# Der Einfluß lateralen Gentransfers auf die Evolution prokaryotischer Genome am Beispiel von Alpha-Proteobakterien und Stämmen von Escherichia coli

Inaugural - Dissertation

zur

Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Christian Eßer
aus Düsseldorf

Düsseldorf
Januar 2010

Aus dem Institut für Botanik III

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

| | |
|---|---|
| Betreuer: | Prof. Dr. William Martin |
| Zweitgutachter: | Prof. Dr. Martin Lercher |

Tag der mündlichen Prüfung: 10.03.2010

# Inhaltsverzeichnis

# Publikationen

M. Wu, L. V. Sun, J. Vamathevan, M. Riegler, R. Deboy, J. C. Brownlie, E. A. McGraw, W. Martin, C. Esser, N. Ahmadinejad, C. Wiegand, R. Madupu, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, A. S. Durkin, J. F. Kolonay, W. C. Nelson, Y. Mohamoud, P. Lee, K. Berry, M. B. Young, T. Utterback, J. Weidman, W. C. Nierman, I. T. Paulsen, K. E. Nelson, H. Tettelin, S. L. O'Neill und J. A. Eisen. **Phylogenomics of the reproductive parasite *wolbachia pipientis* wmel: a streamlined genome overrun by mobile genetic elements.** PLoS Biol, 2:E69, 2004

C. Esser, N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, D. Bryant, M. Steel, P. Lockhart, D. Penny, und W. Martin. **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.** Mol Biol Evol, 21:1643–60, 2004.

C. Esser, W. Martin und T. Dagan. **The origin of mitochondria in light of a fluid prokaryotic chromosome model.** Biol Lett, 3:180–184, 2007.

C. Esser und W. Martin. **Supertrees and symbiosis in eukaryote genome evolution.** Trends Microbiol, 15:435–437, 2007.

# Tagungsbeiträge

**Lateral gene transfer at the infraspecific level: gene distributions among nine *E. coli* strains**
Christian Esser, Tal Dagan, Eric Bapteste and William Martin
SFB TR1 2008, Martinsried (Germany)

**Lateral gene transfer at the infraspecific level: gene distributions among nine *E. coli* strains**
Christian Esser, Tal Dagan, Eric Bapteste and William Martin
SMBE 2008, Barcelona (Spain)

**How $\alpha$-proteobacterial are $\alpha$-proteobacterial Genoms**
Christian Eßer, William Martin and Tal Dagan
Vereinigung für Allgemeine und Angewandte Mikrobiologie VAAM 2007, Osnabrück (Germany)

**The origin of mitochondria in light of a fluid prokaryotic chromosome model**
Christian Esser, Tal Dagan and William Martin
SMBE 2007, Halifax (Canada)

**A genome phylogeny for mitochondria and eubacterial ancestry of nuclear yeast proteins**
Christian Eßer, Nahal Ahmadinejad and William Martin
FEBS Advanced Lecture Course 2005, Wildbad Kreuth (Germany)

# 1 Zusammenfassung

Lateraler Gentransfer (LGT) ist ein wichtiger Mechanismus der natürlichen Variation in Prokaryoten. In dieser Arbeit wurde die Häufigkeit und die Charakteristik des Genaustauschs zwischen verschiedenen Prokaryoten und der Einfluß von LGT auf die Genomevolution von Prokaryoten am Beispiel von $\alpha$-Proteobakterien und verschiedenen Stämmen des $\gamma$-Proteobakteriums *Escherichia coli* untersucht.

Die Ableitung von lateralem Gentransfer hat auch Auswirkungen auf die Abschätzung der Herkunft von kernkodierten Genen bei Eukaryoten. Daher wurden die Auswirkungen des LGT auf die nächsten freilebenden Verwandten der heutigen Mitochondrien, die $\alpha$-Proteobakterien, analysiert. Um den Anteil $\alpha$-proteobakterieller Gene in $\alpha$-Proteobakterien zu bestimmen wurden die nächsten Nachbarn aller Gene aus 18 Genomen von $\alpha$-Proteobakterien über Sequenzähnlichkeit und phylogenetische Analysen bestimmt. Zwei Drittel der 47.143 untersuchten Gene hatten einen nächsten Nachbarn in Prokaryoten außerhalb der eigenen Gattung. Von diesen wiesen 7 % bis 36 % einen nächsten Nachbarn außerhalb der Klasse der $\alpha$-Proteobakterien auf. Diese Ergebnisse wurden mit der taxonomischen Verteilung nächster Nachbarn in sechs Mitochondriengenomen verglichen. Aufgrund ihrer $\alpha$-proteobakteriellen Herkunft wird von Genen mitochondrialer Abstammung oft ein nächster Nachbar innerhalb der $\alpha$-Proteobakterien erwartet. Die Untersuchung zeigte, daß die Zusammensetzung der nächsten Nachbarn beider Datensätze sehr ähnlich ist. Der Anteil von Genen mit nächsten Nachbarn außerhalb der $\alpha$-Proteobakterien deutet darauf hin, daß lateraler Gentransfer einen großen Einfluß auf die Genomevolution von Prokaryoten hat. Das spricht dafür, daß es sich bei Prokaryotengenomen nicht um beständige Einheiten, sondern um variable Gensammlungen handelt.

Nachdem gezeigt wurde, wie hoch der Einfluß von LGT auf der Ebene der taxonomischen Klasse war, wurde anhand von neun Stämmen von *Escherichia coli* ermittelt, wie sich Gentransfers innerhalb einer Art auswirken. Das geschah mithilfe von LGT-Netzwerken, die Gentransfers basierend auf Genverteilungsmustern und

einer Baumtopologie als Referenz ableiten. Alle Gene der neun Stämme wurden dabei nach Sequenzähnlichkeit in 6.688 Genfamilien eingeteilt. Die Abschätzung der Genomgröße der Vorfahren wurde dabei als Kriterium für die Häufigkeit von Gentransfers während der Evolution von *E. coli* benutzt. Für 27 % aller Proteinfamilien wurde ein einzelner Gentransfer abgeleitet. Ergebnisse die mit dieser Methode ermittelt werden sind stark abhängig von der Korrektheit der der Berechnung zugrundeliegenden Genfamilien. Dafür wurde ein neuer Algorithmus entwickelt der Gene in syntenisch orthologe Gene (SOGs) einteilt. Dabei wird nicht nur Sequenzähnlichkeit, sondern auch die Position des Genes innerhalb des Genoms berücksichtigt. Dieser Ansatz wurde benutzt um „wahre“ Orthologe zu identifizieren. Als Häufigkeit von Gentransfers in *E. coli* wurde mithilfe der SOGs ein Wert von einem Gentransfer in 23 % aller Proteinfamilien ermittelt.

Abschließend wurde der relative Einfluß von Punktmutationen und Rekombinationen auf die Genevolution in 14 Stämmen von *E. coli* abgeschätzt. Für einen Konsensusbaum aus *maximum-likelihood*-Bäumen für 2.665 universelle Proteinfamilien wurde ein Verhältnis von 3:1 für Substitutionen gegenüber Rekombination abgeleitet. Um die Robustheit dieser Analyse zu überprüfen wurde die Analyse mit anderen Referenz-Bäumen wiederholt. Für einen Baum basierend auf rRNA-Sequenzen änderte sich das Verhältnis auf 3:2 und mit einem zufälligen Baum wurde das umgekehrte Verhältnis wie für den Konsensusbaum bestimmt, nämlich 1:3. Das deutet darauf hin, daß die Ableitung stark abhängig von der Verwendung der Referenztopologie war. Aufgrund unterschiedlicher Ansätze sind Literaturvergleiche schwierig. Der Einfluß der Rekombination wurde hier wesentlich niedriger eingeschätzt.

# 2 Abstract

Lateral gene transfer (LGT) is an important mechanism of natural variation among prokaryotes. In this thesis the frequency and characteristics of gene exchange among different prokaryotes and the impact of LGT on prokaryotic genome evolution have been studied using the examples of $\alpha$-proteobacteria and several strains of *Escherichia coli*.

The inference of lateral gene transfer in prokaryotes also has implications on phylogenetic reconstruction of the origin of eukaryotic nuclear encoded genes. Therefore the effect of LGT on inference of the closest free-living relative of the mitochondrial ancestors, namely the modern $\alpha$-proteobacteria was studied. To estimate the proportion of $\alpha$-proteobacterial genes within alphaproteobacterial species, nearest neighbors for each gene in 18 $\alpha$-proteobacterial genomes were inferred by a similarity approach and by phylogenetic reconstruction. About two thirds of the 47,143 genes inspected had their nearest neighbor within the prokaryotes, outside of their own genus. Of these, between 7 % and 36 % of the proteins in each genome had their nearest neighbor outside of the class of $\alpha$-proteobacteria. These results were compared to the taxonomic distribution of the nearest neighbors to mitochondrial encoded proteins of six eukaryotes. Because of its $\alpha$-proteobacterial ancestry, mitochondrial genes are often expected to have an $\alpha$-proteobacterial nearest neighbor. The analysis showed that the distribution of nearest neighbors is very similar in both datasets. The amount of genes that showed a nearest neighbor outside their own class, and even outside their own phylum suggest that LGT has a big effect on prokaryotic genome evolution. This indicates that prokaryotic genomes are fluid and not static over time.

After showing the impact of LGT on the level of the taxonomic class, the influence of gene transfers within the species level was inspected by analyzing nine strains of *E. coli*. This was done using LGT-Networks which infer gene transfers based on gene distribution patterns and a reference tree topology. All genes within the nine

strains were clustered by similarity into 6.688 protein families. Ancestral genome size was used as a criterion for the frequency of gene transfers during *E. coli* evolution. For 27% of all protein families a single transfer event was inferred. LGT inference using this approach depends on the accuracy of the protein clustering algorithm in identifying orthologous genes. Here a novel algorithm is presented to sort the proteins into families by syntenic orthologues genes (SOGs). In this algorithm, not only sequence similarity is taken into account but also the physical position of the gene within the genome. This approach was used to sort all proteins into families including only „true“ orthologs. The inference of the frequency of gene transfers within E. coli for the SOGs resulted in 23% of the families whose evolution included at least one transfer event.

Finally the relative impact of point mutation and recombination on gene evolution within 14 strains of *E. coli* was estimated. Based on a consensus tree calculated from maximum-likelihood trees for 2.665 universal gene-families as reference, a ratio of 3:1 for substitution vs. recombination was inferred. To test the robustness of this inference the analysis was repeated using different reference tree topologies. Using a tree reconstructed from rRNA the ratio changed to 3:2 and using a random tree topology the ratio reversed to 1:3. This indicates that the inference is highly dependent on the reference tree topology. Comparison to other studies in the literature is difficult due to differences in the approaches in use. The influence of recombination in this study was inferred at lower levels.

# 3 Einleitung

## 3.1 Darwin und der Stammbaum des Lebens

Auch wenn das Bild eines Stammbaums schon vor Darwin existierte, so ist Darwins „*Tree of Life*“ doch das bekannteste Beispiel, wenn es um die Einteilung der lebenden Welt in systematische Einheiten geht (Darwin, 1859). Bis heute ist der Baum das System, das intuitiv verwendet wird, um Verwandtschaften und Abstammungen der Arten darzustellen. Seine Denkweise darüber fasste Charles Darwin in dem Satz zusammen:

> „As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.“

Die bifurzierende Struktur eines Stammbaumes zerteilt die Diversität der belebten Natur in immer kleinere Einheiten und macht somit deren Komplexität begreifbar. Für Darwin eignete sich der Stammbaum optimal, da sich die Aspekte seiner Evolutionstheorie nämlich Variation und Selektion in diesem Modell sehr gut darstellen lassen.

Die Konzepte Darwins werden weiterhin grundlegend akzeptiert, sind aber immer weiter den Ergebnissen der Forschung angepasst worden. So wird die Vorstellung das jede Variation einen Vor- oder Nachteil hat weitgehend entkräftet (Kimura, 1968; King und Jukes, 1969). Das Modell der neutralen Theorie besagt, daß der größte Anteil von Mutationen durch einen zufälligen Effekt fixiert wird, und keinerlei Auswirkungen auf den Organismus hat. Ein sehr kleiner Teil hat einen schwachen vorteilhaften Effekt, was aber insgesamt vernachlässigt werden kann. Allerdings haben ebenfalls seltene Mutationen die einen negativen Effekt auf die Anpassung des Organismus haben wesentlich stärkere Auswirkungen auf den Selektionsdruck

(Kimura, 1991). Der wesentliche Unterschied zu Darwins Modell ist die Position: Nicht der fitteste überlebt, sondern der am wenigsten fitte überlebt nicht („*purifying selection*“).

Die Baum-Metapher ist grundsätzlich auch dazu geeignet, biologische Prozesse wiederzugeben. Schließlich geht die Vererbung nur in eine Richtung, und die Nachkommen eines Individuums, ob Tier, Pflanze oder Einzeller bilden jeweils neue Einheiten, in denen durch Mutationen Variationen auftreten können, die schließlich wieder zu neuer Variation führen können. Der Stammbaum, meistens in der Variante eines bifurzierenden Baumes, hat sich bis heute in der Biologie als das Mittel der Wahl zur Darstellung von Verwandtschaftsbeziehungen erhalten. Die tiefste Verzweigung in diesem Baum des Lebens ist sicherlich die zwischen den Archaebakterien und den Eukaryoten (Doolittle, 1998; Mayr, 1998).

Während Einigkeit darüber herrscht, daß die Einflüsse, die zu Mutationen und Lesefehlern während der Replikation von Chromosomen führen, zwischen Eukaryoten und Prokaryoten keine wesentlichen Unterschiede aufweisen, ist das bei den unterschiedlichen Formen der Rekombination anders. Bei den Eukaryoten sorgt die Meiose für einen Austausch und eine Vermischung von Allelen zwischen den Individuen einer Art. Bei den Prokaryoten ist das nicht der Fall, der Genaustausch erfolgt hier häufig auch über die Artgrenzen hinweg (Bapteste et al., 2009).

## 3.2 Wenn der Baum kein Baum mehr ist: Endosymbiose

Solange nur die Vererbung mittels Zellteilung, beziehungsweise die Abstammung direkt verwandter Individuen betrachtet wird, reicht die Baum-Metapher für die Darstellung von Verwandtschaften aus. Allerdings ist die Natur wesentlich komplexer. Beispielsweise haben alle bekannten eukaryotischen Organismen eine Gemeinsamkeit, nämlich Organellen endosymbiontischer Herkunft: Mitochondrien und bei den Pflanzen zusätzlich Plastiden (Embley und Martin, 2006).

Bei der Endosymbiose nimmt eine Wirtszelle eine andere Zelle auf. Diese wird jedoch nicht verdaut, sondern bildet eine funktionelle Einheit mit dem Wirt. Über den genauen Hergang dieses Prozesses gibt es im Zusammenhang mit dem Ursprung

der Mitochondrien vielfältige Theorien (Martin et al., 2001). Auch darüber, welche Organellen auf Endosymbiosen zurückgehen, herrscht nicht immer Einigkeit. Während Theorien, nach denen Geißeln endosymbiontischen Ursprungs sind (Sagan, 1967; Margulis et al., 2000; Margulis, 1981) weitgehend als widerlegt gelten, ist die Endosymbiose für Mitochondrien und Chloroplasten weitgehend akzeptiert. Bei den Mitochondrien und deren Verwandten, den Hydrogenosomen und den Mitosomen wird weitgehend akzeptiert, daß sie auf ein einzelnes Endosymbioseereignis zurückgehen.

Die Chloroplasten der Pflanzen und vieler Algen gehen auf die Endosymbiose eines eukaryotischen Wirts mit einem Cyanobakterium zurück (Reyes-Prieto et al., 2007). Allerdings gibt es nicht nur Endosymbiosen bei denen es sich bei dem Symbionten um einen Prokaryoten handelt. Zahlreiche Algengruppen besitzen im Gegensatz zu den eben beschriebenen primären Plastiden sogenannte sekundäre Plastiden, bei denen der Wirt eine Alge aufgenommen hat und diese sich im Laufe der Zeit zum Organell umgewandelt hat (Taylor, 1974; Reyes-Prieto et al., 2007). Während die Entstehung der primären Plastide ein einmaliges Ereignis während der Evolution war, entstanden sekundäre Plastiden in mehreren Verwandtschaftsbereichen unabhängig voneinander (Gould et al., 2008). Im Laufe der Entwicklung sekundärer Plastiden werden die Bestandteile der endosymbiontischen Wirtszelle reduziert. Der Zellkern wurde zurückgebildet (Archibald, 2007) oder verschwand ganz (Gould et al., 2008). Es existieren ebenfalls tertiäre Plastiden, die durch die Aufnahme einer Alge mit sekundärer Plastide entstanden (Yoon et al., 2005).

Allerdings ist es in allen Fällen ein langer Weg von der Endosymbiose bis zum Organell. Die ursprünglich freilebenden Bakterien verlieren Gene, die in einem Prozeß, der endosymbiontischer Gentransfer (EGT) genannt wird, in das Genom des Wirts übertragen werden (Timmis et al., 2004; Martin, 2003). Die von diesen Genen kodierten Proteine werden nun vom Wirt exprimiert und müssen über komplexe Transportmechanismen in das Organell transportiert werden (Neupert und Herrmann, 2007; Mokranjac und Neupert, 2009). Als Signal, daß sie in das Organell transportiert werden müssen, tragen solche Proteine N-terminale Transitpeptide, die als Adresse fungieren (Chacinska et al., 2009; Kleine et al., 2009).

Die Endosymbiose ist ein Prozeß, der grundsätzlich den gängigen Vorstellungen der baumartigen Abstammung widerspricht. Insbesondere der EGT hat dabei einen Einfluß auf das Modell des phylogenetischen Baumes für die Verwandtschafts-

beziehung lebender Organismen. Die Verschmelzung mehrerer ursprünglich unabhängiger Organismen zu Einem und die Zusammenführung der Genome haben zur Folge, daß wir in eukaryotischen Genomen Gene unterschiedlicher Herkunft nebeneinander vorfinden (Esser et al., 2004). Trotz der langen Zeit, die seit der Etablierung der Endosymbiosen vergangen ist, lassen sich die Spuren von Genen, die durch endosymbiontischen Gentransfer in das Kerngenom gelangten, auch noch in heutigen Genomen wiederfinden (Deusch et al., 2008).

Diese Tatsache hat dazu geführt, daß neue Denkmodelle entwickelt wurden, um diese Vorgänge in der frühen Evolution der Eukaryoten darzustellen. Zumeist wurde der weiterhin zugrundeliegende Baum durch Querverbindungen zwischen $\alpha$-Proteobakterien und Archaeen auf dem Weg zu den Eukaryoten, sowie zwischen Cyanobakterien und Eukaryoten auf dem Weg zu den Pflanzen ergänzt (Martin et al., 2003; Doolittle, 1999). Zum anderen wurde der „*ring of life*“ vorgeschlagen, um die Aufspaltung in Archaea und Bakterien mit einer anschließenden Verschmelzung zu den Eukaryoten in einem einfachen Modell zu verbinden (Rivera und Lake, 2004)

## 3.3 Vom Baum zum Netzwerk: Lateraler Gentransfer bei Prokaryoten

Die Entstehung von Mitochondrien und Plastiden hatte einen beträchtlichen Einfluß auf die Zusammensetzung eukaryotischer Genome. Allerdings handelt es sich hierbei nur um äußerst seltene – einmalige Ereignisse. Bei den Prokaryoten gibt es Mechanismen, die ständig auf die Genome einwirken. Transformation (die Aufnahme freier DNA), Transduktion (die über Viren vermittelte Übertragung von Genen) sorgen für einen Austausch von Genen zwischen verschiedenen (nicht direkt verwandten) Arten (Frost et al., 2005). Bei der Konjugation erfolgt der Gentransfer zwischen Donor und Empfänger über eine Plasmabrücke (Thomas und Nielsen, 2005).

