon merolae* Genomprojekts[3] (Stand Februar 2005).

1 http://www.ncbi.nlm.nih.gov/RefSeq/
2 http://genome.jgi-psf.org/Chlre3/Chlre3.home.html
3 http://merolae.biol.s.u-tokyo.ac.jp

4.1.2 Plastidäre Proteinsequenzen

Für die Identifizierung von transferierter Plastiden-DNA (NUPTs, von engl. *nuclear plastid DNA*, Richly und Leister (2004)) in den Genomen der vier Algen und Pflanzen (Abschnitt 4.5.7) wurden die Translationen aller Gene der entsprechenden Plastidengenome aus der RefSeq-Datenbank (Stand Januar 2008) heruntergeladen. Die Zugangsnummern (engl. *accession numbers*) der Genome sind NC_000932 (*Arabidopsis*), NC_001320 (*Oryza*), NC_005353 (*Chlamydomonas*) und NC_004799 (*Cyanidioschyzon*).

4.1.3 Referenzsequenzen

Als Referenzsequenzen wurden die Proteome von neun Cyanobakterien, 215 weiteren prokaryotischen Organismen sowie 13 nicht-photosynthetischen Eukaryoten heruntergeladen. Sämtliche prokaryotische Sequenzen wurden aus der RefSeq-Datenbank bezogen. Sequenzen für *Entamoeba histolytica* (Loftus et al., 2005) und *Trichomonas vaginalis* (Carlton et al., 2007) stammen vom TIGR[4][5] (The Institute for Genomic Research). Die Sequenzen für *Trichoderma reesei* (Martinez et al., 2008) und *Phanerochaete chrysosporium* (Martinez et al., 2004) wurden vom JGI[6][7] heruntergeladen. Die Sequenzen der restlichen neun Eukaryoten wurden der Refseq-Datenbank (Stand Januar 2006) entnommen. Eine Liste aller Referenztaxa inklusive Quelle, Version und Phylumeinteilung ist dem Anhang beigefügt (Tabelle 6.1).

4.1.4 Die KEGG-Datenbank

KEGG[8] (Kyoto Encyclopedia of Genes and Genomes) stellt eine Sammlung von Datenbanken dar, die Informationen zu biologischen Systemen enthalten (Kanehisa und Goto, 2000; Kanehisa et al., 2006, 2008). Aus dieser Sammlung wurden die GENES- und die PATHWAY-Datenbank (Version 47.0) verwendet, um Pflanzenproteinen cyanobakteriellen Ursprungs eine Funktion zuzuweisen (Abschnitt 4.5.8).

4 http://www.tigr.org/tdb/e2k1/eha1/
5 http://www.tigr.org/tdb/e2k1/tvg/
6 http://genome.jgi-psf.org/Trire2/Trire2.home.html
7 http://genome.jgi-psf.org/Phchr1/Phchr1.home.html
8 http://www.genome.jp/kegg/

4.2 Rechner und Betriebssysteme

Berechnungen wurden auf zwei Servern an der Heinrich-Heine-Universität Düsseldorf durchgeführt. Der Opteron-Cluster[9] des Zentrums für Informations- und Medientechnologie (ZIM) ist durch seine drei Server mit je vier Dual-Core-Prozessoren gut für den Batch-Betrieb geeignet. Weitere Berechnungen und Datenbankanwendungen (Abschnitt 4.3.5) wurden auf dem Server „Jukes“ des Instituts für Ökologische Pflanzenphysiologie durchgeführt. Dabei handelt es ich um das Modell ProLiant DL360 der Firma Hewlett-Packard, das mit zwei Quad-Core-Prozessoren ausgerüstet ist. Als Arbeitsplatzrechner diente ein MacBook® Pro der Firma Apple®. Das Betriebssystem MacOS® X ermöglicht sowohl das Ausführen von Unix-Programmen als auch die Nutzung kommerzieller Programme wie Microsoft® Office® oder Adobe® Illustrator®.

4.3 Programme

4.3.1 Der Editor TextWrangler

Der Editor TextWrangler[10] wurde in dieser Arbeit verwendet, um Skripte zu schreiben. TextWrangler ist eine kostenlose Ausgabe des kommerziellen Programms BBEdit der Firma Bare Bones Software. TextWrangler unterstützt den Programmierer durch eine Vielzahl an Funktionen, die das Programmieren komfortabler machen. Dazu zählen das Hervorheben der Syntax für verschiedene Programmiersprachen und eine Suche-und-Ersetzen-Funktion, die reguläre Ausdrücke beherrscht.

4.3.2 Die Skriptsprache Perl

In dieser Arbeit wurde die Skriptsprache Perl [11] verwendet, um verschiedene Aufgaben zu bewältigen. Ein Beispiel für die Anwendungen von Perl ist der Arbeitsablauf zur Identifizierung cyanobakterieller Gene mit automatisierter Homologiesuche mittels BLAST, Erstellen von FASTA-Dateien der homologen Sequenzen, alignieren der Sequenzen und Rekonstruktion der phylogenetischen Bäume (inklusive Auswertung). Die Anwendung der HoT-Methode (Abschnitt 4.3.14) sowie

9 http://www.zim.uni-duesseldorf.de/hpc/hpc-infrastruktur/opteron-cluster/
10 http://www.barebones.com/products/textwrangler/
11 http://www.perl.org/

der Arbeitsablauf zur Rekonstruktion des Metabolismus (Abschnitt 4.5.8) wurden ebenfalls in Perl implementiert. In Perl können komfortabel Dateien gelesen, verändert sowie geschrieben werden. Eine Implementierung von regulären Ausdrücken sowie die Eigenschaft, dass Skripte nach Änderungen nicht erst kompiliert werden müssen, machen Perl ideal für die Arbeit an bioinformatischen Fragestellungen.

4.3.3 Die Programme sed und awk

Die Programme sed[12] (engl. *Stream EDitor*) und awk[13] (zusammengesetzt aus den Anfangsbuchstaben der Namen der Autoren Aho, Weinberger und Kernighan) sind sehr nützliche Werkzeuge zur Manipulation von Dateien und gehören zu jeder Unix-Distribution. Sed liest Dateien zeilenweise, verändert diese entsprechend vorgegebener Regeln und gibt die veränderten Zeilen wieder aus. Die Regeln für die Dateimanipulation können dabei reguläre Ausdrücke enthalten. Awk ist eigentlich eine Skriptsprache, da es alle Elemente einer Programmiersprache wie Schleifen und Kontrollstrukturen beinhaltet. In der Bioinformatik wird awk oft verwendet, um Tabellen zu verändern, da sehr einfach verschiedene Spalten einer Datei verglichen oder manipuliert werden können. Beide Programme lesen Dateien und schreiben auf die Standardausgabe. Durch Umleiten der Standardausgabe mittels des Operators „>“ können Dateien geschrieben werden.

4.3.4 BLAST

BLAST (engl. *Basic Local Alignment Search Tool*, Altschul et al. (1997)) ist das gängigste Werkzeug zur Homologiesuche in Sequenzdatenbanken und hat das ältere Programm FASTA (Pearson und Lipman, 1988) größtenteils abgelöst. FASTA wird nur noch für Spezialanwendungen wie die Suche in EST-Datenbanken verwendet. Beide Programme arbeiten mit heuristischen Methoden, die die Rechenzeit gegenüber optimalen Methoden extrem reduzieren, wobei die Ergebnisse Näherungen an ein optimales Ergebnis darstellen. Die Berechnung von optimalen lokalen oder globalen Sequenzalignments nach dem Smith-Waterman- (Smith und Waterman, 1981) oder Needleman-Wunsch-Algorithmus (Needleman und Wunsch, 1970) ist zur Homologiesuche in exponentiell wachsenden Sequenzdatenbanken – die Anzahl der Sequenzen in GenBank® vedoppelt sich in etwa alle 18 Monate (Benson et al., 2008) – aus Laufzeitgründen ungeeignet.

12 http://sed.sourceforge.net
13 http://www.gnu.org/software/gawk/gawk.html

BLAST erstellt aus einer Suchsequenz zunächst einen Index aus Einträgen der Wortlänge w. Die Referenzdatenbank wird anschließend in einer Initialsuche nach Wörtern dieser Länge durchsucht. Sowohl für Aminosäure- als auch für Nukleotidsequenzen verwendet BLAST dabei eine Ähnlichkeitsmatrix. Wird bei der Suche ein Schwellenwert T innerhalb eines Worts überschritten, so liegt ein Treffer (engl. *hit*) vor. Dieser Treffer wird nur dann weiter untersucht, wenn in direkter Nachbarschaft ein zweiter Treffer gefunden wird, wobei der maximale Abstand durch die Fensterlänge A vorgegeben wird. Diese sogenannte *Two-Hit*-Methode ist für BLAST charakteristisch. Anschließend werden die Trefferpaare bidirektional ausgedehnt bis sich die Qualität (engl. *score)* des Alignments nicht weiter erhöhen lässt. Wird ein bestimmter Schwellenwert S überschritten, so wird die Treffersequenz am Ende der Suche als HSP (engl. *high-scoring segment pair*) ausgegeben. FASTA arbeitet nach einem ähnlichen Algorithmus jedoch werden bei der Initialsuche nur identische Sequenzfragmente berücksichtigt. Darüber hinaus wird bei FASTA aus den besten Alignments noch ein optimales lokales Alignment nach dem Smith-Waterman-Algorithmus erstellt. Aufgrund dieser Unterschiede ist BLAST bei nahezu gleicher Sensitivität schneller als FASTA und wurde in dieser Arbeit verwendet, um Homologe zu 83.138 Pflanzenproteinen in einer Referenzdatenbank von 851.607 Proteinen zu suchen.

Von BLAST existieren mehrere Varianten, die sich in der Art der Such- und Referenzsequenzen unterscheiden. BLASTN sucht in einer Nukleotiddatenbank nach Homologen zu Nukleotidsequenzen. In dieser Arbeit wurde BLASTP verwendet, das analog mit Proteinsequenzen verfährt. BLASTX sowie TBLASTN werden verwendet, wenn mit Nukleotidsequenzen gegen Proteinsequenzen beziehungsweise umgekehrt gesucht werden soll. Mit TBLASTX wird mit einer translatierten Nukleotidsequenz in einer translatierten Nukleotiddatenbank gesucht, wobei alle sechs Leseraster berücksichtigt werden.

Dem Programm BLAST dient eine Suchsequenz im FASTA-Format sowie eine Referenzdatenbank im BLAST-Format als Eingabe. Diese Datenbank kann durch Umformatierung von FASTA-Sequenzen mittels des Programms formatdb erzeugt werden. Die Ausgabe einer BLAST-Suche enthält die Treffersequenzen sowie Angaben zu deren Identität zur Suchsequenz. Ebenfalls wird für jede Treffersequenz ein Erwartungswert als Indikator für die statistische Signifikanz des Treffers angegeben. Diese Ausgabe kann dann unter Anwendung bestimmter Kriterien für diese Parameter nach Homologen zur Suchsequenz durchsucht werden.

4.3.5 MySQL

MySQL[14] ist ein Programm, um relationale Datanbanken anzulegen, zu speichern und zu verwalten, sowie effizient Anfragen an diese zu stellen. In dieser Arbeit wurde MySQL genutzt, um die Ergebnisse der BLAST-Suche zu speichern und mit den Informationen der taxonomischen Einteilung in Phyla des NCBI zu verknüpfen. Aus dieser Datenbank wurden alle Treffer abgefragt, die die Homologiekriterien erfüllen (Abschnitt 4.5.1). Im weiteren Verlauf der Arbeit wurde MySQL genutzt, um Anfragen an eine Datenbank zu stellen, die diverse Parameter wie Alignmentlänge, Anzahl an Taxa, CS-, SPS-, SPS-Wert über die berechneten Alignments sowie die daraus abgeleiteten Bäume enthält.

4.3.6 Clustal W

Clustal W (Thompson et al., 1994) gehört zu den bekanntesten Programmen, um multiple Alignments zu berechnen. Die Berechnung optimaler multipler Alignments von k Sequenzen hat eine Komplexität aus $O(2^k n^k)$, wobei n die längste zu alignierende Sequenz ist (Abschnitt 3.7). Clustal W verfolgt einen heuristischen Ansatz, da es in der Praxis nicht möglich ist, optimale multiple Sequenzalignments zu erstellen. Clustal W implementiert ein progressives Verfahren. Zunächst wird zwischen jedem Sequenzpaar die Distanz berechnet und in einer Matrix verzeichnet. Aus dieser Distanzmatrix wird nach dem Verfahren des *Neighbor-Joining* (Abschnitt 4.3.10) ein Initialbaum berechnet. Dieser gibt die Reihenfolge vor, mit der die Sequenzen nacheinander unter Verwendung von dynamischer Programmierung zu einem multiplen Alignment zusammengefügt werden. Zunächst wird das am engsten verwandte Sequenzpaar aligniert. In den folgenden Schritten werden weitere Sequenzen zu dem Profil dieses Alignments hinzugefügt. Je nach Topologie des Initialbaums können im Verlauf des Algorithmus auch zwei Sequenzen oder zwei Alignmentprofile aneinander ausgerichtet werden. Bei jedem Alignmentschritt wird eine neue Substitutionsmatrix ausgewählt, die an den Grad der Divergenz der Sequenzen angepasst ist. Jeder Sequenz wurde bei der Berechnung des Initialbaums eine Gewichtung zugeordnet (niedrige Werte für Gruppen von eng verwandten Sequenzen und hohe Werte für divergente Sequenzen), die als Multiplikator in die Berechnung der Bewertungsfunktion eingeht. Darüberhinaus werden positionsabhängige Strafpunkte für das Einführen und Verlängern

14 http://mysql.com

```
>Arabidopsis
MVLPSSTPLQTTGKKTISSPEYNFPVIDFSLNDRSKLSEKIVKACEVNGFFK
>Schizosaccharomyces
MGSLEVPCIDLSENDTSIVVKELLDACKNWGFVS
>Anabaena
MTVLQLPIIDISGLTCQRNNSSDVVAQQIKQACQDYGFFY
>Acinetobacter
MTQHIPIISISGLFSPLLEDRLQVAMQMKQACEDNGFFY
>Streptomyces
MTNTATTPSYQRYQQLPIIDLSAADRGPQARALLHAQLHSAAHDVGFFQ
```

Abbildung 4.1: Proteinsequenzen im FASTA-Format. Für dieses Sequenzformat ist das „>“-Zeichen charakteristisch. Nach diesem Zeichen folgt der Bezeichner für die Sequenz, die nach dem nächsten Zeilenumbruch beginnt.

von Lücken verwendet, die unter anderem davon abhängen, ob an einer Position bereits Lücken vorhanden sind.

In der O-Notation der asymptotischen Laufzeitanalyse, die eine obere Schranke für die Komplexität eines Algorithmus angibt, liegen die Anforderungen von Clustal W für N Sequenzen der Länge L an den Speicherplatz bei $O(N^2 + L^2)$ und für die Rechenzeit bei $O(N^4 + L^2)$. Die Anforderungen von Clustal W sind demnach polynomial und wesentlich leichter zu erfüllen als die exponentiellen Anforderungen optimaler multipler Alignments.

Das Programm Clustal W erzeugt multiple Alignments im Clustal-Format (Abbildung 4.2) aus Sequenzdaten im FASTA-Format (Abbildung 4.1). Clustal W kann auch dazu verwendet werden, fertige Alignments in andere Formate wie beispielsweise das PHYLIP-Format (Abbildung 4.3) zu konvertieren.

```
CLUSTAL W (1.83) multiple sequence alignment

Anabaena               ----------------MTV--LQLPIIDISGLTCQRNNSSDVVAQQIKQACQDYGFFY
Acinetobacter           ----------------MT---QHIPIISISGLFSPLLEDRLQVAMQMKQACEDNGFFY
Arabidopsis            MVLPSSTPLQTTGKKTISSPEYNFPVIDFS------LNDRSKLSEKIVKACEVNGFFK
Schizosaccharomyces     ----------------MGS--LEVPCIDLS------ENDTSIVVKELLDACKNWGFVS
Streptomyces           -------MTNTATTPSYQR-YQQLPIIDLSAADRG-PQARALLHAQLHSAAHDVGFFQ
                                           ..* *.:*      :    :  :: .*..  **.
```

Abbildung 4.2: Multiples Sequenzalignment im Clustal-Format. Die Darstellung erfolgt typischerweise in Blöcken zu 60 Zeichen. Unter den Blöcken ist die Konservierung der Alignmentspalten mittels Symbolen dargestellt: (*) Alle Zeichen einer Spalte sind identisch. (:) Es kommen konservierte Austausche vor. (.) Es kommen semi-konservierte Austausche vor.

4.3.7 Muscle

Muscle (Edgar, 2004*a*,*b*) ist ein weiteres Programm zur Berechnung von multiplen Alignments. Der Name leitet sich aus dem Englischen von *MUltiple Sequence Comparison by Log-Expectation* ab. Im ersten Schritt verwendet Muscle wie Clustal W (Abschnitt 4.3.6) ein progressives Verfahren zur Berechnung eines multiplen Alignments. In einem zweiten Schritt wird jedoch zusätzlich versucht, das Alignment zu verbessern. Dazu wird aus dem multiplen Alignment ein Baum abgeleitet und mit dem Initialbaum verglichen. Für die unterschiedlichen Teilbäume wird anschließend das Alignment neu berechnet und aus dem neuen Alignment ein weiterer Baum abgeleitet. Dieser Vorgang wird so lange wiederholt, bis die Anzahl der unterschiedlichen Knoten nicht weiter abnimmt. Im dritten Schritt findet eine iterative Verbesserung des Alignments statt. Dabei wird der Baum – angefangen mit der von der Wurzel am weitesten entfernten Kante – in zwei Teilbäume und damit zwei nicht-überlappende multiple Alignments zerlegt. Die Profile beider multipler Alignments werden aneinander ausgerichtet und für das erhaltene Alignment die Güte berechnet. Diese berechnet sich aus einem Wert aus einer Substitutionsmatrix für jedes alignierte Sequenzpaar abzüglich Strafpunkte für Lücken. Ist die Güte des neuen Alignments höher als die des alten, so wird das neue Alignment beibehalten. Anderenfalls wird das neue Alignment verworfen. Bei dieser Strategie werden zuerst individuelle Sequenzen und dann eng verwandte Sequenzen neu aligniert. Der dritte Schritt endet, wenn die beschriebene Methode auf alle Kanten angewendet und keine Änderung beibehalten wurde oder die maximale Anzahl an Iterationen (engl. *maximum number of iterations*) erreicht ist, die beim Programmaufruf angegeben wird.

Obwohl Muscle die oben erwähnten Schritte zur Alignmentverbesserung implementiert, sind seine maximalen asymptotischen Anforderungen an den Speicherplatz mit denen von Clustal W identisch ($O(N^2 + L^2)$). Die Rechenzeit ist mit $O(N^4 + NL^2)$ gegenüber Clustal W leicht erhöht. Da die O-Notation eine obere Grenze für die Laufzeit darstellt, ist in der Praxis die Laufzeit von Muscle der von Clustal W sehr ähnlich, wobei Muscle oft schneller ist als Clustal W (Edgar, 2004 *a*).

Muscle liest Sequenzdaten im FASTA- (Abbildung 4.1) und gibt Alignments im Clustal-Format (Abbildung 4.2) oder anderen Alignmentformaten aus.

```
     5     58
Anabaena   ---------- ------MTV- -LQLPIIDIS GLTCQRNNSS DVVAQQIKQA
Acinetobac ---------- ------MT-- -QHIPIISIS GLFSPLLEDR LQVAMQMKQA
Arabidopsi MVLPSSTPLQ TTGKKTISSP EYNFPVIDFS ------LNDR SKLSEKIVKA
Schizosacc ---------- ------MGS- -LEVPCIDLS ------ENDT SIVVKELLDA
Streptomyc -------MTN TATTPSYQR- YQQLPIIDLS AADRG-PQAR ALLHAQLHSA

           CQDYGFFY
           CEDNGFFY
           CEVNGFFK
           CKNWGFVS
           AHDVGFFQ
```

Abbildung 4.3: Multiples Sequenzalignment im PHYLIP-Format. Gezeigt ist das selbe Alignment wie in Abbildung 4.2. Taxonnamen können maximal zehn Zeichen lang sein. Das PHYLIP-Format zeigt explizit keine Informationen zur Sequenzkonservierung, enthält jedoch Angaben über die Anzahl an Taxa sowie die Gesamtlänge des Alignments.

4.3.8 Das PHYLIP-Programmpaket

Das Programmpaket PHYLIP (engl. *PHYlogeny Inference Package*, (Felsenstein, 1989)) ist eine Sammlung von Programmen, um phylogenetische Bäume nach einer Vielzahl von Methoden abzuleiten, zu analysieren und darzustellen. Dieses Programmpaket nutzt ein eigenes Alignmentformat, das PHYLIP-Format (genauer PHYLIP interleaved (Abbildung 4.3)). Dieses Sequenzformat kann beispielsweise mit dem Programm Clustal W (Abschnitt 4.3.6) aus Sequenzen im Clustal-Format (Abbildung 4.1) generiert werden.

4.3.9 Protdist (PHYLIP)

Das Programm protdist erstellt aus Alignments von Proteinsequenzen Distanzmatrizen anhand verschiedener Modelle und Substitutionsmatrizen. Zur Auswahl stehen das einfache Kimura-Modell (Kimura, 1983), die PAM-Matrix (Dayhoff et al., 1978), die JTT-Matrix (Jones et al., 1992) und andere Methoden. In dieser Arbeit wurde das JTT-Modell von Jones et al. (1992) verwendet, um Distanzen zwischen Proteinsequenzen zu berechnen. Protdist liest ein Alignment im PHYLIP-Format (Abbildung 4.3) und erzeugt eine Distanzmatrix im PHYLIP-Format (Abbildung 4.4).

4.3.10 Neighbor (PHYLIP)

Das Programm neighbor leitet ungewurzelte phylogenetische Bäume aus Distanzmatrizen anhand der UPGMA- (engl. *unweighted pair-group method using arithmetic averages*) oder der *Neighbor-Joining*-Methode (NJ, Saitou und Nei (1987)) ab.

```
   5
Anabaena    0.000000  0.791076  1.467175  1.206101  1.698759
Acinetobac  0.791076  0.000000  1.133821  1.787111  1.780763
Arabidopsi  1.467175  1.133821  0.000000  1.575617  2.152059
Schizosacc  1.206101  1.787111  1.575617  0.000000  2.213663
Streptomyc  1.698759  1.780763  2.152059  2.213663  0.000000
```

Abbildung 4.4: Distanzmatrix im PHYLIP-Format. Taxonnamen können maximal zehn Zeichen lang sein.

Die UPGMA-Methode geht von einer konstanten Evolutionsrate, also einer molekularen Uhr (engl. *molecular clock*), aus. Bei dieser Methode wird das Sequenzpaar mit der geringsten Distanz zu einem Knoten zusammengefasst und die Distanz aller übrigen Taxa zu diesem Knoten anhand ihres arithmetischen Mittels neu berechnet. Zu diesem Knoten wird dann nach und nach die Sequenz mit der geringsten Distanz hinzugefügt. Danach wird wieder das Paar mit der geringsten Distanz gesucht, zu einem Knoten zusammengefügt und die Distanzen der übrigen Taxa zu diesem Knoten neu berechnet. Ein solches Paar kann aus zwei Taxa, einem Taxon und einem Knoten oder zwei Knoten bestehen. Der Algorithmus endet, wenn ein vollständig bifurzierender Baum entstanden ist.

Die NJ-Methode (Saitou und Nei, 1987) geht vom Kriterium der minimalen Evolution aus und nicht von einer konstanten Evolutionsrate. Es wird also versucht, unter allen Bäumen denjenigen mit der geringsten Gesamtastlänge zu berechnen. Bei dieser Methode werden – ausgehend von einer Sterntopologie – zunächst die zwei am nächsten verwandten Sequenzen zu einem Knoten zusammengefügt. Anschließend werden die Distanzen aller übrigen Taxa zu diesem Knoten nach einem speziellen Verfahren neu berechnet. Im Unterschied zu UPGMA ist die neue Distanz nicht das arithmetische Mittel, sondern berücksichtigt den Verwandtschaftsgrad zwischen den Sequenzen. Danach wird das Sequenzpaar gesucht, das zu dem Baum mit der geringsten Gesamtastlänge führt, zu einem Knoten zusammengefügt und die Distanzen der übrigen Taxa zu diesem Knoten wieder neu berechnet. Dieser Teil des Algorithmus endet, wenn alle Taxa zu Knoten zusammengefasst wurden. Abschließend werden die Knoten wieder durch die darin gespeicherten Taxa ersetzt, bis ein vollständig bifurzierender Baum entstanden ist. Es ist nicht praktikabel, alle möglichen Bäume zu untersuchen, da deren Anzahl mit der Anzahl an Taxa exponentiell steigt (Abschnitt 3.7). Es wird daher eine heuristische Methode benutzt, um die Ergebnisse von Teilbäumen zu verwerfen, die höchstwahrscheinlich zu keiner optimalen Lösung führen. Es ist daher theoretisch möglich, dass der berechnete Baum nicht der mit der geringsten Gesamtastlänge ist. In der Praxis sind die Bäume jedoch annähernd optimal.

```
(((Acinetobac:0.42759,Arabidopsi:0.70623):0.14081,Streptomyc:1.25869):0.09447,
Schizosacc:0.86984,Anabaena:0.33626);
```

Abbildung 4.5: Phylogenetischer Baum im Newick-Format

Das Programm neighbor erzeugt aus einer Distanzmatrix (Abbildung 4.4) einen phylogenetischen Baum im Newick-Format (Abbildung 4.5).

4.3.11 Consense (PHYLIP)

Consense ist ein Programm zur Erstellung eines Konsensus-Baums aus einer Kollektion von Bäumen. Der Konsensus-Baum enthält die Bipartitionen, die in der Mehrzahl der Eingabebäume auftreten. Typischerweise wird ein Konsensus-Baum am Ende einer Bootstrap-Analyse (Felsenstein, 1985) durchgeführt. Bei einer solchen Analyse werden aus einem Alignment durch „Ziehen mit Zurücklegen“ von Spalten zufällige Alignments erzeugt. Aus diesen 1.000 oder 10.000 Alignments werden dann phylogenetische Bäume abgeleitet und zu einem Konsensus-Baum zusammengefügt, der aus den Bipartitionen besteht, die in den Einzelbäumen am häufigsten auftraten. Die Bootstrap-Werte des Konsensusbaums sind ein Maß für die Eindeutigkeit der Daten.

Consense werden als Eingabe Bäume im Newick-Format überreicht. Der erzeugte Konsensus-Baum wird ebenfalls in diesem Format gespeichert. Wird consense mit nur einem Baum als Eingabedatei aufgerufen, so zeigt das Programm alle in diesem Baum enthaltenen Bipartitionen an (Abschnitt 4.5.4).

4.3.12 Treedist (PHYLIP)

Das Programm treedist berechnet die Distanz zwischen zwei im Newick-Format vorliegenden phylogenetischen Bäumen nach zwei verschiedenen Methoden. Die Robinson-Foulds-Metrik (Robinson und Foulds, 1981) berechnet die symmetrische Distanz, die als $A + B$ definiert ist, wobei A die Anzahl der Bipartitionen beschreibt, die im ersten aber nicht im zweiten Baum auftreten. B ist analog als die Anzahl der Bipartitionen definiert, die im zweiten aber nicht im ersten Baum auftreten. Die *branch score distance* (Kuhner und Felsenstein, 1994) misst die Distanz zweier Bäume auf eine ähnliche Weise, bezieht jedoch Astlängen in die Berechnung mit ein. In dieser Arbeit wurde die symmetrische Distanz verwendet, um die zwei nach der HoT-Methode (Abschnitt 4.3.14) berechneten alternativen Bäume des selben Datensatzes zu vergleichen. Die symmetrische Distanz dient als Grundlage

zur Berechnung des PPS-Werts (engl. *phylogenetic partitions score*), der ein Maß für die Ähnlichkeit zweier Bäume darstellt (Abschnitt 4.6.4).

4.3.13 PHYML

Das Programm PHYML (Guindon und Gascuel, 2003) leitet phylogenetische Bäume nach der Methode der maximalen Wahrscheinlichkeit (engl. *maximum likelihood*, ML) ab. Bei dieser Methode wird der Baum berechnet, der mit der größten Wahrscheinlichkeit zu den gegebenen Daten (Alignment) und einem eingestellten Evolutionsmodell (z.B. PAM-Matrix (Dayhoff et al., 1978)) passt. Für ein Alignment mit n Spalten wird für jede Spalte x_1, x_1, ..., x_n die Wahrscheinlichkeit L berechnet, mit der eine bestimmte Topologie unterstützt wird. Aus diesen Wahrscheinlichkeiten wird nach der Formel

$$lnL(\Theta) = ln(\prod_{i=1}^{n} f(x_i|\Theta)) = \sum_{i=1}^{n} lnf(x_i|\Theta)$$

das Produkt der Einzelwahrscheinlichkeiten – beziehungsweise die Summe aus deren Logarithmen – berechnet, mit der diese Topologie zu dem Alignment passt. Wurden die Wahrscheinlichkeiten für alle möglichen Topologien berechnet, so wird der Baum ausgegeben, für dessen Topologie die maximale Wahrscheinlichkeit berechnet wurde. Ähnlich wie bei der *Neighbor-Joining*-Methode (Abschnitt 4.3.10) ist es nicht praktikabel, alle möglichen Bäume zu untersuchen, da deren Anzahl mit der Anzahl an Taxa exponentiell steigt (Abschnitt 3.7). Beim ML-Verfahren wird daher zunächst ein Distanzbaum abgeleitet. Für diesen Initialbaum wird anschließend berechnet, wie gut er zu den Daten und dem Evolutionsmodell passt. Diese Wahrscheinlichkeit wird danach durch Änderungen an der Topologie versucht zu erhöhen, bis der Baum mit der maximalen Wahrscheinlichkeit erreicht ist. Das Programm PHYML führt diese Schritte auf eine schnelle, aber akkurate Weise durch und ist daher gut für die Analyse großer Datensätze geeignet (Guindon und Gascuel, 2003). Zusätzlich kann PYHML eine Heterogenität der Mutationsraten berücksichtigen, indem für die Verteilung der Mutationsraten eine Γ-Verteilung angenommen wird. Die Γ-Verteilung ist eine stetige Verteilung deren Berechnung sehr aufwändig ist. Daher verwendet PHYML ein diskretes Γ-Modell (Yang, 1994), für das die Anzahl an Kategorien spezifiziert werden kann. Der Formparameter α kann vorgegeben oder aus den Daten geschätzt werden. Darüberhinaus kann der Anteil invarianter Positionen abgeschätzt und berücksichtigt werden. Andernfalls

können solche Positionen zu einer Unterbewertung der tatsächlichen Distanz führen (Shoemaker und Fitch, 1989).

In dieser Arbeit wurde als Evolutionsmodell die JTT-Matrix (Jones et al., 1992) verwendet. Für die Mutationsraten wurde eine diskretes Γ-Modell mit acht Kategorien angenommen, deren α-Parameter aus den Daten geschätzt wurde. Invariante Positionen wurden abgeschätzt und berücksichtigt. PHYML akzeptiert Alignments im PHYLIP-Format (Abbildung 4.3) als Eingabe und berechnet einen phylogenetischen Baum im Newick-Format (Abbildung 4.5).

4.3.14 Die *heads-or-tails*-Methode (HoT)

Die HoT-Methode (von engl. *heads-or-tails*, Landan und Graur (2007)) ist ein Verfahren, um die Verlässlichkeit von multiplen Sequenzalignments zu identifizieren und zu quantifizieren. Die Methode geht von der *a priori*-Annahme aus, dass ein multiples Sequenzalignment unabhängig von der Orientierung sein sollte, in der die Sequenzen an das Alignmentprogramm überreicht werden. In der Praxis unterscheiden sich das normale und das reverse Alignment jedoch unterschiedlich stark (Landan und Graur, 2007). Die Diskrepanz der alternativen Alignments kann als Maß für die Verlässlichkeit eines multiplen Alignments herangezogen werden. Abschnitt 4.5.5 beschreibt die Anwendung der HoT-Methode in dieser Arbeit. In Abschnitt 4.6.2, Abschnitt 4.6.3 und Abschnitt 4.6.4 sind der CS-, der SPS- und der PPS-Wert definiert, die in dieser Arbeit als Maße für die Verlässlichkeit der Alignments und Bäume verwendet wurden.