Diese Mechanismen sorgen dafür, daß es in Prokaryoten zu einer Vermischung von Genen abseits der vertikalen Vererbung kommt. Diese Erkenntnis hat dafür gesorgt, daß der Stammbaum als Sinnbild für die Evolution Konkurrenz bekommen hat. Für eine Darstellung der evolutionären Prozesse, die hierbei vonstatten gehen, werden die Bäume durch die Hinzufügung von Querverbindungen immer weiter zu Netzwerken ausgebaut.

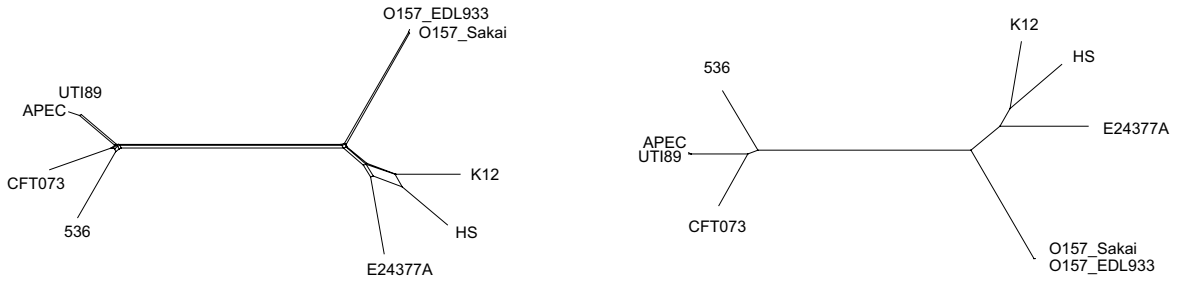



**Abbildung 3.1:** Vergleich zwischen einer Netzwerk- und einer Baumdarstellung am Beispiel einer Phylogenie von neun *E. coli* Genomen. Beide Darstellungen beruhen auf den selben Daten (Einem konkatenierten Alignment für universelle Proteinfamilien), und beide wurden mit einem *neighbor-joining*-Algortihmus berechnet (Huson und Bryant, 2006). Während in der `NeighborNet` Darstellung auf der linken Seite deutlich wird das es sowohl eine Unterstützung für einen Split zwischen *E. coli* HS und *E. coli* E24377A gibt als auch für den zwischen *E. coli* HS und *E. coli* K12, wird bei der Erstellung des Baumes eine Reihenfolge der Verzweigungen festgelegt. Der Widerspruch in den Daten wird nicht deutlich.

Vorgeschlagen wurden daher komplexere Modelle. Dabei werden sehr baumähnliche Topologien mit wenigen Querverbindungen vorgeschlagen (Brown, 2003; Doolittle, 1999). Alternativ präsentierte Martin 1999 eine fächerartige Veranschaulichung mit zahlreichen netzartigen Verbindungen zwischen den Prokaryoten aus dem die Eukaryoten nach Abschluß der Endosymbiosen baumähnlich hervorgehen. Bei dem „*net of life*“ (Huang und Gogarten, 2006) handelt es sich ebenfalls um einen Baum mit überlagerten Querverbindungen, allerdings legen die Verzweigungen hier nahe, daß Aufspaltungen erst recht spät erfolgen. Methoden die statt der Baumdarstellung, die eines Netzwerkes benutzen (siehe Abb. 3.1) sind in der Lage auch widersprüchliche phylogenetische Signale darzustellen.

## 3.4 Genomdaten und Gentransfer

Trotz der bereits beschriebenen Problematik, evolutionäre Prozesse in die Form eines bifurzierenden Baumes zu pressen, ist die Vorstellung eines „*tree of life*“ immer noch sehr verlockend. Während dies zunächst auf Basis der 16S rRNA geschah (Woese et al., 1990), wurden mit der Veröffentlichung von immer mehr vollständig sequenzierten Genomen Versuche gestartet, den Genbaum des Lebens durch einen Genombaum des Lebens zu ersetzen. Dabei existieren zwei Ansätze.

Zum einen wurde versucht universelle Proteine zur Erstellung eines „Genombaumes“ zu verwenden (Ciccarelli et al., 2006). Dabei trat das Problem auf, daß mit einer zunehmenden Anzahl berücksichtigter Genome die Menge der Gene, die homologe Sequenzen in allen Organismen aufwiesen, schrumpfte, so daß der endgültigen Phylogenie nur eine Datenbasis von 31 Genen zugrunde lag (Dagan und Martin, 2006). Ein weiterer Ansatz die Form des Baumes beizubehalten, ist der Ansatz sogenannte *Supertrees* zu verwenden. Dabei handelt es sich um Konsensusbäume, bei denen es allerdings nicht erforderlich ist, daß jeder Baum alle untersuchten Taxa enthält (Pisani et al., 2007). Diese Bäume beruhen dementsprechend auf einer wesentlich größeren Datenbasis, sind allerdings nur in der Lage das stärkste Signal darzustellen (Esser und Martin, 2007).

Das führt dazu, daß nach Methoden lateralen Gentransfer zu identifizieren gesucht wurde, um seinen Einfluß auf die Genomevolution zu bestimmen, oder ihn bei phylogenetischen Analysen zu berücksichtigen. Es gibt dabei drei prinzipielle Ansätze die auf unterschiedlichen Merkmalen übertragener Sequenzen beruhen (Gogarten et al., 2008; McInerney et al., 2008).

Jedes Genom hat charakteristische Eigenschaften, beispielsweise einen bestimmten Gehalt an Guanin und Cytosin (Nakamura et al., 2004) oder unterschiedliche Codonpreferenz (Ochman et al., 2000). Daher gibt es als erste Gruppe von Methoden zur Identifizierung von lateral transferierten Genen solche, die nicht auf Phylogenien beruhen. Diese untersuchen entweder die Sequenz-Zusammensetzung auf bestimmte Merkmale (Lawrence und Ochman, 1997, 1998) oder aber die Genverteilung in sehr eng verwandten Arten (Welch et al., 2002). Diese Methoden sind nur in der Lage kürzliche Gentransfers zu identifizieren, da die aufnehmenden Genome dazu tendieren aufgenommenen Gene ihre Merkmale anzueignen (Lawrence und Ochman, 1997). Da sich solche Merkmale aber auch innerhalb eines Genoms unterscheiden können, führen solche Methoden teilweise zu falsch positiven beziehungsweise falsch negativen Identifikationen (Azad und Lawrence, 2005; Koski et al., 2001).

Ein weiterer Ansatz für die Identifikation von lateralem Gentransfer eignet sich auch für weiter zurückliegende Gentransferereignisse und bedient sich dabei Genverteilungsmustern von Genfamilien über verschiedene Arten. Diese werden mit einer Referenzphylogenie verglichen, und Genverteilungen die sich nicht mit der Phylogenie erklären lassen, werden als Indikator für einen Gentransfer betrachtet (Snel et al., 2002; Mirkin et al., 2003). Eine derartige Methode wurde zur Untersuchung

von 51 Genomen angewendet und fand unter Verwendung des 16S rRNA-Baumes für 18 - 41 % aller Genfamilien einen Hinweis auf lateralen Gentransfer (Kunin und Ouzounis, 2003*b*,*a*). Da hierbei Transfers nur erfasst werden, wenn es sich bei dem transferierten Gen um eines handelt, das im Zielorganismus noch nicht vorhanden war, bildet die Abschätzung eher eine untere Schranke (Gogarten et al., 2008).

Die dritte Methode benutzt nicht genetische Verteilungen sondern Vergleiche zwischen Genbäumen. Dabei werden ebenfalls Genfamilien benötigt, für die jeweils phylogenetische Bäume erzeugt werden. Anschließend werden die Topologien der einzelnen Bäume verglichen – zumeist mit einer Referenztopologie. Hier gelten widersprüchliche Topologien als Indikator für lateralen Gentransfer (Lerat et al., 2005; Beiko et al., 2005). Für einen Datensatz aus cyanobakteriellen Genfamilien wurde auf diese Weise für 61 % der untersuchten Genfamilien ein Gentransfer abgeleitet (Zhaxybayeva et al., 2006).

## 3.5 Zielsetzung

Vor diesem Hintergrund war das Ziel der vorliegenden Arbeit, den Einfluß des lateralen Gentransfers auf unterschiedlichen Ebenen der Genomevolution zu beleuchten. Im Zusammenhang mit der Analyse von Mitochondrien und Chloroplasten stellt sich oft die Frage, welche Gene durch endosymbiontischen Gentransfer in das Wirtsgenom gelangt sind. Da widersprüchliche Interpretationen der Ergebnisse von Sequenzvergleichen existieren, soll in Kapitel 5 „Wie $\alpha$-proteobakteriell sind $\alpha$-Proteobakterien?“ analysiert werden, wie realistisch die Annahmen sind, die solchen Analysen zugrunde liegen. Insbesondere soll dabei der Einfluß von lateralem Gentransfer (LGT) zwischen den verschiedenen Prokaryoten untersucht werden.

Im Laufe der Evolution spielen Genduplikationen eine wichtige Rolle. In Kapitel 6 „Identifizierung konservierter Genreihenfolgen bei Escherichia coli und Verwandten“ soll darauf eingegangen werden, ob es möglich ist bessere Rekonstruktionen von Proteinfamilien zu erreichen, indem neben Sequenzvergleichen auch Positionsinformationen in die Berechnung einbezogen werden, um Differenzierungen zwischen Paralogen zu ermöglichen.

Nachdem der LGT bereits auf der Ebene von verschiedenen prokaryotischen Gattungen untersucht wurde, soll in Kapitel 7 „Der Einfluß der Genomgröße auf die

Ableitung von Genaustauschen bei Escherichia coli“, am Beispiel von neun Stämmen von *Escherichia coli*, eine Abschätzung gemacht werden, ob lateraler Gentransfer auch zwischen verschiedenen Stämmen einer Art eine Rolle spielt. Mit Hilfe von LGT-Netzwerken sollen bevorzugte Transferpartner sichtbar gemacht werden. Hierbei soll untersucht werden, ob sich die Ergebnisse für die mit der in Kapitel 6 entwickelten Methode erstellten Genfamilien von denen mit einer klassischen Methode erstellten, unterscheiden.

Abschließend sollen für 14 Stämme des $\gamma$-Proteobakteriums *Escherichia coli* im Kapitel 8 „Rekombination in *E. coli*“ Substitutions- und Rekombinationsraten bestimmt werden, um zu ermitteln, wie groß der Einfluß von Rekombination auf die Genomevolution ist.

Die Arbeit ist in die vier oben umrißenen Kapitel gegliedert in denen jeweils zusätzlicher Hintergrund, Übersichten zur Methodik und die Ergebnisse erläutert werden. Publizierte Methoden die die Grundlagen der Analysen bilden werden dagegen in Kapitel 4 aufgeführt. Abschließend folgt die Diskussion in Kapitel 9.

# 4 Material und Methoden

Im folgenden Teil werden die grundlegenden Programme beschrieben, die in mehreren Abschnitten der Arbeit benutzt werden, sowie allgemeine Informationen zu den Daten, die die gesamte Arbeit betreffen. Die Beschreibung der selbst erstellten Programme und Skripte folgt im entsprechenden thematischen Kontext.

## 4.1 Daten

Für die vorliegende Arbeit wurden prokaryotische Genomdaten vom National Center for Biotechnology Information (NCBI) bezogen. Diese sind dort frei verfügbar und auf der FTP Seite des Institutes[1] herunterzuladen. Da sich die verfügbaren Sequenzdaten sehr schnell entwickeln, wurden zu bestimmten Zeitpunkten aktualisierte Versionen hinzugezogen (Tabelle 4.1).

## 4.2 Programme

### 4.2.1 BLAST

Als Grundlage für zahlreiche Arbeitsschritte wurden die Ergebnisse von automatisierten Datenbanksuchen verwendet. Diese wurden mit dem `BLAST`-Algorithmus durchgeführt (Altschul et al., 1997).

Der Algorithmus geht in mehreren Schritten vor und ermöglicht eine sehr schnelle Datenbanksuche. Exemplarisch soll hier der `blastp`-Algorithmus für die Suche einer Proteinsequenz in einer Proteindatenbank erläutert werden, da dieser am häufigsten verwendet wurde:

[1] `ftp://ftp.ncbi.nih.gov/Genoms/Bacteria`

**Tabelle 4.1:** Die verschiedenen in dieser Arbeit genutzten Datensätze. Details dazu werden in den einzelnen Kapiteln erwähnt und finden sich zusätzlich im Anhang auf Seite 104.

| Daten | Anzahl der Genome | Sequenzen | Stand | Verwendung |
|---|---|---|---|---|
| Prokaryoten | 231 | 676.476 | 04/2005 | Referenzgenome und Datenbank für Kapitel 5 |
| Prokaryoten | 616 | 616 Genom-sequenzen | 04/2008 | Datenbank für Genom-verankerte Identitätsplots in Kapitel 6 |
| *Escherichia coli* | 9 | 43.358 | 01/2008 | Synteniedatensatz für Kapitel 6 und 7 |
| *Escherichia coli* | 14 | 65.819 | 10/2008 | Synteniedatensatz für Kapitel 8 |

Die BLAST-Suche beginnt mit einer Indexsuche, bei der die Suchsequenz in Abschnitte der Länge w (für Proteinsequenzen standardmäßig drei) aufgeteilt wird. Für jedes dieser w-mere wird nach einer Ähnlichkeitsmatrix (bei Nukleotidsequenzen nach einer Identitätsmatrix) eine Bewertung errechnet (engl. *score*). Die Bewertungen geben die Qualität der untersuchten Treffer wieder. Alle *scores* über einem Schwellenwert T werden als Treffer gewertet. Es werden nur Sequenzen weiter betrachtet, für die innerhalb eines Fensters der Länge A zwei solche Treffer auf einer Diagonalen (in einem imaginären Dotplot) liegen (engl. *two-hit*-Methode).

Findet BLAST zwei Treffer auf einer Diagonalen, werden diese jeweils in beide Richtungen soweit ausgedehnt, bis sich der aufsummierte *score* nicht mehr erhöhen läßt. Liegt der *score* dann über einem Schwellenwert S, so werden sie als hochbewertetes Paar (engl. *high scoring pairs*, kurz HSP) bezeichnet. Im weiteren Verlauf wird versucht den *score* zu erhöhen, indem mehrere HSPs verknüpft werden. Ist es nicht möglich mehrere HSPs zu verknüpfen, so gibt BLAST mehrere Treffer innerhalb einer Sequenz aus.

Die Schnelligkeit von BLAST beruht auf einer Heuristik, die bei der Indexsuche Verwendung findet. Dadurch, daß hier mit sehr kurzen Sequenzabschnitten gearbeitet wird, muß die Datenbank für die Indexsuche nicht vollständig gelesen werden. Für die Standardlänge von drei Aminosäuren existieren also nur $20^3 = 8000$ mögliche w-mere. Für diese kann gespeichert werden, welche Kombinationen zu einem positiven *score* führen würden, und an welchen Stellen sie in der Datenbank vorkommen. Werden zwei dieser Treffer, die sowohl in der Referenz als auch in der Datenbank denselben

Abstand haben, gefunden (sie liegen also in einem gedachten Dotplot auf einer Diagonale), bilden sie den Ausgangspunkt für eine genauere Suche. Der Großteil der Datenbank, für den dieses Kriterium nicht zutrifft, muß also nicht betrachtet werden.

### 4.2.2 `ClustalW`

Multiple Alignments wurden mit dem Programm `ClustalW` erstellt (Thompson et al., 1994). `ClustalW` erstellt zunächst einen Führungsbaum (engl. *guiding tree*). Dieser wird durch die Berechnung paarweiser Distanzen zwischen den Sequenzen und die anschließende Anwendung eines schnellen Algorithmus (UPGMA) berechnet. Anhand dieses Baumes werden die Sequenzen aligniert. Zunächst wird mit den Sequenzen mit der geringsten Distanz begonnen. Aus diesen wird ein Profil gebildet. Eine erneute Berechnung der Distanzen bestimmt das nächste zu alignierende Paar. Dabei kann das zuvor erstellte Profil ebenso wie eine Sequenz benutzt werden.

In einem Zwischenschritt eingefügte Lücken (engl. *gaps*) werden in die Profile eingetragen und verbleiben über alle Zwischenschritte bis zum endgültigen Alignment, auch wenn sich später herausstellen sollte, daß die ursprünglich gewählte Position unvorteilhaft ist. Dieses Vorgehen macht den Algorithmus anfällig für die falsche Positionierung von Lücken (Golubchik et al., 2007; Landan und Graur, 2007, 2009). Im Gegensatz zu den Algorithmen für paarweise Alignments (Needleman und Wunsch, 1970; Smith und Waterman, 1981) handelt es sich um einen heuristischen Algorithmus, bei dem die optimale Lösung nicht garantiert werden kann. Daher empfahlen die Autoren, `ClustalW` als Ausgangspunkt für ein manuelles Alignment zu benutzen.

### 4.2.3 `HoT`-Methode

Da ein manuelles Alignment in Zeiten, in denen die verfügbare Datenmenge exponentiell ansteigt nicht mehr tragbar ist, ist es wichtig, die Verlässlichkeit automatisch erstellter Alignments zu überprüfen.

Eine Methode, die dabei zunehmend Beachtung findet, ist die sogenannte `HoT`-Methode (engl. *heads or tails*) (Roettger et al., 2009; Landan und Graur, 2009), bei der die gewählte Methode zur automatischen Alignmenterstellung zweimal angewendet wird, einmal normal und einmal mit den Sequenzen in umgekehrter

Reihenfolge. Die beiden so erstellten Alignments werden miteinander verglichen. Als Bewertungsmaß wird ausgegeben, welcher Anteil der Alignmentpositionen sich korrekt (das heißt in beiden Alignments gleich) rekonstruieren ließ (engl. *column score*). Eine zweite Bewertung gibt den Anteil der korrekt rekonstruierten Sequenzpositionspaare (Nukleotide oder Aminosäuren) an (engl. *sum of pairs score*)

Im Gegensatz zu dem sogenannten *benchmarking*, bei dem die Daten einer begrenzten Datenbank (Thompson et al., 1999), für die zuvor optimale Alignments bestimmt beziehungsweise errechnet worden waren, mit einer bestimmten Methode berechnet werden, wird hier nicht ein grundsätzliches Urteil darüber gefällt, wie gut oder schlecht eine Methode ist. Stattdessen wird die Leistung einer Methode für den verwendeten Datensatz ermittelt. Das ermöglicht die Verwendung schneller Algorithmen für große Datensätze. Anschließend kann überprüft werden ob die Daten ausreichend verlässlich sind, um konkrete Aussagen zu treffen.

### 4.2.4 `PHYLIP`

Das `PHYLIP-Paket` (engl. *phylogeny inference package*) enthält zahlreiche Programme zur Erstellung phylogenetischer Bäume (Felsenstein, 2004). Hier wurden hauptsächlich die Programme `seqboot`, `protdist`, `neighbor` und `consense` benutzt, um *neighbor-joining*-Bäume mit und ohne Bootstrapping zu erstellen.

#### `seqboot`

Das Programm dient zur Erstellung von Datensätzen für die statistische Methode des Bootstrappens. Dabei wird ein einzelnes Alignment vervielfacht. Es werden m Replikate erstellt, die die gleiche Länge haben wie das Originalalignment, deren Inhalt aber durch zufälliges Ziehen von Spalten mit Zurücklegen aus dem ursprünglichen Alignment gebildet wird. Diese Methode kann benutzt werden, um statistische Aussagen über die Verlässlichkeit des Ergebnisses einer durchgeführten Methode zu treffen.

Nachdem mit jedem dieser Alignments die zu überprüfende phylogenetische Analyse durchgeführt wurde, wird mit dem Programm `consense`, überprüft wie oft das gleiche Ergebnis erzielt werden konnte. Dabei wird ein Konsensusbaum erstellt, in dem für jeden Ast angegeben wird, wie oft er in den erstellten Bäumen auftrat. Wird dieser

Wert als Anteil angegeben, wird er als Bootstrap-Wert bezeichnet. Ist dieser Wert hoch, so handelt es sich um ein Ergebnis, bei dem es unwahrscheinlich ist, daß es durch Zufall entstanden ist.

#### dnadist

Das Programm `dnadist` erstellt zu einem gegebenen Alignment eine Distanzmatrix mit den paarweisen Distanzen der beteiligten Sequenzen.

#### protdist

Das Programm `protdist` erstellt zu einem gegebenen Alignment eine Distanzmatrix mit den paarweisen Distanzen der beteiligten Sequenzen. Standardmäßig wird dabei die Jones-Taylor-Thornton-Matrix (Jones et al., 1992) verwendet, bei der es sich um eine Weiterentwicklung der Dayhoff / PAM Matrizen handelt (Dayhoff et al., 1978). Sie wurde mit einer ähnlichen Methode erstellt, allerdings basierend auf einem größeren Datensatz. Die so erstellten Distanzmatrizen dienen als Ausgangspunkt für distanzbasierte phylogenetische Methoden (zum Beispiel UPGMA und *neighbor-joining*)

#### protml

Ausgehend von einem multiplen Sequenzalignment errechnet `protml` einen phylogenetischen Baum basierend auf der *Maximum-Likelihood* Methode (Felsenstein, 1981). Im Gegensatz zu der oben beschriebenen Distanzberechnung wird hier für jede einzelne Position eines Alignments eine Abschätzung gemacht wie wahrscheinlich ein gewisser Baum für diese Position ist. Hierbei wird von einer Starttopologie ausgegangen, für die eine Wahrscheinlichkeit berechnet wird. Durch den Austausch von Teilbäumen wird die Starttopologie abgewandelt. Dies geschieht solange bis die wahrscheinlichste Topologie (ausgehend von der Starttopologie) gefunden wurde. Diese Berechnungen sind in der Regel langwieriger als distanzbasierte Methoden. Die Funktionsweise von `protml` ist ähnlich der von `PHYML`.

**neighbor**

Das Programm `neighbor` implementiert die *neighbor-joining* Methode (Saitou und Nei, 1987). Hierbei wird basierend auf einer Distanzmatrix ein Baum erstellt, in dem alle Sequenzen zunächst in einer Sterntopologie angeordnet werden (alle Sequenzen sind direkt mit einem Knoten verbunden). Nun werden nacheinander alle Paare von Sequenzen aus dieser Topologie herausgenommen und an einen eigenen Ast gehängt. Die beiden Sequenzen, die bei diesem Verfahren zu der geringsten Gesamtdistanz des Baumes führen, werden als nächste Nachbarn bezeichnet und bilden den ersten Ast. Das Prozedere wird solange wiederholt bis jeder interne Knoten nur noch mit drei anderen Knoten verbunden ist (bifurzierender Baum).

**consense**

Das Programm `consense` erstellt Konsensus-Bäume (Adams III, 1972) mit Hilfe der erweiterten Mehrheitsregel (engl. *extended majority rule*). Hierbei werden zahlreiche einzelne Bäume als Eingabe akzeptiert, solange sie dieselben Speziesnamen enthalten. Das Ergebnis ist der sogenannte Konsensus-Baum, der möglichst viele der in den Einzelbäumen enthaltenen Daten repräsentiert. Anstatt Distanzen erhält jeder Ast als Länge die Anzahl der Bäume, in denen er vorkommt. In der zusätzlich erstellten Datei „*outfile*“ ist eine Liste enthalten, die für jeden vorkommenden Split (das heißt die Aufteilung der enthaltenen Arten in zwei Gruppen) auflistet, wie oft er in den Daten vorkam, und ob er in dem erstellten Konsensus-Baum berücksichtigt worden ist.

### 4.2.5 PHYML

Ähnlich dem `PHYLIP`-Programm `protml` dient `PHYML` (Guindon und Gascuel, 2003) der Erstellung von *maximum-likelihood* Bäumen (Felsenstein, 1973). Es verfügt über eine sehr ähnliche Oberfläche, läßt sich jedoch auch im Stapelverarbeitungsmodus (engl. *batch mode*) benutzen, was die Arbeit mit großen Datensätzen erleichtert. Außerdem ist es möglich weitere Funktionen direkt innerhalb des Programms auszuführen, zum Beispiel das Erstellen eines Startbaums oder die Anwendung von Bootstrapping (siehe Seite 16).

### 4.2.6 Perl

Die Programmiersprache Perl[2] ist eine Skriptsprache, die aufbauend auf Elementen von awk, C und Shellskripten von Larry Wall entwickelt wurde. Sie ist in der Bioinformatik sehr beliebt, da sie neben einer im Vergleich zu anderen Programmiersprachen flexiblen Syntax auch Möglichkeiten bietet, die viele Programmiersprachen vermissen lassen. So ist bei der Deklaration von Variablen keine Größe anzugeben, was es ermöglicht mit sehr großen Datenmengen zu arbeiten, ohne sich um die Verwaltung derselben kümmern zu müssen. Des Weiteren sind insbesondere die Ein- und Ausgaberoutinen optimal auf das Einlesen und Verarbeiten großer Textdateien zugeschnitten. Dies war der Zweck zu dem Perl ursprünglich entwickelt wurde.

Ein weiterer Vorteil ist, daß die Verwendung der Skripte keine Kompilierung erfordert. Dadurch ist die Geschwindigkeit der PERL-Skripte in der Regel langsamer als in anderen Programmiersprachen. Allerdings wird die Entwicklung von Programmen beziehungsweise die Anpassung existierender Programme an neue Aufgaben beschleunigt.

### 4.2.7 EMBOSS

Das EMBOSS-Paket (engl. *European Molecular Biology Open Software Suite*) ist eine frei verfügbare Programmsammlung, die Programme für viele bioinformatische Anwendungen enthält (Rice et al., 2000). Unter anderem sind auch die Programme des (in Kapitel 4.2.4 beschriebenen) PHYLIP-Paketes in einer an die Bedienungsweise der EMBOSS-Programme angepassten Version verfügbar, was die Verarbeitung großer Datenmengen erheblich erleichtert.

#### seqretsplit

Seqretsplit ist ein Programm das zur Aufspaltung großer (multi-) Fasta-Dateien in einzelne Dateien dient. Diese erhalten dabei einen auf der Annotation der Sequenz basierenden Dateinamen.

[2] http://www.perl.org

### 4.2.8 Markov-Cluster Algorithmus (MCL)

Der MCL Algorithmus wurde benutzt um Proteinfamilien zu bilden. Der Algorithmus wurde von Stijn van Dongen im Rahmen seiner Doktorarbeit entwickelt (Enright et al., 2002; van Dongen, 2000).

Das Programm mcl ist auch in der Lage die Identitäten aus einer BLAST- Datenbanksuche einzulesen und als Eingabe zu verarbeiten. Da BLAST jedoch lokale Identitäten ausgibt und ein Treffer auch mit 100 % bewertet werden kann, wenn er nur wenige Aminosäuren oder Nukleotide lang ist, wurde als Zwischenschritt ein paarweises globales Alignment durchgeführt. Dieser Schritt wurde mit ClustalW bzw. mit der für diesen Einsatz speziell modifizierten Variante des EMBOSS-Programmes needle durchgeführt. Diese so ermittelten paarweisen Identitäten bilden die Kanten in einem ungerichteten partiellen Graphen, in dem die Gene beziehungsweise Proteine enthalten sind.

### 4.2.9 LDDist

LDDist (Thollesson, 2004) berechnet aus multiplen Alignments Distanzen, die auf den Logarithmen der Determinanten der Divergenzmatrizen beruhen. Für diese Methode ist eine Abschätzung der invarianten Positionen empfehlenswert, die durch die Methode von Sidow et al. (1992) oder die von Steel et al. (2000) erfolgen kann.

Es wurde mit einer modifizierten Programmversion gearbeitet, die nicht nur mit den einzelnen Aminosäuren arbeitet, sondern diese auch in die Dayhoff-Klassen rekodieren kann. Die sechs Dayhoff-Klassen (im IUPAC Einbuchstabencode: ASTGP, DNEQ, RKH, MVIL, FYW, C) sind Aminosäureklassen, zwischen denen Aminosäureaustausche sehr wahrscheinlich sind. Die Rekodierung dient dazu sehr alte Verwandtschaften besser auflösen zu können. In dieser Arbeit wurde sowohl die Abschätzung der invarianten Positionen nach Sidow *et al.* als auch die Rekodierung in die Dayhoff-Klassen durchgeführt.

Als Eingabe wird ein multiples Sequenzalignment im CLUSTAL-Format benötigt. Die Ausgabe der Distanzen erfolgt im NEXUS-Format.

### 4.2.10 Splitstree

Splitstree (Huson und Bryant, 2006) ist ein Programm, das zahlreiche phylogenetische Daten zur Berechnung von phylogenetischen Netzwerken und Bäumen nutzen kann. Es existiert sowohl in einer Version für die Kommandozeile als auch mit einer javabasierten graphischen Oberfläche. In der graphischen Version erlaubt es auch komplette Analysen ausgehend von einem multiplen Alignment in einem Arbeitsschritt durchzuführen. Hier wurde Splitstree zur Darstellung und Aufarbeitung von NeighborNet Graphen verwendet.

### 4.2.11 NeighborNet

NeighborNet (Bryant und Moulton, 2004) ist ein Programm, das phylogenetische Netzwerke basierend auf der *neighbor-joining*-Methode erstellen kann. Die Vorgehensweise ist dabei identisch mit der bei *neighbor-joining* (siehe 4.2.4 auf Seite 18). Allerdings ist das Ergebnis kein Baum sondern ein phylogenetisches Netzwerk. Während in einem Baum nur ein einziges Signal dargestellt werden kann (in der Regel das stärkste) ermöglichen es Netzwerke auch widersprüchliche Signale darzustellen. Das ermöglicht es einen realistischeren Eindruck der Daten zu erlangen, weil bereits ohne statistische Analyse Widersprüche und Unsicherheiten offensichtlich werden (siehe Abb. 3.1).

### 4.2.12 Sortal

Das Programm sortal dient zum Sortieren von Alignments. Es weist jedoch auch zahlreiche Funktionen für die Umformatierung und andere Funktionen für die Bearbeitung von Alignments auf. In dieser Arbeit wurden insbesondere die Funktionen zur Erstellung von konkatenierten Alignments und zur Entfernung von Lückenpositionen verwendet. Des Weiteren war für einige phylogenetische Programme die Umformatierung der verwendeten Alignments notwendig.

### 4.2.13 MUMmer

Das Programm MUMmer ist ein Werkzeug für den schnellen Vergleich von ganzen Genomen (Delcher et al., 2003; Page, 2003). Es führt paarweise Vergleiche zwischen

zwei Genomen durch und sucht nur nach exakten Übereinstimmungen. Dadurch dauert das Alignment zweier Bakteriengenome beispielsweise nur wenige Sekunden.

### 4.2.14 MATLAB

Das kommerzielle Programm MATLAB bietet viele Möglichkeiten mathematische Berechnungen durchzuführen und große Datenmengen zu visualisieren.

## 4.3 Verwendete Dateiformate

### 4.3.1 Eingabeformate

Für das Einlesen von Ausgangsdaten wurden hauptsächlich Daten verwendet die aus den Datenbanken des *National Center for Biotechnolgy Information* (NCBI)[3] bezogen werden können, und in der Regel ohne Umformatierung in der weiteren Analyse verwendet werden können.

#### Das FASTA-Format

Das FASTA-Format ist eines der einfachsten Formate für die Speicherung von Sequenzdaten. Jeder Eintrag beginnt dabei mit einer Kopfzeile die mit einem „>“ eingeleitet wird. Darauf folgt der Bezeichner der Sequenz. In den folgenden Zeilen steht bis zum Beginn des nächsten Eintrages die Sequenzinformation. Das FASTA-Format kann ebenfalls für die Speicherung von Alignments benutzt werden.

FASTA-Dateien wurden in zwei Varianten benutzt, zunächst in der Formatierung wie sie in der Genbank-Datenbank[4] des NCBI verwendet wird, und in einer vereinfachten Formatierung, in der nur der Name des Organismus und die Identifikationsnummer der Sequenz in der Kopfzeile steht.

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

```
Escherichia coli str. K-12 substr. MG1655, complete genome - 1..4639675 4132 proteins
Location Strand Length PID Gene Synonym Code COG Product
190..255    +   21  16127995  thrL  b0001  -   -                  thr operon leader peptide
337..2799   +  820  16127996  thrA  b0002  -   COG0460E,COG0527E  fused aspartokinase I <...>
2801..3733  +  310  16127997  thrB  b0003  -   COG0083E           homoserine kinase
3734..5020  +  428  16127998  thrC  b0004  -   COG0498E           threonine synthase
5234..5530  +   98  16127999  yaaX  b0005  -   -                  predicted protein
5683..6459  -  258  16128000  yaaA  b0006  -   COG3022S           conserved protein
6529..7959  -  476  16128001  yaaJ  b0007  -   COG1115E           predicted transporter
```

**Abbildung 4.1:** Ausschnitt aus einer Proteintabelle von *E. coli K12*

### Das Newick-Format

Das Newick-Format ist das Standardformat zur Darstellung phylogenetischer Bäume in Textform. Die Teilbäume werden dabei durch Klammern gruppiert. Alle HTU's und OTU's werden durch Kommata getrennt. Zusätzlich zu der reinen Topologie (Reihenfolge der Verzweigungen) können auch Informationen über die Astlängen und eventuelle Bootstrap-Werte eingefügt werden.

### Die Proteintabellen

Das NCBI stellt neben den oben bereits erwähnten Sequenzen im FASTA-Format auch weiterführende Informationen zur Verfügung. Für diese Arbeit wurden dabei die Proteintabellen (Abb. 4.1) verwendet. Diese stehen in der Regel für alle vollständig sequenzierten und annotierten Genome zur Verfügung und enthalten grundlegende Informationen über jedes Protein wie etwa die Länge, die Funktion, unterschiedliche Bezeichnungen, sowie die Position auf dem Chromosom.

### Das Nexus-Format

Das Nexus-Format (Maddison et al., 1997) ist sehr variabel. In ihm lassen sich zahlreiche Informationen abspeichern. Einzelne Blöcke werden dabei durch BEGIN ... END Anweisungen getrennt. Die Blöcke können Sequenzen, Alignments, Distanzen, phylogenetische Bäume (angelehnt an das Newick-Format), Splits, sowie Daten zur Formatierung enthalten. Auch Daten zu Arbeitsabläufen können in entsprechende Abschnitte integriert werden. Das Nexus-Format ist das Standardausgabeformat von `Splitstree`.

#### Die BLAST-Ausgabe

Eine wichtige Datenquelle stellen die Ergebnisse von BLAST-Datenbanksuchen dar. Diese wurden in der tabellarischen Variante benutzt, da hier alle relevanten Ergebnisse sehr kompakt zusammengefasst sind (Abb 4.2). Die Ausgabe ist in Spalten gegliedert, wobei die ersten beiden Spalten die Sequenznamen für die Suchsequenz und den Treffer enthalten. Darauf folgen die Identität (in %), die Alignmentlänge, die Anzahl der Positionen, die keinen Treffer bilden und die Anzahl der eingefügten Lücken. Bei den folgenden vier Spalten handelt es sich um die Start- und Endposition in der Suchsequenz und dem Treffer. Abschließend folgen der Erwartungswert und der *bitscore*. Die Standardausgabe enthält zusätzlich das paarweise lokale Alignment der beiden Sequenzen und ist daher interessant für Fälle, in denen wenige Suchen durchgeführt, und manell untersucht werden sollen.

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gi\|90654904\|gb\|ABD96051.1\| | 110807394 | 29.87 | 231 | 135 | 7 | 71 | 288 | 4 | 220 | 1e-16 | 85.1 |
| gi\|90654904\|gb\|ABD96051.1\| | 74314416 | 30.77 | 208 | 121 | 6 | 90 | 288 | 7 | 200 | 2e-16 | 84.0 |
| gi\|90654904\|gb\|ABD96051.1\| | 82778917 | 30.77 | 208 | 121 | 6 | 90 | 288 | 7 | 200 | 2e-16 | 84.0 |
| gi\|90654904\|gb\|ABD96051.1\| | 82546265 | 30.77 | 208 | 121 | 6 | 90 | 288 | 7 | 200 | 2e-16 | 84.0 |
| gi\|90654904\|gb\|ABD96051.1\| | 56480481 | 30.77 | 208 | 121 | 6 | 90 | 288 | 1 | 194 | 3e-16 | 83.2 |

**Abbildung 4.2:** Beispiel einer BLAST-Ausgabe im tabellarischen Format, das in dieser Arbeit als Eingabe verwendet wurde. Die Spalten bedeuten: Suchsequenz (1), Treffersequenz (2), Identität (3), Alignmentlänge (4), Nichttreffer (engl. *mismatch*) (5), Lückenbereiche (6), Anfang und Ende der Suchsequenz (7,8) beziehungsweise der Treffersequenz (9,10), Erwartungswert (11) sowie der *bitscore* (12)

### 4.3.2 Ausgaben / Zwischenformate

Die meisten Ausgaben werden in der Form von Textdateien mit Tabulator bzw. Leerzeichen getrennten Feldern oder einfachen Listen gespeichert. Die Beschreibungen der einzelnen Formate erfolgt in den Methodenbeschreibungen bei den entsprechenden Programmen.

## 4.4 Rechner

Ein großer Teil der Berechnungen für diese Arbeit wurde auf Standardrechnern mit Linux beziehungsweise MAC OSX Betriebssystem durchgeführt. Größere

Berechnungen wurden teilweise auf den Servern Jukes und Horst (jeweils acht Prozessorkerne) beziehungsweise auf dem Hybrid-Cluster Gauss[5] (derzeit 524 Prozessorkerne) des Zentrums für Informations und Medientechnologie (ZIM) der Heinrich-Heine-Universität durchgeführt.

[5] http://www.zim.uni-duesseldorf.de/hpc/hpc-infrastruktur/bull-cluster

# 5 Wie $\alpha$-proteobakteriell sind $\alpha$-Proteobakterien?

Teile der Analysen, die in diesem Kapitel präsentiert werden, wurden bereits veröffentlicht in:

- C. Esser, W. Martin and T. Dagan. **The origin of mitochondria in light of a fluid prokaryotic chromosome model.** Biol Lett, 3(2):180–184, Apr 2007.

## 5.1 $\alpha$-Proteobakterien und der Vorfahr der Mitochondrien

In der Biologie herrscht Einigkeit darüber, daß die Mitochondrien von freilebenden Prokaryoten abstammen, und daß dieses Organell nur einmal im Laufe der Evolution entstanden ist (Gray et al., 1999; Dolezal et al., 2006). Die Frage nach dem Verwandtschaftsbereich und den Eigenschaften des Vorfahren der Mitochondrien wird dagegen weiterhin diskutiert (Martin et al., 2001; Cavalier-Smith, 2002, 2009). Schon seit längerem liegt der Fokus dabei auf den Purpurbakterien (John und Whatley, 1975), die inzwischen als $\alpha$-Proteobakterien bekannt sind (Stackebrandt et al., 1988).

Eine methodische Vorgehensweise um dieses Problem anzugehen ist, Gene die im Mitochondrium kodiert sind mit denen freilebender Prokaryoten zu vergleichen. Mit einer solchen Methode wurde mit Hilfe der 16S rRNA *Agrobacterium tumefaciens* als nächster Verwandter des Mitochondriums identifiziert (Yang et al., 1985). Neuere Studien ordnen den Vorfahren der Mitochondrien in die Ordnung der Rickettsiales ein, die zum großen Teil aus parasitischen Organismen mit stark reduzierten Genomen besteht (Lang et al., 1999; Emelyanov, 2003). Andere Studien deuteten auf *Rickettsia*

*prowazekii* (Andersson et al., 1998) hin, beziehungsweise einen Vorläufer von *Rickettsia* und *Wolbachia* (Wu et al., 2004), während andere Quellen freilebende Arten mit größeren Genomen mit dem Vorfahren in Verbindung bringen (Esser et al., 2004).