4.3.15 MATLAB®

MATLAB® ist eine kommerzielle Software der Firma The MathWorks ™, die numerische Berechnungen anhand von Vektoren und Matrizen durchführt und deren Ergebnisse visualisiert. Der Name leitet sich aus dem Englischen von *MATrix LABoratory* ab. MATLAB® wurde in dieser Arbeit verwendet, um große Datenmengen zu erfassen, zu analysieren und graphisch darzustellen. Die statistischen Analysen der Korrelation zwischen Alignment- und Baumverlässlichkeit (vgl. Abschnitt 4.6.5) wurden ebenfalls mit MATLAB® durchgeführt.

4.4 TargetP

TargetP (Emanuelsson et al., 2000, 2007) ist ein Programm, um die subzelluläre Lokalisierung von Proteinen mittels neuronaler Netze vorauszusagen, indem Transitsequenzen für die entsprechenden Zellkompartimente identifiziert werden. Das Programm bewertet wie ähnlich N-terminale Sequenzen zu Signalpeptiden sind, die für Chloroplasten, Mitochondrien oder den Sekretionsweg charakteristisch sind und gibt dafür Wahrscheinlichkeiten aus. Ausgegeben wird das Kompartiment, für welches das entsprechende Signalpeptid die höchste Wahrscheinlichkeit aufweist. Die Vorhersagen werden in fünf Verlässlichkeitsklassen eingeteilt, welche sich aus der Differenz Δ der Wahrscheinlickeiten für das vorhergesagte Zielkompartiment und der zweithöchsten Wahrscheinlichkeit ergeben. Für $\Delta > 0{,}8$ gilt $RC = 1$, für $0{,}6 > \Delta \geq 0{,}8$ gilt $RC = 2$ und so weiter. $RC = 1$ entspricht der höchsten und $RC = 5$ der niedrigsten Verlässlichkeitsklasse. Die Charakteristika der verschiedenen Signalsequenzen wurden TargetP zuvor mit einem Datensatz aus Proteinen antrainiert, für welche die Lokalisierug experimentell belegt wurde. Diese Proteine entstammen der kuratierten Datenbank SWISS-PROT (Boeckmann et al., 2003).

TargetP akzeptiert Proteinsequenzen im FASTA-Format (Abbildung 4.1) als Eingabe und generiert eine tabellarische Ausgabe der Vorhersagen der Zielkompartimente (Abbildung 4.6).

4.5 Arbeitsabläufe

4.5.1 Homologiesuche

Die Translationen aller kernkodierten Gene der Planzen *Arabidopsis thaliana* und *Oryza sativa*, der Grünalge *Chlamydomonas reinhardtii* sowie der Rotalge *Cyanidioschyzon merolae* wurden aus verschiedenen öffentlichen Datenbanken heruntergeladen (sieht Abschnitt 4.1 (Daten)). Multiple Einträge identischer Proteinsequenzen innerhalb eines Organismus wurden zu einzelnen Einträgen zusammengefasst. Die Proteinsequenzen der 237 Referenztaxa (siehe Tabelle 6.1) wurden aus Refseq (Pruitt et al., 2005) und anderen Datenbanken heruntergeladen (Abschnitt 4.1) und mit dem Programm formatdb (Abschnitt 4.3.4) in eine BLAST-Datenbank mit 851.607 Einträgen geschrieben. Mit jeder der 83.138 Suchsequenzen wurde in dieser Datenbank eine Homologiesuche mit BLASTP (Abschnitt 4.3.4) durch-

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  8
Cleavage site predictions not included.
Using PLANT networks.

Name                 Len     cTP    mTP     SP  other  Loc  RC
-----------------------------------------------------------------------
15217431             423   0.205  0.093  0.059  0.838   _    2
15217450             933   0.036  0.468  0.182  0.055   M    4
15217465             784   0.001  0.067  0.983  0.045   S    1
15217492             379   0.556  0.434  0.023  0.071   C    5
15217498             412   0.193  0.242  0.046  0.528   _    4
15217507             324   0.024  0.868  0.045  0.098   M    2
15217529             322   0.004  0.152  0.777  0.251   S    3
15217547             487   0.867  0.048  0.071  0.027   C    2
```

Abbildung 4.6: Ausgabe von TargetP (gekürzt). Für jedes Protein werden die Länge (*Len*) sowie die Wahrscheinlickeiten der Übereinstimmung mit Signalpeptiden für Chloroplasten (*cTP*), Mitochondrien (*mTP*), den Sekretionsweg (*SP*) oder das Cytosol (kein Signalpeptid, *other*) angegeben. Das Kompartiment für welches das entsprechende Signalpeptid die größte Wahrscheinlichkeit bekam, wird als Zielkompartiment (*Loc*) angegeben. Die Verlässlichkeit der Vorhersage (*RC*) ist in fünf Abstufungen von eins (höchste Verlässlichkeit) bis fünf (geringste Verlässlichkeit) angegeben.

geführt. Eine prozentuale Aminosäureidentität der lokalen HSP-Alignments von $\geq 25\,\%$ sowie ein Erwartungswert $E \leq 10^{-10}$ dienten als Homologiekriterien. Wurden Homologe ausschließlich in Cyanobakterien gefunden, so wurde das Pflanzengen als cyanobakteriell klassifiziert. Für Suchsequenzen, die Homologe in Cyanobakterien und einem weiteren Phylum aufwiesen, wurde die Anzahl der verschiedenen Spezies mit Homologen in den beiden Phyla verglichen. Die Phylumeinteilung wurde vom TaxBrowser[15] des NCBI übernommen. Pflanzenproteine, die in mindestens vier cyanobakteriellen Spezies Homologe aufwiesen, wurden als cyanobakteriell klassifiziert. Wurden Homologe in Cyanobakterien und mindestens zwei weiteren Phyla gefunden, so wurden für die Suchsequenz und ihre Homologen phylogenetische Analysen durchgeführt. Pro Phylum wurden die besten drei Homologen für die Berechnung multipler Alignments und das Ableiten von Stammbäumen verwendet. Zu diesem Zweck wurde ein Qualitätskriterium Q definiert, welches die prozentuale Aminosäureidentität der Suchsequenz zur Treffersequenz im HSP-Alignment ID_{HSP}, die Länge der Suchsequenz L_S und die Länge des HSP-Alignments L_{HSP} berücksichtigt. Das Qualitätskriterium Q wurde als

$$Q = ID_{HSP}\frac{L_{HSP}}{L_S}$$

15 http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root

definiert. Dadurch soll die Tatsache berücksichtigt werden, dass BLAST dafür bekannt ist, die Treffersequenzen nur schlecht zu sortieren (Koski und Golding, 2001). Darüber hinaus können so kurze Treffer mit hoher und lange Treffer mit niedriger prozentualer Aminosäureidentität verglichen werden.

4.5.2 Multiple Alignments

Die Suchsequenz aus Pflanzen bzw. Algen wurde mit ihren Homologen (Abschnitt 4.5.1) in eine FASTA-Datei geschrieben und mit dem Programm Muscle (Abschnitt 4.3.7) aligniert. Für den von Muscle implementierten abschließenden Algorithmus zur Verbesserung der Alignmentqualität wurden maximal 16 Iterationen festgelegt. Spalten mit Lücken (engl. *gaps*) wurden aus den Alignments vor Beginn der phylogenetischen Analysen entfernt. Analog wurden multiple Alignments aus Sequenzen in umgekehrter Orientierung nach der HoT-Methode berechnet (Abschnitt 4.5.5).

4.5.3 Phylogenetische Bäume

Stammbäume wurden in dieser Arbeit mit einer Distanz- und einer Wahrscheinlichkeitsmethode abgeleitet. Für den Distanzansatz wurden die Programme protdist (Abschnitt 4.3.9) und neighbor (Abschnitt 4.3.10) des PHYLIP-Programmpakets verwendet. Distanzmatrizen wurden anhand des JTT-Modells (Jones et al., 1992) erstellt und Bäume nach der *Neighbor-Joining*-Methode (Saitou und Nei, 1987) abgeleitet. Für den Wahrscheinlichkeitsansatz wurde das Programm PHYML (Abschnitt 4.3.13) verwendet. Proteindistanzen wurden ebenfalls anhand des JTT-Modells erstellt. Für die Variabilität der Mutationsraten wurde ein diskretes Γ-Modell mit acht Kategorien angenommen. Der α-Parameter wurde aus den Daten geschätzt. Analog wurden phylogenetische Bäume nach den beschriebenen Distanz- und Wahrscheinlichkeitsmethoden aus den alternativen Alignments der HoT-Methode abgeleitet (Abschnitt 4.5.5). Die phylogenetischen Bäume wurden mit einem Perl-Skript analysiert und solche Bäume als Indiz für einen cyanobakteriellen Ursprung des Pflanzenproteins gewertet, in denen die Pflanzensequenz einen kleinsten Klan (vgl. Tabelle 3.1) mit Cyanobakterien bildet (Abschnitt 4.5.4).

4.5.4 Auswertung phylogenetischer Bäume

Phylogenetische Bäume wurden in dieser Arbeit auf die Nachbargruppen der Pflanzensequenzen hin analysiert. Diese Nachbargruppen wurden über den klein-

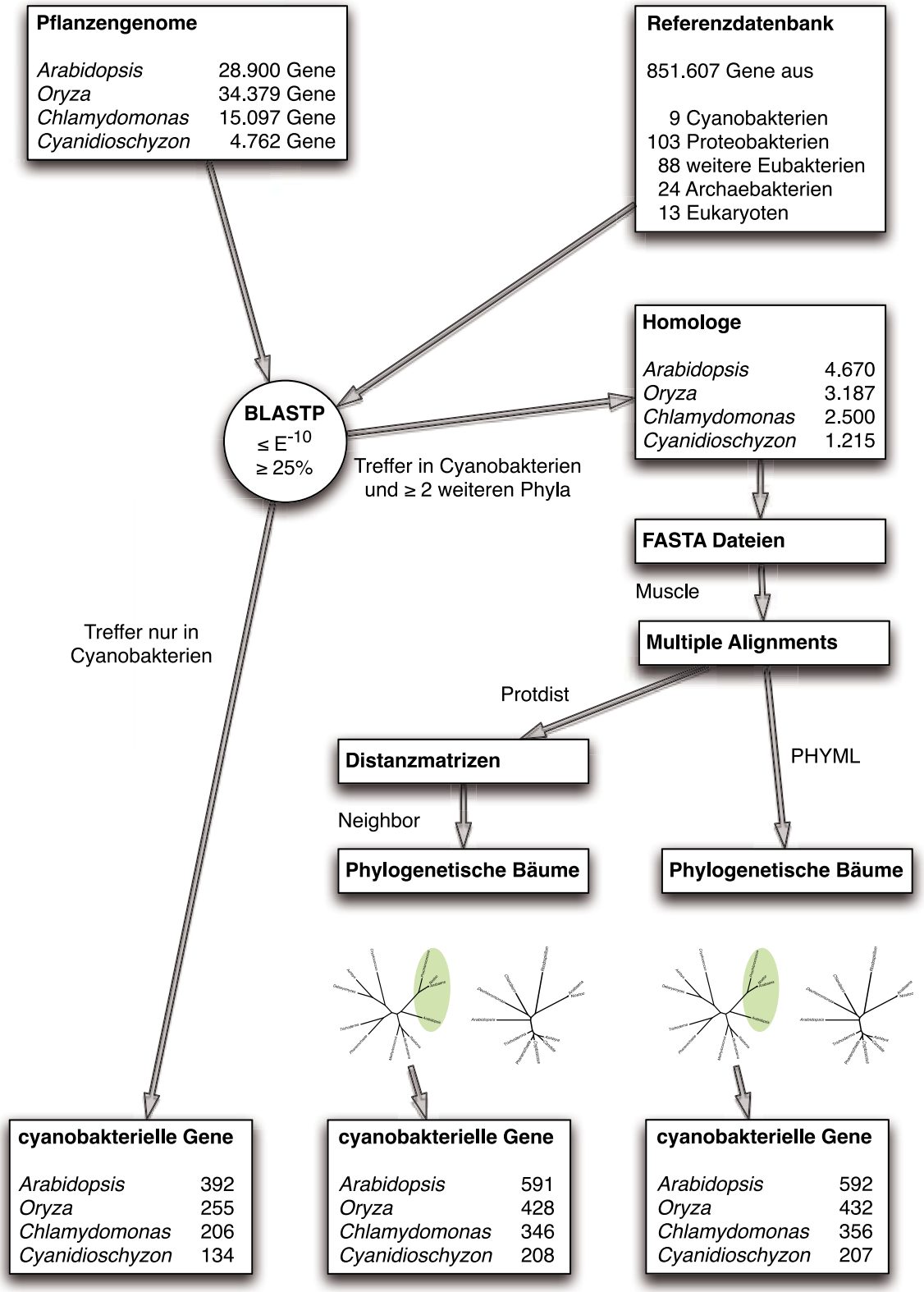

Abbildung 4.7: Arbeitsablauf zur Identifizierung cyanobakterieller Gene. Der Datensatz der Pflanzenproteine, die durch die Homologiesuche als cyanobakteriell klassifiziert wurden, enthält Proteine, die nur in Cyanobakterien Homologe aufweisen, und solche, die hauptsächlich in Cyanobakterien Homologe aufweisen (vgl. Abschnitt 4.5.1).

sten Klan definiert, den eine Pflanzensequenz mit Taxa anderer Gruppen bildet. Enthielt dieser kleinste Klan ausschließlich Homologe aus Cyanobakterien, so wurde dieser Baum als Indiz für einen cyanobakteriellen Ursprung des Pflanzengens gewertet. Der Einfachheit halber wurden die restlichen Phyla zu größeren Gruppen wie Proteobakterien, weitere Eubakterien, Archaebakterien und Eukaryoten zusammengefasst. Gemischte Gruppen enthalten Taxa mehrerer dieser Gruppen. Die Auswertung der phylogenetischen Bäume erfolgte unter Zuhilfenahme des Programms consense, welches normalerweise für die Erstellung eines Konsensus-Baums aus mehreren Bäumen genutzt wird (Abschnitt 4.3.11). Dieses Programm generiert eine Ausgabedatei mit Informationen über die Häufigkeit des Auftretens der verschiedenen Bipartitionen der einzelnen Bäume in einer Notation aus Punkten und Sternchen. Wird consense mit nur einem Baum aufgerufen, so enthält die Ausgabe alle Bipartitionen, die in diesem auftreten (Abbildung 4.8). Die Ausgabe kann leicht mit einem Perl-Skript (Abschnitt 4.3.2) gelesen und auf die Nachbargruppe der Pflanzensequenz untersucht werden.

Der prozentuale Anteil cyanobakterieller Gene P_{cyano} eines Organismus ergibt sich aus der Anzahl an Genen N_{cyano}, für die nach der oben beschriebenen Methode ein cyanobakterieller Ursprung abgeleitet wurde, und der Anzahl an berechneten Bäumen N_{trees} als

$$P_{cyano} = \frac{N_{cyano}}{N_{trees}} \times 100.$$

Für den Anteil cyanobakterieller Pflanzengene wurde darüber hinaus eine normalisierte Darstellung verwendet, die berücksichtigt, wie groß der prozentuale Anteil an Cyanobakteriengenen P_{data} an der Referenzdatenbank ist. Der (dimensionslose) normalisierte Anteil P_{norm} ist

$$P_{norm} = \frac{P_{cyano}}{P_{data}}.$$

4.5.5 Die Anwendung der HoT-Methode

Die HoT-Methode (Abschnitt 4.3.14) ist eine Methode zur Bestimmung der Alignmentverlässlichkeit. Die Anwendung dieser Methode in der vorliegenden Arbeit ist im Folgenden näher beschrieben. Zunächst wird ein konventionelles multiples Sequenzalignment aus Sequenzen berechnet (Abbildung 4.9, 1). Danach werden die nicht-alignierten Sequenzen invertiert (Abbildung 4.9, 2). Nukleotidsequenzen werden dabei von der 5′-3′- in die 3′-5′-Orientierung, Aminosäuresequenzen von

```
Consensus tree program, version 3.67

Species in order:

  1. Acinetobac
  2. Arabidopsi
  3. Streptomyc
  4. Schizosacc
  5. Anabaena

Sets included in the consensus tree

Set (species in order)     How many times out of    1.00

..***                      1.00
...**                      1.00

CONSENSUS TREE:

         +--------------------Arabidopsi
  +------|
  |      |      +-------------Streptomyc
  |      +--1.0-|
  |             |      +------Anabaena
  |             +--1.0-|
  |                    +------Schizosacc
  |
  +---------------------------Acinetobac