Ein anderer Ansatz versucht herauszufinden welche Stoffwechselwege im Mitochondrium ablaufen ohne die Frage nach dem nächsten Nachbarn zu stellen (Gabaldón und Huynen, 2003). Eine weitere Methode untersucht kernkodierte Gene die in Mitochondrien und Hydrogenosomen gemeinsam vorkommen und leiten physiologische Eigenschaften des gemeinsamen Vorfahren ab (Embley und Martin, 2006; van der Giezen und Tovar, 2005).

Der nächste Nachbar des Mitochondriums unter den freilebenden $\alpha$-Proteobakterien ist weiterhin unbekannt (Lang et al., 1999; Esser et al., 2004). Die Zusammensetzung der Genome von Bakterien ist variabel und ändert sich im Laufe der Zeit durch Mutationen, den Verlust von Genen und lateralen Gentransfer (Lawrence und Ochman, 1998; Martin, 1999; Kunin et al., 2005; Lerat et al., 2005). Im Folgenden sollen 47.143 Proteine von 18 vollständig sequenzierten $\alpha$-Proteobakterien auf ihre Verwandtschaft zu anderen Prokaryoten untersucht werden.

## 5.2 Vorgehensweise der vorliegenden Genomanalyse

Die Genome von 18 $\alpha$-Proteobakterien[1], inklusive des als unklassifiziertes Proteobakterium geführten Stammes *Magnetococcus MC-1*, und zusätzlich sechs Genome von Mitochondrien[2] wurden als Referenzen beim NCBI heruntergeladen. Für jedes Protein wurde eine Suche in einer Datenbank bestehend aus 231 prokaryotischen Genomen durchgeführt (Altschul et al., 1997). Details zu den verwendeten Genomen finden sich im Anhang auf Seite 104 und in Tabelle 4.1.

Als Ausgangspunkt für die Analyse dienen die Ergebnisse einer `BLAST`- Datenbanksuche von 18 $\alpha$-Proteobakterien-Genomen in den Sequenzen von 231 vollständig sequenzierten prokaryotischen Genomen in tabellarischer Form (Siehe Seite 24).

Um die `BLAST`-Ausgaben einlesen und verarbeiten zu können wurde das PERL-Skript `tbpt.pl` (engl. *tabular blast parsing tool*) erstellt. Es liest die Ausgabe dieser

[1] `ftp://ftp.ncbi.nih.gov/Genoms/Bacteria/`
[2] `ftp://ftp.ncbi.nih.gov/Genoms/MITOCHONDRIA/`

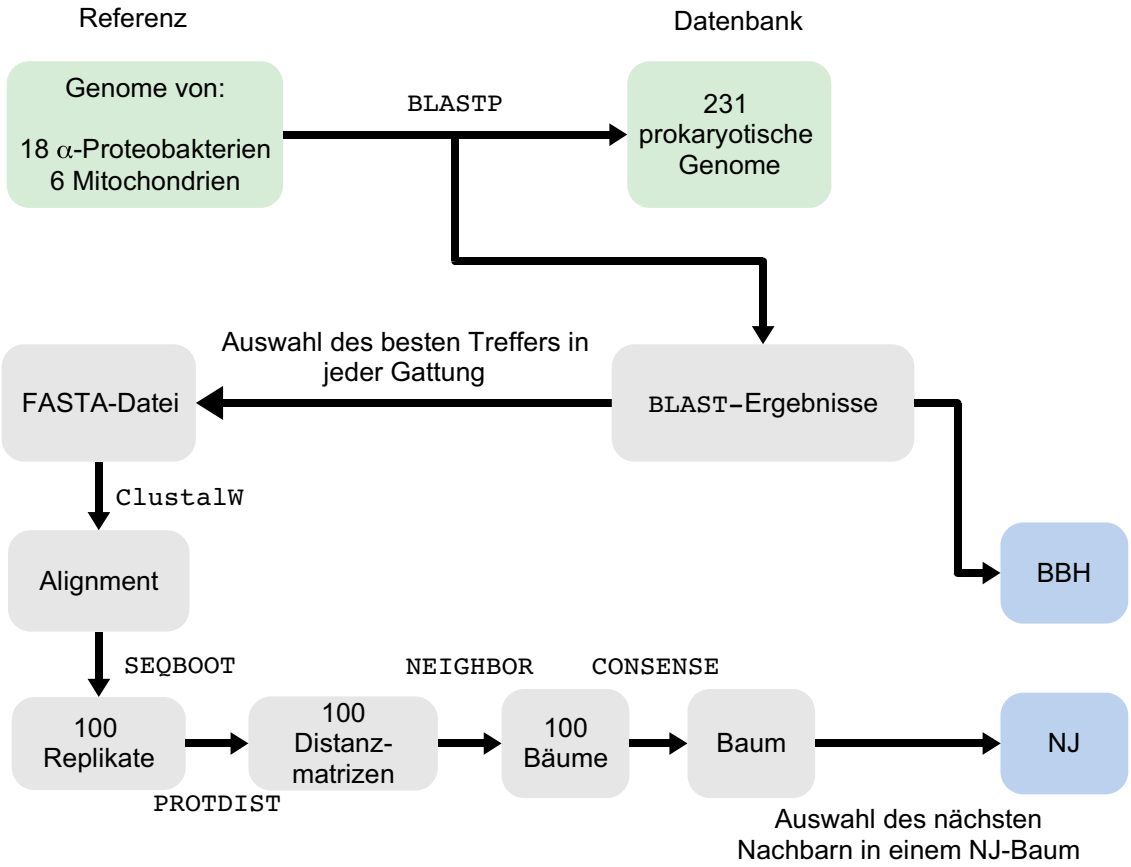



**Abbildung 5.1:** Schema des Arbeitsablaufes für die Analyse der Struktur $\alpha$-proteobakterieller Genome. Die einzelnen Schritte werden im Text erläutert.

**Tabelle 5.1:** Erforderliche Eingaben für `tbpt.pl`

| | Eingaben |
|---|---|
| BLAST-Datei | Ausgabe der Datenbanksuche in tabellarischer Form, bei Benutzung von Standardsequenzdaten wie sie auf der Webseite des National Center for Biotechnology Information (NCBI) erhältlich sind |
| GI-Liste | Eine Liste über die den GI-Nummern der einzelnen Proteine das Genom aus dem sie stammen zugeordnet wird. |
| TAXFILE | Eine Datei die die taxonomischen Daten für jede im Datensatz enthaltene Spezies enthält |

BLAST-Datenbanksuche ein und gleicht sie mit taxonomischen Daten, die über eine Taxonomiedatei zur Verfügung gestellt werden, ab. Die Ausgabe ist eine Matrix, in der für jedes Referenzgenom aufgeführt ist, wieviele nächste Nachbarn in den in der Taxonomiedatei spezifizierten taxonomischen Gruppen liegen.

1. Einlesen der Taxonomie-Informationen (.tax-File)
2. Einlesen der GI-Liste
3. Einlesen der tabellarischen BLAST-Ergebnisse
   a) Überprüfen der Schwellenwerte
   b) Auswahl des nächsten Nachbarn außerhalb der Gattung der Suchsequenz
   c) Überprüfung der taxonomischen Einordnung des besten Treffers
4. Ausgabe einer Tabelle mit den nächsten Nachbarn für jedes Suchgenom in den betrachteten taxonomischen Einheiten

Für die hier präsentierten Ergebnisse erfolgte die Aufteilung der Datenbank in Stämme (lat. *Phyla*), die Proteobakterien wurden in ihre Klassen aufgeteilt. Die genaue Aufteilung der Gruppen ist im Anhang auf Seite 110 beschrieben.

Schritt 2 (b) wurde mit unterschiedlichen Methoden durchgeführt:

BBH (engl. *best-blast-hit*) Als nächster Nachbar wird der beste Treffer in einer BLAST-Datenbanksuche ausgewählt.

NJ Für jede Gattung wurden die besten Treffer in einer BLAST- Datenbanksuche ermittelt und deren Sequenzen in eine Fastadatei geschrieben.

```
CO    #    Definitions of Taxa
CO    ID        Organism                              Group  Active

AC    ID_001    Acinetobacter sp. ADP1                 0315    1
AC    ID_002    Aeropyrum pernix K1                    5100    1
AC    ID_003    Agrobacterium tumefaciens str. C58     0311    1
AC    ID_004    Anaplasma marginale str. St. Maries    0311    1
AC    ID_005    Aquifex aeolicus VF5                   3210    1
AC    ID_006    Archaeoglobus fulgidus DSM 4304        5110    1
AC    ID_007    Azoarcus sp. EbN1                      0312    1
...

CO    #    Definitions of Taxonomic groups
CO    Group            Identifier

FA    Crenarchaeota      5100
FA    Euryarchaeota      5110
FA    Nanoarchaeota      5120
FA    Actinobacteria     3200
FA    Aquificae          3210
FA    Bacteroidetes      3220
FA    Chlamydiae         3230
FA    Chlorobi           3240
FA    Chloroflexi        3250
FA    Cyanobacteria      3260
...
FA    unidentified       9000
FA    DUPLICATION        0001
FA    NO_HITS            9991
FA    TOO_SHORT          9992
FA    ERRORS             9993
FA    PROTEINS           9999
```

**Abbildung 5.2:** Taxonomiedatei für die Verwendung mit `tbpt.pl`. Die mit AC beginnenden Zeilen definieren Taxoneinträge, in denen jedem Genom eine Gruppe zugeordnet wird. Die Spalte „Active“ ermöglicht es eine Auswahl an Genomen zu treffen. Die Genome, bei denen hier eine 0 eingetragen ist, werden automatisch bei der Bearbeitung ignoriert. Diese Funktion wurde für die statistische Überprüfung der Daten benutzt. Die entsprechenden Gruppen werden in den mit FA beginnenden Zeilen definiert. Alle übrigen Zeilen werden von dem Programm beim einlesen ignoriert und können für Kommentare benutzt werden. Die Anfangsbuchstaben CO wurden jedoch für den Fall einer zukünftigen Erweiterung für Kommentare reserviert. Die Spalten sind durch Tabulatorzeichen getrennt.

Mit dieser wurde mit `ClustalW` ein multiples Alignment erstellt aus dem mit den Programmen des `PHYLIP`-Paketes *neighbor-joining* Bäume erstellt wurden. In diesen wurde der nächste Nachbar bestimmt, indem das Taxon mit der geringsten Anzahl an Kanten zwischen sich und der Suchsequenz und, falls dieses Kriterium nicht ausreicht, mit der geringsten phylogenetischen Distanz gesucht wird.

NJ 90 % Diese Methode entspricht der NJ-Methode, allerdings wurden die phylogenetischen Analysen mit Bootstrapping und 100 Replikaten durchgeführt. Nächste Nachbarn wurden nur berücksichtigt wenn in mindestens 90 % der Bäume der Nachbar innerhalb derselben taxonomischen Gruppe zu finden war.

ML Es werden ebenfalls Alignments erstellt, aus denen dann *maximum-likelihood*-Bäume mit dem Programm `PhyML` erstellt werden. In diesen Bäumen wurde wie oben beschrieben der nächste Nachbar ermittelt.

Wie im Arbeitsablauf (Abb. 5.1 auf Seite 28) ersichtlich ist, wurde die Analyse mit zwei Methoden durchgeführt. Zunächst in einem einfachen Ansatz, in dem der nächste Nachbar als der beste Treffer einer `BLAST`-Datenbanksuche definiert wurde, der nicht aus derselben Gattung stammt wie die Suchsequenz. Da frühere Analysen (Koski und Golding, 2001) zeigten, daß dieser Ansatz nicht immer zu korrekten Ergebnissen führt, wurde zur Kontrolle eine weitere Analyse durchgeführt. In diesem komplexeren Ansatz wurden die besten Treffer jeder Gattung gesammelt. Für den Fall, daß es mindestens zwei Treffer gab, wurden mit diesen Treffern Alignments berechnet, aus denen mit Hilfe der *neighbor-joining*-Methode (Saitou und Nei, 1987) phylogenetische Bäume erstellt wurden. Dafür wurden die Programme `seqboot`, `protdist`, `neighbor` und `consense` aus dem PHYLIP-Paket (4.2.4) mit den Standardeinstellungen verwendet. Für die Bootstrapanalyse wurden 100 Replikate erstellt.

Um eine Automatisierung der phylogenetischer Analysen zu ermöglichen wurde das PERL-Skript tree-o-mat.pl entwickelt. Es dient zur Automatisierung standardisierter Abläufe für die Erstellung phylogenetischer Bäume. Der Ausgangspunkt dafür ist in der Regel eine Liste von FASTA-Dateien, die die zu bearbeitenden Sequenzen enthalten. Diese werden mit dem Programm `ClustalW` aligniert. Anschließend werden entweder mit Hilfe der Programme des `PHYLIP`-Paketes *neighbor-joining*

```
          T B P T
        Tabular Blast Parsing Tool

The program ./tbpt_008.pl has been started on
Fri Nov  2 10:05:18 2007 with the following options:

g       cyano.gi
G       standard.group

Using settings from standard.group

 *** Settings ***

E-Value < 1e-20  (default)
Length  > 49     (default)
Ident.  > 24     (default)
Score  > 0       (default)
Best hit selected on ident (default)

Reading GI-List cyano.gi ...
Processing Inputfile(s) :
file 1 / 374     results/CB_0000.fasta.blastp      100 Sequences
file 2 / 374     results/CB_0001.fasta.blastp      100 Sequences
file 3 / 374     results/CB_0002.fasta.blastp      100 Sequences
file 4 / 374     results/CB_0003.fasta.blastp      100 Sequences
file 5 / 374     results/CB_0004.fasta.blastp      100 Sequences
...
file 372 / 374   results/CB_0372.fasta.blastp      100 Sequences
file 373 / 374   results/CB_0373.fasta.blastp      99 Sequences
file 374 / 374   results/CB_0375.fasta.blastp      54 Sequences


                                  CB1     CB2     CB3     CB4     CB5     CB6
DUPLICATION              0001     0       0       0       0       0       0
Proteobacteria           0310     3       2       1       3       0       0
Alphaproteobacteria      0311     31      19      22      36      32      19
Betaproteobacteria       0312     5       13      2       11      8       7
Deltaproteobacteria      0313     4       4       2       7       5       4
Epsilonproteobacteria    0314     1       6       3       5       0       4
Gammaproteobacteria      0315     22      26      25      29      24      30
Actinobacteria           3200     4       6       6       3       6       5
Aquificae                3210     2       0       1       0       0       3
Bacteroidetes            3220     0       6       0       5       1       3
Chlamydiae               3230     4       2       3       0       2       3
Chlorobi                 3240     1       1       1       0       2       2
Chloroflexi              3250     1       1       1       2       1       1
Cyanobacteria            3260     1155    1157    1177    1669    1170    1142
Deinococcus-Thermus      3270     0       1       0       0       0       1
Firmicutes               3280     18      24      17      5       15      18
Fusobacteria             3290     1       0       0       0       1       2
Planctomycetes           3300     2       2       2       6       2       4
Spirochaetes             3320     4       4       4       5       4       4
Thermotogae              3330     1       0       1       0       1       0
Crenarchaeota            5100     1       1       1       0       1       1
Euryarchaeota            5110     3       4       6       4       4       2
Nanoarchaeota            5120     0       0       0       1       0       0
NO_HITS                  9991     567     605     546     863     521     570
PROTEINS                 9999     1830    1884    1821    2654    1800    1825

This job was finished at Fri Nov  2 10:07:16 2007.


Thank you for using TBPT !
```

**Abbildung 5.3:** Gekürztes Beispiel einer Ausgabe des Programms `tbpt.pl` zur Bestimmung des nächsten Nachbarn mit Hilfe des BBH Kriteriums (Der beste Treffer in einer `BLAST`-Datenbanksuche der nicht aus derselben Gattung stammt wie die Suchsequenz). Zunächst werden die verwendeten Parameter (beziehungsweise falls nicht gesetzt die Standardwerte (engl. *default*)) aufgelistet. Nach einer Auflistung der Eingabedateien folgt eine Tabelle mit den Ergebnissen der Suche. Die zweite Spalte gibt die „Gruppenkürzel“ wieder (siehe Seite 110).

```
./query/
./data/
   running/
   results/
   done/
```

**Abbildung 5.4:** Empfohlene Verzeichnisstruktur für die Benutzung des `PERL`-Skriptes `tree-o-mat.pl`

(Saitou und Nei, 1987) oder mit `PhyML` (Guindon und Gascuel, 2003) *maximum-likelihood*-Bäume berechnet. Alle Methoden können beliebig kombiniert werden und ebenfalls durch Bootstrapping ergänzt werden. Wahlweise ist auch die Erstellung phylogenetischer Netzwerke mit Hilfe von `LDDist`, `NeighborNet` und `Splits` möglich.

Der Hauptvorteil ergibt sich durch die Möglichkeit parallel mehrere Prozesse starten zu können. Dadurch müssen größere Datenmengen nicht vor der Abarbeitung aufgeteilt werden, was längere Wartezeiten bei ungleichmäßiger Aufteilung vermeidet. Eine geschickte Verwaltung ermöglicht es dabei auch noch nachträglich Prozesse zu starten und so die Geschwindigkeit an die aktuelle Serverauslastung beziehungsweise die Geschwindigkeitsbedürfnisse anzupassen. Dies wird durch eine Ordnerstruktur ermöglicht in der alle abzuarbeitenden Daten als Einzeldateien in einem Verzeichnis („query“) vorliegen. Diese werden während der Bearbeitung in ein anderes Verzeichnis verschoben („running“). Nach der Fertigstellung werden die Ergebnisse in den entsprechenden Ordner („results“) verschoben und die Quelldateien in einen anderen („done“). Dadurch ist es möglich auch nachträglich Prozesse zu starten und zu beenden falls dies erforderlichsein sollte. Ist das Programm erfolgreich durchgelaufen, so ist das Verzeichnis mit den Quelldateien und das Verzeichnis mit den momentan bearbeiteten Dateien leer.

Aus Effizienzgründen ist es wichtig darauf zu achten, daß die Genome nicht in zu viele Teile zerlegt werden, da sonst die Effizienzsteigerung durch die Nutzung mehrerer Prozessoren durch eine Überlastung des Dateisystems zunichte gemacht wird.