  remember: this is an unrooted tree!
```

Abbildung 4.8: Ausgabe von Consense für einen einzelnen Baum (gekürzt).

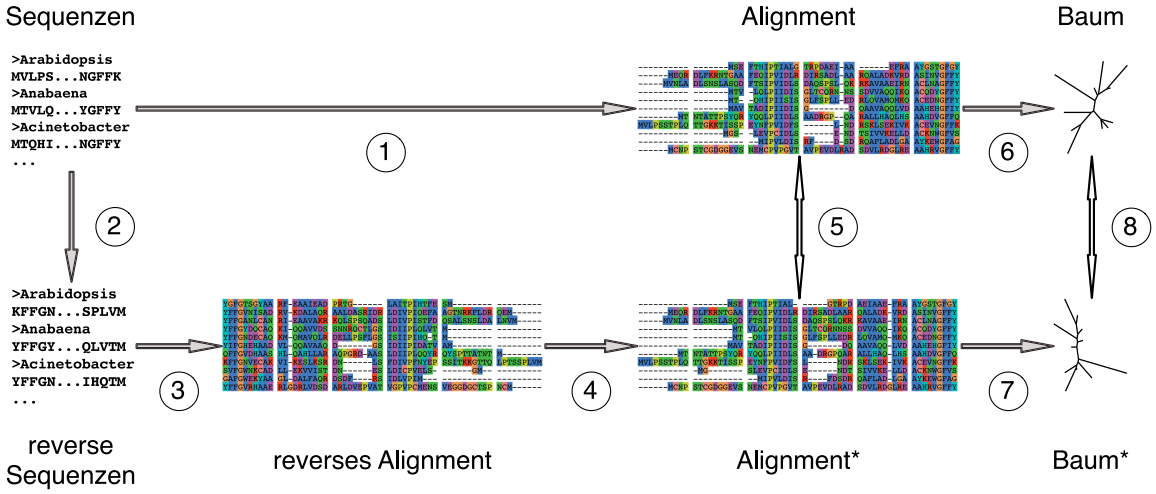

Abbildung 4.9: Anwendung der HoT-Methode (Landan und Graur, 2007). Der Arbeitsablauf zur Ermittlung der Verlässlichkeit von multiplen Sequenzalignments und phylogenetischen Bäumen umfasst acht Schritte: 1. Berechnung des „normalen“ Sequenzalignments aus Sequenzen in N-C-Orientierung; 2. Inversion der unalignierten Sequenzen in die C-N-Orientierung; 3. Berechnung eines reversen Alignments aus diesen Sequenzen; 4. Inversion des reversen Alignments; 5. Vergleich der alternativen Alignments; 6. und 7. Ableitung phylogenetischer Bäume aus den alternativen Alignments; 8. Vergleich der alternativen Bäume.

der N-C- in die C-N-Orientierung überführt. Aus diesen reversen Sequenzen wird anschließend mit dem selben Programm und den selben Parametern ein Alignment berechnet (Abbildung 4.9, 3). Dieses reverse Alignment wird invertiert und stellt eine Alternative zu dem konventionellen Alignment dar (Abbildung 4.9, 4). Die Diskrepanz zwischen den alternativen Alignments (Abbildung 4.9, 5) ist ein Maß für die Verlässlichkeit derselben. Um diese zu messen, wurden zwei Kriterien für die Alignmentverlässlichkeit verwendet. Der CS-Wert (engl. *column score*, Thompson et al. (1999*a*)) ist der Anteil identischer Spalten zwischen zwei Alignments (Definition in Abschnitt 4.6.2). Der SPS-Wert (engl. *sum of pairs score*, Thompson et al. (1999*a*)) ist der Anteil identisch alignierter Aminosäure- bzw. Nukleotidpaare zwischen zwei Alignments (Definition in Abschnitt 4.6.3). Beinhaltet eine Analyse das Erstellen von Stammbäumen, so kann aus beiden Alignments jeweils ein phylogenetischer Baum abgeleitet werden (Abbildung 4.9, 6 und 7), deren Unterschied ebenfalls quantifiziert werden kann. Für die Bewertung der Verlässlichkeit der phylogenetischen Bäume (Abbildung 4.9, 8) wurde der PPS-Wert (engl. *phylogenetic partitions score*) verwendet, der den Anteil identischer Bipartitionen zweier Bäume beschreibt (Definition in Abschnitt 4.6.4).

4.5.6 Paarweise Distanzen zwischen orthologen Pflanzenproteinen

Orthologe zwischen *Arabidopsis* und *Oryza* sowie *Arabidopsis* und *Chlamydomonas* wurden über den besten reziproken BLAST-Treffer (Rivera et al., 1998) identifiziert. Ein Erwartungswert $E \leq 10^{-10}$ diente bei der BLASTP-Suche als Homologiekriterium. Sequenzpaare wurden mit Clustal W (Abschnitt 4.3.6) aligniert und Proteindistanzen mit dem Programm protdist (Abschnitt 4.3.9) unter Verwendung des JTT-Modells (Jones et al., 1992) berechnet.

4.5.7 Identifizierung von NUPTs

Die vier Genome der Pflanzen und Algen wurden auf transferierte Plastiden-DNA (NUPTs, von engl. *nuclear plastid DNA*, Richly und Leister (2004)) untersucht. Dazu wurde zu allen plastidenkodierten Proteinen (Abschnitt 4.1) mittels BLAST-Suche (Abschnitt 4.3.4) nach Homologen unter den zugehörigen kernkodierten Proteinen gesucht. Alle Treffer mit 100 % Aminosäureidentität wurden als NUPTs eingestuft.

4.5.8 Funktionelle Charakterisierung

Zu Pflanzenproteinen, für die ein cyanobakterieller Ursprung vorhergesagt worden war (Abschnitt 4.5.1 bis Abschnitt 4.5.4), wurde unter den 3.305.720 Proteinen der GENES-Datenbank (KEGG, Abschnitt 4.1.4) mittels BLASTP nach Homologen gesucht. Als Homologiekriterium diente ein Erwartungswert E $\leq 10^{-10}$. Einträge dieser Datenbank sind in 10.990 Proteinfamilien (engl. *cluster*) unterteilt und für viele dieser Familien sind Informationen zur biochemischen Funktion anhand von EC-Nummern (engl. *enzyme commission*) hinterlegt. Auf diese Weise wurde für jeden Organismus eine Liste von EC-Nummern erstellt, die den Funktionen der cyanobakteriellen Enzyme entspricht.

Anhand der Listen von EC-Nummern wurde auf der Internetseite von KEGG mit der Funktion *search objects in pathways*[16] nach Stoffwechselwegen gesucht, in denen diese Enzyme auftreten. Die Datenbank enthält Stoffwechselkarten für die in dieser Arbeit analysierten Organismen *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae*. Auf diese Weise wurden aus den Listen die EC-Nummern für diejenigen Enzyme entfernt, über die der entsprechende Organismus nicht verfügt. Die Enzyme der so bereinigten Listen wurden in neun Kategorien von Stoffwechselwegen (vgl. Tabelle 5.5 und Abbildung 6.1) eingeteilt.

16 http://www.genome.jp/kegg/tool/search_pathway.html

Enzyme können dabei in mehr als einem Stoffwechselweg auftreten. Ebenfalls über die Funktion *search objects in pathways* konnten die cyanobakteriellen Enzyme in einzelnen Stoffwechselwegen visualisiert werden.

4.6 Maße

Die in dieser Arbeit verwendeten Maße zur Quantifizierung der Sequenzdivergenz, sowie Alignment- und Baumverlässlichkeit sind im Folgenden näher erläutert.

4.6.1 Der MBL-Wert

Als Maß für die Sequenzdivergenz von Proteinfamilien wurde die mittlere Astlänge (gemessen in Substitutionen pro Position) berechnet. Die Anzahl b der Äste in einem ungewurzelten Baum von n Taxa ergibt sich aus $b = 2n - 3$. Die mittlere Astlänge (MBL, engl. *mean branch length*) wurde definiert als

$$MBL = \frac{1}{b} \sum_{i=1}^{b} a_i$$

und ist die Summe aller Astlängen a_i dividiert durch die Anzahl der Äste b.

4.6.2 Der CS-Wert

Der CS-Wert (engl. *column score*) beschreibt den Anteil identischer Spalten zwischen zwei multiplen Sequenzalignments und wurde von Thompson et al. (1999 *a*) definiert. In dieser Arbeit wurden nach der HoT-Methode für alle multiplen Alignments alternative Alignments berechnet, die sich ergeben, wenn die Sequenzen in umgekehrter Orientierung (C-N- anstatt N-C-Orientierung für Aminosäuresequenzen) aligniert werden (Abschnitt 4.5.5). Der CS-Wert des Vergleichs eines Alignments mit dem entsprechenden alternativen Alignment wurde als Maß für die Alignmentverlässlichkeit verwendet. Der CS-Wert eines Alignments ergibt sich aus

$$CS = \frac{1}{M} \sum_{i=1}^{M} C_i \quad \text{mit} \quad \begin{cases} C_i = 1, & \text{wenn alle Zeichen in Spalte } i \text{ auch im} \\ & \text{Referenzalignment aligniert sind,} \\ C_i = 0, & \text{sonst.} \end{cases}$$

M ist die Anzahl an Spalten des Standardalignments (N-C-Orientierung). Für die i-te Alignmentspalte ist $C_i = 1$, wenn alle Zeichen in Spalte i im N-C-Alignment

auch im C-N-Alignment aligniert sind. Ist ein Alignment mit dem alternativen Alignment in allen Spalten identisch, so ergibt sich ein $CS = 1$, was zuverlässigen Alignments entspricht.

4.6.3 Der SPS-Wert

Der SPS-Wert (engl. *sum of pairs score*) beschreibt den Anteil identisch alignierter Aminosäure- bzw. Nukleotidpaare zwischen zwei Alignments und wurde erstmalig von Thompson et al. (1999*a*) beschrieben. In dieser Arbeit wurden nach der HoT-Methode für alle multiplen Alignments alternative Alignments berechnet, die sich ergeben, wenn die Sequenzen in umgekehrter Orientierung aligniert werden (Abschnitt 4.5.5). Der SPS-Wert des Vergleichs eines Alignments mit dem entsprechenden alternativen Alignment wurde als Maß für die Alignmentverlässlichkeit verwendet. In einem Alignment mit N Taxa und M Spalten besteht die i-te Spalte aus den Aminosäuren A_{i1}, A_{i2}, ..., A_{iN}. Für jedes Aminosäurepaar A_{ij} und A_{ik} wird ein Variable p_{ijk} definiert, wobei $p_{ijk} = 1$ ist, wenn A_{ij} und A_{ik} auch im Referenzalignment aligniert sind. Sonst gilt $p_{ijk} = 0$. Dabei kann ein Paar aus zwei Zeichen oder einem Zeichen und einem Lückensymbol, nicht aber aus zwei Lückensymbolen bestehen. Für den spaltenweisen Vergleich eines Alignments mit einem Referenzalignment gilt für die i-te Spalte

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

S_i entspricht der doppelten Anzahl identisch alignierter Aminosäurepaare, da sowohl der Wert für den Vergleich von A_{ij} mit A_{ik} und den Vergleich von A_{ik} mit A_{ij} in die Berechnung einfließt. Der SPS-Wert für ein Alignment ist als

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

definiert, wobei Mr die Anzahl an Spalten im Referenzalignment und S_{ri} der Wert S_i für die i-te Spalte des Referenzalignments ist. Der Term $\sum_{i=1}^{Mr} S_{ri}$ beschreibt den Vergleich des Referenzalignments mit sich selbst und ergibt so die doppelte Anzahl aller möglichen Aminosäurepaare (siehe oben). Da in der Formel für den SPS-Wert sowohl der Zähler als auch der Nenner des Bruchs doppelt gezählt werden, gibt der SPS-Wert den Anteil identisch alignierter Aminosäurepaare korrekt wieder. Ist ein Alignment mit dem alternativen Alignment in allen Aminosäurepaaren identisch, so ergibt sich ein $SPS = 1$, was zuverlässigen Alignments entspricht.

4.6.4 Der PPS-Wert

Der PPS-Wert (engl. *phylogenetic partitions score*) wurde in dieser Arbeit definiert, um die Verlässlichkeit phylogenetischer Bäume zu messen. Dazu wird ein Stammbaum mit einem alternativen Baum verglichen, der sich ergibt, wenn aus Sequenzen in umgekehrter Orientierung ein Alignment berechnet und daraus ein phylogenetischer Baum abgeleitet wird (Abschnitt 4.5.5). Der PPS-Wert basiert auf der Robinson-Foulds-Metrik (Robinson und Foulds, 1981), wie sie von dem Programm treedist des PHYLIP-Programmpakets (Abschnitt 4.3.8) berechnet wird. Dieses Maß wird auch symmetrische Distanz genannt und ist als $A + B$ definiert, wobei A die Anzahl der Bipartitionen beschreibt, die im ersten aber nicht im zweiten Baum auftreten. B ist analog als die Anzahl der Bipartitionen definiert, die im zweiten aber nicht im ersten Baum auftreten. Die Werte für die symmetrische Distanz zweier Stammbäume mit je n Taxa liegt zwischen null (identische Topologie) und $2n - 6$ (alle nicht trivialen Bipartitionen unterschiedlich). Obwohl zwei ungewurzelte Bäume $4n - 6$ Bipartitionen aufweisen, können Unterschiede nur in den $2n - 6$ internen Kanten und nicht in den $2n$ externen Kanten auftreten. Der PPS-Wert beschreibt die Ähnlichkeit zweier Bäume und wurde als

$$PPS = 1 - \frac{A+B}{2n-6}$$

definiert, um die Werte für Stammbäume unterschiedlicher Taxonzahlen vergleichbar zu machen. Identische – im HoT-Vergleich vollständig reproduzierbare Bäume – erhalten einen PPS-Wert von eins und Stammbäume, die sich an allen internen Kanten unterscheiden, einen Wert von null.

4.6.5 Die Pearson-Korrelation

Mit der Pearson-Korrelation wird ein Korrelationskoeffizient berechnet, der ein dimensionsloses Maß für den linearen Zusammenhang zwischen zwei Merkmalen ist. Er kann Werte zwischen -1 (vollständige negative Korrelation) und +1 (vollständige positive Korrelation) annehmen. Bei einem Wert von null liegt keine lineare Korrelation zwischen den Merkmalen vor. Dennoch können die Merkmale nicht-linear korreliert sein. Der Korrelationskoeffizient für zwei quadratisch integrierbare Zufallsvariablen X und Y ist als

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)} \times \sqrt{Var(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{Var(X)} \times \sqrt{Var(Y)}}$$

definiert, sofern für die Varianz $Var(X) > 0$ und $Var(Y) > 0$ gilt. Dabei bezeichnet E den Erwartungswert und $\sqrt{Var(X)}$ die Standardabweichung von X. Liegen für die Zufallsvariablen lediglich zwei Messreihen x_1, x_1, ..., x_n und y_1, y_1, ..., y_n vor, so ist der empirische Korrelationskoeffizient als

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\times\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

definiert. Dabei sind $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ die empirischen Erwartungswerte (Mittelwerte) der Messreihen. In dieser Arbeit wurde die Pearson-Korrelation mit MATLAB® (Abschnitt 4.3.15) berechnet, wobei neben dem Korrelationskoeffizienten r auch der p-Wert berechnet wird, welcher ein Maß für die Signifikanz der Korrelation ist. Nur wenn der Korrelationskoeffizient signifikant ist, kann von einem statistischen Zusammenhang der Messreihen gesprochen werden. In dieser Arbeit wurde ein Signifikanzniveau von $p < 0{,}05$ verwendet.

5 Ergebnisse

5.1 Homologiesuche zu 83.138 Pflanzengenen in 237 Referenzgenomen

Um den cyanobakteriellen Anteil der vier Genome der Pflanzen und Algen zu bestimmen, wurde zunächst nach Homologen zu den darin kodierten Genen gesucht (Abschnitt 4.5.1). Die aus den Referenzdatenbanken (Tabelle 6.1) heruntergeladenen Listen der Translationen der vorhergesagten Gene (hier synonym mit Proteine benutzt) enthielten zum Teil mehrere Einträge der selben Aminosäuresequenzen. Diese doppelten Einträge wurden entfernt. Zu dem so erhaltenen nicht-redundanten Datensatz von 83.138 Pflanzenproteinen wurde mit BLAST (Altschul et al., 1997) in 237 komplett sequenzierten Referenzgenomen – darunter neun Cyanobakterien, 103 Proteobakterien, 97 weitere Eubakterien, 15 Archaebakterien und 13 nicht-photosynthetische Eukaryoten (Tabelle 6.1) – nach Homologen gesucht (vgl. Abschnitt 4.5.1).

Für *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae* wurden 5.199, 3.524, 2.767 beziehungsweise 1.363 Homologe (Tabelle 5.1) in den Cyanobakteriengenomen gefunden, welche die in Abschnitt 4.5.1 beschriebenen Kriterien erfüllen. Davon wurde für 4.670, 3.186, 2.500 beziehungsweise 1.213 Pflanzenproteine zusätzlich noch mindestens ein Homolog in zwei weiteren Phyla gefunden (Tabelle 5.1). Für diesen Datensatz wurden im Folgenden phylogenetische Bäume berechnet, um zu klären, ob diese Pflanzenproteine cyanobakteriellen Ursprungs sind. 341, 232, 166 beziehungsweise 109 Pflanzenproteine wiesen Homologe ausschließlich in Cyanobakterien auf. Diese Proteine wurden ohne phylogenetische Analyse als cyanobakteriell klassifiziert. Weitere 188, 105, 101 beziehungsweise 38 Proteine aus Pflanzen und Algen hatten Homologe in Cyanobakterien und einem weiteren Phylum. Davon wurden 51, 23, 40 beziehungsweise 25 Proteine als cyanobakteriell charakterisiert, weil Homologe hauptsächlich

Tabelle 5.1: Anzahl an Homologen zu 83.138 Pflanzenproteinen in 237 Referenzgenomen aus 21 Phyla. Es ist jeweils angegeben, wieviele den in Abschnitt 4.5.1 beschriebenen Kriterien entsprechende Homologe insgesamt gefunden wurden (Treffer) und wieviele davon in Bäumen repräsentiert sind (Bäume). Treffer sind nicht in Bäumen repräsentiert, wenn für das entsprechende Pflanzenprotein keine Homologe in Cyanobakterien und zwei weiteren Phyla gefunden wurden.

Phylum	*A. thaliana*		*O. sativa*		*C. reinhardtii*		*C. merolae*	
(Anzahl Genome)	Treffer	Bäume	Treffer	Bäume	Treffer	Bäume	Treffer	Bäume
Cyanobacteria (9)	5.199	4.670	3.524	3.186	2.767	2.500	1.363	1.213
Actinobacteria (17)	5.265	3.760	3.934	2.617	3.437	2.143	1.298	1.056
Aquificae (1)	1.550	1.387	911	811	857	792	603	564
Bacteroidetes (4)	3.116	2.431	2.024	1.558	1.702	1.392	988	849
Chlamydiae (6)	1.884	1.638	1.124	982	946	828	2.126	858
Chlorobi (3)	2.264	2.105	1.404	1.283	1.325	1.204	813	758
Chloroflexi (2)	1.232	1.157	882	691	874	742	485	464
Deinococcus-Thermus (2)	2.598	2.158	1.912	1.455	2.061	1.478	890	770
Firmicutes (45)	5.188	3.545	3.402	2.412	2.608	1.860	1.214	996
Fusobacteria (1)	1.328	1.200	742	675	773	700	481	441
Planctomycetes (1)	2.730	2.177	1.882	1.479	1.483	1.266	788	705
Proteobacteria (103)	6.857	4.378	4.958	3.009	4.385	2.392	1.585	1.160
Spirochaetes (5)	2.766	2.317	1.633	1.380	1.415	1.211	851	735
Thermotogae (1)	1.615	1.368	952	817	854	757	582	530
Crenarchaeota (5)	1.941	1.366	1.175	815	974	694	693	491
Euryarchaeota (18)	4.100	2.902	2.716	1.911	1.982	1.450	1.122	801
Nanoarchaeota (1)	429	208	248	122	217	131	202	106
Ascomycota (8)	10.734	3.850	7.853	2.624	4.881	1.819	2.303	922
Basidiomycota (2)	10.047	3.545	7.685	2.497	4.151	1.619	2.126	858
Microsporidia (1)	2.849	836	1.594	516	1.182	415	769	284
Protisten (2)	7.187	2.158	4.432	1.509	3.155	1.139	1.415	525

in Cyanobakterien auftraten (vgl. Abschnitt 4.5.1). Die meisten auf diese Weise als cyanobakteriell klassifierten Proteine hatten Homologe in allen neun Cyanobakterien und nur wenigen Spezies eines weiteren Phylums. Insgesamt wurden so für *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* 392, 255, 206 beziehungsweise 134 Pflanzenproteine über die Homologiesuche ein cyanobakterieller Ursprung abgeleitet.

5.2 Identifizierung cyanobakterieller Proteine mittels phylogenetischer Analysen

4.670, 3.186, 2.500 beziehungsweise 1.213 Pflanzenproteine aus *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae* wiesen Homologe in Cyanobakterien und mindestens zwei weiteren Phyla auf (Abschnitt 5.1). Diese Proteine wurden mit dem Programm Muscle (Edgar, 2004 *a*,*b*) mit ihren Homologen aligniert. Aus den multiplen Sequenzalignments (MSAs) wurden Spalten entfernt, die Lücken (engl. *gaps*) enthalten (siehe Abschnitt 4.5.2). Phylogenetische

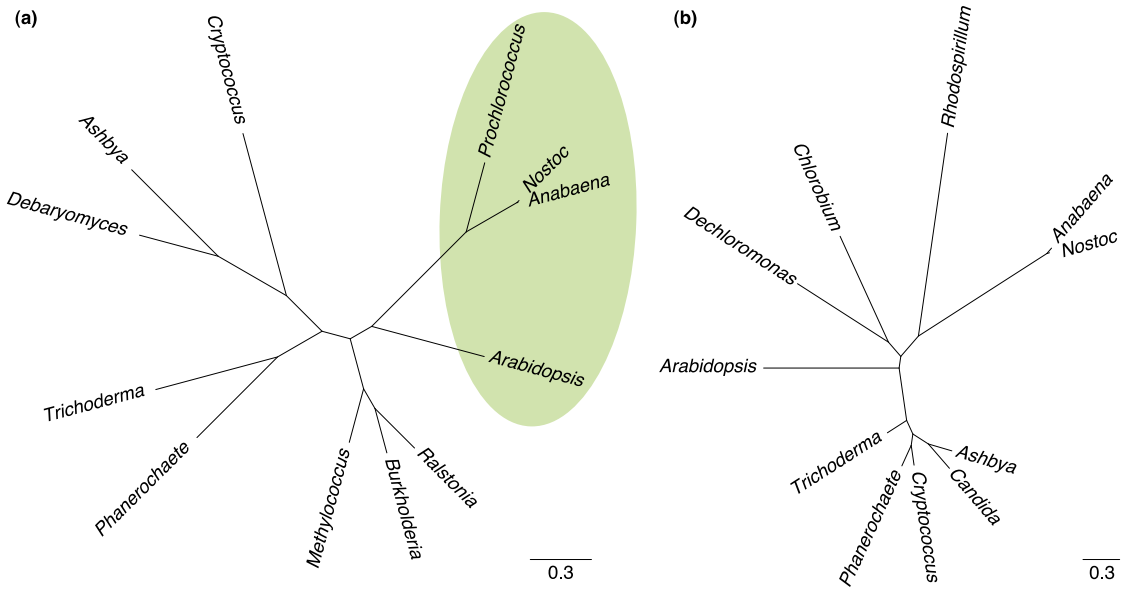

Abbildung 5.1: Zwei Wahrscheinlichkeitsbäume für Proteine aus *Arabidopsis thaliana*: (a) Thioredoxin-abhängige Peroxidase 2, TPX2, GI:15218876. Das Pflanzenprotein bildet einen Klan mit Homologen aus Cyanobakterien der Gattungen *Anabaena*, *Nostoc* und *Prochlorococcus*. Für dieses Protein wurde ein cyanobakterieller Ursprung angenommen. (b) Metacaspase 6, ATMC6, GI:15219340. Das Pflanzenprotein bildet einen Klan mit Homologen aus Pilzen oder einen gemischten Klan mit Proteobakterien, grünen Schwefelbakterien und Cyanobakterien. Für dieses Protein wurde ein cyanobakterieller Ursprung abgelehnt.

Bäume wurden sowohl mit der Methode der maximalen Wahrscheinlichkeit (engl. *maximum likelihood*, ML) als auch mit einem Distanzverfahren und der *Neighbor-Joining*-Methode (NJ) rekonstruiert (siehe Abschnitt 4.5.3).

Die phylogenetischen Bäume wurden ausgewertet und ihre Topologien auf einen möglichen cyanobakteriellen Ursprung des Pflanzenhomologs untersucht. Gab es einen kleinsten Klan (für eine Erläuterung der verwendeten Termini „Klan“ und „Nachbargruppe“ siehe Abschnitt 3.7), der nur das Pflanzenprotein sowie ein oder mehrere Homologe aus Cyanobakterien enthielt, so wurde dieser Baum als ein Indiz für einen cyanobakteriellen Ursprung des Pflanzenproteins gewertet (Abbildung 5.1a). Enthielt dieser kleinste Klan neben dem Pflanzenprotein Homologe aus Proteobakterien, weiteren Eubakterien, Archaebakterien oder Eukaryoten so wurde ein cyanobakterieller Ursprung des entsprechenden Pflanzenproteins abgelehnt (Abbildung 5.1b). Ebenso wurde mit gemischten Klans verfahren, die Taxa mehrerer dieser Gruppen enthalten.

Für *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae* wurde für 592, 432, 356 beziehungsweise 207 Wahrscheinlichkeitsbäume

Tabelle 5.2: Nachbargruppen der Pflanzenhomologen in 11.569 Wahrscheinlichkeitsbäumen. Multiple Sequenzalignments wurden mit Muscle (Edgar, 2004 *a*,*b*) berechnet. Phylogenetische Bäume wurden mit Phyml (Guindon und Gascuel, 2003) anhand des JTT-Modells (Jones et al., 1992) aus Alignmentspalten ohne Lücken rekonstruiert. Eine detaillierte Beschreibung aller Parameter ist unter Abschnitt 4.5.1 bis Abschnitt 4.5.4 gegeben. Gemischte Klans enthalten Taxa verschiedener Phyla. Bei der Rekonstruktion mit einer Distanzmethode wurden ähnliche Ergebnisse erzielt (Tabelle 6.2).

Nachbargruppe	*A. thaliana* (%)	*O. sativa* (%)	*C. reinhardtii* (%)	*C. merolae* (%)	Datenbankgröße (%)
Cyanobakterien	592 (12,7%)	432 (13,6%)	356 (14,2%)	207 (17,1%)	31.940 (3,8%)
Proteobakterien	522 (11,2%)	308 (9,7%)	458 (18,3%)	104 (8,6%)	360.234 (42,3%)
Weitere Eubakterien	882 (18,9%)	588 (18,4%)	524 (21,0%)	257 (21,1%)	226.314 (26,6%)
Archaebakterien	45 (1,0%)	38 (1,2%)	25 (1,0%)	18 (1,5%)	56.513 (6,6%)
Eukaryoten	2.156 (46,2%)	1.498 (47,0%)	895 (35,8%)	523 (43,1%)	176.606 (20,7%)
Gemischt	473 (10,1)	323 (10,1%)	242 (9,7%)	106 (8,8%)	

ein cyanobakterieller Ursprung des Pflanzenproteins abgeleitet (Tabelle 5.2). Bezogen auf die Gesamtzahl an berechneten Bäumen pro Spezies waren dies 12,7 %, 13,6 %, 14,2 % beziehungsweise 17,1 %. Proteobakterien bildeten in 8,6 % (*Cyanidioschyzon*) bis 18,3 % (*Chlamydomonas*) und weitere Eubakterien in 18,4 % (*Oryza*) bis 21,1 % (*Cyanidioschyzon*) der Bäume eine Nachbargruppe zu dem untersuchten Pflanzenprotein. In 1,0 % (*Arabidopsis* und *Chlamydomonas*) bis 1,5 % (*Cyanidioschyzon*) war die Nachbargruppe archaebakteriell und in 35,8 % (*Chlamydomonas*) bis 47,0 % (*Oryza*) bestand sie aus eukaryotischen Homologen. Gemischte Klans traten in 8,8 % (*Cyanidioschyzon*) bis 10,1 % (*Arabidopsis* und *Oryza*) der Bäume auf. In solchen Fällen enthielt die Nachbargruppe Homologe aus verschiedenen Phyla. Die Proteine der Referenzdatenbank stammen zu 3,8 % aus Cyanobakterien, zu 42,3 % aus Proteobakterien, zu 26,6 % aus weiteren Eubakterien, zu 6,6 % aus Archaebakterien und zu 20,7 % aus Eukaryoten (Tabelle 5.2).

Tabelle 5.2 zeigt den prozentualen Anteil cyanobakterieller Pflanzengene, wie er aus 11.569 Wahrscheinlichkeitsbäumen abgeleitet wurde. Mit Distanzbäumen wurden in der Summe ähnliche Ergebnisse erzielt (vgl. Tabelle 6.2). Auch für individuelle Pflanzenproteine stimmten die Vorhersagen cyanobakteriellen Ursprungs mit den beiden Ansätzen der phylogenetischen Rekonstruktion weitgehend überein (Abbildung 5.2). Für *Arabidopsis* wurden mittels Wahrscheinlichkeitsbäumen 592 (Tabelle 5.2) und mittels Distanzbäumen 591 (Tabelle 6.2) cyanobakterielle Proteine identifiziert. 528 Proteine ($\hat{=}$80,1 %) wurden übereinstimmend mit beiden Rekonstruktionsmethoden vorhergesagt (Abbildung 5.2). Auch für *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* stimmten die Vorhersagen mit beiden Methoden für 79,5 %, 79,5 % beziehungsweise 86,9 % der Proteine überein (Abbildung 5.2).

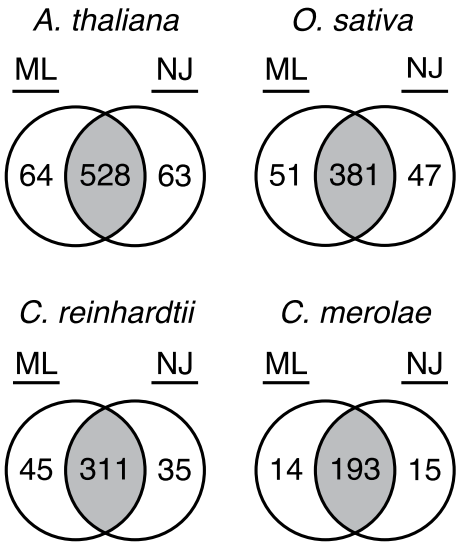

Abbildung 5.2: Venndiagramme der Übereinstimmung der Vorhersage cyanobakterieller Proteine mittels Wahrscheinlichkeits- (ML) und Distanzbäumen (NJ) für individuelle Gene.

In 12,7 % bis 17,1 % der Wahrscheinlichkeitsbäume wurde eine cyanobakterielle Nachbargruppe beobachtet (Tabelle 5.2). 3,8 % der Proteine der Referenzdatenbank stammen aus Cyanobakterien (Tabelle 5.2). Abbildung 5.3 zeigt eine normalisierte Darstellung der Ergebnisse aus Tabelle 5.2, bei der der prozentuale Anteil der Nachbargruppe durch den prozentualen Anteil dividiert wurde, den die jeweilige Gruppe an der Größe der Referenzdatenbank hat (vgl. Abschnitt 4.5.4). In dieser (dimensionslosen) Darstellung zeigen Cyanobakterien mit 3,3 (*Arabidopsis*) bis 4,5 (*Cyanidioschyzon*) vor Eukaryoten mit 1,7 (*Chlamydomonas*) bis 2,3 (*Oryza*) das stärkste Signal. Das drittstärkste Signal liefern Eubakterien mit 0,7 (*Arabidopsis* und *Oryza*) bis 0,8 (*Chlamydomonas* und *Cyanidioschyzon*) vor Proteobakterien mit 0,2 (*Oryza* und *Cyanidioschyzon*) bis 0,4 (*Chlamydomonas*) und Archaebakterien (0,2 für alle vier photosynthetischen Eukaryoten).

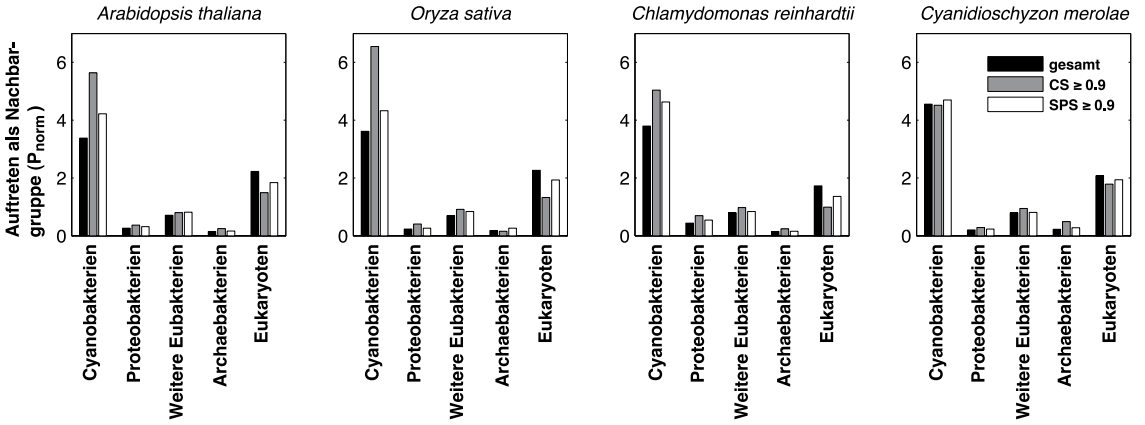

Abbildung 5.3: Normalisierte Darstellung von Nachbargruppen zu Pflanzenproteinen in 11.569 Wahrscheinlichkeitsbäumen. Der normalisierte Anteil der Nachbargruppen P_{norm} erfolgte mittels Division des prozentualen Auftretens der Nachbargruppen aus Tabelle 5.2 durch den prozentualen Anteil der entsprechenden Gruppe an der Datenbank (vgl. Abschnitt 4.5.4). Schwarze Balken zeigen die normalisierten Werte für den gesamten Datensatz. Graue und weiße Balken zeigen die Ergebnisse für zuverlässigere Daten mit CS $\geq$ 0,9 bzw. SPS $\geq$ 0,9 (vgl. Abschnitt 4.5.5).

5.3 Abhängigkeit des Gentransfers von der Sequenzkonservierung

Für die Genome der zwei Pflanzen und zwei Algen wurde ein Anteil an kernkodierten Proteinen cyanobakteriellen Ursprungs zwischen 12,7 % (*Arabidopsis*) und 17,1 % (*Cyanidioschyzon*) abgeleitet. Dieser Anteil war jedoch für Klassen unterschiedlich konservierter Proteine stark verschieden (Abbildung 5.4). Die Sequenzdivergenz – gegeben als mittlere Astlänge des phylogenetischen Baums einer Proteinfamilie (Definition in Abschnitt 4.6.1) – zeigte eine Normalverteilung um einen Mittelwert von 0,3 Substitutionen pro Position. Aus phylogenetischen Bäumen konservierter Sequenzen wurde öfter ein cyanobakterieller Ursprung des Pflanzenhomologs abgeleitet als aus Stammbäumen variabler Sequenzen. Für die am stärksten konservierten Proteine aus *Arabidopsis thaliana* (mittlere Astlänge des phylogenetischen Baums $\leq$ 0,2 Substitutionen pro Position) wurde ein cyanobakterieller Anteil von 22 % abgeleitet. Aus den am schwächsten konservierten Proteinen (mittlere Astlänge des phylogenetischen Baums $\geq$ 1 Substitution pro Position) wurde für diesen Organismus ein Anteil von 6 % abgeleitet. Für *Oryza* und *Chlamydomonas* wurde ebenfalls dieser Trend gefunden. Für *Cyanidioschyzon* war keine Abhängigkeit des abgeleiteten endosymbiontischen Gentransfers von der Sequenzkonservierung zu erkennen. Die Werte schwankten um den Mittelwert

von 17,1 %. Für alle vier untersuchten Genome von Pflanzen und Algen war die gefundene Abhängigkeit des abgeleiteten Gentransfers von der Wahl der Methode zur Baumrekonstruktion unabhängig. Wahrscheinlichkeitsbäume (schwarze Linie in Abbildung 5.4) und Distanzbäume (graue Linie in Abbildung 5.4) zeigten sehr ähnliche Ergebnisse.

Für konservierte Proteinfamilien wurde ein höherer Anteil cyanobakterieller Proteine abgeleitet als für variablere (Abbildung 5.4). Analysen von Pflanzenproteinen cyanobakteriellen und eukaryotischen Ursprungs zeigten keinen Unterschied in deren Sequenzkonservierung (Abbildung 5.5). Die Verteilung der Sequenzkonservierung – gemessen in Substitutionen pro Position – war für Orthologe zwischen *A. thaliana* und *O. sativa* für cyanobakterielle und eukaryotische Proteine identisch (durchgezogene und gestrichelte schwarze Linie in Abbildung 5.5). Der analoge Vergleich zeigte für Orthologe zwischen *A. thaliana* und *C. reinhardtii* geringfügige Unterschiede in der Verteilung der Sequenzkonservierung (durchgezogene und gestrichelte graue Linie in Abbildung 5.5).

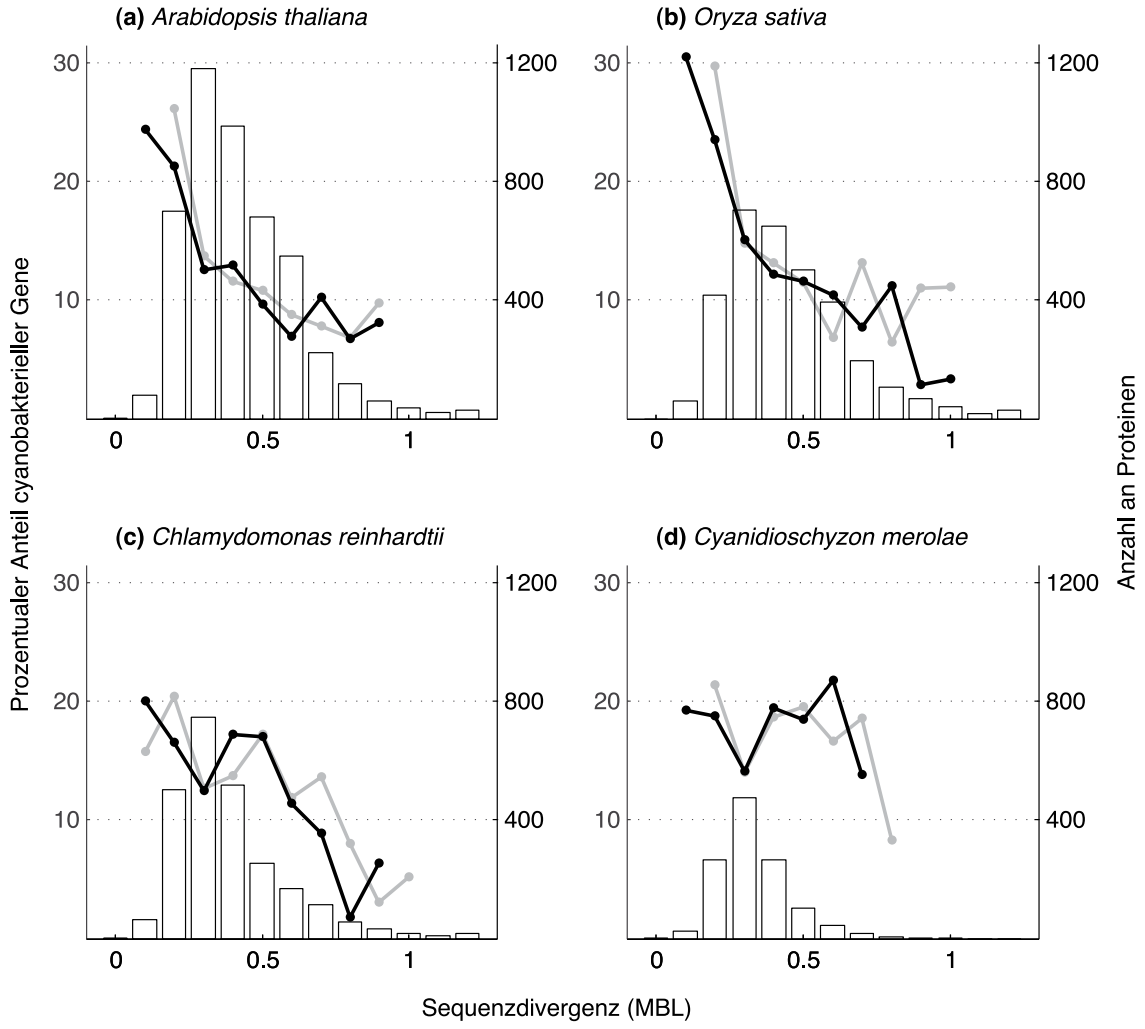

Abbildung 5.4: Prozentualer Anteil an abgeleitetem endosymbiontischen Gentransfer (EGT) in Abhängigkeit von der Sequenzdivergenz. Die Verteilung der Sequenzkonservierung (Histogramm im Hintergrund) ist durch den MBL-Wert gegeben, der die mittlere Astlänge in ML-Bäumen gemessen in Substitutionen pro Position beschreibt (definiert in Abschnitt 4.6.1). Eine ähnliche Verteilung wurde für NJ-Bäume beobachtet (Daten nicht gezeigt). Die schwarzen Linien zeigen die Ableitung des cyanobakteriellen Ursprungs aus ML-, die grauen aus NJ-Bäumen.

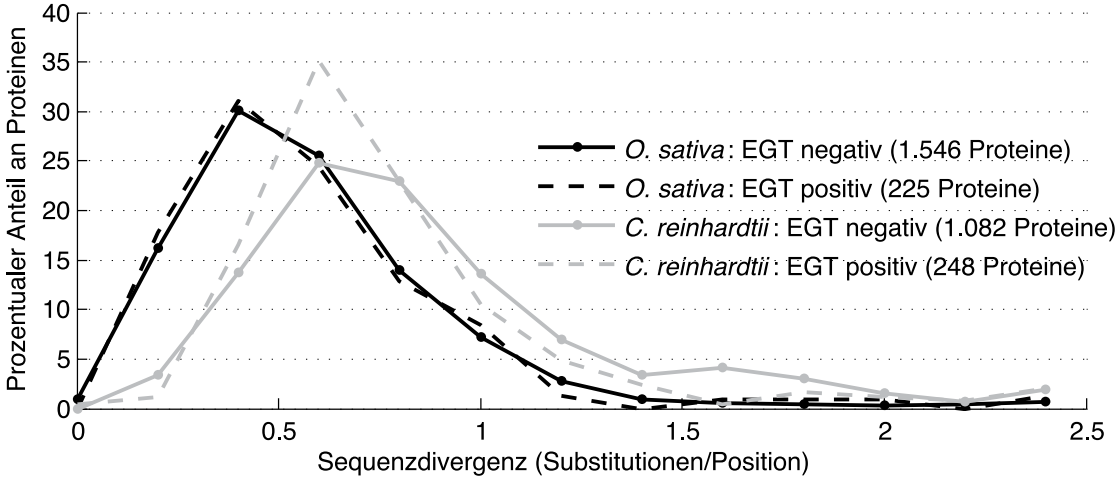

Abbildung 5.5: Divergenz von Proteinen in Abhängigkeit von deren evolutionärem Ursprung. Die Distanzen zwischen 1.771 orthologen Proteinen aus *Arabidopsis* und *Oryza* sowie 1.330 Orthologen aus *Arabidopsis* und *Chlamydomonas* ist gegeben durch JTT-Distanzen im globalen paarweisen Alignment (vgl. Abschnitt 4.5.6). Der Graph zeigt die Verteilungen nach evolutionärem Ursprung aufgeschlüsselt: Für Proteine, für die ein cyanobakterieller Ursprung abgeleitet wurde (EGT positiv), und für Proteine, für die ein solcher abgelehnt wurde (EGT negativ).

5.4 Aus divergenten Proteinfamilien werden unzuverlässige Alignments berechnet

Für konservierte Proteinfamilien wurde ein höherer Anteil cyanobakterieller Proteine abgeleitet (Abbildung 5.4). Proteine cyanobakteriellen Ursprungs zeigen jedoch keine höhere Sequenzkonservierung (Abbildung 5.5). Um diese Beobachtungen zu erklären, wurden die den phylogenetischen Bäumen zugrunde liegenden multiplen Sequenzalignments näher untersucht. Die HoT-Methode (Landan und Graur, 2007) vergleicht ein multiples Alignment mit einem alternativen reversen Alignment, das sich ergibt, wenn die entsprechenden Sequenzen in umgekehrter Orientierung in den Prozess des Alignierens eingebracht werden. Für das reverse Alignment werden Proteinsequenzen in C-N-Orientierung (Standard: N-C) und Nukleotidsequenzen in 3′-5′-Orientierung (Standard: 5′-3′) aligniert. Ein Vergleich dieser zwei alternativen Alignments derselben Sequenzen ermöglicht Aussagen über die Verlässlichkeit des multiplen Alignments (Abschnitt 4.3.14).

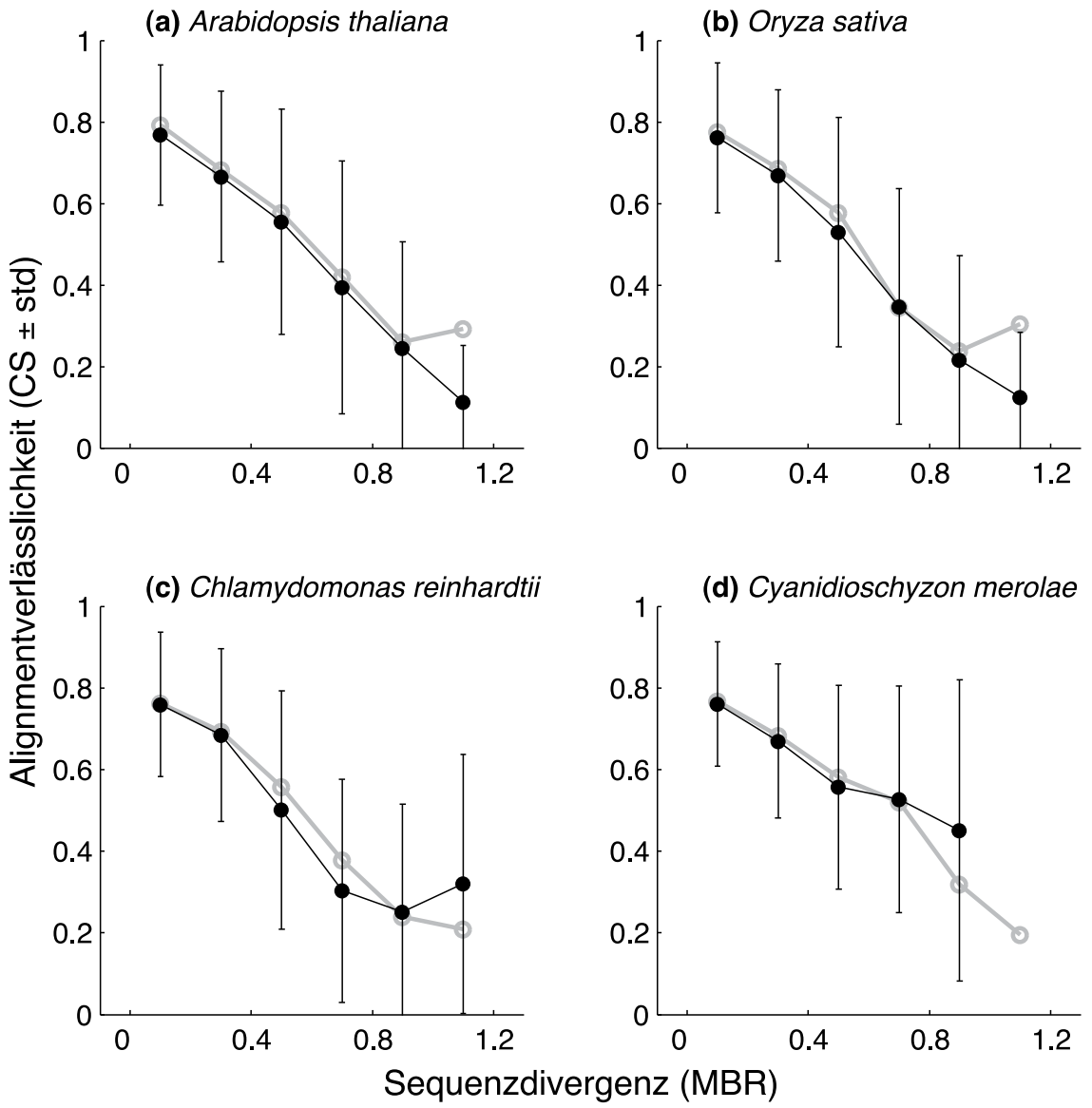

Abbildung 5.6: Alignmentverlässlichkeit in Abhängigkeit von der Sequenzdivergenz. Die Alignmentverlässlichkeit ist durch den CS-Wert (definiert in Abschnitt 4.6.2) gegeben, der den Anteil identischer Spalten zwischen Standardalignment und alternativem Alignment nach der HoT-Methode beschreibt (vgl. Anwendung der HoT-Methode in Abschnitt 4.5.5). Die Verteilung der Sequenzdivergenz ist durch den MBL-Wert gegeben, der die mittlere Astlänge in phylogenetischen Bäumen gemessen in Substitutionen pro Position beschreibt (definiert in Abschnitt 4.6.1). Schwarze Linien zeigen die Verteilung für ML-Bäume, graue Linien für NJ-Bäume. Schwarze Balken zeigen die Standardabweichung der Alignmentverlässlichkeit für ML-Bäume.

Für Alignments der am stärksten konservierten Proteinfamilien (mittlere Astlänge im Wahrscheinlichkeitsbaum $\leq$ 0,2 Substitionen pro Position) wurden zwischen 76 % und 77 % identische Spalten (CS-Wert, definiert in Abschnitt 4.6.2) im HoT-Vergleich gefunden (Abbildung 5.6). Mit abnehmender Sequenzkonservierung nahm der Anteil der identischen Spalten immer mehr ab. In den Alignments der divergentesten Proteinfamilien (mittlere Astlänge $\geq$ 1 Substition pro Position) wurden zwischen 11 % (*Arabidopsis*) und 45 % (*Cyanidioschyzon*) identische Spalten zwischen den alternativen Alignments beobachtet. Diese Beobachtungen wurden auch gemacht, wenn die Sequenzkonservierung aus Distanzbäumen berechnet wurde (Abbildung 5.6, graue Linie). Wurde die Verlässlichkeit über den Anteil identisch alignierter Aminosäuren (SPS, engl. *sum-of-pairs score*) berechnet, so wurde ebenfalls die Beobachtung gemacht, dass Alignments konservierter Proteinfamilien verlässlicher sind als die variablerer Proteinfamilien (Daten nicht gezeigt).

Die Verlässlichkeit wurde für das selbe Alignmentpaar über den Anteil identisch alignierter Aminosäuren (SPS-Wert, definiert in Abschnitt 4.6.3) fast immer höher eingeschätzt als über den Anteil identisch alignierter Spalten (CS-Wert, Abbildung 5.7). Diese Alignments sind durch Datenpunkte unterhalb der Diagonalen repräsentiert. Mindestens 80 % der Alignments hatten eine Verlässlichkeit von $\geq$ 0,8 SPS. Insbesondere unter den unzuverlässigsten Alignments (SPS $\leq$ 0,1) kamen solche vor, für welche die Verlässlichkeit über den CS- höher als über den SPS-Wert eingeschätzt wurde (Datenpunkte oberhalb der Diagonalen). In diesen Alignments kommen Bereiche vor, welche sehr lückenreich sind (Daten nicht gezeigt, vgl. Abschnitt 6.3).

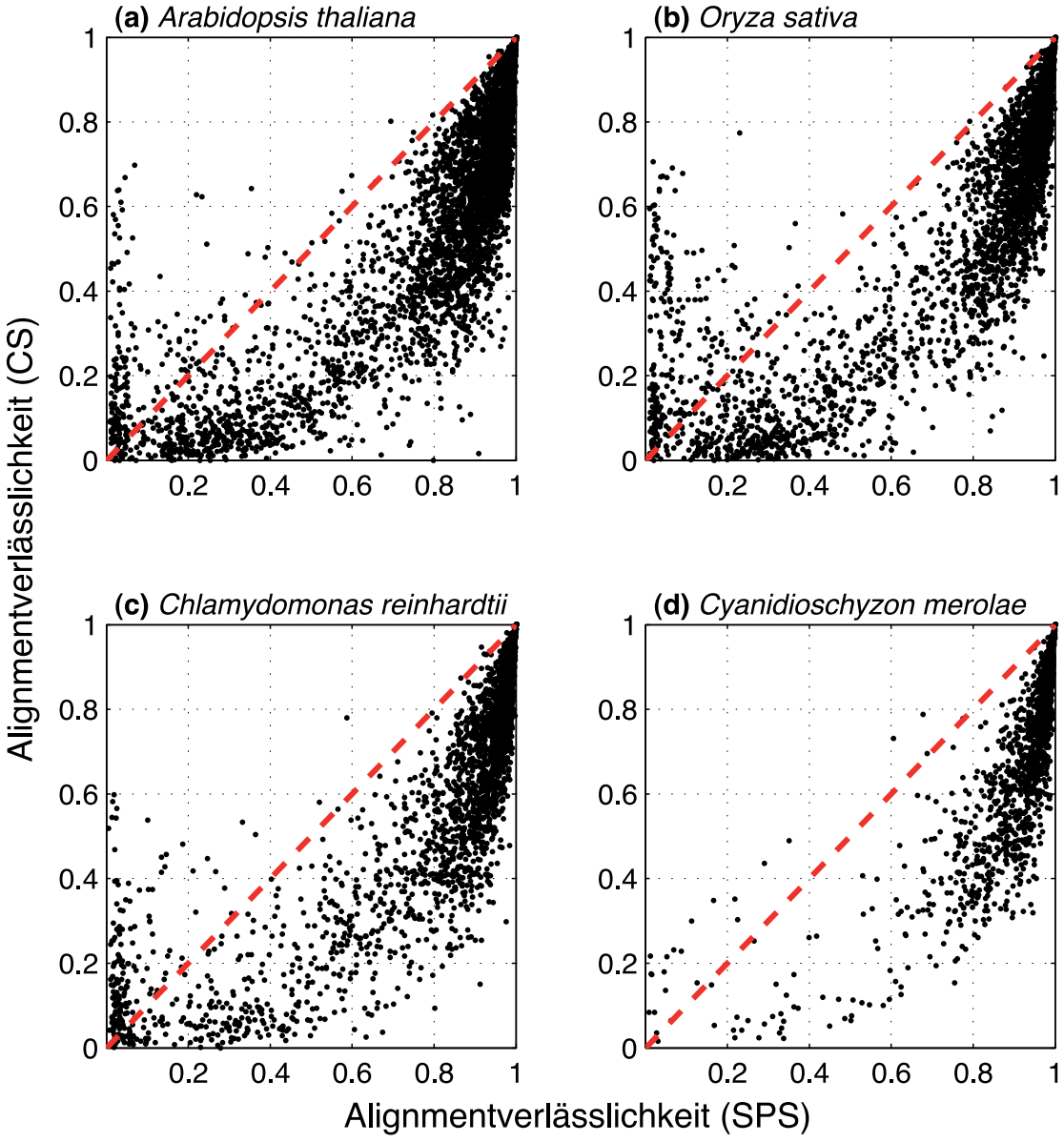

Abbildung 5.7: Beziehung zwischen den zwei Maßen zur Bewertung der Alignmentverlässlichkeit. Der SPS-Wert (Anteil identisch alignierter Aminosäurepaare, definiert in Abschnitt 4.6.3) war fast immer höher als der CS-Wert (Anteil identisch alignierter Spalten, definiert in Abschnitt 4.6.2).

5.5 Aus unzuverlässigen Alignments werden unzuverlässige Bäume abgeleitet

Die Alignments divergenter Proteinfamilien waren im HoT-Vergleich unzuverlässiger als die konservierter Proteinfamilien (Abschnitt 5.4). Aus den reversen Alignments wurden Distanz- und Wahrscheinlichkeitsbäume berechnet, um die Auswirkungen der Verlässlichkeit der Alignments auf die der phylogenetischen Bäume zu untersuchen (Abschnitt 5.2). Ein Vergleich der Topologien der alternativen phylogenetischen Bäume ermöglicht Aussagen über deren Verlässlichkeit (Abschnitt 4.3.14). Ein Maß hierfür ist der PPS-Wert (engl. *phylogenetic partitions score*, definiert in Abschnitt 4.6.4), der den Anteil an identisch rekonstruierten Bipartitionen (Äste) beschreibt (Abschnitt 4.5.5).

Tabelle 5.3: Korrelation der Verlässlichkeiten von Alignments und phylogenetischen Bäumen. Die Verlässlichkeit der Alignments ist gegeben durch den Anteil identischer Spalten (CS) sowie identischer Paare (SPS) zwischen normalem und reversen Alignment. Beide Maße für die Verlässlichkeit der Alignments korrelieren signifikant mit der Verlässlichkeit der Bäume (Pearson-Korrelation, vgl. Abschnitt 4.6.5, $p < 0{,}05$) gegeben durch den Anteil identischer Bipartitionen (PPS). Gezeigt sind die Korrelationen für Wahrscheinlichkeits- (ML) und Distanzbäume (NJ) für alle vier analysierten Genome.

	A. thaliana		*O. sativa*		*C. reinhardtii*		*C. merolae*	
	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)
CS	0,60	0,63	0,61	0,66	0,61	0,65	0,42	0,45
SPS	0,71	0,75	0,77	0,79	0,70	0,75	0,48	0,53

Wurde zur Bewertung der Alignments der Anteil identischer Spalten (CS-Wert) herangezogen, so lag der Korrelationskoeffizient zwischen Baum- und Alignmentverlässlichkeit zwischen 0,42 (*Cyanidioschyzon*) und 0,61 (*Oryza* und *Chlamydomonas*) in ML- und zwischen 0,45 (*Cyanidioschyzon*) und 0,66 (*Oryza*) in NJ-Bäumen (Tabelle 5.3). Wurde der Anteil identischer Aminosäurepaare (SPS-Wert) verwendet, so lag der Korrelationskoeffizient zwischen 0,48 (*Cyanidioschyzon*) und 0,77 (*Oryza sativa*) in ML- und zwischen 0,53 (*Cyanidioschyzon*) und 0,79 (*Oryza*) in NJ-Bäumen. Insgesamt war die Korrelation in einem Datensatz für NJ-Bäume immer höher als für ML-Bäume und für den SPS- immer höher als für den CS-Wert. Mit Ausnahme von *Cyanidioschyzon merolae* wurden gute Korrelationen beobachtet.

Das Punktdiagramm in Abbildung 5.8 zeigt den Zusammenhang von Alignment- und Baumverlässlichkeit für einzelne Alignments bzw. Bäume. Für die Mehrzahl an Datenpunkten wurde eine Abhängigkeit der Baum- von der Alignmentver-

lässlichkeit beobachtet. Jedoch gab es auch Fälle, in denen aus unzuverlässigen Alignments (z.B. CS $\leq$ 0,5) zuverlässige Bäume mit PPS $\geq$ 0,9 abgeleitet wurden (Abbildung 5.8). Diese Ausreisser konnten weder auf eine geringe Anzahl an Taxa, eine hohe Anzahl an Lücken, noch durch andere Gemeinsamkeiten zurückgeführt werden. Abbildung 5.8 zeigt nochmal auf, dass die Alignmentverlässlichkeit über den Anteil identischer Aminosäurepaare höher eingeschätzt wurde, als über den Anteil identischer Alignmentspalten (vgl. Abbildung 5.7). Die Diagramme für Wahrscheinlichkeitsbäume (Abbildung 5.8a–h) waren denen für Distanzbäume sehr ähnlich (Abbildung 5.8i–p).

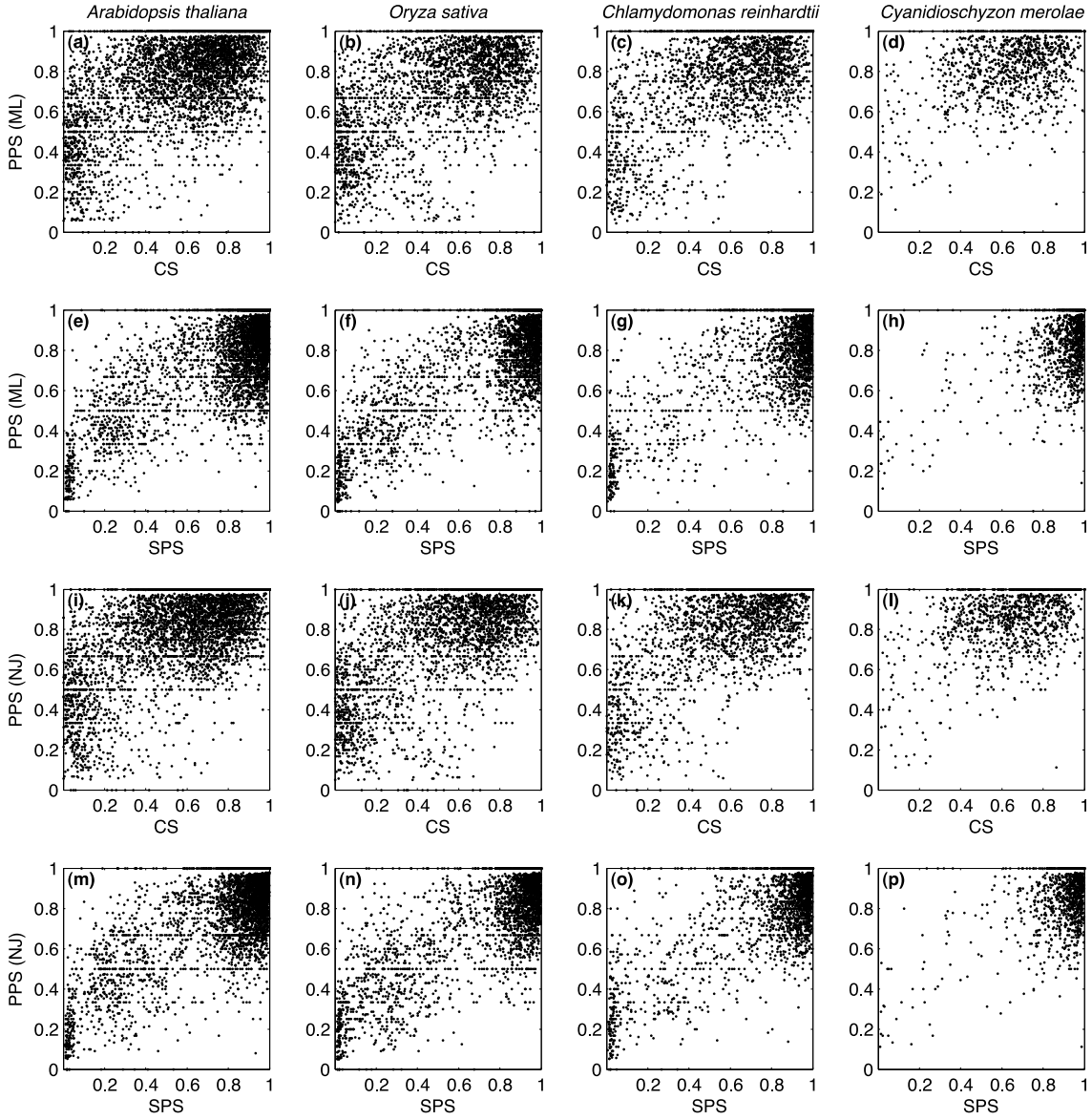

Abbildung 5.8: Abhängigkeit der Verlässlichkeit phylogenetischer Bäume von der Alignmentverlässlichkeit. Die Verlässlichkeit der Bäume ist durch den PPS-Wert (Anteil identischer Bipartitionen, definiert in Abschnitt 4.6.4), die der Alignments durch den CS- (Anteil identischer Spalten, definiert in Abschnitt 4.6.2, a–d und i–l) und den SPS-Wert (Anteil identischer Aminosäurepaare, definiert in Abschnitt 4.6.3, e–h und m–p) gegeben (vgl. Anwendung der HoT-Methode in Abschnitt 4.5.5). Die Abhängigkeiten sind für Wahrscheinlichkeits- (ML, a–h) und Distanzbäume (NJ, i–p) gezeigt.

5.6 Verlässlichere Alignments ergeben höhere Schätzer für den Anteil cyanobakterieller Gene

Bislang wurde gezeigt, dass die Sequenzkonservierung die Alignment- (Abschnitt 5.4) und damit auch die Baumverlässlichkeit beeinflusst (Abschnitt 5.5). Im Folgenden wurde untersucht, wie der prozentuale Anteil abgeleiteter cyanobakterieller Gene von der Verlässlichkeit der Daten abhängt. Abbildung 5.9 zeigt den abgeleiteten Gentransfer in Abhängigkeit von den drei Parametern CS, SPS, und PPS zur Quantifizierung der Qualität von Alignments bzw. phylogenetischen Bäumen. Darüberhinaus zeigt ein Histogramm wie das entsprechende Maß die Daten aufgeteilt hat.

Wurde die Anzahl identischer Spalten der alternativen Alignments (CS-Wert) zur Bewertung der Alignmentverlässlichkeit herangezogen, so war diese für alle vier analysierten Genome annähernd normalverteilt (Abbildung 5.9a–d). Das Intervall mit den meisten Elementen lag um 0,75. Der abgeleitete Anteil cyanobakterieller Gene war mit ca. 5 % im Intervall der geringsten Verlässlichkeit am niedrigsten und im Intervall der höchsten Verlässlichkeit mit ca. 20 % am höchsten. Dazwischen stieg der prozentuale Anteil cyanobakterieller Gene mit einem lokalen Maximum um 0,4 an.

Die Alignmentqualität zeigte eine andere Verteilung, wenn der Anteil identisch alignierter Aminosäurepaare (SPS-Wert) als Kriterium herangezogen wurde (Abbildung 5.9i–l). Die meisten Elemente lagen im Intervall der höchsten und der zweithöchsten Verlässlichkeit. Die Elemente in den restlichen Intervallen waren annähernd gleichverteilt. Auch hier wurde mit ca. 15 % der größte Anteil cyanobakterieller Gene im Intervall der größten Verlässlichkeit gemessen. Der prozentuale Anteil abgeleiteter cyanobakterieller Gene nahm tendenziell in den Intervallen geringerer Verlässlichkeit ab. Der Trend war aufgrund der geringen Anzahl an Elementen jedoch nicht so deutlich wie bei der Anwendung des CS-Kriteriums. In einer Darstellungsweise mit einer geringeren Anzahl an Intervallen wurde dieser Trend jedoch beobachtet (Daten nicht gezeigt).

Bei Verwendung des Anteils identischer Bipartitionen der alternativen Bäume (PPS-Wert) zur Bewertung der Baumverlässlichkeit lagen die meisten Elemente im Intervall der höchsten Verlässlichkeit und nahmen in den Intervallen geringerer Verlässlichkeit stetig ab (Abbildung 5.9e–h). Mit ca. 15 % war der Anteil abgeleiteter cyanobakterieller Gene im Intervall der größten Verlässlichkeit am höchsten und nahm in den Intervallen geringerer Baumverlässlichkeit ab. Der Trend war nicht so

deutlich wie bei der Verwendung des CS-, jedoch besser als bei der Verwendung des SPS-Wertes.

Die bislang beschriebene Abhängigkeit des endosymbiontischen Gentransfers bezog sich auf die Vorhersage cyanobakterieller Gene aus phylogenetischen Bäumen, die mit der Methode der maximalen Wahrscheinlichkeit abgeleitet wurden (schwarze Linie in Abbildung 5.9). Die Vorhersage cyanobakterieller Gene anhand von Distanzbäumen war auf sehr ähnliche Weise von der Verlässlichkeit der Alignments und Stammbäume abhängig (graue Linie in Abbildung 5.9). In manchen Fällen waren die Ergebnisse derart ähnlich, dass die Graphen für die Ergebnisse der beiden Methoden kaum zu unterscheiden waren (vgl. Abbildung 5.9i).

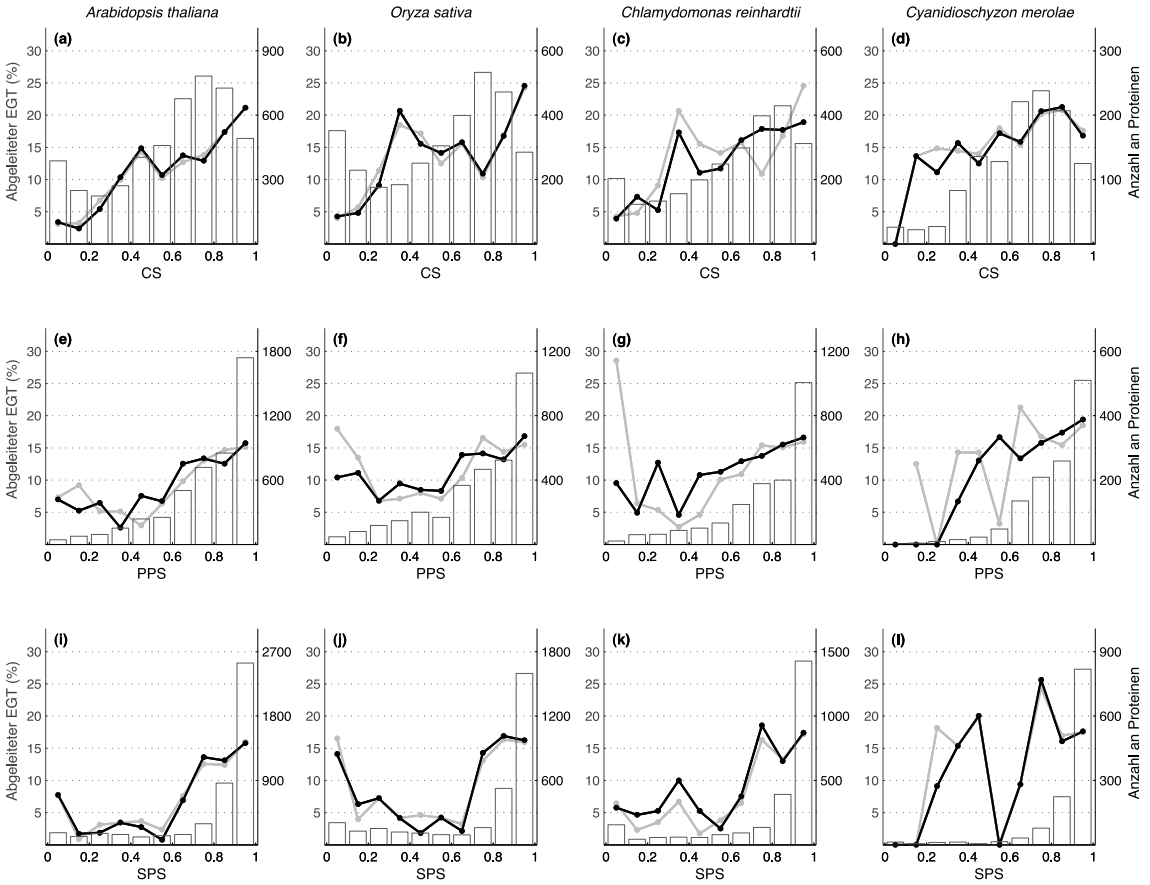

Abbildung 5.9: Abhängigkeit des abgeleiteten Anteils cyanobakterieller Gene von der Verlässlichkeit der Daten. Die Verteilung der Verlässlichkeit (Histogramm im Hintergrund) ist gegeben durch drei Einheiten, die sich aus dem Vergleich eines Alignments mit dem zugehörigen reversen Alignment ergeben (HoT-Methode (Landan und Graur, 2007)): (a–d) Anteil identischer Spalten der multiplen Sequenzalignments (engl. *column score*, CS (Thompson et al., 1999*a*), definiert in Abschnitt 4.6.2). (e–h) Anteil identischer Bipartitionen der abgeleiteten phylogenetischen Bäume (engl. *phylogenetic partitions score*, PPS, definiert in Abschnitt 4.6.4). (i–l) Anteil identischer Paare der multiplen Sequenzalignments (engl. *sum of pairs score*, SPS (Thompson et al., 1999*a*), definiert in Abschnitt 4.6.3). Die Verteilung der identischen Bipartitionen wurde aus ML-Bäumen berechnet. Eine ähnliche Verteilung wurde für NJ-Bäume beobachtet (Daten nicht gezeigt). Die schwarzen Linien zeigen die Ableitung des cyanobakteriellen Ursprungs aus ML-, die grauen aus NJ-Bäumen.

5.7 Ähnlichkeit cyanobakterieller Pflanzengene zu neun rezenten Cyanobakterien

Bei der Auswertung von 11.569 phylogenetischen Bäumen von Pflanzenproteinen und ihren Homologen in 237 Referenzspezies wurde für 1.587 Proteine (592 für *Arabidopsis*, 432 für *Oryza*, 356 für *Chlamydomonas* und 207 für *Cyanidioschyzon*) ein cyanobakterieller Ursprung abgeleitet (Abschnitt 5.2). Weitere 987 Proteine (392 für *Arabidopsis*, 255 für *Oryza*, 206 für *Chlamydomonas* und 134 für *Cyanidioschyzon*) wurden als cyanobakteriell klassifiziert, weil sie nur in Cyanobakterien Homologe aufwiesen (Abschnitt 5.1). Anhand dieser 2.574 Proteine wurde die Fragestellung untersucht, welches rezente Cyanobakterium eine Kollektion an Genen besitzt, die diesen Pflanzenproteinen cyanobakteriellen Ursprungs am ähnlichsten ist.

Abbildung 5.10 zeigt die Aminosäureidentität dieser Proteine im lokalen BLAST-Alignment (HSP-Alignment, vgl. Abschnitt 4.3.4) zu ihren Homologen aus den Genomen neun rezenter Cyanobakterien anhand einer Farbkodierung. Pflanzenproteine wurden nach absteigender prozentualer Identität zu ihren Homologen sortiert. Cyanobakterielle Referenzspezies wurden nach absteigender Gesamt-Aminosäureidentität sortiert. *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 waren unter den neun analysierten Cyanobakterien diejenigen mit der größten Gesamt-Aminosäureidentität. Diese Spezies haben mit 5.657 bzw. 6.130 Genen unter den neun analysierten Cyanobakterien die größten Genome (Abbildung 5.11c). *Prochlorococcus marinus* MIT 9313 zeigte die geringste Gesamt-Aminosäureidentität und hat mit 2.265 Genen das kleinste Genom. *Gloeobacter violaceus* PCC 7421 hat mit 4.430 Genen das drittgrößte Genom und zeigte eine unterdurchschnittliche Aminosäureidentität. Die Analyse zeigte bei *Oryza sativa* vergleichsweise viele Proteine mit sehr hoher Ähnlichkeit zu den Homologen aus Cyanobakterien (Aminosäureidentität $\geq 80\,\%$, dunkelrote Felder). 51 dieser Gene wurden als in den Kern transferierte Plastiden-DNA (NUPTs) charakterisiert (Abschnitt 4.5.7).

Anabaena und *Nostoc* waren unter den neun analysierten Cyanobakterien in den Stammbäumen am häufigsten in der Nachbargruppe des Pflanzenhomologs vertreten, wenn ein cyanobakterieller Ursprung desselben abgeleitet wurde (Abbildung 5.11a). Diese Beoachtung wurde auch gemacht, wenn nur Teilmengen der Daten mit höherer Verlässlichkeit (Alignments mit SPS $\geq 0{,}8$ oder SPS $\geq 0{,}9$) betrachtet wurden. Auch für die Pflanzenproteine, die bereits bei der Homologiesuche als cyanobakteriell klassifiziert wurden, traten *Anabaena* und *Nostoc* am häufigsten auf (Abbildung 5.11b). Diese Spezies haben unter den neun analysier-

ten Cyanobakterien die größten Genome (Abbildung 5.11c). *Gloeobacter violaceus* PCC 7421 ist das Cyanobakterium mit dem drittgrößten Genom und zeigte beim Vergleich der Sequenzähnlichkeit zu 2.574 Pflanzenproteinen cyanobakteriellen Ursprungs die sechsthöchste Gesamtähnlichkeit (Abbildung 5.10). In den phylogenetischen Bäumen war *Gloeobacter* unterdurchschnittlich (Abbildung 5.11a) häufig repräsentiert und kam bei der Homologiesuche durschschnittlich häufig vor (Abbildung 5.11b). *Prochlorococcus marinus* MIT 9313 hat das kleinste Genom unter den neun Referenzgenomen und zeigte die niedrigste Gesamtidentität zu den cyanobakteriellen Pflanzenproteinen (Abbildung 5.10a). In den phylogenetischen Bäumen, aus denen ein cyanobakterieller Ursprung des Pflanzenhomologs abgeleitet wurde, war *Prochlorococcus* unterdurschnittlich oft vertreten (Abbildung 5.11a). In der Homologiesuche kam *Prochlorococcus* am seltensten vor.

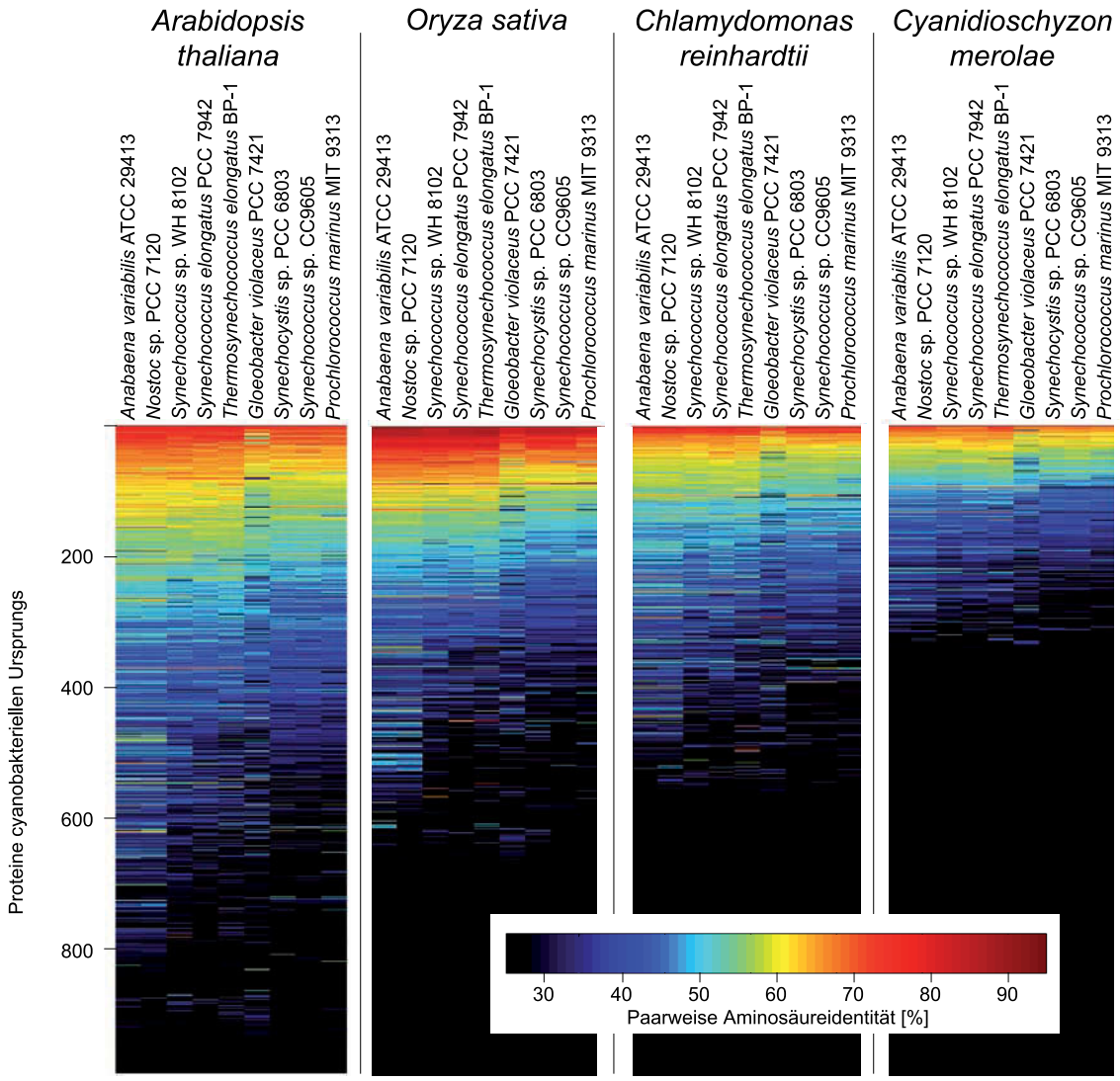

Abbildung 5.10: Ähnlichkeit der 2.574 als cyanobakteriell klassifizierten kernkodierten Proteine aus Pflanzen und Algen zu deren Homologen aus neun cyanobakteriellen Referenzgenomen. Zeilen entsprechen den kernkodierten Proteinen für die ein cyanobakterieller Ursprung abgeleitet wurde. Spalten entsprechen den cyanobakteriellen Genomen. Die Elemente der Matrix geben die prozentuale Aminosäureidentität der Proteine zu ihren cyanobakteriellen Homologen anhand einer Farbkodierung wieder. Zeilen (cyanobakterielle Pflanzenproteine) sind nach absteigender Ähnlichkeit, Spalten (Referenzgenome) nach absteigender Gesamtähnlichkeit sortiert.

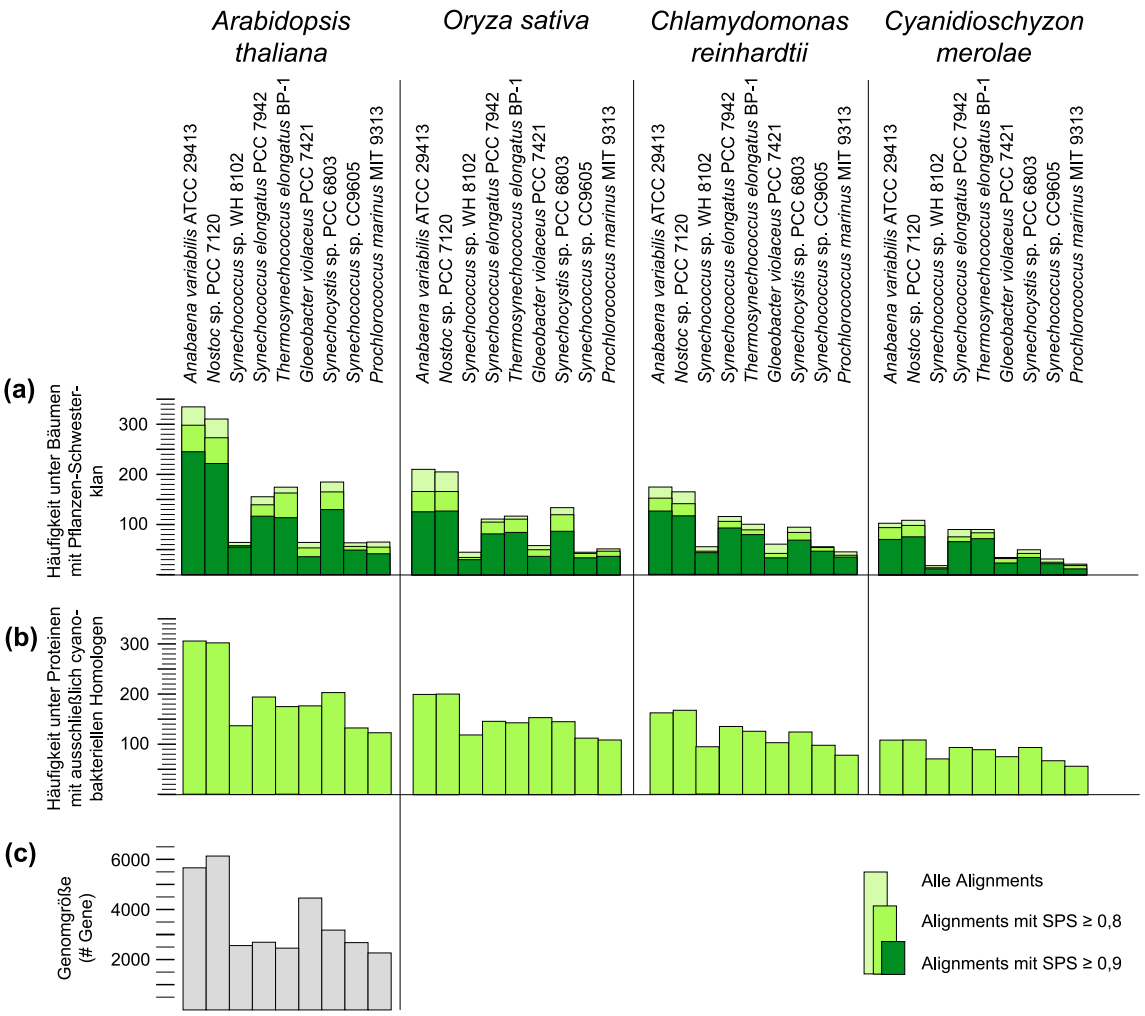

Abbildung 5.11: Häufigkeiten der neun cyanobakteriellen Referenzspezies: (a) in den Wahrscheinlichkeitsbäumen, aus denen ein cyanobakterieller Ursprung des kernkodierten Pflanzenproteins gefolgert wurde und (b) in den Ergebnissen der Homologiesuche für die Proteine, für die ausschließlich cyanobakterielle Homologe gefunden wurden. (c) Anzahl der in den neun cyanobakteriellen Genomen kodierten Proteine.

5.8 Vorhersage des Zielkompartiments cyanobakterieller Proteine

Diejenigen Pflanzenproteine, für die phylogenetische Bäume abgeleitet wurden, wurden einer Vorhersage der zellulären Lokalisierung unterzogen. Für 4.670 (*Arabidopsis*), 3.186 (*Oryza*), 2.500 (*Chlamydomonas*) beziehungsweise 1.213 (*Cyanidioschyzon*) Proteine wurde das Zielkompartiment mit dem Programm TargetP (Abschnitt 4.4) vorhergesagt und für Proteine cyanobakteriellen und nichtcyanobakteriellen Ursprungs getrennt dargestellt.

Es wurde für 351 Proteine cyanobakteriellen Ursprungs aus *Arabidopsis thaliana* eine Lokalisierung in Plastiden vorhergesagt (Abbildung 5.12a). In 67 % der Fälle lagen die Vorhersagen in den zwei Intervallen der höchsten Verlässlichkeiten. Für 62 Proteine wurden Mitochondrien (77 % der Daten in den zwei Intervallen der niedrigsten Verlässlichkeiten) und für 51 Proteine der Sekretionsweg abgeleitet (69 % der Vorhersagen in den beiden Intervallen der höchsten Verlässlichkeiten). 128 Proteine wiesen nur geringe Übereinstimmungen mit Signalpeptiden auf und das Cytosol wurde als Zielkompartiment vorhergesagt. 63 % der Vorhersagen lagen im Intervall der zweithöchsten und der mittleren Verlässlichkeit. Für 703 Proteine aus *Arabidopsis*, für die ein cyanobakterieller Ursprung abgelehnt worden war, wurde eine Lokalisierung in Plastiden vorhergesagt (Abbildung 5.12a). Die Verlässlichkeit der Vorhersagen war über alle Intervalle annähernd gleichverteilt. Für 532 Proteine wurden Mitochondrien (55% der Daten in den zwei Intervallen der niedrigsten Verlässlichkeiten), für 845 Proteine der Sekretionsweg (61 % der Vorhersagen in den zwei Intervallen der niedrigsten Verlässlichkeiten) und für 1.998 Proteine das Cytosol als Zielkompartiment vorhergesagt (63 % der Daten im Intervall der zweithöchsten und der mittleren Verlässlichkeit).

Für 165 Proteine cyanobakteriellen Ursprungs aus *Oryza sativa* wurde eine Lokalisierung in Plastiden vorhergesagt (Abbildung 5.12b). Für 61 Proteine wurden Mitochondrien, für 41 Proteine der Sekretionsweg und für 165 Proteine wurde das Cytosol als Zielkompartiment vorhergesagt. Für 422 Proteine aus *Oryza*, für die ein cyanobakterieller Ursprung abgelehnt worden war, wurde eine Lokalisierung in Plastiden vorhergesagt (Abbildung 5.12b). Für 509 Proteine wurden Mitochondrien, für 597 Proteine der Sekretionsweg und für 1.226 Proteine das Cytosol als Zielkompartiment vorhergesagt. Die Verteilungen der Verlässlichkeiten der Vorhersagen für *Oryza* war denen von *Arabidopsis* ähnlich. Für Proteine cyanobakteriellen Ursprungs, für die Chloroplasten als Zielkompartiment vorhergesagt wurden, lagen verglichen mit *Arabidopsis* weniger Vorhersagen in dem Intervall

der höchsten Verlässlichkeit. Die Ergebnisse für die Algen *Chlamydomonas* und *Cyanidioschyzon* unterschieden sich stark von denen der Pflanzen (Abbildung 5.13a+b). Für Proteine, für die ein cyanobakterieller Ursprung vorhergesagt worden war, wurde in etwa zu gleichen Teilen eine plastidäre und mitochondriale Lokalisierung vorhergesagt. Insbesondere waren die Verlässlichkeiten für die Algen niedriger als für die Pflanzen.