Für die phylogeniebasierten Ansätze wurde der nächste Nachbar, durch die minimale Distanz zwischen Suchsequenz und einer beliebigen anderen Sequenz in einem phylogenetischen Baum im Newick-Format bestimmt. Das Perl-Skript `ntfp.pl` (engl. *Newick Tree File Parser*) liest einen phylogenetischen Baum im Newick-Format ein und ist in der Lage, unterschiedliche Werte aus dem Baum auszulesen. Der

**Tabelle 5.2:** Benötigte Eingaben für das Skript `ntfp.pl` das nächste Nachbarn in Newickbäumen identifiziert. Neben der Taxonomiedatei und der GI-Liste, die bereits im vorherigen Schritt bei der Bearbeitung der `BLAST`-Ausgaben benötigt wurden, ist das eine spezielle Ausgabe aus dem vorherigen Schritt, Bäume im Newickformat und die Eingabe einer Referenz.

| Datei | Inhalt |
|---|---|
| TBPT Ausgabe | Ausgabe der Datenbanksuche in tabellarischer Form, bei Benutzung von Standardsequenzdaten wie sie auf der Webseite des National Center for Biotechnology Information (NCBI) erhältlich sind |
| TAXFILE | Eine Datei, die die taxonomischen Daten für jede im Datensatz enthaltene Spezies enthält |
| GI-Liste | Eine Liste über die den GI-Nummern der einzelnen Proteine das Genom aus dem sie stammen zugeordnet wird |
| | Des Weiteren muß eine Referenz angegeben werden, zu der der nächste Nachbar bestimmt werden soll |

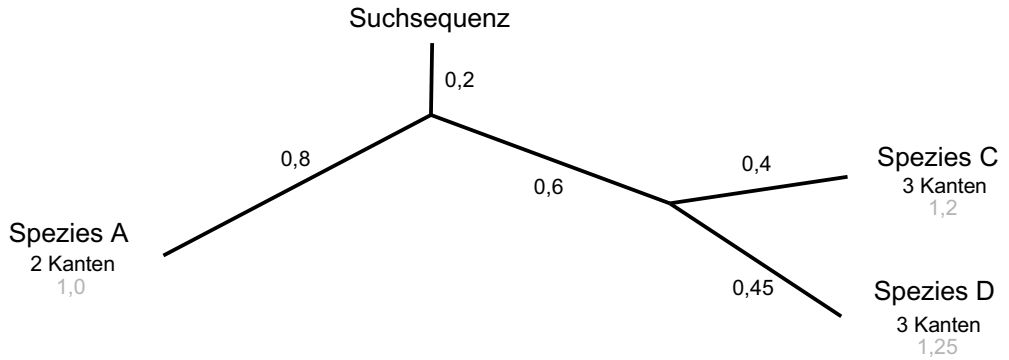
Hauptzweck ist die Ermittlung des nächsten Nachbarn einer Sequenz. Dabei wird eine Matrix erstellt, in der die Anzahl an Kanten zwischen den einzelnen Sequenzen eingetragen wird und eine weitere, in der die Distanzen auf dem Weg von einem OTU zu einem anderen addiert werden. Der nächste Nachbar ist dabei der OTU mit der geringsten Anzahl von Kanten zu der Suchsequenz. Sollte es mehrere Möglichkeiten geben, wird die Distanz als zweites Kriterium herangezogen. Dieses Vorgehen wird in Abbildung 5.5 illustriert.

Als Eingabe benutzt das Skript neben Bäumen im Newick-Format die gleiche Sorte Taxonomiedateien wie das Skript zum Einlesen von `BLAST` Ausgaben. Die Ausgabe enthält für jeden Baum eine Liste der enthaltenen Taxa und den Abstand in Kanten und in phylogenetischer Distanz zum Referenzgenom.

## 5.3 Bester Treffer in einer `BLAST`-Datenbanksuche

Den größten Anteil an $\alpha$-proteobakteriellen nächsten Nachbarn (basierend auf besten `BLAST`-Treffern (BBH – engl. *best blast hit*) wies mit 91,9 % *Sinorhizobium meliloti* auf (Abb. 5.6), der niedrigste Anteil fand sich mit 63,9 % bei *Magnetospirillum magnetotacticum*. *Magnetococcus* sp. wies mit 32,6 % zwar einen noch niedrigeren

a)




b)

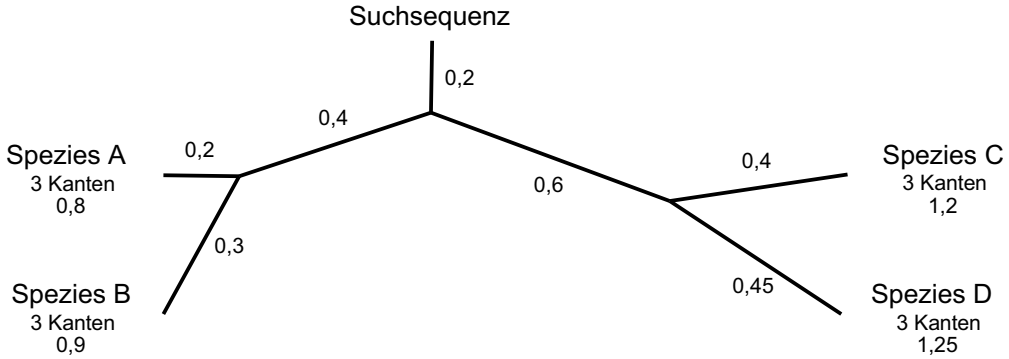



**Abbildung 5.5:** Beispiel zur Bestimmung des nächsten Nachbarn in einem phylogenetischen Baum. Es werden zunächst für jede Sequenz im Baum die Anzahl der Kanten zwischen ihr und der Suchsequenz bestimmt. Zusätzlich wird die Distanz des kürzesten Weges zwischen den betrachteten Sequenzen addiert. In Beispiel a) wird das erste Kriterium angewendet. Spezies A wird als nächster Nachbar bestimmt, da der kürzeste Weg die niedrigste Anzahl an Kanten zur Suchsequenz hat. Die Distanz wird nicht betrachtet. In Beispiel b) weisen vier Sequenzen die gleiche Anzahl an Kanten auf, daher wird auf die Distanzen als zweites Kriterium zurückgegriffen. Sequenz A wird hier ebenfalls ausgewählt, da es die kürzeste Gesamtdistanz zu der Suchsequenz hat.

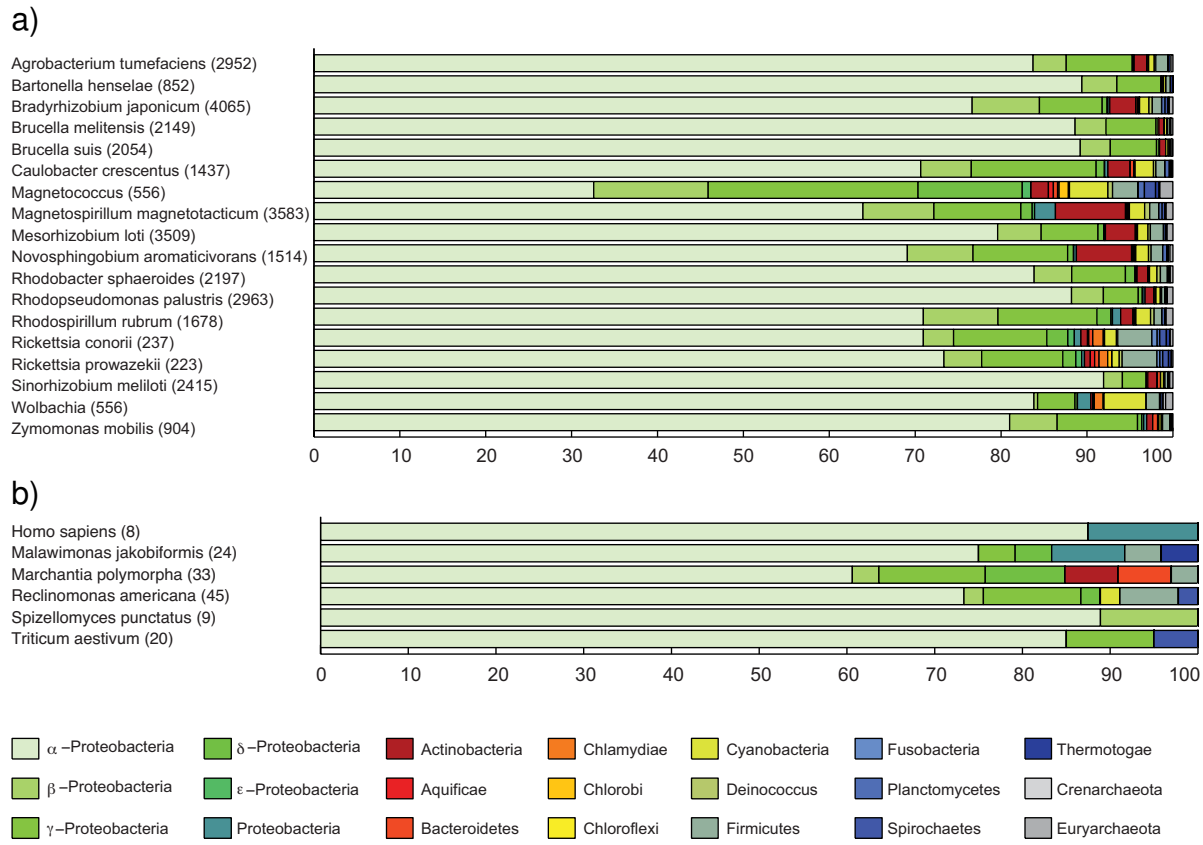



**Abbildung 5.6:** Verteilung der nächsten Nachbarn von a) $\alpha$-proteobakteriellen Genen und b) mitochondrial kodierten Genen auf der Basis von reziproken besten Treffern in einer `BLAST`-Datenbanksuche.

Anteil an nächsten Nachbarn (BBH) innerhalb der $\alpha$-Proteobakterien auf, erhielt jedoch einen Sonderstatus, weil es als unklassifiziertes Proteobakterium gilt. Im Schnitt wiesen 77,1 % der Proteine aller untersuchten Genome einen nächsten Nachbarn innerhalb der $\alpha$-Proteobakterien auf. Die übrigen nächsten Nachbarn lagen hauptsächlich in den anderen Klassen der Proteobakterien: 0,4–8,7 % in den $\beta$-, 2,8–14,5 % in den $\gamma$-, 0,0–2,4 % in den $\delta$- und unter 0,8 % in den $\varepsilon$-Proteobakterien. 1,3–13,7 % der nächsten Nachbarn lagen sogar außerhalb des Phylums. Von den übrigen Phyla fanden sich die meisten nächsten Nachbarn in den Actinobakterien (2,0%), den Cyanobakterien (1,4%) und den Firmicutes (1,4 %).

Für *Magnetococcus* war die Verteilung insbesondere bei der Aufteilung der proteobakteriellen Gruppen anders als bei den anderen untersuchten Genomen: Der Anteil an nächsten Nachbarn innerhalb des eigenen Phylums liegt bei 32,6 % ($\alpha$-), 13,3 % ($\beta$-), 24,4 % ($\gamma$-), 12,1 % ($\delta$-) und 1,0 % in den $\varepsilon$-Proteobakterien. Auch bei den nächsten Nachbarn wurde *Magnetococcus* in einer eigenen Kategorie als (unklassifiziertes) Proteobakterium geführt. Während im Mittel nur 0,5 % der Genome ihren nächsten Nachbarn innerhalb von *Magnetococcus* fand, ist dies bei *Magnetospirillum* für 2,4 % der Gene der Fall.

Neben nächsten Nachbarn innerhalb der Eubakterien wiesen im Schnitt 0,5 % der untersuchten Genome auch nächste Nachbarn in der Domäne der Archaebakterien auf. Die größte Anzahl archaebakterieller Treffer wies mit 54 *Magnetospirillum* auf, den höchsten Anteil aufgrund des kleineren Genoms hatte *Magnetococcus* mit 1,6 % (36 Gene) gefolgt von *Wolbachia* mit 1,1 %.

Diese Aufteilungen wichen signifikant von der taxonomischen Zusammensetzung der Datenbank ab, das heißt der p-Wert war kleiner als 0,05 bei einem $\chi^2$ -Test mit Bonferroni-Korrektur, und waren daher als nicht zufällig anzusehen. Die Zusammensetzung der Datenbank ist in Abbildung 5.8 b) zu sehen. Sie bestand zu 7 % aus Sequenzen von Archaebakterien und zu 93 % aus eubakteriellen Sequenzen. Davon entstammten 48,6 % Proteobakterien.

Abbildung 5.6 zeigt die Daten nur für den Anteil der Daten, für die eine Aussage über den nächsten Nachbarn getroffen werden kann. In Abbildung 5.7 wird gezeigt, für wieviele Gene das der Fall ist. Während sich für *Rickettsia conori* nur für 47,6 % der Gene ein nächster Nachbar, der den Anforderungen genügt, finden läßt, weisen bei *Sinorhizobium meliloti* 88,1 % einen nächsten Nachbarn auf. Insgesamt läßt sich

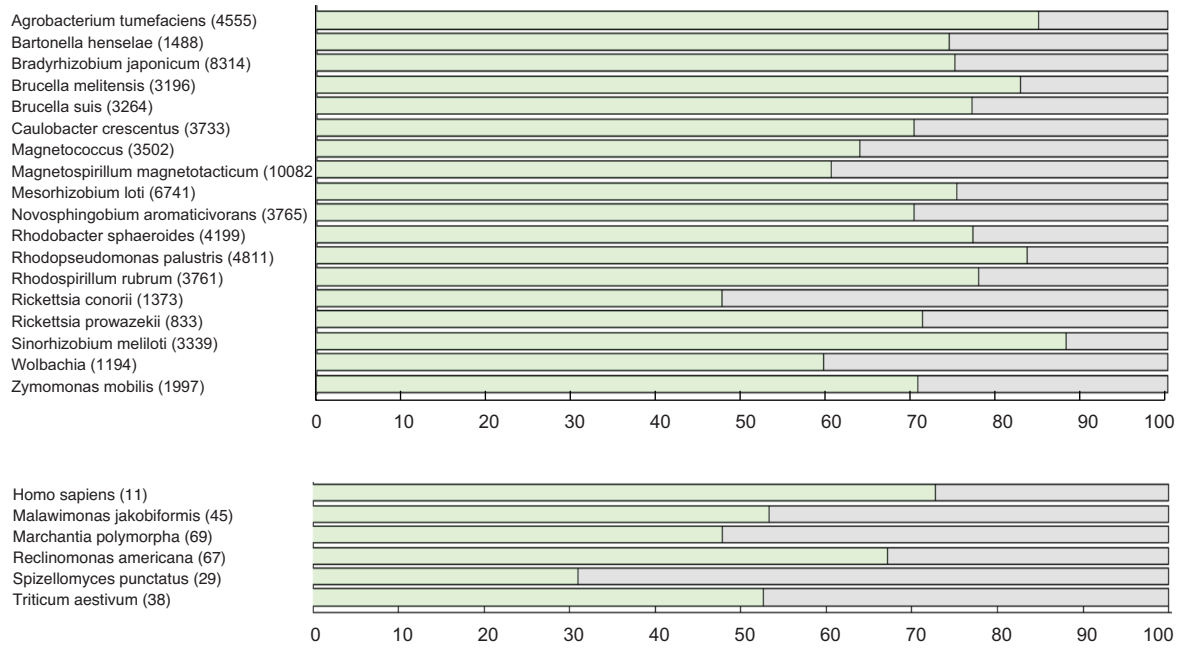



**Abbildung 5.7:** Anteil der $\alpha$-proteobakteriellen Gene mit nächstem Nachbarn (BBH-Methode). Die Gene die keinen nächsten Nachbarn aufweisen (grauer Anteil) erfüllen die Schwellenwerte nicht.

für 17.953 (26,6 %) der untersuchten Gene kein nächster Nachbar finden. Hier liegt *Magnetococcus* mit 36,2 % über dem Durchschnitt.

Die Verteilung der nächsten Nachbarn für mitochondrial kodierte Gene mit der BBH Methode, die in Abb. 5.6 b) gezeigt wird, ist ähnlich wie die der $\alpha$-Proteobakterien. Da hier nur Gene, die im Mitochondrium kodiert werden, berücksichtigt wurden und nicht auch solche, deren Genprodukte später in dieses transportiert werden, war die Größe des Datensatzes hier sehr gering. Nur elf bis 69 Gene stehen hier für die Untersuchung zur Verfügung. 90,6 % der mitochondrialen Gene wiesen proteobakterielle nächste Nachbarn auf, 74,1 % deuteten auf ein $\alpha$-Proteobakterium hin. Die größte Gruppe unter den nicht-Proteobakterien bildeten die Firmicutes, in denen mit fünf Genen 3,6 % aller untersuchten Mitochondriengene ihren nächsten Nachbarn hatten.

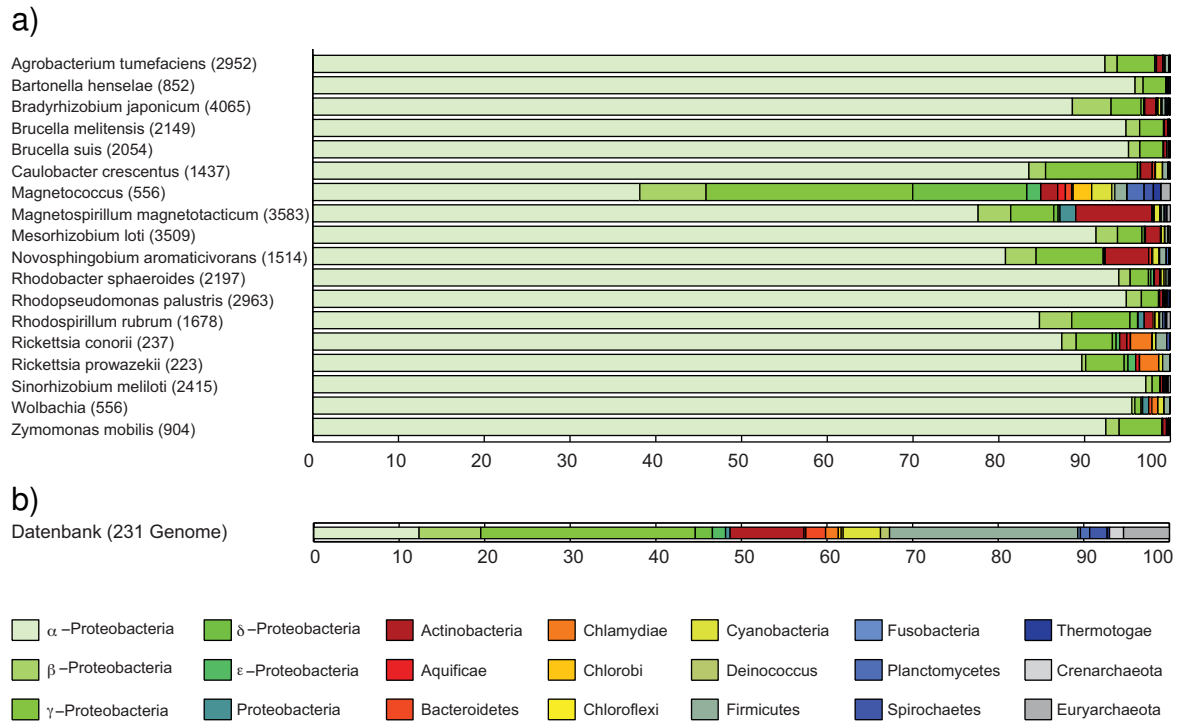



**Abbildung 5.8:** a) Verteilung der nächsten Nachbarn von $\alpha$-proteobakteriellen Genen auf der Basis von nächsten Nachbarn in *neighbor-joining*-Bäumen. Die Bäume wurden aufgrund von 100 Bootstrapreplikaten erstellt. Es wurden nur Ergebnisse gewertet, in denen der nächste Nachbar in mindestens 90 der Bootstrapbäume zu derselben Gruppe gehörte. In b) ist die Zusammensetzung der verwendeten Datenbank dargestellt.