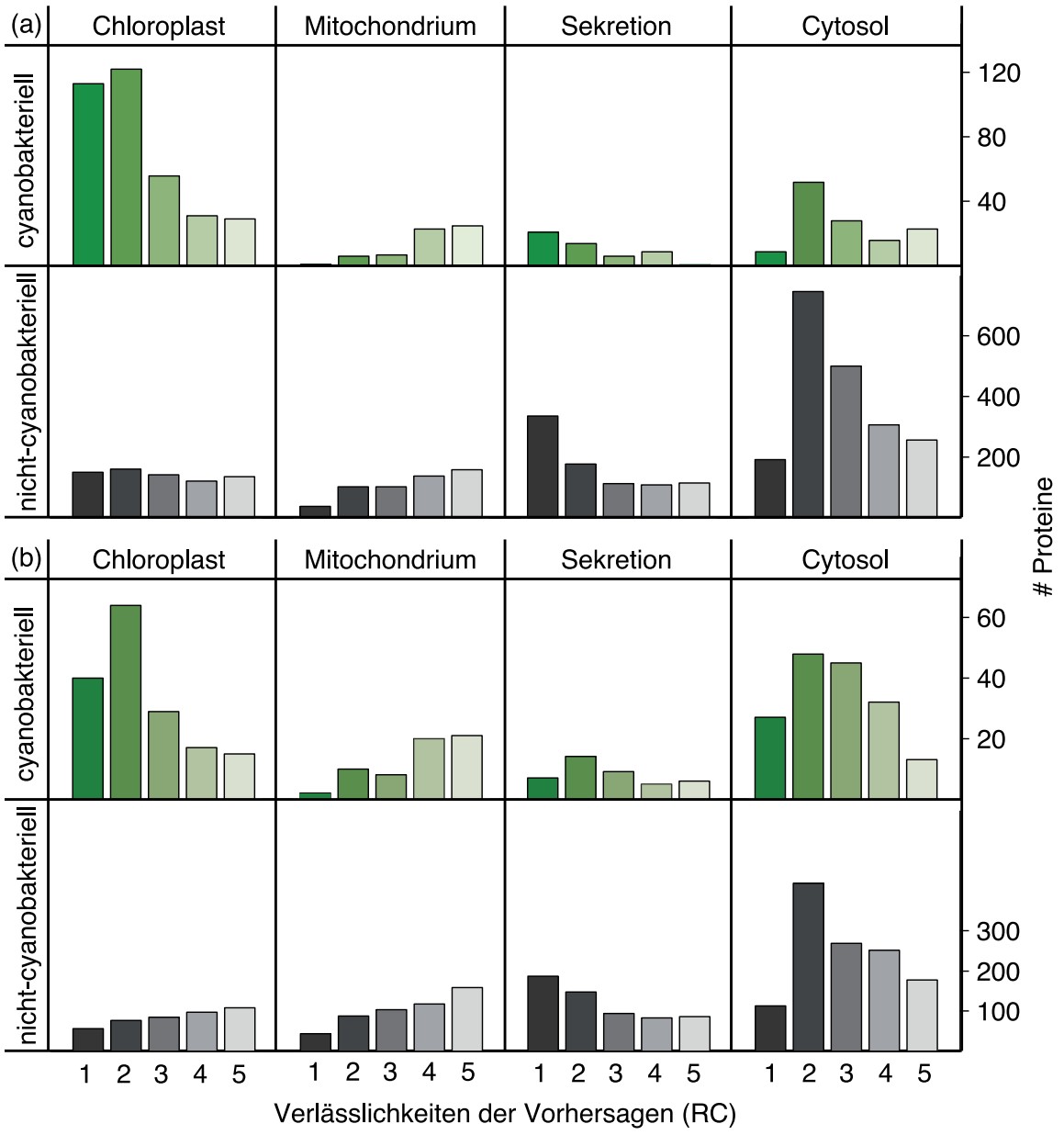

Abbildung 5.12: Vorhersage des Zielkompartiments für 4.670 Proteine aus *Arabidopsis thaliana* (a) und 3.187 Proteine aus *Oryza sativa* (b) mittels TargetP (Abschnitt 4.4). Die Darstellung erfolgt getrennt für 592 (*Arabidopsis*) bzw. 432 (*Oryza*) Proteine, für die aus ML-Bäumen ein cyanobakterieller Ursprung abgeleitet wurde (oben), und für 4.078 (*Arabidopsis*) bzw. 2.755 (*Oryza*) Proteine, für die ein solcher abgelehnt wurde (unten). Die Vorhersagen sind für jedes der Kompartimente in Verlässlichkeitsklassen eingeteilt ($RC = 1$ entspricht der höchsten und $RC = 5$ der niedrigsten Verlässlichkeit, definiert in Abschnitt 4.4).

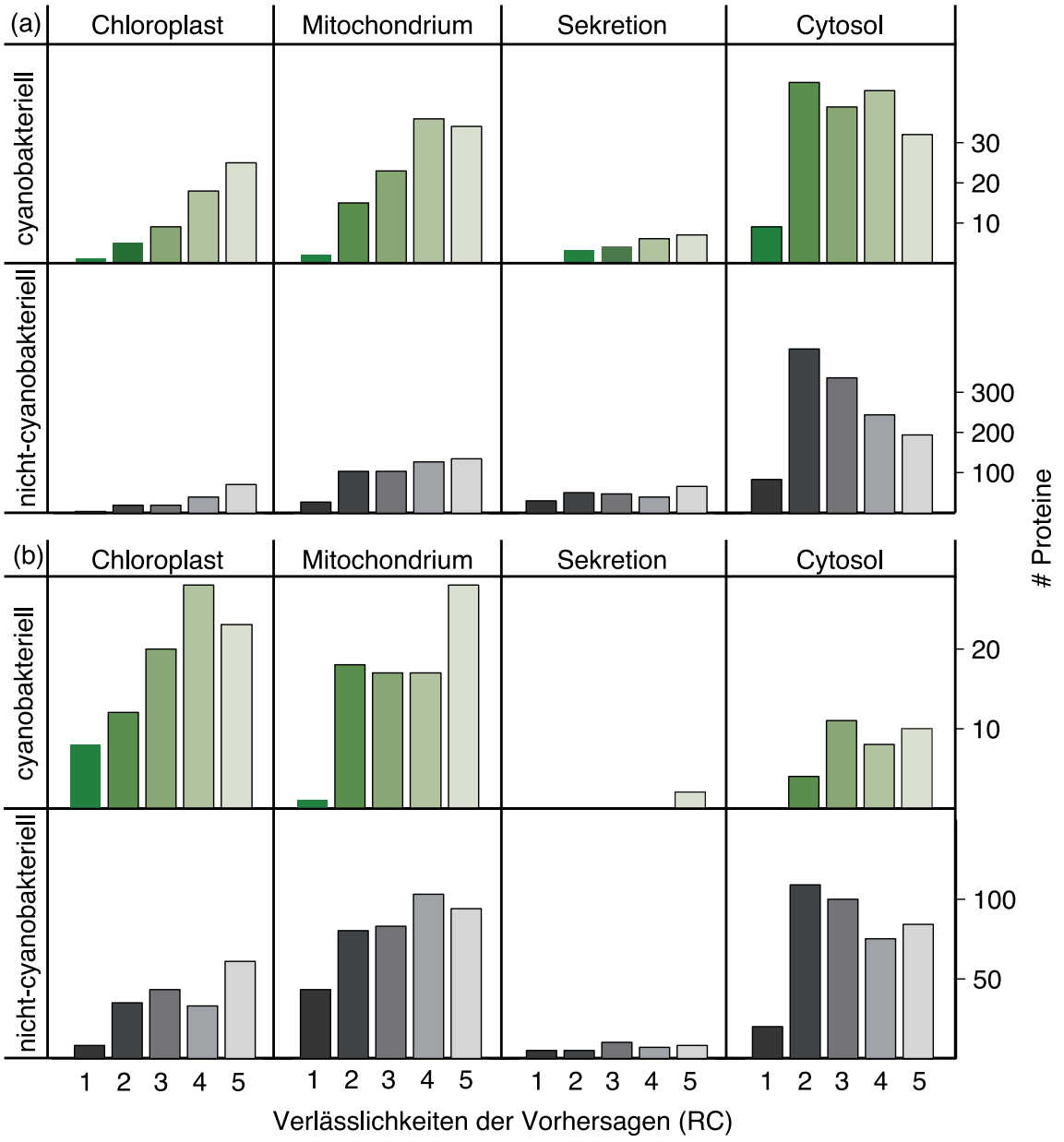

Abbildung 5.13: Vorhersage des Zielkompartiments für 2.500 Proteine aus *Chlamydomonas reinhardtii* (a) und 1.215 Proteine aus *Cyanidioschyzon merolae* (b) mittels TargetP (Abschnitt 4.4). Die Darstellung erfolgt getrennt für 356 (*Chlamydomonas*) bzw. 207 (*Cyanidioschyzon*) Proteine, für die aus ML-Bäumen ein cyanobakterieller Ursprung abgeleitet wurde (oben), und für 2.144 (*Chlamydomonas*) bzw. 1.008 (*Cyanidioschyzon*) Proteine, für die ein solcher abgelehnt wurde (unten). Die Vorhersagen sind für jedes der Kompartimente in Verlässlichkeitsklassen eingeteilt ($RC = 1$ entspricht der höchsten und $RC = 5$ der niedrigsten Verlässlichkeit, definiert in Abschnitt 4.4).

Tabelle 5.4: Zusammenfassung der funktionellen Charakterisierung cyanobakterieller Pflanzenproteine. Für jeden Organismus ist die Anzahl an Proteinen angegeben, für die ein cyanobakterieller Ursprung abgeleitet worden war (vgl. Abschnitt 5.1 und Abschnitt 5.2), sowie in wievielen Proteinfamilien der KEGG-Datenbank Homologe dazu gefunden wurden (vgl. Abschnitt 4.5.8). In Klammern ist die Anzahl an Proteinen angegeben, die in mehr als einem Cluster Homologe aufwies. Desweiteren enthält die Tabelle Informationen darüber, wievielen Proteinen eine enzymatische Funktion in Form einer EC-Nummer (engl. *enzyme commission*) zugeordnet werden konnte. In der letzten Spalte ist die bereinigte Anzahl an Enzymen angegeben. Aus diesen Listen wurden alle Enzyme entfernt, die in den Stoffwechselwegen der KEGG-Datenbank für den entsprechenden Organismus nicht gelistet sind.

Organismus	Proteine	Familien	Enzyme	Enzyme *
A. thaliana	984	685 (309)	313	156
O. sativa	687	476 (230)	258	138
C. reinhardtii	562	396 (173)	291	114
C. merolae	341	235 (81)	175	92

5.9 Funktionelle Charakterisierung cyanobakterieller Pflanzengene

Um den cyanobakteriellen Pflanzenproteinen eine Funktion zuzuordnen, wurden mit ihnen Datenbanksuchen mit BLASTP gegen die KEGG-Datenbank durchgeführt (Abschnitt 4.5.8). Einige der 3.305.720 Proteine dieser Datenbank sind in 10.990 Proteinfamilien unterteilt, von denen 4.556 Familien eine Funktion in Form einer EC-Nummer (engl. *enzyme commission*) zugeordnet ist, die für ein bestimmtes Enzym steht. 78 Familien sind zwei, einer sind drei und einer weiteren sind vier EC-Nummern zugewiesen.

Für die 984 (*Arabidopsis*), 687 (*Oryza*), 562 (*Chlamydomonas*) beziehungsweise 341 (*Cyanidioschyzon*) Proteine cyanobakteriellen Ursprungs wurden Homologe in 685, 476, 396 beziehungsweise 235 Proteinfamilien gefunden (Tabelle 5.4). 309, 230, 173 beziehungsweise 81 Proteine wiesen Homologe in mehreren Familien auf. Auf diese Weise wurden 313, 258, 291 beziehungsweise 175 enzymatische Funktionen ermittelt. Die Listen der EC-Nummern wurden mit den Stoffwechselwegen der KEGG-Datenbank für *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* abgeglichen und diejenigen Enzyme entfernt, über die der jeweilige Organismus nicht verfügt. Die so bereinigten Listen ergaben 156, 138, 114 beziehungsweise 92 Enzyme mit abgeleitetem cyanobakteriellem Ursprung.

Die 156, 138, 114 beziehungsweise 92 Enzyme cyanobakteriellen Ursprungs wurden den Kategorien von Stoffwechselwegen zugeordnet, in denen sie Funktionen

Tabelle 5.5: Einteilung der 156, 138, 114 beziehungsweise 92 cyanobakteriellen Enzyme aus *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* (vgl. Tabelle 5.4) in Stoffwechselwege. Die Einteilung entspricht der der Übersichtskarten der KEGG-Datenkank (vgl. Abschnitt 4.5.8 und Abbildung 6.1). Unter weitere Stoffwechselwege sind die Enzyme zusammengefasst, die nicht in den neun angegebenen Stoffwechselwegen vorkommen. Da ein Enzym in mehreren Stoffwechselwegen vorkommen kann, sind die Spaltensummen größer als die Anzahlen cyanobakterieller Enzyme.

Stoffwechselwege	*A. thaliana*	*O. sativa*	*C. reinhardtii*	*C. merolae*
Aminosäuresynthese	59	54	52	46
Kohlenhydratstoffwechsel	39	41	33	25
Biosynthese von Sekundärmetaboliten	31	26	12	12
Lipidstoffwechsel	28	26	20	14
Energiestoffwechsel	21	22	22	14
Synthese von Vitaminen und Kofaktoren	20	17	18	20
Biosynthese weiterer Aminosäuren	10	8	5	4
Nucleotidstoffwechsel	9	5	10	4
Biosynthese von Glykanen	5	3	7	2
Weitere Stoffwechselwege	35	36	34	25

ausüben (Tabelle 5.5). Die neun Kategorien entsprechen denen der Übersichtskarten der Stoffwechselwege (vgl. Abbildung 6.1). Die meisten cyanobakteriellen Enzyme haben Funktionen in der Aminosäuresynthese, dem Kohlenhydratstoffwechsel und der Biosynthese von Sekundärmetaboliten. Aufgrund der Tatsache, dass ein Enzym in mehreren Kategorien von Stoffwechselwegen vorkommen kann, wurden in der Summe mehr Kategorien von Stoffwechselwegen als Enzyme gefunden.

Die cyanobakteriellen Enzyme des Calvin-Zyklus von *Arabidopsis thaliana* wurde mit der Funktion *search objects in pathways* der KEGG-Internetseite dargestellt. Für die sechs Enzyme Transketolase, Glycerinaldehyd-3-Phosphat Dehydrogenase (EC 1.2.1.13), 3-Phosphoglycerat Kinase, Ribulose-1,5-Bisphosphat Carboxylase/Oxygenase, Ribulose-5-Phosphat 3-Epimerase und Phosphoribulokinase wurde ein cyanobakterieller Ursprung abgeleitet (Abbildung 5.14). Für die cytosolische Form der Glycerinaldehyd-3-Phosphat Dehydrogenase (EC 1.2.1.12), die ihre Funktion in der Glykolyse ausübt, wurde ebenfalls ein cyanobakterieller Ursprung des Gens abgeleitet (nicht auf Abbildung 5.14 gezeigt). Ein cyanobakterieller Ursprung der fünf Enzyme Fruktose-1,6-Bisphosphatase, Fruktose-1,6-Bisphosphat Aldolase, Triosephosphat Isomerase, Ribose-5-Phosphat Isomerase und Sedoheptulose-1,7-Bisphosphatase wurde verworfen. *Arabidopsis* verfügt nicht über die Gene für die vier Enzyme Fruktose-6-Phosphat Phosphoketolase (EC 4.1.2.22), eine dritte Isoform der Glycerinaldehyd-3-Phosphate Dehydrogenase (EC 1.2.1.59), Phosphoketolase (EC 4.1.2.9) und Sedoheptulokinase (EC 2.7.1.14).

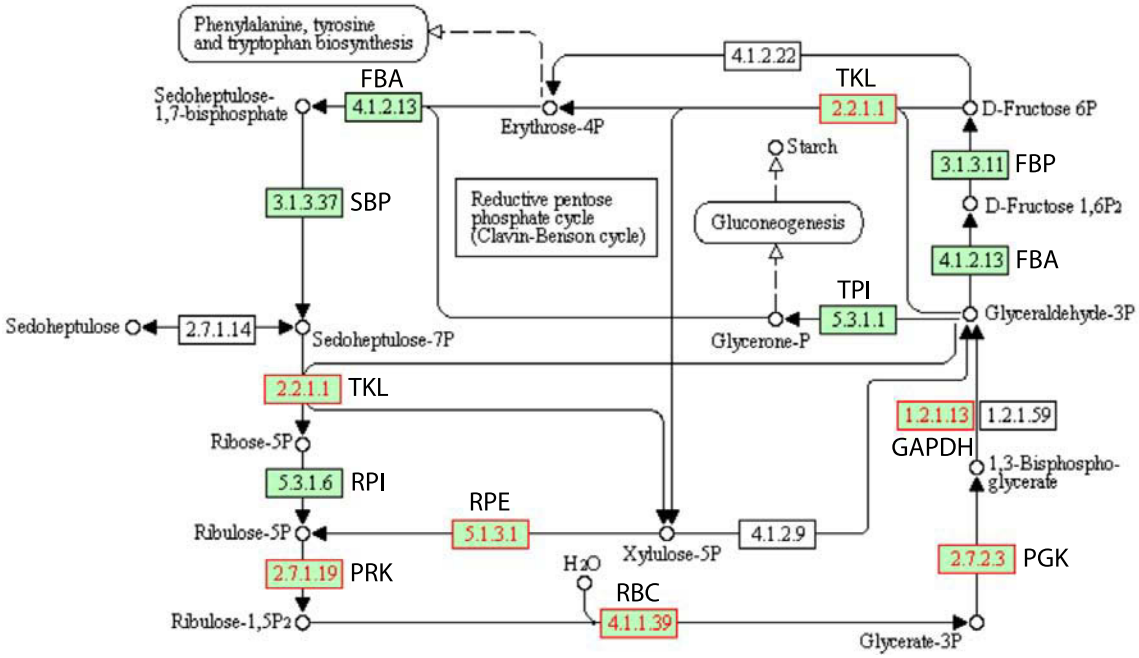

Abbildung 5.14: Evolutionärer Ursprung der Enzyme des Calvin-Zyklus in *Arabidopsis thaliana*. Gezeigt ist ein Ausschnitt aus der Originaldarstellung des Stoffwechselwegs „Kohlenstofffixierung in photosynthetischen Organismen“ (engl. *carbon fixation in photosynthetic organisms*) der KEGG-Internetseite (vgl. Abschnitt 4.5.8). Die EC-Nummern von Enzymen, über die *Arabidopsis* verfügt, sind grün hinterlegt. Enzyme, für die in dieser Arbeit ein cyanobakterieller Ursprung abgeleitet worden war (vgl. Abschnitt 5.1 und Abschnitt 5.2), sind rot markiert. Die Abkürzungen der Enzymnamen wurden nachträglich in die Darstellung eingefügt: FBA Fruktose-1,6-Bisphosphat Aldolase; FBP Fruktose-1,6-Bisphosphatase; GAPDH Glycerinaldehyd-3-Phosphat Dehydrogenase; PGK 3-Phosphoglycerat Kinase; PRI Ribose-5-Phosphat Isomerase; PRK Phosphoribulokinase; RBC Ribulose-1,5-Bisphosphat Carboxylase/Oxygenase; RPE Ribulose-5-Phosphat 3-Epimerase; SBP Sedoheptulose-1,7-Bisphosphatase; TKL Transketolase; TPI Triosephosphat Isomerase; PT Phosphat Translokator.

5.10 Analyse der Introns cyanobakterieller und alter eukaryotischer Gene

In einer Zusammenarbeit mit der Arbeitsgruppe von Eugene V. Koonin vom National Center for Biotechnology Information (NCBI[1]) wurden Unterschiede in den Introns cyanobakterieller und alter eukaryotischer Pflanzengene untersucht. Die im Folgenden beschriebenen Ergebnisse sind ein Auszug aus der Veröffentlichung, die im Rahmen dieser Zusammenarbeit entstanden ist (Basu et al., 2008).

1 http://www.ncbi.nlm.nih.gov/CBBresearch/Koonin/

Proteinkodierende Gene aus *Arabidopsis thaliana* sowie deren Orthologe aus *Oryza sativa* und *Populus trichocarpa* wurden in einen Datensatz cyanobakterieller und alter eukaryotischer Gene (Gene mit Homologen ausschließlich in Pilzen und Tieren) unterteilt. Für alle drei Pflanzen wurden in cyanobakteriellen Genen weniger Introns pro Gen gefunden als für alte eukaryotische Gene. Cyanobakterielle Gene waren durchschnittlich 10 % kürzer als alte eukaryotische Gene. Pro Aminosäure wurden in cyanobakteriellen Genen ebenfalls mehr Introns als in alten eukaryotischen Genen gefunden. In *Arabidopsis* wurden in cyanobakteriellen Genen durchschnittlich 0,0113 (*Populus*: 0,0129, *Oryza*: 0,0135) und in alten eukaryotischen Genen 0,0137 Introns pro Aminosäure (*Populus*: 0,0141, *Oryza*: 0,0149) gefunden. Die Unterschiede waren mit einer Überschreitungswahrscheinlichkeit (*p*-Wert) zwischen $8{,}39 \times 10^{-10}$ (*Oryza*) und $2{,}2 \times 10^{-16}$ (*Arabidopsis*) hochsignifikant (Exakter Test nach Fisher).

Aufgrund der eindeutigen Verwandtschaftsbeziehungen zwischen *Arabidopsis*, *Populus* (dicotyl) und *Oryza* (monocotyl) konnten Ereignisse des Introngewinns und -verlusts untersucht werden. Introns, die in einer der zweikeimblättrigen Spezies und *Oryza* auftraten, wurden als Verlust in der zweiten zweikeimblättrigen Spezies gewertet. Introns, die lediglich in *Arabidopsis* oder *Populus* und nicht in *Oryza* auftraten, wurden als Introngewinn in dieser Spezies gezählt. Nur sehr wenige Introns wurden seit der Aufspaltung in ein- und zweikeimblättrige Pflanzen gewonnen oder verloren. In *Arabidopsis* wurden deutlich mehr Verluste als Gewinne gefunden. Diese Tendenz zeigte sich nicht in *Populus*. Es wurde ein signifikant größerer Anteil an Introngewinn in cyanobakteriellen Genen gegenüber alten eukaryotischen Genen gemessen. Im Bezug auf Intronverlust zeigte sich kein Unterschied zwischen diesen beiden Klassen von Pflanzengenen. Insgesamt gingen in *Arabidopsis* auch in cyanobakteriellen Genen mehr Introns verloren als gewonnen wurden.

Die Analyse der Intronpositionen zeigte, dass mehr Introns in Phase Null (zwischen Codons) als in Phase Eins (hinter der ersten Codonposition) oder Zwei (hinter der zweiten Codonposition) gefunden wurden. In cyanobakteriellen Genen war der Überschuss von Introns in Phase Null signifikant stärker als in alten eukaryotischen Genen.

6 Diskussion

6.1 Relativierung des abgeleiteten Anteils cyanobakterieller Gene in Pflanzen und Algen

In der vorliegenden Arbeit wurden für 11.569 kernkodierte Proteine aus zwei Pflanzen und zwei Algen sowie deren Homologe in 237 Referenzgenomen phylogenetische Bäume abgeleitet und analysiert, um den Anteil cyanobakterieller Gene in deren Genomen abzuschätzen. Für die vier photosynthetischen Eukaryoten wurden für 12,7 % (*Arabidopsis*) bis 17,1 % (*Cyanidioschyzon*) der untersuchten Proteinfamilien ein cyanobakterieller Ursprung abgeleitet. Nur 3,8 % der Proteine aus der Referenzdatenbank stammen aus Cyanobakterien (Tabelle 5.2). Der abgeleitete Anteil cyanobakterieller Proteine lag damit deutlich höher als durch Zufall zu erwarten wäre. Wie beachtlich das cyanobakterielle Signal war, zeigte sich, als der abgeleitete Anteil an endosymbiontischem Gentransfer (EGT) durch die Datenbankgröße normalisiert dargestellt wurde (Abbildung 5.3). In dieser Darstellung bildeten Pflanzen öfter einen Klan mit Cyanobakterien als mit allen anderen Gruppen. Die Dominanz der Cyanobakterien wurde noch stärker, als nur verlässlichere Intervalle der Daten betrachtet wurden, während für die Eukaryoten – die in der normalisierten Darstellung zweitstärkste Gruppe – der Anteil abnahm.

Für die Genome der Spezies der zwei analysierten primären photosyntetischen Linien Rhodophyceae (Rotalgen) und Chloroplastida (Grünalgen und Pflanzen) wurden unterschiedliche Anteile cyanobakterieller Gene abgeleitet. Für Vertreter der Chloroplastida lag er bei 12,7 % (*Arabidopsis*), 13,6 % (*Oryza*) und 14,2 % (*Chlamydomonas*). Für *Cyanidioschyzon*, den einzigen Vertreter der Rhodophyceae in dieser Arbeit, wurde mit 17,1 % ein höherer Anteil abgeleitet. Eine mögliche Interpretation dieser Ergebnisse ist, dass nach der Aufspaltung von Chloroplastida und Rhodophyceae in letzterer Gruppe mehr Gentransfer vom cyanobakteriellen Endosymbionten in das Wirtsgenom stattgefunden hat. Wahrscheinlicher ist

jedoch, dass der Unterschied durch die stark verschiedenen Genomgrößen zu erklären ist. Für Spezies mit vergleichsweise großen Genomen wurde ein geringerer cyanobakterieller Anteil abgeleitet als für Spezies mit kleineren Genomen. *Oryza* und *Arabidopsis* besitzen mit 34.379 beziehungsweise 28.900 Genen die größten Genome in dieser Analyse und für diese beiden Spezies wurde der geringste Anteil cyanobakterieller Gene abgeleitet. *Cyanidioschyzon* hat mit 4.762 Genen das kleinste Genom und den größten prozentualen Anteil abgeleiteten Gentransfers. *Chlamydomonas reinhardtii* liegt von der Genomgröße (15.097 Gene) und dem abgeleiteten Gentransfer (14,2 %) dazwischen. Vermutlich sind auch die beobachteten Unterschiede im Anteil cyanobakterieller Gene zwischen Algen (*Chlamydomonas* und *Cyanidioschyzon*) und Pflanzen (*Arabidopsis* und *Oryza*) auf stark unterschiedliche Genomgrößen zurückzuführen. *Cyanidioschyzon merolae* hat ein sehr kompaktes Genom und ist als einzellige thermo- und acidophile Süßwasseralge ein eher untypischer Vertreter der Rotalgen (Matsuzaki et al., 2004). Derzeit wird ein Sequenzierungsprojekt für das Kerngenom der marinen Alge *Porphyra umbilicalis* durchgeführt (Quelle: JGI[1] (Joint Genome Institute des U.S. Departement of Energy)). Eine Analyse des Anteils cyanobakterieller Gene dieses Genoms könnte die Frage beantworten, ob in den Rhodophyceae tatsächlich mehr Gentransfer vom cyanobakteriellen Bakterienchromosom zum eukaryotischen Kerngenom stattgefunden hat als in den Chloroplastida.

Für divergente Proteinfamilien wurde häufiger ein cyanobakterieller Ursprung des Pflanzenhomologs abgeleitet (Abschnitt 5.2). Dies könnte zwei mögliche Ursachen haben. Entweder evolvieren cyanobakterielle Gene langsamer als andere Gene des selben Genoms oder die phylogenetische Rekonstruktion liefert für divergente Proteine weniger genaue Ergebnisse. Als cyanobakteriell abgeleitete Proteine zeigten in einem Vergleich von Orthologen zwischen *Arabidopsis* und *Oryza*, sowie *Arabidopsis* und *Chlamydomonas*, keine stärkere Sequenzkonservierung als eukaryotische Proteine. Daher kann die erste Hypothese verworfen werden. Es kann angenommen werden, dass die Ursachen in der Natur der phylogenetischen Ableitung zu suchen sind. Für divergente Proteinfamilien wurden öfter als für konservierte Proteinfamilien falsche Topologien berechnet (Abschnitt 5.4). Diese Ergebnisse decken sich mit Veröffentlichungen, die beschreiben, dass phylogenetische Analysen fehleranfällig sind (Phillips et al., 2004), besonders wenn die Sequenzdivergenz mehr als 50 % beträgt (Nei, 1996; Nei et al., 1995). Weil

1 http://www.jgi.doe.gov/genome-projects/pages/projects.jsf

cyanobakterielle Proteine mit 3,8 % nur einen geringen Anteil an den Referenzgenomen haben (Tabelle 5.2), bildet in solchen Fällen öfter ein Pflanzenhomolog einen Klan mit einem nicht-cyanobakteriellen Homologen. Im Verlauf der Arbeit konnte ein Zusammenhang zwischen dem Grad der Sequenzkonservierung, der Alignment- und Stammbaumverlässlichkeit sowie dem abgeleiteten Gentransfer aufgezeigt werden (Abschnitt 5.4, Abschnitt 5.5, Abschnitt 5.6 und Abschnitt 6.2). Wurde ein Teil der unzuverlässigen Alignments von der Analyse ausgeschlossen, so wurde ein höherer prozentualer Anteil cyanobakterieller Gene abgeleitet. Der durchschnittliche abgeleitete Anteil cyanobakterieller Gene von 14 % ist daher höchstwahrscheinlich eine Unterbewertung des Gentransfers, der tatsächlich stattgefunden hat, weil viele Falsch-Negative nicht als cyanobakteriell gezählt werden. Andererseits kann davon ausgegangen werden, dass die als cyanobakteriell klassifizierten Proteine mit relativ hoher Sicherheit auch cyanobakteriellen Ursprungs sind. Ein weiteres Indiz für die Zuverlässigkeit der cyanobakteriellen Vorhersage ist, dass die Übereinstimmung der Vorhersagen mittels Wahrscheinlichkeits- und Distanzbäumen zu etwa 80 % übereinstimmten (Abbildung 5.2). Martin et al. (2002) bezifferten den Anteil cyanobakterieller Gene bei *Arabidopsis thaliana* auf 18%. Diese Arbeit beinhaltet jedoch ebenfalls keine Korrektur für falsch-negative Daten, die auf geringe Alignmentverlässlichkeit zurückgehen. Der, verglichen mit dieser Arbeit, um 5,3 % höher abgeleitete Gentransfer für diesen Organismus ist vermutlich dadurch zu erklären, dass der Datensatz von Martin et al. (2002) mit *Saccharomyces cerevisiae* nur ein eukaryotisches Referenzgenom enthielt.

6.2 Alignmentverlässlichkeit

Das Erstellen von multiplen Alignments ist eine gängige Praxis in vielen Bereichen der Biologie geworden. Ein eindrucksvolles Beispiel für die Relevanz multipler Alignments ist die Tatsache, dass die Veröffentlichung von Thompson et al. (1994), die mit Clustal W das wohl bekannteste Programm zur Berechnung multipler Alignments beschreibt, bis heute (Juli 2009) in 27.770 Artikeln zitiert wurde (Quelle: Web of Science®). Optimale multiple Alignments, die über die Maximierung einer Bewertungsfunktion anhand einer Bewertungsmatrix sowie Strafpunkten für Insertionen bzw. Deletionen berechnet werden, können aufgrund ihrer Komplexität nicht in akzeptabeler Zeit berechnet werden (Abschnitt 3.7). In der Praxis kommen daher durchweg heuristische Methoden zum Einsatz. Am weitesten verbreitet ist das progressive multiple Alignieren von Sequenzen (Feng und Doolittle,

1987), wie es auch in Clustal W (Abschnitt 4.3.6) angewendet wird. Dabei wird zunächst das am nächsten verwandte Sequenzpaar aligniert und sukzessiv weitere Sequenzen an dem resultierenden Alignment ausgerichtet. Auch das in dieser Arbeit verwendete Programm Muscle erstellt Alignments nach dieser Methode. Der Muscle-Algorithmus versucht jedoch, die Alignments in zwei zusätzlichen Arbeitsschritten zu verbessern (vgl. Abschnitt 4.3.7).

Die Autoren von Clustal W merken an, dass mit progressiven Methoden berechnete Alignments in einfachen Fällen von „exzellenter Qualität“ sind, während sie in komplizierteren Fällen „gute Ausgangspunkte für weitere automatische oder manuelle Verfeinerung“ darstellen (Thompson et al., 1994). In der Praxis werden Alignments jedoch größtenteils als gegeben hingenommen. Es wird selten Aufwand betrieben, um die Alignmentqualität zu messen oder zu verbessern. Dafür gibt es im Wesentlichen zwei Gründe: Erstens fehlen objektive Kriterien, um die Qualität von multiplen Alignments einzuschätzen. Zu Publikationszwecken werden Alignments manchmal „nach visueller Prüfung manuell korrigiert“ (Kawasaki et al., 2000; O'Callaghan et al., 1999). Auch wenn das menschliche Gehirn ein sehr effektives Werkzeug zur Mustererkennung ist, so sind manuelle Korrekturen rein subjektiv. Aufgrund der Subjektivität ist auch die Reproduzierbarkeit nicht gegeben, die ein wichtiger Bestandteil der wissenschaftlichen Methode darstellt. Zweitens ist eine visuelle Inspektion von Alignments in vielen Fällen nicht praktikabel. Viele bioinformatische Analysen der Postgenomära – wie auch die vorliegende Arbeit mit 11.569 Alignments – enthalten derart viele Datensätze, dass eine manuelle Überprüfung kaum möglich ist.

Die absolute Korrektheit eines berechneten multiplen Alignments kann nur über den Vergleich dieses Alignments mit einem Referenzalignment erfolgen. Oft werden strukturbasierte Alignments als Referenz verwendet, wie sie in der BAliBASE-Datenbank (engl. *benchmark alignment database*, (Thompson et al., 1999*b*)) gespeichert sind. Derartige Alignments werden oft als Goldstandard herangezogen. Dennoch sind es keine perfekten Alignments, sondern lediglich berechnete Alignments, die unter Zuhilfenahme von Strukturinformation verbessert wurden (Hall, 2008). Das wahre Alignment ist nur bekannt, wenn Analysen mit simulierten Sequenzdaten vorgenommen werden. Ein Vergleich eines berechneten mit dem wahren Alignment erlaubt eine Aussage über die absolute Korrektheit des berechneten Alignments.

Derartige Analysen ermöglichen es, allgemeine Richtlinien für multiple Alignments abzuleiten. Die Korrektheit von Alignments von Aminosäuresequenzen liegt

bei über 80 %, wenn die durschnittliche Sequenzidentität mehr als 25 % beträgt (Thompson et al., 1999*a*). Liegt die durchschnittliche Aminosäureidentität unter 20 %, so liegt die Korrektheit der Alignments bei unter 50%. Für Alignments von DNA werden über 80 % Korrektheit erreicht, wenn die Sequenzidentität mindestens 60 % beträgt, fällt aber unter 50% wenn die Sequenzidentität weniger als 50 % beträgt (Ogden und Rosenberg, 2006). Diese Richtlinien beziehen sich auf durchschnittliche Identitäten zwischen Sequenzen. Die Korrektheit kann jedoch sinken, wenn zu einem Datensatz eng verwandter Sequenzen eine entfernt verwandte Sequenz hinzugefügt wird (Thompson et al., 1999*a*). Ein grundlegender Nachteil dieser Richtlinien ist jedoch, dass sie keine Aussagen über konkrete Alignments erlauben.

Methoden, die die Berechnung der Verlässlichkeit von realen Alignments erlauben, sind extrem selten, wenn überhaupt vorhanden (Kumar und Filipski, 2007). Die HoT-Methode von Landan und Graur (2007) ist eine einfache und schnelle Möglichkeit, eine Aussage über die Verlässlichkeit eines konkreten multiplen Sequenzalignments zu treffen (vgl. Abschnitt 4.3.14). Dabei wird nicht die absolute Korrektheit gemessen (s.o.), sondern von der *a priori*-Annahme ausgegangen, dass Alignments unabhängig von der Orientierung sein sollten, in der die Sequenzen aligniert werden. Unterscheidet sich ein Alignment stark von dem entsprechenden reversen Alignment, so wird dies als Indiz für eine geringe Alignmentverlässlichkeit gewertet.

Hall (2008) zeigte in einer Arbeit mit simulierten Sequenzen eine gute positive Korrelation zwischen absoluter Korrektheit und HoT-Verlässlichkeit. Alignments mit hoher Verlässlichkeit sind demnach solchen mit geringer Verlässlichkeit vorzuziehen. Die HoT-Verlässlichkeiten stellen jedoch Überbewertungen der absoluten Korrektheit dar, wobei der Grad der Überbewertung von Alignmentmethode und -parametern abhängig ist (Hall, 2008). Diese Ergebnisse sind nicht überraschend, da mit der HoT-Methode nur ein sehr einfacher Fall (Orientierung der Sequenzen) analysiert wird. Hall (2008) argumentierte auch, dass mit der HoT-Methode die Reproduzierbarkeit und nicht die Verlässlichkeit von multiplen Sequenzalignments gemessen wird, und zog zur Veranschaulichung die Analogie einer Dartscheibe. Es wird nicht der Abstand der Pfeile zum Mittelpunkt (engl. *bull's eye*) bewertet, sondern die Streuung der Pfeile untereinander.

Die HoT-Methode liefert ein praktikables Werkzeug, um die Verlässlichkeit – oder Reproduzierbarkeit (s.o.) – von multiplen Sequenzalignments zu ermittlen. Da sie nur den einfachsten Fall testet, wird die Verlässlichkeit der Alignments mit die-

ser Methode tendenziell überbewertet. Dennoch korreliert die HoT-Verlässlichkeit positiv mit der Korrektheit der Alignments (Hall, 2008). Die HoT-Methode liefert keinen Schwellenwert für die Alignmentverlässlichkeit, der nicht unterschritten werden sollte. Dennoch hilft sie Fälle zu identifizieren, in denen Alignments und damit auch Phylogenien (Abschnitt 6.3) stark von einem Prozess (hier der Richtung der Sequenzen) abhängig sind, der genauso zufällig wie der Wurf einer Münze ist.

6.3 Die HoT-Methode in dieser Arbeit

Um die Reproduzierbarkeit der Alignments und phylogenetischen Bäume zu beurteilen, wurden drei Maße verwendet: Der Anteil der identischen Spalten (CS, engl. *columns score*) und Aminosäurepaare (SPS, engl. *sum-of-pairs score*) zwischen normalem und reversem Alignment sowie der Anteil der identischen Bipartitionen zwischen den aus diesen Alignments abgeleiteten phylogenetischen Bäumen (PPS, engl. *phylogenetic partitions score*). Welcher dieser drei Quantoren hat in der Praxis die höchste Relevanz und bis zu welchem Wert sollen Alignments überhaupt noch verwendet werden?

Sowohl der CS- (Abschnitt 4.6.2) als auch der SPS-Wert (Abschnitt 4.6.3) sind Maße für die Alignmentverlässlichkeit. Beide zeigten eine Abhängigkeit von der Sequenzkonservierung (Abschnitt 5.4). Der SPS- war für dasselbe Alignmentpaar fast immer höher als der CS-Wert (Abbildung 5.7). Als Konsequenz daraus zeigten die zwei Quantoren sehr unterschiedliche Verteilungen (Abbildung 5.9). Die Verlässlichkeit war über den CS-Wert annähernd normalverteilt mit einem Maximum um 0,75. Beim SPS-Wert als Maß für die Alignmentverlässlichkeit war der Großteil der Daten im Intervall mit der größten und zweitgrößten Verlässlichkeit, während die Anzahl der Elemente in den restlichen Intervallen annähernd gleichverteilt war. Diese Ergebnisse belegen die intuitive Vorstellung, dass der Anteil identischer Spalten ein stringenteres Maß für die Alignmentqualität ist, als der Anteil identisch alignierter Aminosäurepaare. Bei der Berechnung des CS-Wertes wird für jede Alignmentspalte eine Ja/Nein-Entscheidung getroffen, während beim SPS-Wert für jede Spalte der Grad der Ähnlichkeit berechnet wird. Insbesondere bei Alignments vieler Taxa stellt der SPS-Wert das realistischere Maß zur Beurteilung der Alignmentqualität dar. In seltenen Fällen war der CS- höher als der SPS-Wert (Abbildung 5.7). Diese Fälle können durch eine stark ungleichmäßige Verteilung von Lücken in den Alignments erklärt werden. Die Anzahl an möglichen Paaren pro Spalte mit n Taxa ergibt sich aus $\frac{n(n-1)}{2}$. Dabei kann ein Paar

aus zwei Zeichen oder einem Zeichen und einem Lückensymbol, nicht aber aus zwei Lückensymbolen bestehen (Abschnitt 4.6.3). Lückenreiche Spalten enthalten demnach weniger Paare als lückenarme. Den Extremfall stellen Spalten dar, die aus $n-1$ Lückensymbolen und einem Zeichen bestehen und lediglich $n-1$ Paare aufweisen. Da der SPS-Wert des Alignments über die Gesamtzahl des Alignments berechnet wird, tragen derartige Spalten nur in sehr geringem Maße dazu bei, während bei der Berechnung des CS-Wertes jede Spalte das selbe Gewicht erhält.

Die Korrelation von Alignment- und Baumverlässlichkeit zeigt, dass letztere aus ersterer gefolgert werden kann (Abschnitt 5.5). Diese Ergebnisse decken sich mit der Erwartung, dass ein Stammbaum nicht zuverlässiger als die Homologiemuster sein kann, aus denen er abgeleitet wurde. Eine Beurteilung der zu erwartenden Verlässlichkeit eines Stammbaums ist demnach möglich, ohne ihn abzuleiten, was die Effizienz der HoT-Analyse steigert. Wird ein vorher festgelegter Schwellenwert für die Verlässlichkeit unterschritten, kann davon abgesehen werden, den Stammbaum für ein Alignment abzuleiten. Mögliche Ausschlusskriterien für die Alignmentqualität wären beispielsweise ein CS-Wert $\leq$ 0,8 oder ein SPS-Wert $\leq$ 0,9. Die Wahl dieser Parameter ist rein arbiträr. Die Histogramme für CS- und SPS-Wert veranschaulichen jedoch, dass bei stringenterer Wahl der Kriterien ein zu großer Teil der Daten von der Analyse ausgeschlossen wird (Abbildung 5.9). In Analysen, die sich im Grad der Divergenz der Sequenzen stark unterscheiden, könnten diese Ausschlusskriterien anders ausfallen. Bei Hochdurchsatz-Analysen kann dieses Vorgehen helfen, die Verlässlichkeit einer Analyse zu verbessern. Ist ein konkretes Alignment von geringer Verlässlichkeit von besonderem Interesse, so kann versucht werden, die Verlässlichkeit durch das Weglassen bestimmter Sequenzen zu erhöhen. Landan und Graur arbeiten derzeit daran die HoT-Methode derart weiterzuentwickeln, dass die Sequenzen identifiziert werden können, die die Alignmentverlässlichkeit am negativsten beeinflussen (persönliche Korrespondenz).

6.4 Der Einfluss von Alignments und Baumrekonstruktionsmethoden auf die Ergebnisse

Phylogenetische Bäume wurden in dieser Arbeit mit zwei gängigen Methoden abgeleitet. Die Ergebnisse von Distanz- und Wahrscheinlichkeitsbäumen waren bezogen auf die Gesamtzahl abgeleiteter cyanobakterieller Proteine weitgehend gleich (Tabelle 5.2 und Tabelle 6.2). Für *Arabidopsis* wurden beispielsweise mit

Wahrscheinlichkeitsbäumen 592 und mit Distanzbäumen 591 cyanobakterielle Gene vorhergesagt (jeweils 12,7 % der abgeleiteten Bäume). Auf Ebene der einzelnen Proteine lag die Übereinstimmung dieser verschiedenen Methoden zwischen 79,5 % (*Chlamydomonas*) und 86,9 % (*Cyanidioschyzon*) (Abbildung 5.2). Die Verlässlichkeit der Alignments hatte hingegen einen starken Einfluss auf den Schätzer des abgeleiteten Gentransfers (Abbildung 5.9). Für *Arabidopsis* wurde beispielsweise für die verlässlichsten $\geq 20\%$ und für die unzuverlässigsten Alignments $\leq 5\%$ cyanobakterielle Gene abgeleitet.

Bezogen auf die in dieser Arbeit gestellte Frage nach dem Anteil cyanobakterieller Gene in den Kerngenomen von Pflanzen und Algen hatte die Wahl der Methode der phylogentischen Rekonstruktion nur einen geringen Einfluss, während die Verlässlichkeit der Alignments die Ergebnisse stark beeinflussten. In der Literatur (Keane et al., 2006; Kelchner und Thomas, 2007; Phillips et al., 2004) wird oft – neben Phänomenen wie lateralem Gentransfer oder *long branch attraction* – die Methode der phylogenetischen Rekonstruktion als Fehlerquelle genannt, wenn ein Genbaum vom Speziesbaum abweicht. Oft wird als Lösung vorgeschlagen eine fortschrittlichere (bzw. kompliziertere) Methode der Baumrekonstruktion zu verwenden. Doch auch solche Methoden basieren auf den Homologiemustern der Alignments. Sind diese Muster fehlerhaft, so wird auch der daraus abgeleitete phylogenetische Baum fehlerhaft sein. Diese intuitive Annahme konnte durch die Korrelation von Alignment- und Baumverlässlichkeit belegt werden (Abschnitt 5.5). Verschiedene Methoden waren unterschiedlich stark von der Alignmentverlässlichkeit abhängig. So war die Korrelation von Alignment- und Baumverlässlichkeit für Distanz- leicht höher als für Wahrscheinlichkeitsbäume. Daraus lässt sich folgern, dass die Ableitung vom ML-Bäumen weniger stark von der Alignmentverlässlichkeit abhängig ist.

6.5 Der freilebende Vorfahre der Plastiden

In dieser Arbeit wurde versucht, den freilebenden Vorfahren der primären Plastiden zu bestimmen. Zwar haben weder dieser Vorfahre noch seine Gene die Zeit seit der Etablierung der Endosymbiose unverändert überstanden. Dennoch kann die Frage gestellt werden, welches rezente Cyanobakterium die Kollektion an Genen aufweist, die den cyanobakteriellen Pflanzengenen am ähnlichsten ist. *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 zeigten unter den neun analysierten Cyanobakterien die größte und *Prochlorococcus marinus* MIT 9313

die geringste Gesamt-Aminosäureidentität (Abbildung 5.10). Diese Ergebnisse spiegeln die Genomgrößen wieder, da *Anabaena* und *Nostoc* unter den neun analysierten Cyanobakterien die größten Genome besitzen, während *Prochlorococcus* mit 2.265 Genen das kleinste Genom hat. *Gloeobacter violaceus* PCC 7421 hat mit 4.430 Genen das drittgrößte Genom, zeigte jedoch eine sichtbar geringere Aminosäureidentität zu den cyanobakteriellen Pflanzenproteinen. Diese Beobachtungen zeigen, dass die Gesamt-Aminosäureidentität nicht nur die Genomgröße, also die An- oder Abwesenheit von Genen, sondern auch die Sequenzähnlichkeiten widerspiegelt. In Bezug auf *Gloeobacter* stimmen die Ergebnisse mit der Ansicht überein, dass dessen Position unter den Cyanobakterien aufgrund des Fehlens von Thylakoiden und der Position in einigen Stammbäumen (Sato, 2006; Tomitani et al., 2006) als basal angesehen wird.

Die Auswertung von 11.569 Wahrscheinlichkeitsbäumen zeigte, dass *Anabaena* und *Nostoc* in den Stammbäumen, aus denen ein cyanobakterieller Ursprung eines Pflanzengens abgeleitet wurde, am Häufigsten in der Nachbargruppe des Pflanzenhomologs vertreten waren (Abbildung 5.11a). Diese Tendenz blieb erhalten, wenn nur verlässlichere Teilmengen der Daten (Alignments mit SPS $\geq$ 0,8 oder SPS $\geq$ 0,9) betrachtet wurden. Sowohl *Prochlorococcus* mit dem kleinsten Genom und *Gloeobacter* mit dem drittgrößten Genom traten nur sehr selten in der Schwestergruppe auf. *Synechococcus elongatus* PCC 7942, *Thermosynechococcus elongatus* und *Synechocystis sp.* PCC 6803 kamen häufiger vor, obwohl ihre Genome nur unwesentlich größer sind als das von *Prochlorococcus* (Abbildung 5.11c). Der phylogenetische Ansatz zur Bestimmung des freilebenden Vorfahrens der Plastiden war also weitgehend unabhängig von der Genomgröße.

Aufgrund ihres Gengehalts und ihrer Sequenzidentitäten waren *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 den cyanobakteriellen Pflanzengenen – und damit möglicherweise auch dem freilebenden Vorfahren der Chloroplasten – ähnlicher als die restlichen sieben in dieser Arbeit analysierten Cyanobakterien. Auch wenn die Ergebnisse der Ähnlichkeitssuche zu einem gewissen Grad von der Genomgröße beeinflusst wurden, bestätigte der phylogenetische Ansatz *Anabaena* und *Nostoc*. Im Bezug auf *Nostoc* decken sich die Ergebnisse mit denen von Martin et al. (2002) die *Nostoc punctiforme* unter drei Cyanobakterien am ähnlichsten zum freilebenden Vorfahren der Chloroplasten fanden. In Bezug auf *Anabaena* stimmen die Ergebnisse mit den Schlussfolgerungen von Sato (2006) überein. *Anabaena* und *Nostoc* gehören zur Gattung *Nostocales*, die in der Klassifizierung von Rippka et al. (1979) zur Abteilung IV gehört. Zu Beginn dieser Arbeit standen lediglich

neun vollständig sequenzierte Cyanobakteriengenome zur Verfügung, die die Diversität der Cyanobakterien nur unvollständig abdecken. So sind beispielsweise nur Mitglieder der Abteilungen I und IV vertreten. Heute (Stand Juli 2009) sind die vollständig sequenzierten Genome von ca. 40 Cyanobakterien verfügbar. Unter ihnen sind jedoch keine Vertreter der Abteilungen II und V. Liegen vollständig sequenzierte Genome für Repräsentanten aller fünf Abteilung vor, könnte eine Wiederholung der Analyse helfen den freilebenden Vorfahren der Chloroplasten besser zu charakterisieren.

6.6 Die Rolle von Stickstoff bei der Etablierung der Endosymbiose der Chloroplasten

Unter den neun Cyanobakterien in dieser Arbeit wiesen *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 eine Kollektion an Genen auf, die dem Vorfahren der Chloroplasten am ähnlichsten ist. Diese Arten gehören in der Klassifizierung von Rippka et al. (1979) zur Abteilung IV, die eine Gruppe von filamentösen Cyanobakterien bezeichnet. Cyanobakterien aus Abteilung IV und V können Zellen zu Heterozysten differenzieren, in denen ohne Sauerstoffproduktion Stickstoff fixiert wird (Rajaniemi et al., 2005; Rippka et al., 1979). Tomitani et al. (2006) argumentieren, dass die Abteilungen IV und V die am höchsten entwickelten Gruppen von Cyanobakterien sind, und Heterozysten vor dem Anstieg der Sauerstoffkonzentration in der Atmosphäre vor 2,3 Milliarden Jahren entstanden sind. 2,3 Milliarden Jahre können also als Obergrenze für das Alter der Plastiden gesehen werden. Das minimale Alter wurde anhand von fossilen Rotalgen der Gattung *Bangiomorpha* auf 1,2 Milliarden Jahre geschätzt (Butterfield, 2000; Yoon et al., 2004). In dieses Zeitfenster passt die Beobachtung, dass Aktineten (differenzierte Zellen einiger Vertreter der Abteilungen IV und V, vgl. Abschnitt 3.6.1) etwa 400 Millionen Jahre vor *Bangiomorpha* in Fossilien nachgewiesen wurden (Tomitani et al., 2006).

Die Vermutung, dass der freilebende Vorfahre der Plastiden die Fähigkeit zur Stickstofffixierung besessen haben könnte, lässt Spekulationen über die Rolle von Stickstoff bei der Etablierung der Symbiose zu, die letztendlich zu Plastiden in Pflanzen und Algen führte. Unter modernen Endosymbiosen, die Cyanobakterien beeinhalten, gibt es Beispiele, bei denen eine Versorgung mit Stickstoff durch den Symbionten im Vordergrund steht, und teilweise keine organischen Kohlenstoffverbindungen bereitgestellt werden. Gut untersuchte Beispiele sind der Pilz *Geosiphon pyriforme* (Mollenhauer et al., 1996), die Kieselalge *Rhopalodia gibba* (Prechtl et al.,

2004), Algenfarne der Gattung *Azolla* (Prasanna et al., 2006), koralloide (knollenförmige) Wurzeln in Palmfarnen (Costa et al., 2004) und Blütenpflanzen des Genus *Gunnera* (Chiu et al., 2005).

In diesen Beispielen gehören die Symbionten zur Abteilung IV und sind von der Gattung *Nostoc* oder *Anabaena*. Eine Ausnahme bildet *Rhopalodia*, bei der der Symbiont mit der Gattung *Cyanothece* verwandt ist, die zur Abteilung I gehört (Prechtl et al., 2004). In vielen modernen Endosymbiosen mit Cyanobakterien als Symbionten, sind diese in der Lage, Stickstoff zu fixieren und an den Wirt weiterzugeben (Prechtl et al., 2004; Rai et al., 2000; Raven, 2002). Auch in Ektosymbiosen, wie zum Beispiel Flechten, sind Cyanobakterien der Gattung *Nostoc sp.* PCC 7120 beteiligt und Stickstoff spielt eine wichtige Rolle (Rikkinen et al., 2002). Einen Überblick über die Rolle von Stickstoff gibt ein Artikel von Kneip et al. (2007).

In Studien zur Erdgeschichte wurde vorgeschlagen, dass die Ozeane zur Zeit des Erdfrühzeitalters (Proterozoikum) anoxisch und sulfidreich waren, was die Verfügbarkeit von Stickstoff limitiert haben könnte (Anbar und Knoll, 2002). Dieses Erdzeitalter erstreckte sich von vor 2,3 Milliarden bis ca. 580 Millionen Jahren (Canfield et al., 2007; Fike et al., 2006) und fällt damit in den Zeitraum der Entstehung der Plastiden (Butterfield, 2000; Yoon et al., 2004). Zusammenfassend deuten die begrenzte Verfügbarkeit von Stickstoff zum Zeitpunkt der Etablierung der Endosymbiose und die Rolle von Stickstoff in modernen Endosymbiosen, die Cyanobakterien beinhalten, darauf hin, dass die Ergebnisse dieser Arbeit mit der Theorie übereinstimmen, dass Stickstoff zu dem Zeitpunkt, als die Endosymbiose eingegangen wurde, eine wichtige Rolle gespielt haben könnte (Kneip et al., 2007; Raven, 2002).

6.7 Proteinlokalisierung und Stoffwechsel

11.569 Proteine aus Pflanzen und Algen wurden einer Vorhersage des Zielkompartiments mit TargetP unterzogen und die Ergebnisse getrennt für cyanobakterielle und nicht-cyanobakterielle Proteine dargestellt (Abschnitt 5.8). Für 351 von 592 Proteinen (knapp 60 %) cyanobakteriellen Ursprungs aus *Arabidopsis thaliana* wurde der Chloroplast als Zielkompartiment vorhergesagt. Für die restlichen Proteine wurden Mitochondrien (10 %), der Sekretionsweg (9 %) und das Cytosol (21 %) vorhergesagt. Diese Ergebnisse unterstützen die These, dass für die Mehrheit der Proteine Zielkompartiment und Ursprung des Gens übereinstimmen (Horiike et al., 2001). Gleichzeitig liefert die vorliegende Arbeit zahlreiche Gegenbeispiele,

die die Beobachtungen anderer Autoren belegen (Abdallah et al., 2000; Martin und Schnarrenberger, 1997). Die Ergebnisse dieser Arbeit decken sich mit denen von Martin et al. (2002), die für etwa 50 % der cyanobakteriellen Proteine den Chloroplasten als Bestimmungsort fanden.

Der Anteil der Proteine cyanobakteriellen Ursprungs mit Chloroplasten als Bestimmungsort lag für *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* bei 38 %, 16 % beziehungsweise 44 %. Dabei war, verglichen mit den Ergebnissen für *Arabidopsis*, vor allem der Anteil an Proteinen erhöht, für die Mitochondrien als Zielkompartiment vorausgesagt wurden. Insbesondere bei den Algen *Chlamydomonas* und *Cyanidioschyzon* waren viele Vorhersagen von niedriger Verlässlichkeit, weshalb den Ergebnissen für diese Organismen nicht zu viel Bedeutung beigemessen werden sollte. Vermutlich enthielt der Trainingsdatensatz von TargetP (Abschnitt 4.4) hauptsächlich Proteine aus *Arabidopsis thaliana*, da die meisten Einträge von Pflanzenproteinen in der SWISS-PROT-Datenbank vermutlich von diesem Modellorganismus stammen. Daher wurden bei den Vorhersagen für *Arabidopsis* und die andere Pflanze *Oryza* verlässlichere Ergebnisse als für die Algen erzielt. Die Ergebnisse für *Arabidopsis thaliana*, dass 60 % der Proteine cyanobakteriellen Ursprungs in Chloroplasten lokalisiert sind, geben daher auch vermutlich das verlässlichste Bild der Proteinlokalisierung wieder.

Den 984 (*Arabidopsis*), 687 (*Oryza*), 562 (*Chlamydomonas*) beziehungsweise 341 (*Cyanidioschyzon*) Proteinen cyanobakteriellen Ursprungs konnten 313, 258, 291 beziehungsweise 175 Funktionen in Form von EC-Nummern zugeordnet werden (Tabelle 5.4). Ein Abgleich mit den in der KEGG-Datenbank gespeicherten Stoffwechselwegen für die einzelnen Spezies zeigte jedoch, dass die entsprechenden Organismen nur über 156, 138, 114 beziehungsweise 92 dieser Funktionen verfügen. Der Mehrzahl an Pflanzenproteinen, für die ein cyanobakterieller Ursprung abgeleitet worden war, konnte mit dem in dieser Arbeit verwendeten Ansatz (vgl. Abschnitt 4.5.8) keine Funktion zugeordnet werden. Die Mehrzahl an cyanobakteriellen Enzymen hatten Funktionen in der Aminosäuresynthese, dem Kohlenhydratstoffwechsel und der Biosynthese von Sekundärmetaboliten (Tabelle 5.5).

Die Enzyme cyanobakteriellen Ursprungs konnten für einzelne Stoffwechselwege wie den Calvin-Zyklus dargestellt werden (vgl. Abbildung 5.14). Auf diese Weise wurden die Enzyme hervorgehoben, für die ein cyanobakterieller Ursprung abgeleitet worden war. Die in dieser Arbeit getroffenen Vorhersagen für die Enzyme des Calvin-Zyklus stimmen mit denen von Martin und Schnarrenberger (1997)

überein (vgl. Abbildung 3.2). Für alle sechs Enzyme, für die Martin und Schnarrenberger (1997) einen cyanobakteriellen Ursprung abgeleitet hatten, wurde auch in dieser Arbeit ein solcher Ursprung vorhergesagt. Die Ergebnisse unterschieden sich jedoch in Bezug auf die cytosolische Form der GAPDH (EC 1.2.1.12), während sie für die in Plastiden lokalisierte Form (EC 1.2.1.13) übereinstimmten. Martin und Schnarrenberger (1997) leiteten für das cytosolische Isoenzym einen proteobakteriellen Ursprung ab. In der vorliegenden Arbeit wurde es als cyanobakteriell klassifiziert. Obwohl es neben dem beschriebenen Unterschied vermutlich noch weitere Abweichungen zu anderen Methoden der funktionellen Charakterisierung gibt, liegt der Vorteil des in dieser Arbeit angewandten Verfahrens unter Verwendung von EC-Nummern darin, dass die Ergebnisse direkt auf Stoffwechselkarten veranschaulicht werden können.

Auf der Internetseite der KEGG-Datenbank[1] sind Übersichtskarten der Stoffwechselwege vieler Spezies, darunter auch *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon*, hinterlegt. Abbildung 6.1 zeigt eine solche Karte, die alle Stoffwechselwege von *Arabidopsis thaliana* enthält. In dieser Darstellung sind die Edukte und Produkte biochemischer Umsetzungen durch Punkte repräsentiert. Farbige Linien zwischen den Punkten stellen die Enzyme dar, über die der jeweilige Organismus verfügt. Die Enzyme sind in zehn Gruppen von Stoffwechselwegen (vgl. Tabelle 5.5) eingeteilt. Diese Einteilung ist durch eine Farbkodierung veranschaulicht. Graue Linien stellen Enzyme dar, die der jeweilige Organismus nicht besitzt. Auf der Internetseite sind diese Karten interaktiv und es können detaillierte Informationen über die Enzyme, die Edukte sowie die Produkte abgerufen werden. Bis heute (Stand Juli 2009) ist es nicht möglich, die Enzyme, für die in dieser Arbeit ein cyanobakterieller Ursprung vorausgesagt wurde, in einer solchen Übersichtskarte darzustellen. Die Arbeitsgruppe von Professor Minoru Kanehisa[2], welche die KEGG-Datenbank administriert, plant jedoch eine derartige Funktionalität auf ihrer Internetseite zu implementieren (persönliche Mitteilung von Yuriko Matsuura). Dann wäre es möglich, die in dieser Arbeit erstellten Listen mit EC-Nummern zu verwenden, um für *Arabidopsis*, *Oryza*, *Chlamydomonas* und *Cyanidioschyzon* Stoffwechselkarten zu erstellen, auf denen alle Enzyme cyanobakteriellen Ursprungs hervorgehoben sind.

1 http://www.genome.jp/kegg/

2 http://kanehisa.hgc.jp/home_e.html

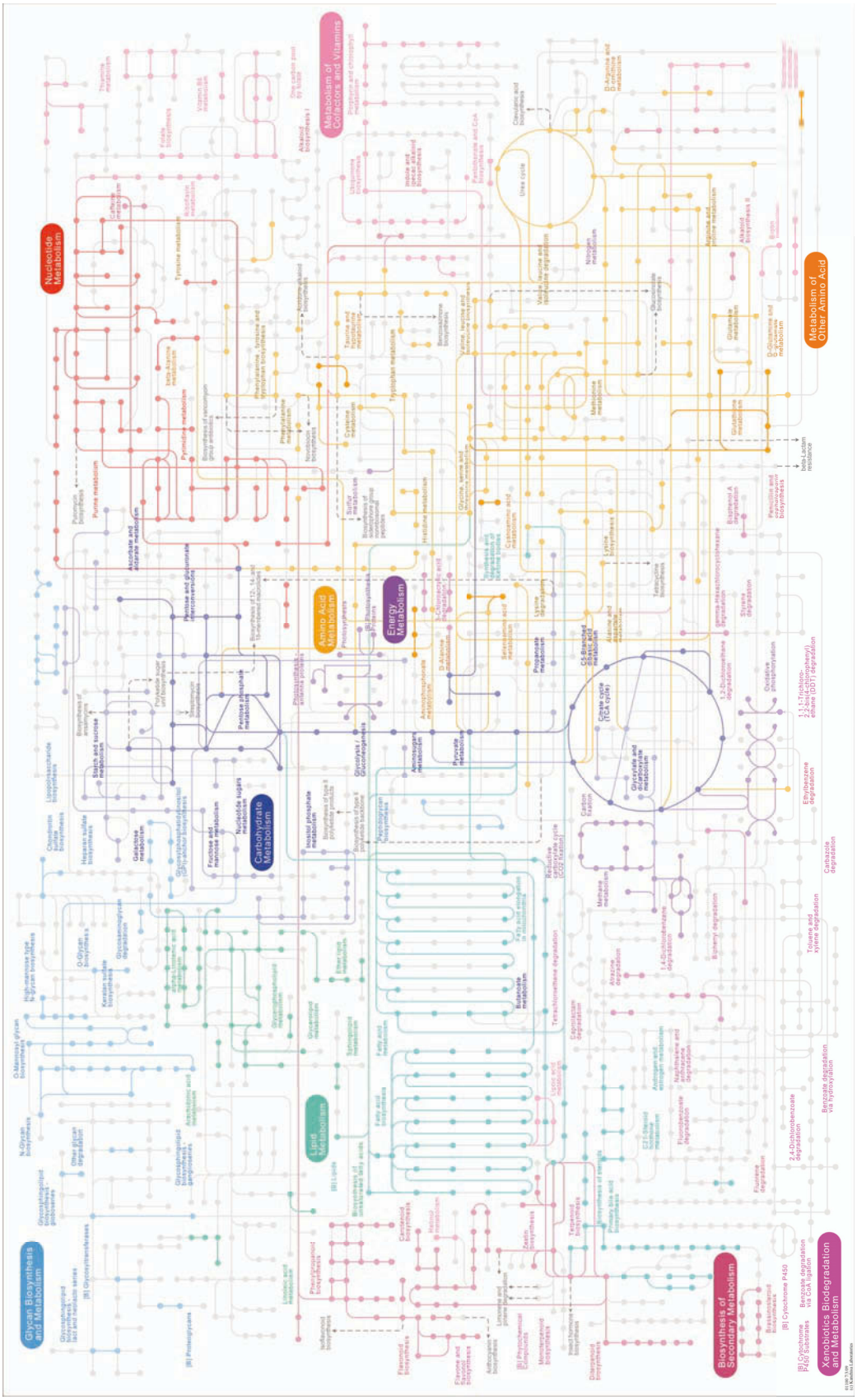

Abbildung 6.1: Stoffwechselkarte für *Arabidopsis thaliana* (Quelle: KEGG [1]). Erläuterung in Abschnitt 6.7.

6.8 Intronevolution

In allen drei analysierten Genomen war die Introndichte in cyanobakteriellen Genen geringer als in alten eukaryotischen Pflanzengenen (Abschnitt 5.10). Die Unterschiede waren gering, aber hochsignifikant (Exakter Test nach Fisher). Nur sehr wenige Introns wurden seit der Divergenz von *Arabidopsis* und *Populus* gewonnen oder gingen verloren. Diese Beobachtung stimmt mit einer früheren Arbeit mit *Arabidopsis* und *Oryza* überein (Roy und Penny, 2007). In *Arabidopsis* wurde mehr Intronverlust als -gewinn beobachtet. Auch diese Ergebnisse decken sich mit denen von Roy und Penny (2007). In *Populus* wurde keine Dominanz des Intronverlusts beobachtet. *Arabidopsis thaliana* besitzt ein wesentlich kompakteres Genom als *Populus* (Tuskan et al., 2006). In diesem Zusammenhang kann der Intronverlust in *Arabidopsis thaliana* als eine Begleiterscheinung einer Entwicklung hin zu einem kompakteren Genom gesehen werden. Gene cyanobakteriellen Ursprungs zeigten einen signifikant höheren Anteil an Introngewinn als alte eukaryotische Gene. Der Introngewinn in cyanobakteriellen Pflanzengenen hält demnach seit der Aufspaltung von *Arabidopsis* und *Populus* an. Der signifikant stärkere Überschuss von Introns in Phase Null (zwischen Codons) in cyanobakteriellen Genen stützt diese Hypothese, da jüngere Introns häufiger in dieser Phase auftreten (Sverdlov et al., 2003). Diese Beobachtungen lassen die Spekulation zu, dass cyanobakterielle Pflanzengene noch keine Sättigung in ihrer Introndichte erreicht haben. Insgesamt kann das Muster des Introngewinns und -verlusts als Zusammenspiel zweier Kräfte gesehen werden. Der Gewinn von Introns hält bis heute an, wenn auch die Geschwindigkeit wesentlich langsamer ist als während der frühen eukaryotischen Evolution. Gleichzeitig findet Intronverlust statt, was sich besonders am Beispiel von *Arabidopsis thaliana* zeigt, da dieser Organismus vermutlich zu einem kompakten Genom hin evolviert.

Fast 96 % der Introns waren zwischen ein- und zweikeimblättrigen Pflanzen konserviert (Abschnitt 5.10), was darauf hindeutet, dass die Mehrzahl der Introns vor der Aufspaltung in diese beiden Gruppen vor 150 bis 200 Millionen Jahren (Chaw et al., 2004; Wolfe et al., 1989) entstanden ist. Daher unterstützen die Ergebnisse dieser Arbeit die Theorie, dass Introns über einen relativ kurzen Zeitraum der frühen eukaryotischen Evolution aufkamen und sich ausgebreitet haben (Babenko et al., 2004; Carmel et al., 2007).