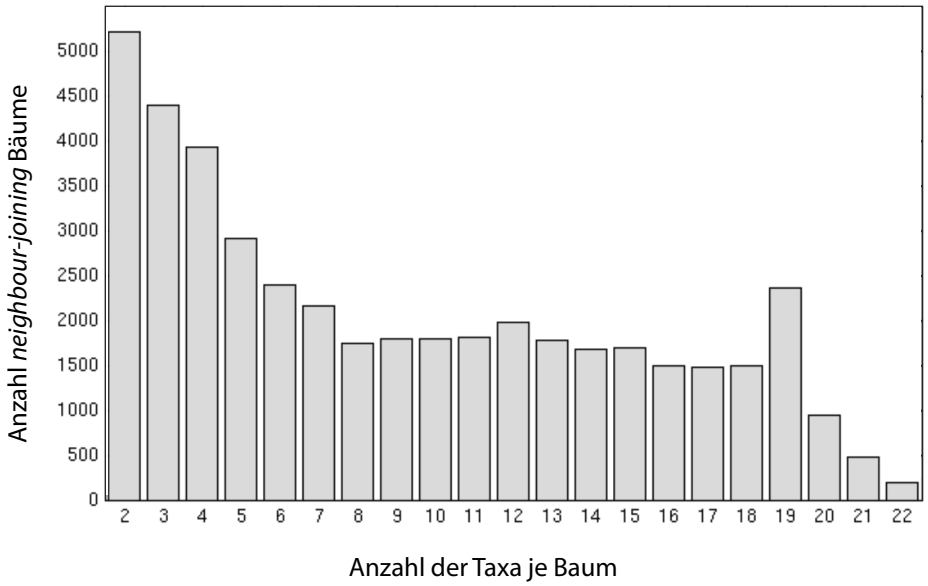



**Abbildung 5.9:** Histogramm der Verteilung der Anzahl der Taxa über die phylogenetischen Bäume.

## 5.4 Der nächste Nachbar in einem *neighbor-joining*-Baum (NJ)

Für die nächsten Nachbarn, die über den Weg der *neighbor-joining*-Bäume mit Bootstrapping erstellt wurden, ließen sich sich ähnliche Ergebnisse ableiten wie für die BBH-Methode (Abb. 5.8). Die große Mehrheit der Gene der $\alpha$-Proteobakterien hatte mit im Schnitt 90,3 % einen nächsten Nachbarn in einem anderen $\alpha$-Proteobakterium (zwischen 77,6 % bei *Magnetospirillum* und 97,1 % bei *Sinorhizobium*). Weitere große Gruppen bildeten $\gamma$-Proteobakterien mit im Schnitt 10,7 % bei den $\alpha$-Proteobakterien und 24,1 % bei *Magnetococcus*, $\alpha$-Proteobakterien mit 2,0 % sowie Actinobakterien mit 1,4 %. Die angewendete Methode ist wesentlich strikter als die oben beschriebene. Im Schnitt lässt sich nur für 49,1 % der Genome ein nächster Nachbar bestimmen. Bei *Rickettsia conorii* konnten nur für 17,3 % der Gene nächste Nachbarn bestimmt werden, bei *Sinorhizobium* waren es 72,3 %. Zusätzliche Tests sollten zeigen, ob es sich bei den Ergebnissen um fundierte Berechnungen handelt oder ob es sich um die Produkte von Zufällen und Artefakten handeln könnte.

Abbildung 5.9 ist ein Histogramm über die Anzahl der Taxa je Baum für die der Abbildung 5.8 zugrunde liegenden Daten. Ein großer Teil der Bäume weist erwartungsgemäß eine geringe Anzahl an Taxa auf. Die Datensätze die hier als zwei

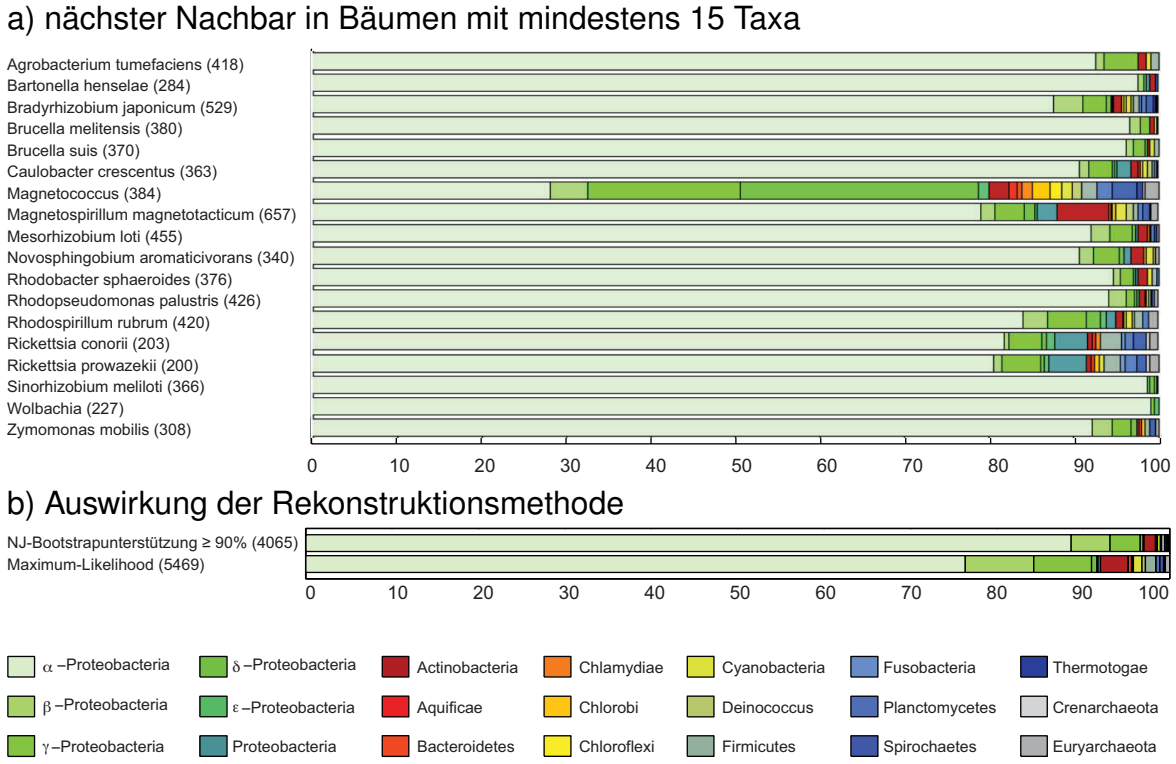



**Abbildung 5.10:** In a) ist ein weiterer Datensatz gezeigt, der mit der NJ-Methode berechnet wurde. Hierbei wurden alle Bäume, die weniger als 15 Taxa enthielten, ausgeschlossen, so daß nur relativ weit verbreitete Gene in die Ergebnisse eingehen. Abbildung b) zeigt die Auswirkung der Baumkonstruktionsmethode auf die Verteilung der nächsten Nachbarn. In der oberen Zeile befindet sich das Ergebnis für *Bradyrhizobium japonicum* wie in Abb. 5.6. In der unteren findet sich die Verteilung bei Verwendung der *maximum-likelihood*-Methode.

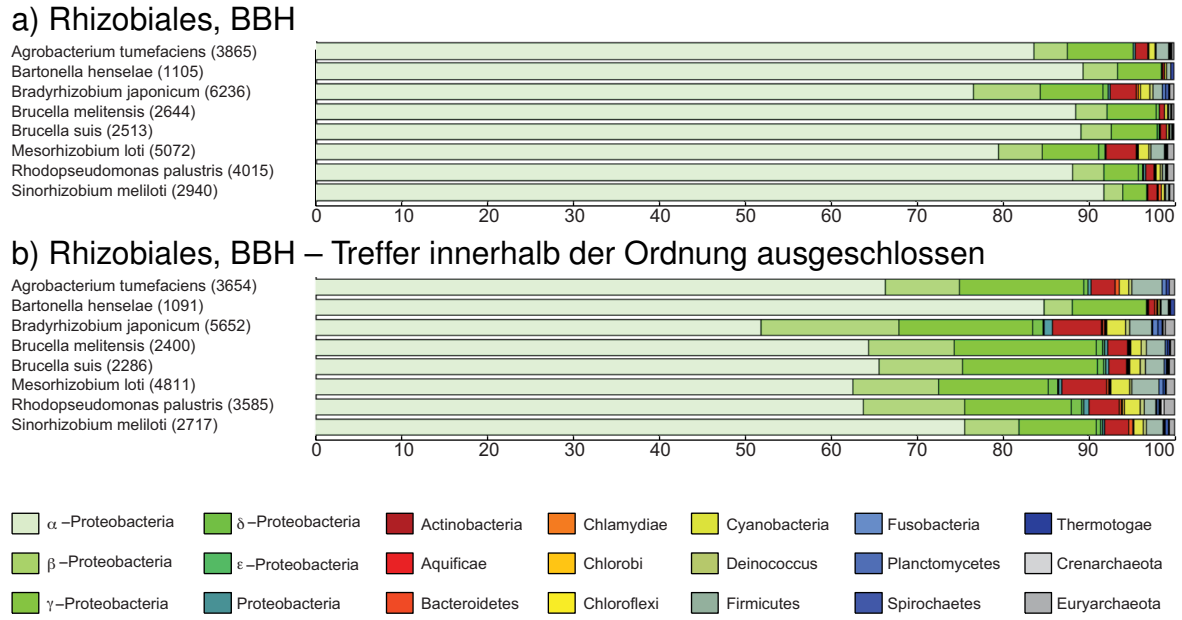



**Abbildung 5.11:** Der Einfluß der Genomauswahl auf die Ergebnisse I: In a) finden sich die Ergebnisse berechnet mit der BBH-Methode, für die in der Analyse enthaltenen Genome der Rhizobiales. Für die Ergebnisse in b) wurden Treffer innerhalb der Rhizobiales nicht berücksichtigt. Das heißt nicht nur Gene der eigenen Gattung wurden ausgeschlossen, sondern auch Gene der eigenen Ordnung.

Taxa aufgeführt werden, wurden im Ansatz mit Bootstrappen nicht verwendet. Bis zu der Zahl von acht Taxa pro Baum fiel die Häufigkeit gleichmäßig ab. Zwischen 8 und 19 Taxa waren die Ergebnisse sehr ähnlich verteilt, um nach einer Spitze bei 19 Taxa stark abzufallen.

Abbildung 5.10 enthält einen Teil des *neighbor-joining*-Datensatzes ohne Bootstrapping. Hierbei wurden nur Gene berücksichtig, bei denen die berechneten Bäume mindestens 15 Taxa enthielten. Der Anteil der nächsten Nachbarn innerhalb der $\alpha$-Proteobakterien lag hier zwischen 79,0 % bei *Magnetospirillum* und 99,1 % bei *Wolbachia*. Für *Magnetococcus* lag der Anteil bei 27,8 % und war damit genauso groß wie der Anteil der $\delta$-Proteobakterien. Insgesamt war die Anzahl der berücksichtigten Gene deutlich geringer als bei den anderen Ansätzen und betrug zwischen 200 Genen bei *Rickettsia conorii* und 657 bei *Magnetospirillum*.

In phylogenetischen Analysen hat neben der Methode zur Berechnung der Alignments auch die Wahl der Rekonstruktionsmethode für den phylogenetischen Baum oft einen Einfluß auf das Ergebnis. Durch die große Anzahl der durchzuführenden Berechnungen fiel die Wahl hier auf die relativ schnelle *neighbor-joining*-Methode

**Tabelle 5.3:** Reihenfolgen, in denen die Genome zu der in Abb. 5.12 gezeigten Analyse hinzugefügt wurden.

| | wachsende Distanz | sinkende Distanz | zufällig |
|---|---|---|---|
| 1 | Rhodopseudomonas | Rhodospirillum | Brucella |
| 2 | Brucella | Rhodobacter | Sinorhizobium |
| 3 | Sinorhizobium | Novosphingobium | Mesorhizobium |
| 4 | Mesorhizobium | Magnetospirillum | Bartonella |
| 5 | Bartonella | Wolbachia | Agrobacterium |
| 6 | Agrobacterium | Ehrlichia | Caulobacter |
| 7 | Caulobacter | Anaplasma | Silicibacter |
| 8 | Silicibacter | Rickettsia | Zymomonas |
| 9 | Zymomonas | Magnetococcus | Gluconobacter |
| 10 | Gluconobacter | Gluconobacter | Magnetococcus |
| 11 | Magnetococcus | Zymomonas | Rickettsia |
| 12 | Rickettsia | Silicibacter | Anaplasma |
| 13 | Anaplasma | Caulobacter | Ehrlichia |
| 14 | Ehrlichia | Agrobacterium | Wolbachia |
| 15 | Wolbachia | Bartonella | Magnetospirillum |
| 16 | Magnetospirillum | Mesorhizobium | Novosphingobium |
| 17 | Novosphingobium | Sinorhizobium | Rhodobacter |
| 18 | Rhodobacter | Brucella | Rhodospirillum |
| 19 | Rhodospirillum | Rhodopseudomonas | Brucella |

(Saitou und Nei, 1987). Um zu testen welchen Einfluß die Wahl der Rekonstruktionsmethode auf diese Art von Analyse hat wurden die Berechnungen für *Bradyrhizobium japonicum* mit der *maximum-likelihood*-Methode (Felsenstein, 1973) unter Verwendung der Standardeinstellungen wiederholt (Abbildung 5.10). Der obere Balken gibt das Ergebnis aus Abbildung 5.6 wieder, der Zweite das Ergebnis der Analyse der *maximum-likelihood*-Bäume.

Des Weiteren ist es möglich, daß die Auswahl der Daten einen Einfluß auf die Ergebnisse hatte. So ist beispielsweise zu erwarten, daß die Wahrscheinlichkeit in einer großen taxonomischen Gruppe einen Treffer vorzufinden größer ist, als in einer kleinen. Zu diesem Zweck wurden zwei Tests durchgeführt. Zunächst wurde eine ganze Ordnung aus dem Ergebnisbereich ausgeschlossen (Abbildung 5.11). Die Rhizobiales sind eine Ordnung innerhalb der $\alpha$-Proteobakterien, zu denen auch einige der untersuchten Genome gehören. Für diese wurde nun nicht nur die Gattung sondern die gesamte Ordnung aus der Ergebnismenge entfernt. Dadurch verschoben sich die Ergebnisse in Richtung der nicht $\alpha$-Proteobakterien.

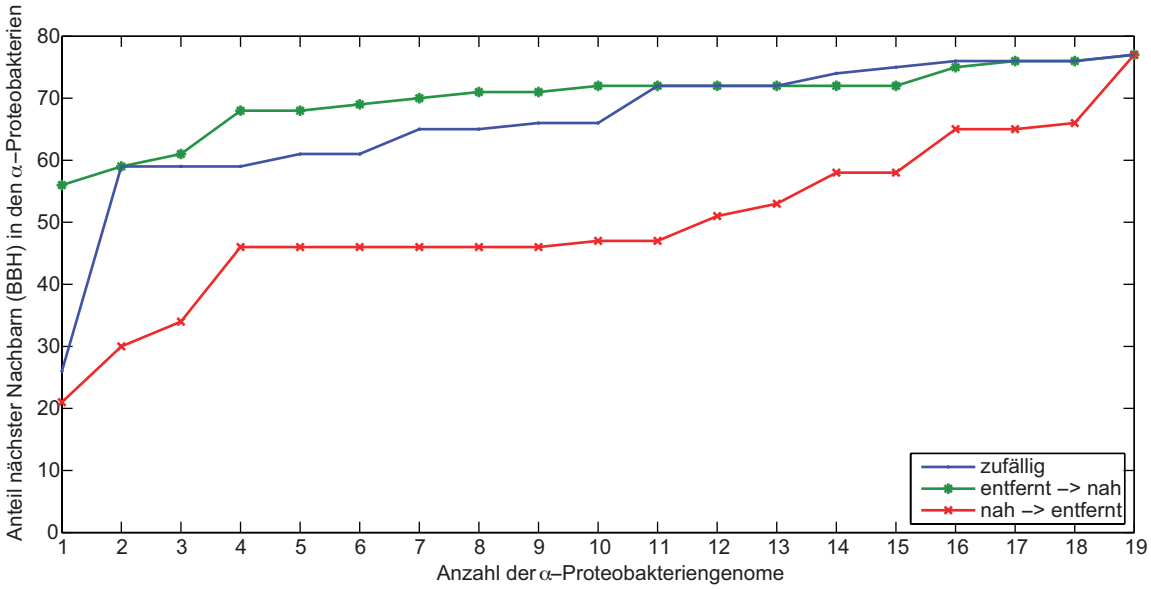



**Abbildung 5.12:** Der Einfluß der Genomauswahl auf die Ergebnisse II: Die Abbildung beschreibt den Anteil der Gene mit $\alpha$-proteobakteriellen Treffer für *Bradyrhizobium japonicum*. Die x-Achse bezeichnet die Anzahl der in der Datenbank enthaltenen $\alpha$-proteobakteriellen Genome. Die Genome wurden in drei verschiedenen Reihenfolgen hinzugefügt. Die rote Kurve beschreibt die Reihenfolge in der zunächst die entferntest verwandten Genome hinzugefügt wurden. Die grüne Kurve beschreibt den gegensätzlichen Ansatz. Für die blaue Kurve wurden die Genome in einer zufälligen Reihenfolge zur Datenbank hinzugefügt. Tabelle 5.3 gibt eine Auflistung der verwendeten Reihenfolgen wieder.

In einem weiteren Ansatz wurde die Datenbank sukzessive um Genome erweitert (Abb. 5.12). Begonnen wurde mit einer Datenbank ohne $\alpha$-Proteobakterien. Die Anzahl der proteobakteriellen Genome wurde in unterschiedlicher Weise erhöht. Zunächst wurden die Genome in der Reihenfolge absteigender Distanz zu *Bradyrhizobium japonicum* hinzugefügt. Für jeden Ansatz wurde der Anteil der Gene mit $\alpha$-proteobakteriellem nächsten Nachbarn (BBH-Methode) bestimmt. Im zweiten Ansatz wurde die gegenläufige Reihenfolge verwendet, und zuletzt eine zufällige Reihenfolge. Die Distanz wurde anhand der paarweisen Distanz der 16S rRNA bestimmt. Bei ansteigender Reihenfolge und bei der zufälligen Auswahl stieg der Anteil schnell auf ein Niveau an, das nur knapp unter dem des vollen Datensatzes lag. Bei absteigender Distanz stieg der Anteil langsamer über einen längeren Bereich an und erreichte erst mit dem letzten hinzugefügten Genom das Niveau des Originaldatensatzes.

## 5.5 *Magnetococcus* Phylogenie

In den Ergebnissen ließ sich erkennen, daß das Genom von *Magnetococcus* eine von $\alpha$-Proteobakterien abweichende Zusammensetzung hatte, in der aber die nächsten Nachbarn der Gene, unabhängig von der Methode, zum Großteil innerhalb der $\alpha$-Proteobakterien lagen. Um die taxonomische Stellung von *Magnetococcus* zu beurteilen, wurde die 16S rRNA verschiedener Proteobakterien mit der von *Magnetococcus* verglichen, indem ein Alignment mit den jeweiligen Sequenzen erstellt wurde. Dieses wurde mit `Splitstree` und der `NeighbourNet`-Methode als Grundlage für ein phylogenetisches Netzwerk verwendet. *Magnetococcus* zweigte dabei basal zu den $\alpha$-Proteobakterien ab. Das heißt es ließ sich ein Split identifizieren, der *Magnetococcus* zusammen mit den (anderen) $\alpha$-Proteobakterien zu einer monophyletischen Gruppe zusammenfaßte. Dieser Split wies eine Bootstrapunterstützung von 61 % auf. Auch mit anderen Methoden ließ sich dieser Split identifizieren und zwar bei *neighbor-joining* mit einer Bootstrapunterstützung von 73 %, bei *maximum-likelihood* mit einer Unterstützung von 45 %.