6.9 Schlussfolgerung und Ausblick

In dieser Arbeit wurde in den Kerngenomen von *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* und *Cyanidioschyzon merolae* ein Anteil an Genen cyanobakteriellen Ursprungs von 12,7 %, 13,6 %, 14,2 % beziehungsweise 17,1 % abgeleitet. Analysen der Verlässlichkeiten der multiplen Alignments nach der HoT-Methode zeigten, dass für verlässlichere Daten ein größerer Anteil cyanobakterieller Gene abgeleitet wurde, als für unzuverlässige. Die in dieser Arbeit abgeleitete cyanobakterielle Komponente von Algen und Pflanzen kann daher als Untergrenze angesehen werden, da für viele Gene cyanobakteriellen Ursprungs dieser Ursprung aufgrund unzuverlässiger Alignments und Stammbäume nicht abgeleitet wurde (Falsch-Negative). Während die Ergebnisse stark von der Alignmentverlässlichkeit abhängig waren, hatte die Wahl der Methode zur Baumrekonstruktion nur einen sehr geringen Effekt auf die Ableitung cyanobakterieller Gene. Bei den Ergebnissen aus Distanz- und Wahrscheinlichkeitsbäumen gab es nur wenige Unterschiede, sowohl für die Gesamtzahl, als auch für einzelne Gene. Die Ergebnisse dieser Arbeit zeigen, dass die Alignmentqualität einen großen Einfluss auf die Resultate phylogenetischer Analysen hat und ihr mehr Beachtung geschenkt werden sollte.

Die identifizierten cyanobakteriellen Gene waren unter den neun in dieser Arbeit vertretenen Cyanobakterien den Genen von *Anabaena variabilis* ATCC 29413 und *Nostoc sp.* PCC 7120 am ähnlichsten. Beide Spezies gehören zu einer Gruppe von filamentösen, heterozysten-bildenden Cyanobakterien, die in der Lage sind Stickstoff zu fixieren. Dies – und die Rolle von Stickstoff in vielen Symbiosen, die Cyanobakterien beinhalten – unterstützt die Hypothese, dass Stickstofffixierung bei der Etablierung der Symbiose, die zu Plastiden in Algen und Pflanzen geführt hat, eine wichtige Rolle gespielt haben könnte. Diese Schlussfolgerungen über die Biologie des freilebenden Vorfahren der Plastiden beruhen auf neun vollständig sequenzierten Cyanobakteriengenomen, die zu Beginn dieser Arbeit verfügbar waren. Diese neun Spezies decken die Diversität der Cyanobakterien nur unvollständig ab. Mittlerweile sind die Genome von ca. 40 Cyanobakterien verfügbar (Stand Juli 2009). Dennoch repräsentieren diese Genome lediglich drei der fünf Abteilungen der Cyanobakterien (Systematik nach Rippka et al. (1979)). Für Mitglieder der Abteilungen II und V liegen keine vollständig sequenzierten Genome vor. Im Institut für Ökologische Pflanzenphysiologie der Heinrich-Heine-Universität Düsseldorf laufen zur Zeit die Vorbereitungen für die Sequenzierung

der Genome von fünf Arten aus Abteilung V und einer Art aus Abteilung IV, die ebenfalls in der Lage sind Stickstoff zu fixieren. Eine dieser Arbeit ähnliche Analyse könnte anhand zusätzlicher cyanobakterieller Genome das Verständnis über die Biologie des freilebenden Vorfahren weiter vertiefen.

Eine derartige Analyse sollte zusätzliche Rotalgengenome beinhalten, um die Frage zu beantworten, ob Rotalgen einen größeren Anteil cyanobakterieller Gene enthalten als Grünalgen und Pflanzen, wie die Ergebnisse dieser Arbeit für *Cyanidioschyzon merolae* nahelegen. Kürzlich wurde unserem Institut das Genom der Rotalge *Galdieria sulphuraria* durch das Institut für Biochemie der Pflanzen der Heinrich-Heine-Universität Düsseldorf zur Verfügung gestellt. *Galdieria* ist wie *Cyanidioschyzon* eine einzellige thermo- und acidophile Süßwasseralge mit einem kompakten Genom von 6.623 Genen. Beide Spezies gehören darüberhinaus zur selben Familie und vermutlich würden die Ergebnisse für *Galdieria* denen von *Cyanidioschyzon* stark ähneln. Zur Zeit läuft ein Sequenzierungsprojekt für das Kerngenom der marinen Rotalge *Porphyra umbilicalis*, die vermutlich besser geeignet ist um das Verständnis des Anteils cyanobakterieller Gene in Rotalgen zu erweitern.

Anhang

Tabelle 6.1: Liste der Referenztaxa

Reich	Phylum	Organismus	Quelle	Version
Eukaryota	Protista	*Entamoeba histolytica*	TIGR	Januar 2006
	Protista	*Trichomonas vaginalis*	TIGR	Januar 2006
	Ascomycota	*Candida glabra*	NCBI FTP	Januar 2006
	Ascomycota	*Debaryomyces hansenii*	NCBI FTP	Januar 2006
	Ascomycota	*Eremothecium gossypii*	NCBI FTP	Januar 2006
	Ascomycota	*Kluyveromyces lactis*	NCBI FTP	Januar 2006
	Ascomycota	*Saccharomyces cerevisiae*	NCBI FTP	Januar 2006
	Ascomycota	*Schizosaccharomyces pombe*	NCBI FTP	Januar 2006
	Ascomycota	*Yarrowia lipolytica*	NCBI FTP	Januar 2006
	Ascomycota	*Trichoderma reesei*	JGI/DOE	Version 1.0
	Basidiomycota	*Phanerochaete chrysosporium*	JGI/DOE	Version 2.0
	Basidiomycota	*Cryptococcus neoformans*	NCBI FTP	Januar 2006
	Microsporidia	*Encephalitozoon cuniculi*	NCBI FTP	Januar 2006
Archaeabacteria	Crenarchaeota	*Aeropyrum pernix*	NCBI FTP	Januar 2006
	Crenarchaeota	*Pyrobaculum aerophilum*	NCBI FTP	Januar 2006
	Crenarchaeota	*Sulfolobus acidocaldarius* DSM 639	NCBI FTP	Januar 2006
	Crenarchaeota	*Sulfolobus solfataricus*	NCBI FTP	Januar 2006
	Crenarchaeota	*Sulfolobus tokodaii*	NCBI FTP	Januar 2006
	Euryarchaeota	*Archaeoglobus fulgidus*	NCBI FTP	Januar 2006
	Euryarchaeota	*Haloarcula marismortui* ATCC 43049	NCBI FTP	Januar 2006
	Euryarchaeota	*Halobacterium sp.*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanobacterium thermoautotrophicum*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanococcus jannaschii*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanococcus maripaludis* S2	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanopyrus kandleri*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanosarcina acetivorans*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanosarcina barkeri fusaro*	NCBI FTP	Januar 2006
	Euryarchaeota	*Methanosarcina mazei*	NCBI FTP	Januar 2006
	Euryarchaeota	*Natronomonas pharaonis*	NCBI FTP	Januar 2006
	Euryarchaeota	*Picrophilus torridus* DSM 9790	NCBI FTP	Januar 2006
	Euryarchaeota	*Pyrococcus abyssi*	NCBI FTP	Januar 2006
	Euryarchaeota	*Pyrococcus furiosus*	NCBI FTP	Januar 2006
	Euryarchaeota	*Pyrococcus horikoshii*	NCBI FTP	Januar 2006
	Euryarchaeota	*Thermococcus kodakaraensis* KOD1	NCBI FTP	Januar 2006
	Euryarchaeota	*Thermoplasma acidophilum*	NCBI FTP	Januar 2006
	Euryarchaeota	*Thermoplasma volcanium*	NCBI FTP	Januar 2006
	Nanoarchaeota	*Nanoarchaeum equitans*	NCBI FTP	Januar 2006

Reich	Phylum	Organismus	Quelle	Version
Eubacteria	Actinobacteria	*Bifidobacterium longum*	NCBI FTP	Januar 2006
	Actinobacteria	*Corynebacterium diphtheriae*	NCBI FTP	Januar 2006
	Actinobacteria	*Corynebacterium efficiens* YS-314	NCBI FTP	Januar 2006
	Actinobacteria	*Corynebacterium glutamicum* ATCC 13032 Bielefeld	NCBI FTP	Januar 2006
	Actinobacteria	*Corynebacterium jeikeium* K411	NCBI FTP	Januar 2006
	Actinobacteria	*Leifsonia xyli subsp. xyli* CTCB0	NCBI FTP	Januar 2006
	Actinobacteria	*Mycobacterium avium subsp. paratuberculosis*	NCBI FTP	Januar 2006
	Actinobacteria	*Mycobacterium bovis*	NCBI FTP	Januar 2006
	Actinobacteria	*Mycobacterium leprae*	NCBI FTP	Januar 2006
	Actinobacteria	*Mycobacterium tuberculosis* CDC1551	NCBI FTP	Januar 2006
	Actinobacteria	*Nocardia farcinica* IFM10152	NCBI FTP	Januar 2006
	Actinobacteria	*Propionibacterium acnes* KPA171202	NCBI FTP	Januar 2006
	Actinobacteria	*Streptomyces avermitilis*	NCBI FTP	Januar 2006
	Actinobacteria	*Streptomyces coelicolor*	NCBI FTP	Januar 2006
	Actinobacteria	*Symbiobacterium thermophilum* IAM14863	NCBI FTP	Januar 2006
	Actinobacteria	*Thermobifida fusca* YX	NCBI FTP	Januar 2006
	Actinobacteria	*Tropheryma whipplei* Twist	NCBI FTP	Januar 2006
	Aquificae	*Aquifex aeolicus*	NCBI FTP	Januar 2006
	Bacteroidetes	*Bacteroides fragilis* YCH46	NCBI FTP	Januar 2006
	Bacteroidetes	*Bacteroides thetaiotaomicron* VPI-5482	NCBI FTP	Januar 2006
	Bacteroidetes	*Porphyromonas gingivalis* W83	NCBI FTP	Januar 2006
	Bacteroidetes	*Salinibacter ruber* DSM 13855	NCBI FTP	Januar 2006
	Chlamydiae	*Chlamydia muridarum*	NCBI FTP	Januar 2006
	Chlamydiae	*Chlamydia trachomatis* A HAR-13	NCBI FTP	Januar 2006
	Chlamydiae	*Chlamydophila abortus* S26 3	NCBI FTP	Januar 2006
	Chlamydiae	*Chlamydophila caviae*	NCBI FTP	Januar 2006
	Chlamydiae	*Chlamydophila pneumoniae* TW 183	NCBI FTP	Januar 2006
	Chlamydiae	*Parachlamydia sp.* UWE25	NCBI FTP	Januar 2006
	Chlorobi	*Chlorobium chlorochromatii* CaD3	NCBI FTP	Januar 2006
	Chlorobi	*Chlorobium tepidum* TLS	NCBI FTP	Januar 2006
	Chlorobi	*Pelodictyon luteolum* DSM 273	NCBI FTP	Januar 2006
	Chloroflexi	*Dehalococcoides sp.* CBDB1	NCBI FTP	Januar 2006
	Chloroflexi	*Dehalococcoides ethenogenes* 195	NCBI FTP	Januar 2006
	Cyanobacteria	*Anabaena variabilis* ATCC 29413	NCBI FTP	Januar 2006
	Cyanobacteria	*Gloeobacter violaceus* PCC 7421	NCBI FTP	Januar 2006
	Cyanobacteria	*Nostoc sp.* PCC 7120	NCBI FTP	Januar 2006
	Cyanobacteria	*Prochlorococcus marinus str.* MIT 9313	NCBI FTP	Januar 2006
	Cyanobacteria	*Synechococcus sp.* CC9605	NCBI FTP	Januar 2006
	Cyanobacteria	*Synechococcus elongatus* PCC 7942	NCBI FTP	Januar 2006
	Cyanobacteria	*Synechococcus sp.* WH8102	NCBI FTP	Januar 2006
	Cyanobacteria	*Synechocystis sp.* PCC 6803	NCBI FTP	Januar 2006
	Cyanobacteria	*Thermosynechococcus elongatus*	NCBI FTP	Januar 2006
	Deinococcus-Thermus	*Deinococcus radiodurans*	NCBI FTP	Januar 2006
	Deinococcus-Thermus	*Thermus thermophilus* HB8	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus anthracis* Ames 0581	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus cereus* ATCC 10987	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus clausii* KSM-K16	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus halodurans*	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus licheniformis* DSM 13	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus subtilis*	NCBI FTP	Januar 2006
	Firmicutes	*Bacillus thuringiensis* konkukian	NCBI FTP	Januar 2006
	Firmicutes	*Carboxydothermus hydrogenoformans* Z-2901	NCBI FTP	Januar 2006

Reich	Phylum	Organismus	Quelle	Version
	Firmicutes	*Clostridium acetobutylicum*	NCBI FTP	Januar 2006
	Firmicutes	*Clostridium perfringens*	NCBI FTP	Januar 2006
	Firmicutes	*Clostridium tetani* E88	NCBI FTP	Januar 2006
	Firmicutes	*Enterococcus faecalis* V583	NCBI FTP	Januar 2006
	Firmicutes	*Geobacillus kaustophilus* HTA426	NCBI FTP	Januar 2006
	Firmicutes	*Lactobacillus acidophilus* NCFM	NCBI FTP	Januar 2006
	Firmicutes	*Lactobacillus johnsonii* NCC 533	NCBI FTP	Januar 2006
	Firmicutes	*Lactobacillus plantarum*	NCBI FTP	Januar 2006
	Firmicutes	*Lactobacillus sakei* 23K	NCBI FTP	Januar 2006
	Firmicutes	*Lactococcus lactis*	NCBI FTP	Januar 2006
	Firmicutes	*Listeria innocua*	NCBI FTP	Januar 2006
	Firmicutes	*Listeria monocytogenes*	NCBI FTP	Januar 2006
	Firmicutes	*Mesoplasma florum* L1	NCBI FTP	Januar 2006
	Firmicutes	*Moorella thermoacetica* ATCC 39073	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma capricolum* ATCC 27343	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma gallisepticum*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma genitalium*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma hyopneumoniae* 232	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma mobile* 163K	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma mycoides*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma penetrans*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma pneumoniae*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma pulmonis*	NCBI FTP	Januar 2006
	Firmicutes	*Mycoplasma synoviae* 53	NCBI FTP	Januar 2006
	Firmicutes	*Oceanobacillus iheyensis*	NCBI FTP	Januar 2006
	Firmicutes	*Onion yellows phytoplasma*	NCBI FTP	Januar 2006
	Firmicutes	*Staphylococcus aureus* Mu50	NCBI FTP	Januar 2006
	Firmicutes	*Staphylococcus epidermidis* RP62A	NCBI FTP	Januar 2006
	Firmicutes	*Staphylococcus haemolyticus*	NCBI FTP	Januar 2006
	Firmicutes	*Staphylococcus saprophyticus*	NCBI FTP	Januar 2006
	Firmicutes	*Streptococcus agalactiae* 2603	NCBI FTP	Januar 2006
	Firmicutes	*Streptococcus mutans*	NCBI FTP	Januar 2006
	Firmicutes	*Streptococcus pneumoniae* TIGR4	NCBI FTP	Januar 2006
	Firmicutes	*Streptococcus pyogenes* MGAS6180	NCBI FTP	Januar 2006
	Firmicutes	*Streptococcus thermophilus* CNRZ1066	NCBI FTP	Januar 2006
	Firmicutes	*Thermoanaerobacter tengcongensis*	NCBI FTP	Januar 2006
	Firmicutes	*Ureaplasma urealyticum*	NCBI FTP	Januar 2006
	Fusobacteria	*Fusobacterium nucleatum*	NCBI FTP	Januar 2006
	Planctomycetacia	*Pirellula sp.*	NCBI FTP	Januar 2006
	Proteobacteria	*Acinetobacter sp.* ADP1	NCBI FTP	Januar 2006
	Proteobacteria	*Agrobacterium tumefaciens* C58 UWash	NCBI FTP	Januar 2006
	Proteobacteria	*Anaplasma marginale* St Maries	NCBI FTP	Januar 2006
	Proteobacteria	*Azoarcus sp.* EbN1	NCBI FTP	Januar 2006
	Proteobacteria	*Bartonella henselae* Houston-1	NCBI FTP	Januar 2006
	Proteobacteria	*Bartonella quintana* Toulouse	NCBI FTP	Januar 2006
	Proteobacteria	*Bdellovibrio bacteriovorus*	NCBI FTP	Januar 2006
	Proteobacteria	*Blochmannia floridanus*	NCBI FTP	Januar 2006
	Proteobacteria	*Bordetella bronchiseptica*	NCBI FTP	Januar 2006
	Proteobacteria	*Bordetella parapertussis*	NCBI FTP	Januar 2006
	Proteobacteria	*Bordetella pertussis*	NCBI FTP	Januar 2006
	Proteobacteria	*Bradyrhizobium japonicum*	NCBI FTP	Januar 2006
	Proteobacteria	*Brucella abortus* 9-941	NCBI FTP	Januar 2006
	Proteobacteria	*Brucella melitensis*	NCBI FTP	Januar 2006

Reich	Phylum	Organismus	Quelle	Version
	Proteobacteria	*Brucella suis* 1330	NCBI FTP	Januar 2006
	Proteobacteria	*Buchnera aphidicola* Sg	NCBI FTP	Januar 2006
	Proteobacteria	*Buchnera sp.*	NCBI FTP	Januar 2006
	Proteobacteria	*Burkholderia sp.* 383	NCBI FTP	Januar 2006
	Proteobacteria	*Burkholderia mallei* ATCC 23344	NCBI FTP	Januar 2006
	Proteobacteria	*Burkholderia pseudomallei* 1710b	NCBI FTP	Januar 2006
	Proteobacteria	*Burkholderia thailandensis* E264	NCBI FTP	Januar 2006
	Proteobacteria	*Campylobacter jejuni* RM1221	NCBI FTP	Januar 2006
	Proteobacteria	*Cand. Blochmannia pennsylvanicus* BPEN	NCBI FTP	Januar 2006
	Proteobacteria	*Cand. Pelagibacter ubique* HTCC1062	NCBI FTP	Januar 2006
	Proteobacteria	*Caulobacter crescentus*	NCBI FTP	Januar 2006
	Proteobacteria	*Chromobacterium violaceum*	NCBI FTP	Januar 2006
	Proteobacteria	*Colwellia psychrerythraea* 34H	NCBI FTP	Januar 2006
	Proteobacteria	*Coxiella burnetii*	NCBI FTP	Januar 2006
	Proteobacteria	*Dechloromonas aromatica* RCB	NCBI FTP	Januar 2006
	Proteobacteria	*Desulfotalea psychrophila* LSv54	NCBI FTP	Januar 2006
	Proteobacteria	*Desulfovibrio desulfuricans* G20	NCBI FTP	Januar 2006
	Proteobacteria	*Desulfovibrio vulgaris* Hildenborough	NCBI FTP	Januar 2006
	Proteobacteria	*Ehrlichia canis* Jake	NCBI FTP	Januar 2006
	Proteobacteria	*Ehrlichia ruminantium* str. Welgevonden	NCBI FTP	Januar 2006
	Proteobacteria	*Erwinia carotovora atroseptica* SCRI1043	NCBI FTP	Januar 2006
	Proteobacteria	*Escherichia coli* CFT073	NCBI FTP	Januar 2006
	Proteobacteria	*Francisella tularensis tularensis*	NCBI FTP	Januar 2006
	Proteobacteria	*Geobacter metallireducens* GS-15	NCBI FTP	Januar 2006
	Proteobacteria	*Geobacter sulfurreducens*	NCBI FTP	Januar 2006
	Proteobacteria	*Gluconobacter oxydans* 621H	NCBI FTP	Januar 2006
	Proteobacteria	*Haemophilus ducreyi* 35000HP	NCBI FTP	Januar 2006
	Proteobacteria	*Haemophilus influenzae* 86 028NP	NCBI FTP	Januar 2006
	Proteobacteria	*Hahella chejuensis* KCTC 2396	NCBI FTP	Januar 2006
	Proteobacteria	*Helicobacter hepaticus*	NCBI FTP	Januar 2006
	Proteobacteria	*Helicobacter pylori* 26695	NCBI FTP	Januar 2006
	Proteobacteria	*Idiomarina loihiensis* L2TR	NCBI FTP	Januar 2006
	Proteobacteria	*Legionella pneumophila* str. Paris	NCBI FTP	Januar 2006
	Proteobacteria	*Magnetospirillum magneticum* AMB-1	NCBI FTP	Januar 2006
	Proteobacteria	*Mannheimia succiniciproducens* MBEL55E	NCBI FTP	Januar 2006
	Proteobacteria	*Mesorhizobium loti*	NCBI FTP	Januar 2006
	Proteobacteria	*Methylococcus capsulatus* str. Bath	NCBI FTP	Januar 2006
	Proteobacteria	*Neisseria gonorrhoeae* FA 1090	NCBI FTP	Januar 2006
	Proteobacteria	*Neisseria meningitidis* MC58	NCBI FTP	Januar 2006
	Proteobacteria	*Nitrobacter winogradskyi* Nb-255	NCBI FTP	Januar 2006
	Proteobacteria	*Nitrosococcus oceani* ATCC 19707	NCBI FTP	Januar 2006
	Proteobacteria	*Nitrosomonas europaea*	NCBI FTP	Januar 2006
	Proteobacteria	*Nitrosospira multiformis* ATCC 25196	NCBI FTP	Januar 2006
	Proteobacteria	*Pasteurella multocida*	NCBI FTP	Januar 2006
	Proteobacteria	*Pelobacter carbinolicus*	NCBI FTP	Januar 2006
	Proteobacteria	*Photobacterium profundum* SS9	NCBI FTP	Januar 2006
	Proteobacteria	*Photorhabdus luminescens*	NCBI FTP	Januar 2006
	Proteobacteria	*Pseudoalteromonas haloplanktis* TAC125	NCBI FTP	Januar 2006
	Proteobacteria	*Pseudomonas aeruginosa*	NCBI FTP	Januar 2006
	Proteobacteria	*Pseudomonas fluorescens* Pf-5	NCBI FTP	Januar 2006
	Proteobacteria	*Pseudomonas putida* KT2440	NCBI FTP	Januar 2006
	Proteobacteria	*Pseudomonas syringae*	NCBI FTP	Januar 2006
	Proteobacteria	*Psychrobacter arcticum* 273-4	NCBI FTP	Januar 2006

Reich	Phylum	Organismus	Quelle	Version
	Proteobacteria	*Ralstonia eutropha* JMP134	NCBI FTP	Januar 2006
	Proteobacteria	*Ralstonia solanacearum*	NCBI FTP	Januar 2006
	Proteobacteria	*Rhodobacter sphaeroides* 2.4.1	NCBI FTP	Januar 2006
	Proteobacteria	*Rhodopseudomonas palustris* CGA009	NCBI FTP	Januar 2006
	Proteobacteria	*Rhodospirillum rubrum* ATCC 11170	NCBI FTP	Januar 2006
	Proteobacteria	*Rickettsia conorii*	NCBI FTP	Januar 2006
	Proteobacteria	*Rickettsia felis* URRWXCal2	NCBI FTP	Januar 2006
	Proteobacteria	*Rickettsia prowazekii*	NCBI FTP	Januar 2006
	Proteobacteria	*Rickettsia typhi str. wilmington*	NCBI FTP	Januar 2006
	Proteobacteria	*Salmonella enterica*	NCBI FTP	Januar 2006
	Proteobacteria	*Salmonella typhi*	NCBI FTP	Januar 2006
	Proteobacteria	*Salmonella typhimurium* LT2	NCBI FTP	Januar 2006
	Proteobacteria	*Shewanella oneidensis*	NCBI FTP	Januar 2006
	Proteobacteria	*Shigella boydii* Sb227	NCBI FTP	Januar 2006
	Proteobacteria	*Shigella dysenteriae*	NCBI FTP	Januar 2006
	Proteobacteria	*Shigella flexneri* 2a	NCBI FTP	Januar 2006
	Proteobacteria	*Shigella sonnei* Ss046	NCBI FTP	Januar 2006
	Proteobacteria	*Silicibacter pomeroyi* DSS-3	NCBI FTP	Januar 2006
	Proteobacteria	*Sinorhizobium meliloti*	NCBI FTP	Januar 2006
	Proteobacteria	*Thiobacillus denitrificans* ATCC 25259	NCBI FTP	Januar 2006
	Proteobacteria	*Thiomicrospira crunogena* XCL-2	NCBI FTP	Januar 2006
	Proteobacteria	*Thiomicrospira denitrificans* ATCC 33889	NCBI FTP	Januar 2006
	Proteobacteria	*Vibrio cholerae*	NCBI FTP	Januar 2006
	Proteobacteria	*Vibrio fischeri* ES114	NCBI FTP	Januar 2006
	Proteobacteria	*Vibrio parahaemolyticus*	NCBI FTP	Januar 2006
	Proteobacteria	*Vibrio vulnificus* YJ016	NCBI FTP	Januar 2006
	Proteobacteria	*Wigglesworthia brevipalpis*	NCBI FTP	Januar 2006
	Proteobacteria	*Wolbachia* endosymbiont of *Drosophila melanogaster*	NCBI FTP	Januar 2006
	Proteobacteria	*Wolinella succinogenes*	NCBI FTP	Januar 2006
	Proteobacteria	*Xanthomonas campestris pv. vesicatoria* 85-10	NCBI FTP	Januar 2006
	Proteobacteria	*Xanthomonas citri*	NCBI FTP	Januar 2006
	Proteobacteria	*Xanthomonas oryzae* KACC10331	NCBI FTP	Januar 2006
	Proteobacteria	*Xylella fastidiosa*	NCBI FTP	Januar 2006
	Proteobacteria	*Yersinia pestis* biovar Mediaevails	NCBI FTP	Januar 2006
	Proteobacteria	*Yersinia pseudotuberculosis* IP32953	NCBI FTP	Januar 2006
	Proteobacteria	*Zymomonas mobilis* ZM4	NCBI FTP	Januar 2006
	Spirochaetes	*Borrelia burgdorferi*	NCBI FTP	Januar 2006
	Spirochaetes	*Borrelia garinii* PBi	NCBI FTP	Januar 2006
	Spirochaetes	*Leptospira interrogans* serovar Lai	NCBI FTP	Januar 2006
	Spirochaetes	*Treponema denticola* ATCC 35405	NCBI FTP	Januar 2006
	Spirochaetes	*Treponema pallidum*	NCBI FTP	Januar 2006
	Thermotogae	*Thermotoga maritima*	NCBI FTP	Januar 2006

Tabelle 6.2: Nachbargruppen der Pflanzenhomologen in 11.569 Distanzbäumen. Multiple Sequenzalignments wurden mit Muscle (Edgar, 2004*a,b*) berechnet. Distanzen wurden anhand des JTT-Modells (Jones et al., 1992) kalkuliert wobei Alignmentspalten mit Lücken von der Berechnung ausgenommen wurden. Phylogenetische Bäume wurden mit der *Neighbor-Joining*-Methode (NJ) (Saitou und Nei, 1987) rekonstruiert. Eine detaillierte Beschreibung aller Parameter ist unter Abschnitt 4.5.1 bis Abschnitt 4.5.4 gegeben. Gemischte Klans enthalten Taxa verschiedener Phyla.

Nachbargruppe	*A. thaliana* (%)	*O. sativa* (%)	*C. reinhardtii* (%)	*C. merolae* (%)	Datenbankgröße (%)
Cyanobakterien	591 (12,7%)	428 (13,4%)	346 (13,9%)	208 (17,2%)	31.940 (3,8%)
Proteobakterien	517 (11,1%)	334 (10,5%)	485 (19,4%)	122 (10,1%)	360.234 (42,3%)
Weitere Eubakterien	890 (19,1%)	612 (19,2%)	536 (21,5%)	273 (22,5%)	226.314 (26,6%)
Archaebakterien	48 (1,0%)	36 (1,1%)	21 (0,8%)	16 (1,3%)	56.513 (6,6%)
Eukaryoten	2.170 (46,5%)	1.472 (46,2%)	881 (35,3%)	494 (40,7%)	176.606 (20,7%)
Gemischt	454 (9,7)	304 (9,5%)	230 (9,2%)	100 (8,2%)	

Literaturverzeichnis

Abdallah, F., Salamini, F. und Leister, D. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci*, 2000. **5**(4):141–142.

Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G., McCourt, R. M., Mendoza, L., Moestrup, O., Mozley-Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A. V., Spiegel, F. W. und Taylor, M. F. J. R. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol*, 2005. **52**(5):399–451.

Allen, J. F. The function of genomes in bioenergetic organelles. *Philos Trans R Soc Lond B Biol Sci*, 2003. **358**(1429):19–37.

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**(17):3389–3402.

Anbar, A. D. und Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science*, 2002. **297**(5584):1137–1142.

Arabidopsis Genome Initiative, T. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 2000. **408**(6814):796–815.

Archibald, J. M. Algal genomics: exploring the imprint of endosymbiosis. *Curr Biol*, 2006. **16**(24):R1033–5.

Archibald, J. M., Longet, D., Pawlowski, J. und Keeling, P. J. A novel polyubiquitin structure in cercozoa and foraminifera: evidence for a new eukaryotic supergroup. *Mol Biol Evol*, 2003. **20**(1):62–66.

Babenko, V. N., Rogozin, I. B., Mekhedov, S. L. und Koonin, E. V. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res*, 2004. **32**(12):3724–3733.

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**(5493):972–977.

Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J. und Doolittle, W. F. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol*, 2008. **25**(1):83–91.

Basu, M. K., Rogozin, I. B., Deusch, O., Dagan, T., Martin, W. und Koonin, E. V. Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues. *Mol Biol Evol*, 2008. **25**(1):111–119.

Belshaw, R. und Bensasson, D. The rise and falls of introns. *Heredity*, 2006. **96**(3):208–213.

Ben Ali, A., De Baere, R., Van der Auwera, G., De Wachter, R. und Van de Peer, Y. Phylogenetic relationships among algae based on complete large-subunit rRNA sequences. *Int J Syst Evol Microbiol*, 2001. **51**(Pt 3):737–749.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. und Wheeler, D. L. Genbank. *Nucleic Acids Res*, 2008. **36**:D25–30.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. und Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 2003. **31**(1):365–370.

Bogorad, L. Evolution of early eukaryotic cells: genomes, proteomes, and compartments. *Photosynth Res*, 2008. **95**(1):11–21.