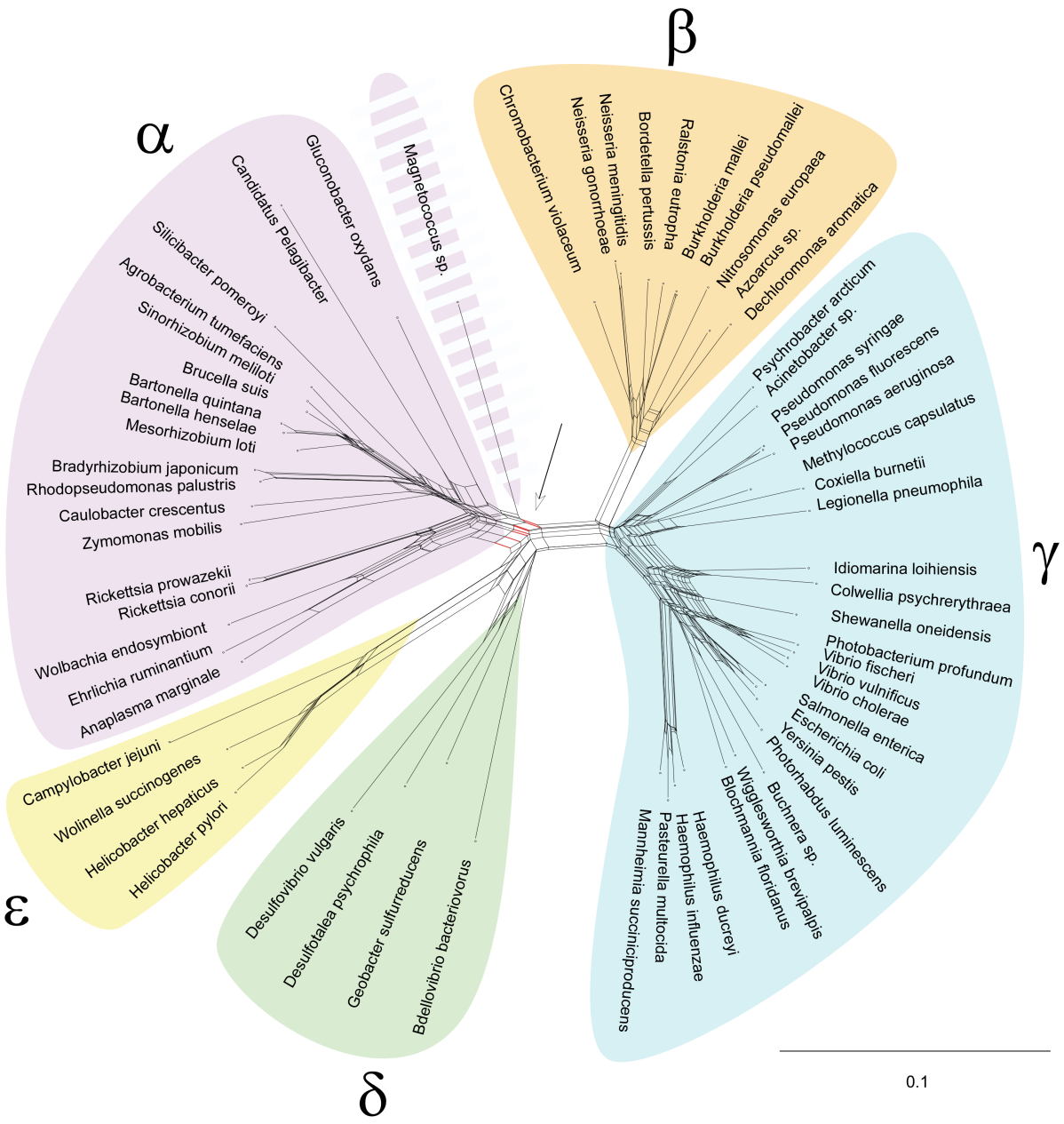



**Abbildung 5.13:** Phylogenetisches Netzwerk von *Magnetococcus* und anderen Proteobakterien erstellt mit der `NeighborNet`-Methode (Huson und Bryant, 2006) aus 16S rRNA-Sequenzen verschiedener Proteobakterien. Die einzelnen Gruppen sind farbig markiert. Der rot hervorgehobene Split der *Magnetococcus* und die $\alpha$-Proteobakterien von den übrigen Klassen trennt, hat eine Bootstrapunterstützung von 73% mit der *neighbor-joining* Methode und 45% mit der *maximum-likelihood*-Methode.

# 6 Identifizierung konservierter Genreihenfolgen bei *E. coli* und Verwandten

## 6.1 Syntenie

Viele Methoden zur Analyse von lateralem Gentransfer benutzen Proteincluster beziehungsweise Proteinfamilien für ihre Abschätzungen (Atkinson et al., 2009). Daher spielt die Qualität dieser Ausgangsdaten eine essentielle Rolle für die Qualität der Ergebnisse. Da die meisten Methoden ausschließlich Sequenzdaten beziehungsweise Sequenzähnlichkeiten für die Erstellung von Proteinfamilien verwendeten, war ein Ziel der vorliegenden Arbeit durch den Einsatz zusätzlicher Daten die Proteinfamilien zu verbessern.

Der Begriff Syntenie wurde ursprünglich aus dem Griechischen (*syn* = zusammen, *tainia* = Band) abgeleitet und bezeichnete Genorte, die auf demselben Chromosom zu finden waren. Inzwischen wird auch von Syntenie gesprochen, wenn Gene in verschiedenen Organismen dieselbe Reihenfolge auf dem Chromosom aufweisen. Synteniekarten stellen dabei Genabschnitte unterschiedlicher Organismen dar, die aneinander ausgerichtet werden (Lin et al., 2008; Welch et al., 2002; Szpirer et al., 1998).

Die Reihenfolge von Genen auf dem Chromosom ist bei eng verwandten Prokaryoten oft sehr gut konserviert. Mit zunehmender Distanz nimmt die Konservierung rapide ab und bleibt nur in bestimmten Bereichen erhalten (Tamames, 2001). Operons sind operative Einheiten, die mehrere funktionell meist zusammengehörende Gene enthalten. Da hier die Regulation über gemeinsame Promotoren erfolgt, ist die Reihenfolge oft auch über enge Verwandtschaftsbereiche hinaus erhalten geblieben.

Bei der Analyse eukaryotischer Genome spielt die Syntenie eine wesentlich größere Rolle, da die Reihenfolge hier weniger störenden Einflüssen ausgesetzt ist. Beim Menschen lassen sich große Änderungen in der Syntenie mikroskopisch durch spezielle Färbetechniken erkennen (Karyogramm). Auf diese Weise lassen sich schwere Erbkrankheiten frühzeitig erkennen (Ichioka et al., 2005).

Bei der Hefe wurde die Syntenie sehr ausgiebig untersucht, da es hier einige Spezies gibt, bei denen im Laufe der Evolution eine Verdopplung des gesamten Genoms stattgefunden hat. Hier existiert ein Browser mit dem sich die Genreihenfolge in den verschiedenen Genomen, sowie bei duplizierten Genomen den einzelnen Kopien betrachten lässt (Byrne und Wolfe, 2005).

Duplikationen von Genen spielen generell eine große Rolle in der Evolution. Durch Duplikationen entstehen zwei Paraloge, von denen eine Kopie als Sicherung fungieren kann, wenn Mutationen in der anderen Kopie auftreten, was den Selektionsdruck unter Umständen erheblich verringern kann. Während Paraloge bei allein auf Sequenzähnlichkeit basierenden Verfahren gleichwertig behandelt werden, können synteniebasierte Verfahren zwischen den Kopien unterscheiden und so die Geschichte unterschiedlicher Kopien getrennt verfolgen.

*E. coli* ist im Bereich der Mikrobiologie einer der wichtigsten prokaryotischen Modellorganismen und sehr gut untersucht (Blattner et al., 1997). Es ist ein Enterobakterium aus dem Stamm der $\gamma$-Proteobakterien, das auch im Darm des Menschen vorkommt. Wegen der großen Bedeutung in medizinischer, wissenschaftlicher und auch wirtschaftlicher Hinsicht wurde und wird viel Energie in die Erforschung dieses Organismus gesteckt. Dadurch stehen eine Vielzahl vollständig sequenzierter Stämme zur Verfügung, was *E. coli* ebenfalls zu einem sehr interessanten Modellorganismus für die Bioinformatik macht. Die Anzahl der verfügbaren Genome ist innerhalb der letzten Jahre von neun (2008) auf 28 (2010) angestiegen. Durch die hohe Sequenzähnlichkeit lassen sich genomweite Analysen durchführen, die bei vielen anderen Organismen nur für wenige Gene möglich wären. Die Reihenfolge der Gene ist sehr konserviert, so daß sich diese Art sehr gut zur Entwicklung von synteniebasierten Algorithmen eignet (siehe Abbildung 6.1).

Um die Syntenie in Genomen der verschiedenen *E. coli* Stämme zu untersuchen, müssen Genfamilien (Cluster) für die sequenzierten Genome erstellt werden. In der Bioinformatik werden Gen- oder Proteinfamilien als Ausgangsmaterial für viele Analysen verwendet. Die Begriffe Gen und Protein werden in der vorliegenden Arbeit

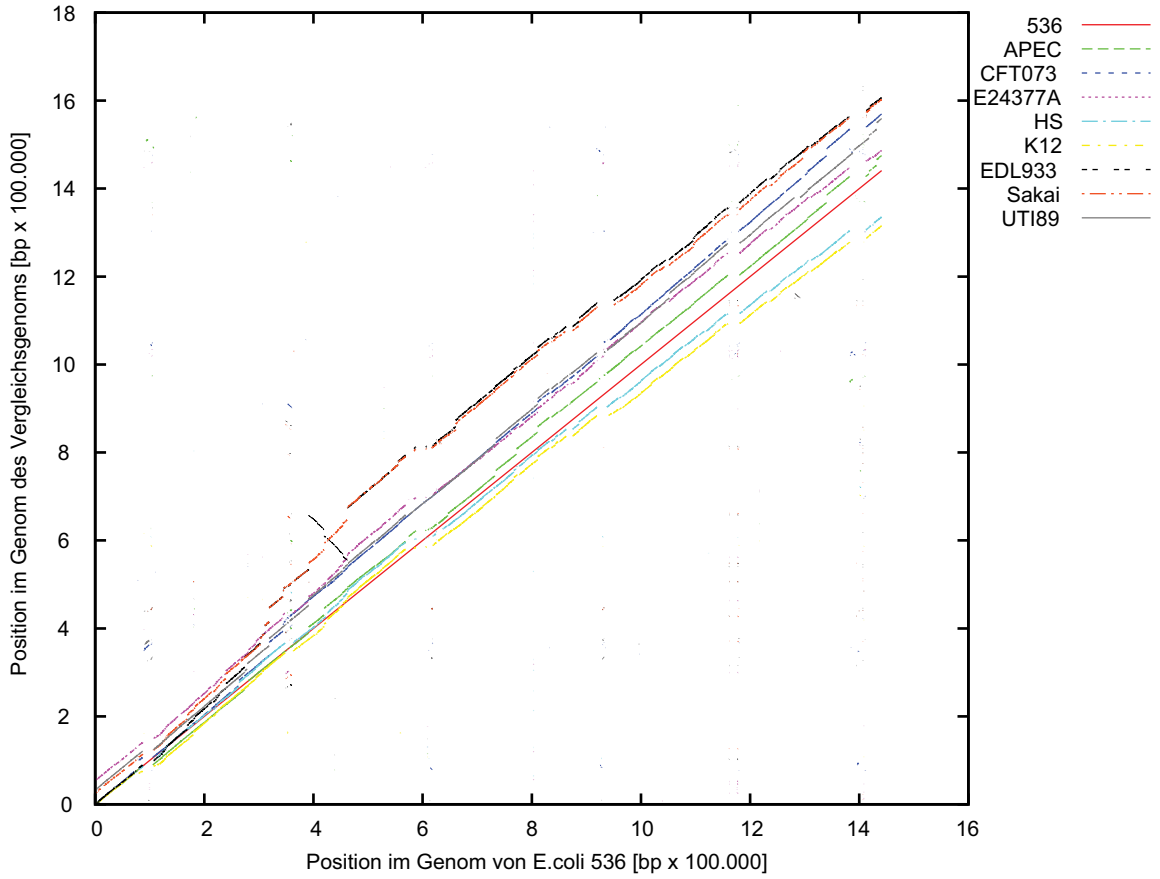



**Abbildung 6.1:** Dotplot für die neun betrachteten *E. coli* Genome beruhend auf Alignments von *E. coli* 536 mit den acht anderen untersuchten *E. coli* Stämmen erstellt mit dem Programm `MUMmer`.

synonym verwendet, da es sich bei den verwendeten Nukleotidsequenzen, falls nicht anders erwähnt, um offene Leseraster (engl. *open reading frames* – ORFs) handelt, die für Proteine kodieren. So werden beispielsweise universelle Proteinfamilien, in denen von jedem Organismus ein Protein enthalten ist, für die Erstellung von Genbäumen und Phylogenien verwendet. Außerdem kann untersucht werden wie sich bestimmte Funktionen über die Vielfalt der Organismen verteilen.

Eine gängige Methode zur Erstellung von Genfamilien ist das den Markov-Cluster-Algorithmus nutzende Programm `mcl` (Enright et al., 2002), das hier in der in Kapitel 4.2.8 beschriebenen abgewandelten Methode verwendet wurde. Für die Erstellung von Positionsorthologen wurden neben Sequenzdaten auch Informationen über die Reihenfolge der Gene auf dem Chromosom einbezogen. Die Methode der Bestimmung sorgt dafür, daß jedes Gen nur in eine Familie eingeordnet werden kann. Während bei mit anderen Methoden erstellten Clustern auch mehrere Kopien eines Gens in eine Gruppe eingeordnet werden können, ist dies hier nicht möglich. Dadurch lassen sich die verschiedenen Kopien eines Gens in den unterschiedlichen Genomen verfolgen.

Durch die zusätzlich enthaltenen Informationen lassen sich die verschiedenen Datensätze positionsbezogen auswerten, was die Entdeckung von lateralem Gentransfer, Rekombination und Inversionen wesentlich erleichtert. So läßt sich bei Genen ohne Homologe im betrachteten Datensatz analysieren, ob es sich um ein durch Mutationen entstandendes Gen (Die Position ist in verwandten Genomen vorhanden, hat aber keine ausreichende Sequenzidentität mehr.) oder um einen lateralen Gentransfer handelt (In den homologen Bereichen ist an dieser Stelle kein Gen vorhanden.). Durch die Einbeziehung von Nukleotiddaten läßt sich ein möglicher Genverlust in den anderen Genomen ausschließen oder bestätigen.

## 6.2 Positionsbezogene Sequenzähnlichkeiten auf Genomebene bei *Salmonella*

Die Idee Proteinfamilien aufgrund von Syntenie zu erstellen setzt voraus, daß es in allen Abschnitten eines Genoms eine ausreichende Homologie gibt. Eine Analyse des Genoms von *Salmonella typhi* sollte zeigen wie die Homologien über das Genom verteilt sind. Zu diesem Zweck wurde das Genom in 500 bp lange Sequenzen zerlegt.

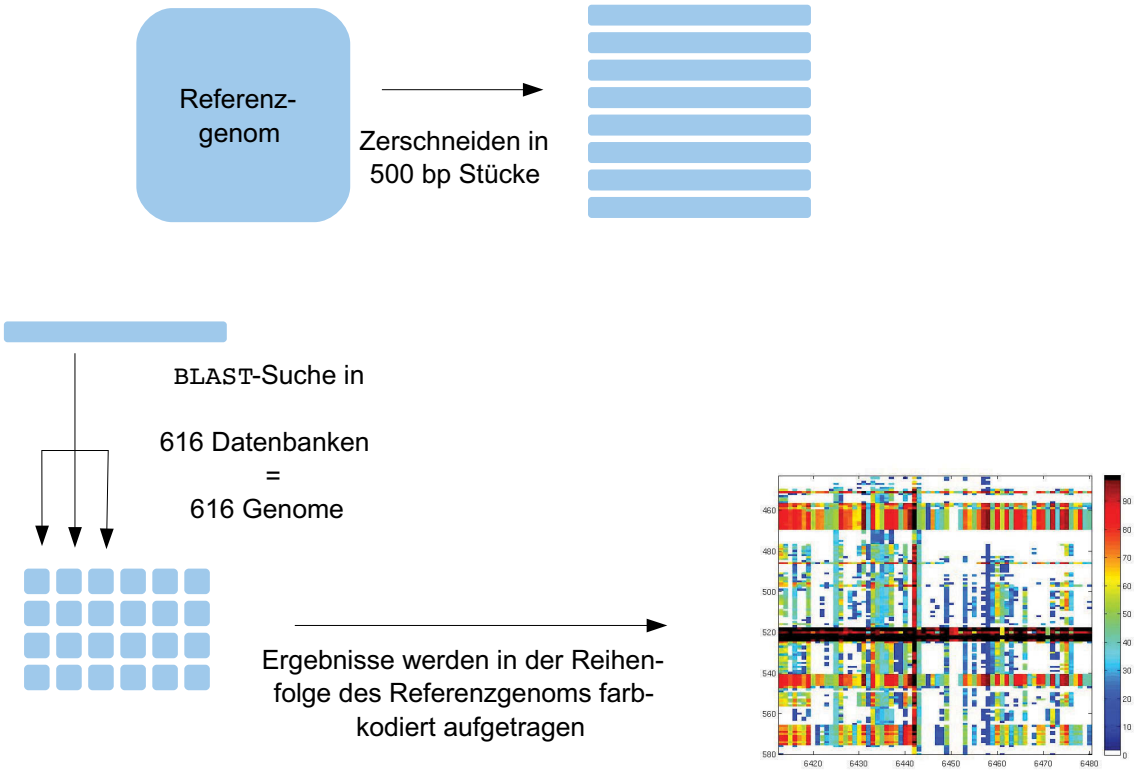



**Abbildung 6.2:** Ablauf der Erstellung eines genomverankerten Identitätsplots. Das Referenzgenom wird in gleichlange Abschnitte zerlegt. Jeder dieser Abschnitte wird in einer Datenbanksuche in allen Genomen der Datenbank gesucht. Die Identität des besten Treffers wird in der durch das Referenzgenom vorgegebenen Reihenfolge eingetragen.

Jede dieser Sequenzen wurde mit dem Programm `blastn` (siehe Seite 13) mit einer Datenbank bestehend aus 616 Genomen verglichen. Dabei entstehen großflächige Plots, bei denen jede waagerechte Linie einem Genom entspricht. Die Reihenfolge der Genome entspricht ihrer taxonomischen Einordnung. Jede senkrechte Linie entspricht einem DNA Abschnitt auf dem Referenzgenom und den besten Treffern in den anderen Genomen. Die Konservierung der Reihenfolge spielt bei diesen Darstellungen keine große Rolle, da nur die Reihenfolge des ReferenzGenoms berücksichtigt wird und nicht ausschlaggebend ist, ob ein Treffer als solcher gewertet wird oder nicht.

Bei dieser Darstellungsform wurde die Nukleotidsequenz eines Organismus in gleichgroße Teile zerschnitten. Mit diesen wird eine Datenbanksuche durchgeführt, und die Identität wird farbkodiert dargestellt. Die Darstellung erfolgt dabei in Abhängigkeit der Reihenfolge der Sequenzabschnitte auf dem Referenzgenom. Die Ergebnisse wurden in einer Grafik zusammengefasst (Abb. 6.3), in der jede Zeile einem Genom entspricht und jede Spalte einem 500 bp langen Abschnitt des Referenzgenoms. Die Qualität des Treffers wurde farbkodiert eingetragen. Kein Treffer oder ein Treffer unterhalb des Schwellenwerts blieben weiß. Andere Treffer wurden entsprechend der normalisierten Identität von blau über grün und gelb bis rot eingefärbt. Treffer mit 100 % Identität wurden in schwarz dargestellt.

Die Ergebnisse zeigen, daß auch für andere Stämme von *Salmonella* eine große Abdeckung mit homologen Sequenzabschnitten detektierbar ist. Für keinen der Stämme läßt sich eine 100 %ige Abdeckung erzielen. Während in einigen nah verwandten Arten (beispielsweise *Escherichia coli*) der $\gamma$-Proteobakterien noch einige Abschnitte zu erkennen sind, beschränkten sich die Bereiche, in denen auch für entferntere Arten homologe Sequenzen entdeckt werden konnten, auf sehr wenige eng begrenzte Bereiche, beispielsweise RNA-kodierende Sequenzen.