Bowe, L. M., Coat, G. und dePamphilis, C. W. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and gnetales' closest relatives are conifers. *Proc Natl Acad Sci U S A*, 2000. **97**(8):4092–4097.

Brochier, C., Forterre, P. und Gribaldo, S. Archaeal phylogeny based on proteins

of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. *Genome Biology*, 2004. **5**(3).

Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes. *Paleobiology*, 2000. **26**(3):386–404.

Canfield, D. E., Poulton, S. W. und Narbonne, G. M. Late-neoproterozoic deep-ocean oxygenation and the rise of animal life. *Science*, 2007. **315**(5808):92–95.

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., Muller, 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. J., 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**(5809):207–212.

Carmel, L., Wolf, Y. I., Rogozin, I. B. und Koonin, E. V. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res*, 2007. **17**(7):1034–1044.

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**(1429):109–133.

Chaw, S.-M., Chang, C.-C., Chen, H.-L. und Li, W.-H. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J Mol Evol*, 2004. **58**(4):424–441.

Chiu, W.-L., Peters, G. A., Levieille, G., Still, P. C., Cousins, S., Osborne, B. und Elhai, J. Nitrogen deprivation stimulates symbiotic gland development in *Gunnera manicata*. *Plant Physiol*, 2005. **139**(1):224–230.

Cho, G. und Doolittle, R. F. Intron distribution in ancient paralogs supports random insertion and not random loss. *J Mol Evol*, 1997. **44**(6):573–584.

Coghlan, A. und Wolfe, K. H. Origins of recently gained introns in *Caenorhabditis*. *Proc Natl Acad Sci U S A*, 2004. **101**(31):11362–11367.

Costa, J. L., Romero, E. M. und Lindblad, P. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. *FEMS Microbiol. Ecol.*, 2004. **49**(3):481–487.

Cowan, D. Genomics - use your neighbour's genes. *Nature*, 2000. **407**(6803):466–467.

Dagan, T., Artzy-Randrup, Y. und Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A*, 2008. **105**(29):10039–10044.

Darnell, J. E. J. Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science*, 1978. **202**(4374):1257–1260.

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 *Atlas of Protein Sequence and Structure*, Band 5. Natl Biomed Res Found, Washington DC, 1978. 345–352.

Delwiche, C. Tracing the thread of plastid diversity through the tapestry of life. *Am Nat*, 1999. **154**(S4):164–177.

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**(4):748–761.

Deutsch, M. und Long, M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*, 1999. **27**(15):3219–3228.

Doolittle, W. F. Genes in pieces: were they ever together? *Nature*, 1978. **272**(5654):581–582.

Douglas, S. E. Plastid evolution: origins, diversity, trends. *Curr Opin Genet Dev*, 1998. **8**(6):655–661.

Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 2004*a*. **5**:113.

Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 2004*b*. **32**(5):1792–1797.

Emanuelsson, O., Nielsen, H., Brunak, S. und von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 2000. **300**(4):1005–1016.

Emanuelsson, O., Brunak, S., von Heijne, G. und Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, 2007. **2**(4):953–971.

Embley, T. M. und Martin, W. Eukaryotic evolution, changes and challenges. *Nature*, 2006. **440**(7084):623–630.

Embley, T. M., van der Giezen, M., Horner, D. S., Dyal, P. L. und Foster, P. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos Trans R Soc Lond B Biol Sci* , 2003. **358**(1429):191–201.

Fedorov, A., Merican, A. F. und Gilbert, W. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci U S A* , 2002. **99**(25):16128–16133.

Fedorov, A., Roy, S., Fedorova, L. und Gilbert, W. Mystery of intron gain. *Genome Res*, 2003. **13**(10):2236–2241.

Felsenstein, J. Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution*, 1985. **39**(4):783–791.

Felsenstein, J. PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 1989. **5**:164–166.

Feng, D. F. und Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 1987. **25**(4):351–360.

Fike, D. A., Grotzinger, J. P., Pratt, L. M. und Summons, R. E. Oxidation of the ediacaran ocean. *Nature*, 2006. **444**(7120):744–747.

Fink, G. R. Pseudogenes in yeast? *Cell*, 1987. **49**(1):5–6.

Gilbert, W. Why genes in pieces? *Nature*, 1978. **271**(5645):501–501.

Gilbert, W. The exon theory of genes. *Cold Spring Harb Symp Quant Biol*, 1987. **52**:901–905.

Gilbert, W. und Glynias, M. On the ancient nature of introns. *Gene*, 1993. **135**(1-2):137–144.

Gilbert, W., de Souza, S. J. und Long, M. Origin of genes. *Proc Natl Acad Sci U S A*, 1997. **94**(15):7698–7703.

Gilson, P. R., Su, V., Slamovits, C. H., Reith, M. E., Keeling, P. J. und McFadden, G. I. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A*, 2006. **103**(25):9566–9571.

Gould, S. B., Waller, R. F. und McFadden, G. I. Plastid evolution. *Annual Review of Plant Biology*, 2008. **59**(1):491–517.

Guindon, S. und Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 2003. **52**(5):696–704.

Hackett, J., Anderson, D., Erdner, D. und Bhattacharya, D. Dinoflagellates: A remarkable evolutionary experiment. *American Journal of Botany*, 2004. **91**(10):1523–1534.

Hall, B. G. How well does the HoT score reflect sequence alignment accuracy? *Mol Biol Evol*, 2008. **25**(8):1576–1580.

Horiike, T., Hamada, K., Kanaya, S. und Shinozawa, T. Origin of eukaryotic cell nuclei by symbiosis of archaea in bacteria is revealed by homology-hit analysis. *Nat Cell Biol*, 2001. **3**(2):210–214.

Huson, D. H. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics*, 1998. **14**(1):68–73.

International Rice Genome Sequencing Project, T. The map-based sequence of the rice genome. *Nature*, 2005. **436**(7052):793–800.

Jeffares, D. C., Mourier, T. und Penny, D. The biology of intron gain and loss. *Trends Genet*, 2006. **22**(1):16–22.

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**(3):275–282.

Kanehisa, M. und Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 2000. **28**(1):27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. und Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 2006. **34**:D354–7.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. und Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 2008. **36**:D480–4.

Kawasaki, K., Minoshima, S. und Shimizu, N. Propagation and maintenance of the 119 human immunoglobulin Vlambda genes and pseudogenes during evolution. *J Exp Zool*, 2000. **288**(2):120–134.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. und Mclnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol*, 2006. **6**:29.

Kelchner, S. A. und Thomas, M. A. Model use in phylogenetics: nine key questions. *Trends Ecol Evol*, 2007. **22**(2):87–94.

Kimura, M. Rare variant alleles in the light of the neutral theory. *Mol Biol Evol*, 1983. **1**(1):84–93.

Kneip, C., Lockhart, P., Voss, C. und Maier, U.-G. Nitrogen fixation in eukaryotes–new models for symbiosis. *BMC Evol Biol*, 2007. **7**:55.

Koonin, E. V. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct*, 2006. **1**:22.

Koski, L. B. und Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 2001. **52**(6):540–2.

Kuhner, M. K. und Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, 1994. **11**(3):459–468.

Kumar, S. und Filipski, A. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res*, 2007. **17**(2):127–135.

Lambowitz, A. M. und Zimmerly, S. Mobile group II introns. *Annu Rev Genet*, 2004. **38**:1–35.

Landan, G. und Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, 2007. **24**(6):1380–1383.

Leander, B. S. Did trypanosomatid parasites have photosynthetic ancestors? *Trends Microbiol*, 2004. **12**(6):251–258.

Loftus, B., Anderson, I., Davies, R., Alsmark, U. C. M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R. P., Mann, B. J., Nozaki, T., Suh, B., Pop, M., Duchene, M., Ackers, J., Tannich, E., Leippe, M., Hofer, M., Bruchhaus, I., Willhoeft, U., Bhattacharya, A., Chillingworth, T., Churcher, C., Hance, Z., Harris, B., Harris, D., Jagels, K., Moule, S., Mungall, K., Ormond, D., Squares, R., Whitehead, S., Quail, M. A., Rabbinowitsch, E., Norbertczak, H., Price, C., Wang, Z., Guillen, N., Gilchrist, C., Stroup, S. E., Bhattacharya, S., Lohia, A., Foster, P. G., Sicheritz-Ponten, T., Weber, C., Singh, U., Mukherjee, C., El-Sayed, N. M., Petri, W. A. J., Clark, C. G., Embley, T. M., Barrell, B., Fraser, C. M. und Hall, N. The genome of the protist parasite *Entamoeba histolytica*. *Nature*, 2005. **433**(7028):865–868.

Logsdon, J. M. J. The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev*, 1998. **8**(6):637–648.

Logsdon, J. M. J. und Palmer, J. D. Origin of introns–early or late? *Nature*, 1994. **369**(6481):526; author reply 527–8.

Logsdon, J. M. J., Tyshenko, M. G., Dixon, C., D-Jafari, J., Walker, V. K. und Palmer, J. D. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci U S A*, 1995. **92**(18):8507–8511.

Logsdon, J. M. J., Stoltzfus, A. und Doolittle, W. F. Molecular evolution: recent cases of spliceosomal intron gain? *Curr Biol*, 1998. **8**(16):R560–3.

Lynch, M. und Conery, J. S. The origins of genome complexity. *Science*, 2003. **302**(5649):1401–1404.

Lynch, M. und Richardson, A. O. The evolution of spliceosomal introns. *Curr Opin Genet Dev*, 2002. **12**(6):701–710.

Marin, B., Nowack, E. C. M. und Melkonian, M. A plastid in the making: evidence for a second primary endosymbiosis. *Protist*, 2005. **156**(4):425–432.

Martin, W. und Embley, T. M. Evolutionary biology: Early evolution comes full circle. *Nature*, 2004. **431**(7005):134–137.

Martin, W. und Herrmann, R. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol*, 1998. **118**(1):9–17.

Martin, W. und Koonin, E. V. Introns and the origin of nucleus-cytosol compartmentalization. *Nature*, 2006. **440**(7080):41–45.

Martin, W. und Muller, M. The hydrogen hypothesis for the first eukaryote. *Nature*, 1998. **392**(6671):37–41.

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):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 U S A*, 1993. **90**(18):8692–8696.

Martin, W., Hoffmeister, M., Rotte, C. und Henze, K. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol Chem*, 2001. **382**(11):1521–1539.

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 U S A*, 2002. **99**(19):12246–12251.

Martinez, D., Larrondo, L. F., Putnam, N., Gelpke, M. D. S., Huang, K., Chapman, J., Helfenbein, K. G., Ramaiya, P., Detter, J. C., Larimer, F., Coutinho, P. M., Henrissat, B., Berka, R., Cullen, D. und Rokhsar, D. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat Biotechnol*, 2004. **22**(6):695–700.

Martinez, D., Berka, R. M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S. E., Chapman, J., Chertkov, O., Coutinho, P. M., Cullen, D., Danchin, E. G. J., Grigoriev, I. V., Harris, P., Jackson, M., Kubicek, C. P., Han, C. S., Ho, I., Larrondo, L. F., de Leon, A. L., Magnuson, J. K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A. A., Schmoll, M., Terry, A., Thayer, N., Westerholm-Parvinen, A., Schoch, C. L., Yao, J., Barabote, R., Nelson, M. A., Detter, C., Bruce, D., Kuske, C. R., Xie, G., Richardson, P., Rokhsar, D. S., Lucas, S. M., Rubin, E. M., Dunn-Coleman, N., Ward, M. und Brettin, T. S. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat Biotechnol*, 2008. **26**(5):553–560.

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**(6983):653–657.

Mattick, J. S. Introns: evolution and function. *Curr Opin Genet Dev*, 1994. **4**(6):823–831.

McFadden, G. I. und van Dooren, G. G. Evolution: red algal genome affirms a common origin of all plastids. *Curr Biol*, 2004. **14**(13):R514–6.

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Marechal-Drouard, L., Marshall, W. F., Qu, L.-H., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S. M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.-L., Cognat, V., Croft, M. T., Dent, R., Dutcher, S., Fernandez, E., Fukuzawa, H., Gonzalez-Ballester, D., Gonzalez-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P. A., Lemaire, S. D., Lobanov, A. V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J. V., Moseley, J., Napoli, C., Nedelcu, A. M., Niyogi, K., Novoselov, S. V., Paulsen, I. T., Pazour, G., Purton, S., Ral,

J.-P., Riano-Pachon, D. M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S. L., Allmer, J., Balk, J., Bisova, K., Chen, C.-J., Elias, M., Gendler, K., Hauser, C., Lamb, M. R., Ledford, H., Long, J. C., Minagawa, J., Page, M. D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A. M., Yang, P., Ball, S., Bowler, C., Dieckmann, C. L., Gladyshev, V. N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R. T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y. W., Jhaveri, J., Luo, Y., Martinez, D., Ngau, W. C. A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I. V., Rokhsar, D. S. und Grossman, A. R. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, 2007. **318**(5848):245–250.

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

Mollenhauer, D., Mollenhauer, R. und Kluge, M. Studies on initiation and development of the partner association in *Geosiphon pyriforme* (Kütz) v. Wettstein, a unique endocytobiotic system of a fungus (*Glomales*) and the cyanobacterium *Nostoc punctiforme* (Kütz) Hariot. *Protoplasma*, 1996. **193**:3–9.

Mourier, T. und Jeffares, D. C. Eukaryotic intron loss. *Science*, 2003. **300**(5624):1393.

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**(3):443–453.

Nei, M. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet*, 1996. **30**:371–403.

Nei, M., Takezaki, N. und Sitnikova, T. Assessing molecular phylogenies. *Science*, 1995. **267**(5195):253–254.

Nguyen, H. D., Yoshihama, M. und Kenmochi, N. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol*, 2005. **1**(7):e79.

O'Callaghan, D., Cazevieille, C., Allardet-Servent, A., Boschiroli, M. L., Bourg, G., Foulongne, V., Frutos, P., Kulakov, Y. und Ramuz, M. A homologue of the *Agrobacterium tumefaciens* VirB and *Bordetella pertussis* Ptl type IV secretion

systems is essential for intracellular survival of *Brucella suis*. *Mol Microbiol*, 1999. **33**(6):1210–1220.

Ogden, T. H. und Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol*, 2006. **55**(2):314–328.

Pearson, W. R. und Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 1988. **85**(8):2444–2448.

Phillips, M. J., Delsuc, F. und Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*, 2004. **21**(7):1455–1458.

Prasanna, R., Kumar, R., Sood, A., Prasanna, B. M. und Singh, P. K. Morphological, physiochemical and molecular characterization of *Anabaena* strains. *Microbiol Res*, 2006. **161**(3):187–202.

Prechtl, J., Kneip, C., Lockhart, P., Wenderoth, K. und Maier, U.-G. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Mol Biol Evol*, 2004. **21**(8):1477–1481.

Pruitt, K. D., Tatusova, T. und Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 2005. **33**:501–504.

Qiu, W.-G., Schisler, N. und Stoltzfus, A. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*, 2004. **21**(7):1252–1263.

Rai, A., Soderback, E. und Bergman, B. Cyanobacterium-plant symbioses. *New Phytologist*, 2000. **147**(3):449–481.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J., Bork, P. und Arendt, D. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science*, 2005. **310**(5752):1325–1326.

Rajaniemi, P., Hrouzek, P., Kastovska, K., Willame, R., Rantala, A., Hoffmann, L., Komarek, J. und Sivonen, K. Phylogenetic and morphological evaluation of the genera *Anabaena*, *Aphanizomenon*, *Trichormus* and *Nostoc* (*Nostocales*, *Cyanobacteria*). *Int J Syst Evol Microbiol*, 2005. **55**:11–26.

Raven, J. A. Evolution of cyanobacterial symbioses. In **Rai, A. N., Bergman, B.**

und Rasmussen, U. (Herausgeber) *Cyanobacteria in symbiosis*. Kluwer Academic Publishers, Dordrecht (The Netherlands), 2002. 326–1246.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. und Blankenship, R. E. Whole-genome analysis of photosynthetic prokaryotes. *Science*, 2002. **298**(5598):1616–1620.

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**(23):2320–2325.

Richly, E. und Leister, D. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene*, 2004. **329**:11–16.

Rikkinen, J., Oksanen, I. und Lohtander, K. Lichen guilds share related cyanobacterial symbionts. *Science*, 2002. **297**(5580):357.

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**(3):1–61.

Rivera, M. C. und Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 2004. **431**(7005):152–155.

Rivera, M. C., Jain, R., Moore, J. E. und Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A*, 1998. **95**(11):6239–6244.

Robinson, D. und Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences*, 1981. **53**(1-2):131–147.

Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H. J., Philippe, H. und Lang, B. F. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol*, 2005. **15**(14):1325–1330.

Rogers, M. B., Gilson, P. R., Su, V., McFadden, G. I. und Keeling, P. J. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol*, 2007. **24**(1):54–62.

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. und Koonin, E. V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, 2003. **13**(17):1512–1517.

Roy, S. W. Intron-rich ancestors. *Trends Genet*, 2006. **22**(9):468–471.

Roy, S. W. und Gilbert, W. Complex early genes. *Proc Natl Acad Sci U S A* , 2005*a*. **102**(6):1986–1991.

Roy, S. W. und Gilbert, W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci U S A* , 2005*b*. **102**(16):5773–5778.

Roy, S. W. und Penny, D. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol Biol Evol*, 2006. **23**(12):2259–2262.

Roy, S. W. und Penny, D. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. *Mol Biol Evol*, 2007. **24**(1):171–181.

Sagan, L. On the origin of mitosing cells. *J Theor Biol*, 1967. **14**(3):225–274.

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

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

Shoemaker, J. S. und Fitch, W. M. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol Biol Evol*, 1989. **6**(3):270–289.

Smith, T. F. und Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol*, 1981. **147**(1):195–197.

Stoltzfus, A. Origin of introns–early or late. *Nature*, 1994. **369**(6481):526–527.

Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. J. und Doolittle, W. F. Testing the exon theory of genes: the evidence from protein structure. *Science*, 1994. **265**(5169):202–207.

Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. und Koonin, E. V. Evidence of splice signal migration from exon to intron during intron evolution. *Curr Biol*, 2003. **13**(24):2170–2174.

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**(22):4673–4680.

Thompson, J. D., Plewniak, F. und Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 1999*a*. **27**(13):2682–2690.

Thompson, J. D., Plewniak, F. und Poch, O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 1999*b*. **15**(1):87–88.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. und Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*, 2004. **5**(2):123–135.

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**(14):5442–5447.

Toro, N., Jimenez-Zurdo, J. I. und Garcia-Rodriguez, F. M. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev*, 2007. **31**(3):342–358.

Turner, S., Pryer, K., Miao, V. und Palmer, J. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small submit rRNA sequence analysis. *J Eukaryot Microbiol*, 1999. **46**(4):327–338.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger,

B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. und Rokhsar, D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 2006. **313**(5793):1596–1604.

Weeden, N. F. Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. *J Mol Evol*, 1981. **17**(3):133–139.

Wilkinson, M., McInerney, J. O., Hirt, R. P., Foster, P. G. und Embley, T. M. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol*, 2007. **22**(3):114–115.

Wolfe, K. H., Gouy, M., Yang, Y. W., Sharp, P. M. und Li, W. H. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci U S A*, 1989. **86**(16):6201–6205.

Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*, 1994. **39**(3):306–314.

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**(5):809–818.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. und Papke, R. T. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*, 2006. **16**(9):1099–1108.

Zimmerly, S., Hausner, G. und Wu, X. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Research*, 2001. **29**(5):1238–1250.

Danke

Ich danke Herrn Prof. Dr. William Martin für die Vergabe des interessanten Themas, sowie die Möglichkeit in einem Institut von exzellenter Infrastruktur arbeiten zu können. Die internationalen Konferenzen, an denen er mir die Teilnahme ermöglichte, haben meine Arbeit bereichert. Insbesondere möchte ich mich jedoch für den Neuseelandaufenthalt bedanken, der mir neue Perspektiven aufgezeigt hat.

Ich danke Herrn Prof. Dr. Martin Lercher für das meiner Arbeit entgegengebrachte Interesse und seine Bereitschaft das Korreferat zu übernehmen.

Darüberhinaus möchte ich mich bei Frau Dr. Tal Dagan für die Betreuung meiner Arbeit bedanken. Die von ihr vermittelten Matlab-Kenntnisse waren für die Datenanalyse in dieser Arbeit ebenfalls von großem Wert.

Frau Dr. Katrin Henze danke ich für ihre Diskussionsbereitschaft und ihren Einsatz für eine zeitnahe Fertigstellung dieser Arbeit.

I thank Professor Peter Lockhart and his wife Trish McLenachan for their hospitality during my stay at the Allan Wilson Centre in Palmerston North, New Zealand. I am looking forward to future collaborations. I thank the guys from the Allan Wilson Centre's "boffin lounge" for sharing their space.

Ich danke Dr. Gabriel Gelius-Dietrich recht herzlich für das Korrekturlesen dieser Arbeit und die damit verbundenen guten Vorschläge für Verbesserungen und Ergänzungen sowie nützliche Tipps zu LaTeX.

Auch Frau Dr. Nahal Ahmadinejad danke ich für gute Tipps zu LaTeX. Mein Dank gilt ebenfalls Christian Eßer. Als Meister der regulären Ausdrücke hatte er oft den richtigen Einzeiler parat. Verena Zimorski danke ich für das sorgfältige Korrekturlesen dieser Arbeit.

Mein ganz besonderer Dank gilt Nicole Grünheit. Sie gab gute Tipps zu LaTeX, half durch mehrfaches Korrekturlesen dieser Arbeit und trug dazu bei das Druckbild zu optimieren. Darüberhinaus ertrug sie mich während der Fertigstellung dieser Arbeit. Danke! Danke! Danke!

Bei Mayo Röttger und Xavier Pereira bedanke ich mich für die angenehme Atmosphäre in unserem Büro. Bei Mayo bedanke ich mich zusätzlich für viele nützliche Tipps zu Perl und dafür, dass er die Implementation der HoT-Methode verbessert und erweitert hat.

Darüberhinaus danke ich allen Mitarbeitern des Instituts für Botanik III, die direkt oder indirekt beim Gelingen dieser Arbeit geholfen haben.

Peter Stegt danke ich dafür, dass er den ein oder anderen Fehlerteufel gefunden und ausgetrieben hat.

Nicht zuletzt gilt mein Dank auch meiner Familie, insbesondere meinen Eltern Heidemarie und Gottfried Deusch, für ihre Unterstützung und ihr Verständnis. Danke für Alles!

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 20. Juli 2009

.....

(Oliver Deusch)