Die gleiche Analyse wurde ebenfalls mit einer anderen Variante des Programms `BLAST` durchgeführt: `tblastx`. Dabei wird eine Nukleotidsequenz mit einer Nukleotiddatenbank verglichen, beide werden aber vorher in alle sechs Leseraster übersetzt. Diese Methode ist unempfindlicher gegen Mutationen und auch bei Leserasterverschiebungen können Treffer erzielt werden (Abb 6.4). Es zeigte sich, daß sich die Nutzung der in Aminosäuren übersetzten Sequenzen durch eine stark erhöhte Anzahl von Treffern bemerkbar macht. Für die anderen Stämme von

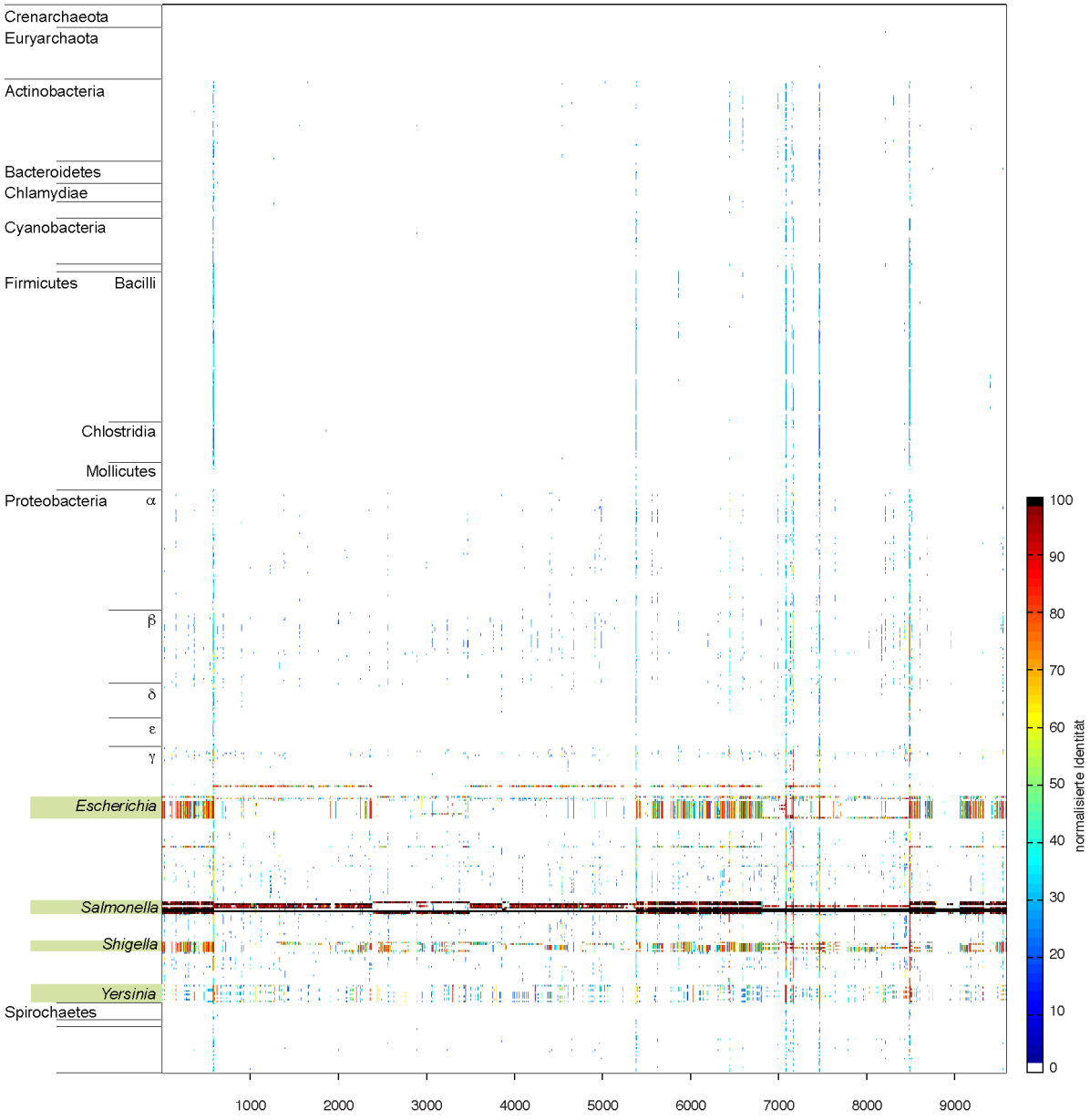



**Abbildung 6.3:** Genomverankerter Identitätsplot für *Salmonella typhi* basierend auf einer Suche von 500 bp langen Sequenzstücken mit Hilfe von `blastn`. Auf der Y-Achse finden sich die Genome sortiert nach ihrer taxonomischen Einordnung. Auf der X-Achse finden sich die 500 bp langen Sequenzstücke in der gleichen Reihenfolge wie auf dem Referenzgenom. Die Reihenfolge der Treffer ist unabhängig von der Position im Vergleichsgenom. Die Farbkodierung stellt die (auf 500 bp normalisierte) Identität zwischen der Suchsequenz und dem besten Treffer in dem jeweiligen Genom dar.

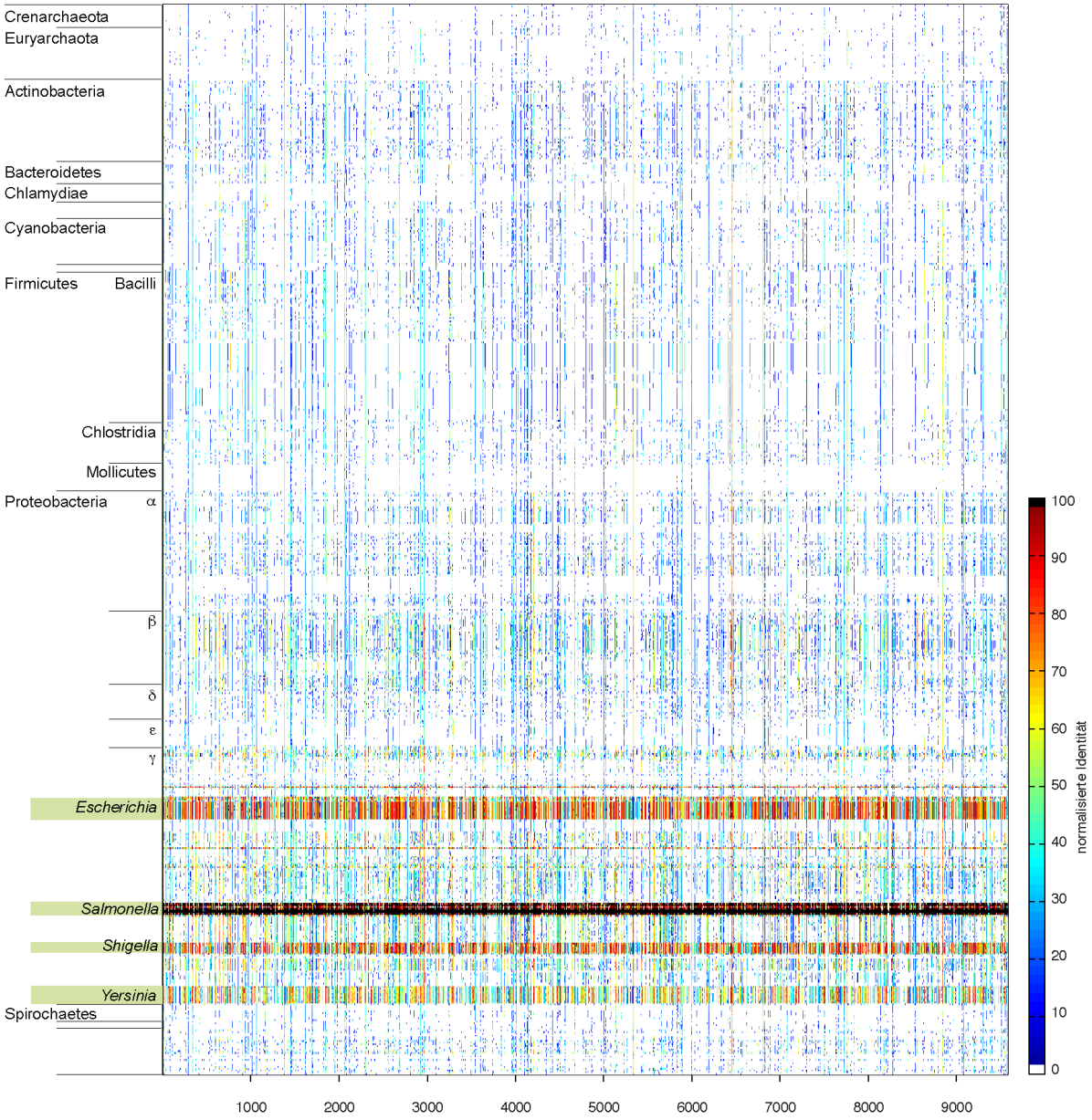



**Abbildung 6.4:** Genomverankerter Identitätsplot für *Salmonella typhi* basierend auf einer Suche von 500 bp langen Sequenzstücken mit Hilfe von `tblastx`. Auf der Y-Achse finden sich die Genome sortiert nach ihrer taxonomischen Einordnung. Auf der X-Achse finden sich die 500 bp langen Sequenzstücke in der gleichen Reihenfolge wie auf dem Referenzgenom. Die Reihenfolge der Treffer ist unabhängig von der Position im Vergleichsgenom. Die Farbkodierung stellt die (auf 500 bp normalisierte) Identität zwischen der übersetzten Suchsequenz und dem besten übersetzten Treffer in dem jeweiligen Genom dar.

*Salmonella* lassen sich fast über das komplette Genom Abschnitte mit sehr hohen Identitäten finden.

Innerhalb der $\gamma$-Proteobakterien existieren zahlreiche Gattungen, die über sehr große Abschnitte eine hohe Sequenzidentität aufweisen. Es lassen sich auch in anderen Stämmen der Eubakterien und sogar bei den Archaea Treffer finden.

## 6.3 Synteniebasierte orthologe Gene (SOGs)

Die bisher beschriebenen Ergebnisse geben die Menge der Homologie zu bestimmten Abschnitten eines Referenzgenoms an. Das `PERL`-Skript `synti.pl` dagegen identifiziert Genfamilien aufgrund ihrer Position im Genom. Dabei handelt es sich um sogenannte synteniebasierte Orthologe (SOG's). Die Beurteilung erfolgt dabei im Gegensatz zu anderen Methoden (Rodelsperger und Dieterich, 2008) nicht auf der Gensequenz verschiedener Organismen sondern auf der Basis von Proteinen. Die Proteine werden zunächst aufgrund ihrer Sequenzähnlichkeit in einer Datenbanksuche mit dem Programm `blastp` (Altschul et al., 1997) einander zugeordnet. Diese Proteine werden jeweils als einzelne Zeichen betrachtet, die innerhalb ihres Genoms eine bestimmte Position haben. Der Algorithmus versucht nun die Auswahl der Orthologen so zu treffen, daß möglichst viele Proteine in derselben Reihenfolge in dem Referenz- und dem Zielgenom vorliegen. Auf diese Weise sollen bei der Erstellung von Proteinfamilien solche Proteine bevorzugt werden, die nicht unbedingt die größte Sequenzähnlichkeit sondern bei einer ausreichenden Sequenzähnlichkeit die ähnlichste Position haben.

Synteniekarten dienen zur Veranschaulichung der Reihenfolge von Genen auf den Chromosomen unterschiedlicher Organismen. Bei Eukaryoten sind sie sehr verbreitet, um die Abfolge bestimmter Abschnitte auf den Chromosomen zu vergleichen (Szpirer et al., 1998). Hier soll versucht werden mittels Syntenie sogenannte Positionsorthologe zu bestimmen. Dabei wird davon ausgegangen, daß die Position von Genen in einigen Fällen konservierter ist als die Sequenz selber. Dies ist zum Beispiel dann der Fall wenn ein Gen durch eine Duplikation doppelt vorliegt, so daß der Selektionsdruck auf die einzelne Kopie sinkt. Ob dabei auf lange Sicht die Kopie oder das Original die Funktion übernimmt, ist bei vollständigen Kopien (inklusive regulatorischer Abschnitte) ungewiß. Durch die Verwendung des Synteniekriteriums

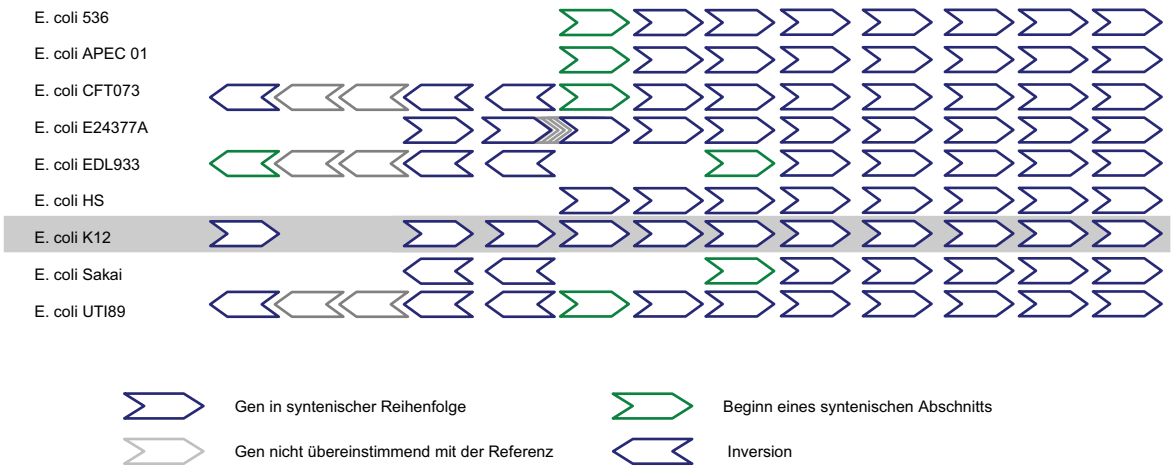



**Abbildung 6.5:** Beispielhafte Darstellung eines Abschnittes des *E. coli* Genoms als Synteniekarte. Grau hinterlegt ist das Referenzgenom, an dem die anderen Genome ausgerichtet werden.

soll dabei sichergestellt werden, daß es sich bei den erhaltenen Genfamilien um „wahre“ Orthologe handelt.

Um synteniebasierte Genfamilien zu erstellen, erfolgt eine Berechnung in drei Schritten. Im ersten Schritt werden die Protein-Tabellen eingelesen und alle benötigten Informationen abgespeichert. Für die eigentliche Berechnung werden die Positionsinformationen für die Gene verarbeitet. Jedem Gen wird eine eindeutige Positions-ID zugeordnet.

Im zweiten Schritt werden die Ergebnisse einer `BLAST`-Datenbanksuche in den Arbeitsspeicher gelesen, sofern sie die Mindestkriterien erfüllen. Hier wird zusätzlich überprüft, ob es sich um reziproke Treffer handelt. In der interaktiven Version findet hier jedoch zunächst nur eine Speicherung statt, da die eigentliche Berechnung später erfolgt. Der einzige Parameter, der nach dem Start des Programms nicht mehr beeinflußt werden kann, ist der Schwellenwert für die Einbeziehung der Daten in die Ergebnisse. Dieser wird am Anfang festgelegt, so daß nicht unnötig viele Daten im Arbeitsspeicher eingelagert werden müssen. Das Kriterium, das für den Ausschluß von Daten verwendet wird, ist die normalisierte Identität, das heißt die Anzahl der identischen Positionen geteilt durch die Länge des kürzeren der beiden verglichenen Proteine. Für die *E. coli* Daten wird hier 70 % als Schwellenwert verwendet. Dieser Wert läßt sich über den Parameter -n ändern. Alternativ kann über den Parameter -e auch ein Schwellenwert für den Erwartungswert (engl. *e-value*) aus der `BLAST`-Ausgabe angegeben werden.

Im dritten Schritt findet die eigentliche Berechnung statt. Diese kann mit mehreren Methoden erfolgen (Abb. 6.2). Hier soll die Standardmethode („*5-Point long run*“) erklärt werden. In der chromosomalen Reihenfolge des gewählten Referenzgenoms wird für jedes Gen versucht in jedem anderem Genom parallele Abschnitte zu finden. Dabei wird zunächst von allen Treffern der Datenbanksuche ausgegangen. Lassen sich mehrere aufeinanderfolgende Sequenzen mit ihren Homologen aneinander ausrichten, so wird diese Reihe solange weiterverfolgt bis sich kein weiterer Treffer mehr finden läßt. Alle beteiligten Gene werden für die beiden betrachteten Gene als bereits verwendet markiert, so daß sie in zukünftige Suchen nicht mehr einbezogen werden. Sollte sich herausstellen, daß sich weniger als fünf Gene am Stück aneinander ausrichten lassen, so werden die Treffer verworfen und die Suche am nächsten Gen fortgesetzt.

**Tabelle 6.1:** Eingaben `synti.pl`

| | Eingaben |
|---|---|
| BLASTFILE | Ausgabe der Datenbanksuche in tabellarischer Form, bei Verwendung von Standardsequenzdaten wie sie auf der Webseite des National Center for Biotechnology Information (NCBI) erhältlich sind |
| PTTFILES | Protein-Tabellen für jedes Genom in der Analyse. Diese enthalten neben weiteren Informationen die Positionsdaten für jedes einzelne Gen des Genoms |

Die Ausgabe des Skripts ist eine Tabelle, die für jedes Genom eine Spalte enthält und eine zusätzliche Spalte in der die Größe der Proteinfamilien angegeben ist. Jede Zeile repräsentiert dabei eine Proteinfamilie. Wenn nur ein Genom als Referenz ausgewählt wird, werden die Proteinfamilien an dieser ausgerichtet. Die Einträge sind die Identifikationsnummern (GI) der zugehörigen Proteine. Werden alle Genome als Referenz ausgewählt, so werden die einzelnen Datensätze nacheinander in die Ausgabedatei geschrieben.

a)

```
# 2009505_1031_xp_008_5_point_R_long_run_NI70_K12.long

536
APEC
CFT073
E24377A
EDL933
HS
K12        *     128936 16127995  thrL     thr operon leader peptide
O157H7
UTI89
1

536       100 99 100003 110640215 thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
APEC      100 98 104632 117622296 thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
CFT073     99 97 109092 26245920  thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
E24377A    99 99 114470 157158747 thrA     aspartokinase/homoserine dehydrogenase I
EDL933     99 99 119227 15799682  thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
HS         99 99 124552 157159469 thrA     aspartokinase/homoserine dehydrogenase I
K12        *     128937 16127996  thrA     fused aspartokinase I and homoserine dehydrogenase I
O157H7     99 99 133071 15829256  thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
UTI89     100 99 138325 91209057  thrA     bifunctional aspartokinase I/homeserine dehydrogenase I
9

536        99 98 100004 110640216 ECP_0003 homoserine kinase
APEC       99 99 104633 117622297 thrB     homoserine kinase
CFT073     99 98 109093 26245921  thrB     homoserine kinase
E24377A    99 99 114471 157158096 thrB     homoserine kinase
EDL933     99 98 119228 15799683  thrB     homoserine kinase
HS         99 99 124553 157159470 thrB     homoserine kinase
K12        *     128938 16127997  thrB     homoserine kinase
O157H7     99 98 133072 15829257  ECs0003 homoserine kinase
UTI89      99 99 138326 91209058  thrB     homoserine kinase
9
```

b)

```
-         -         -         -         -         -         16127995 -        -         1
110640215 117622296 26245920  157158747 15799682 157159469 16127996 15829256 91209057  9
110640216 117622297 26245921  157158096 15799683 157159470 16127997 15829257 91209058  9
110640217 117622298 26245922  157157451 15799684 157159471 16127998 15829258 91209059  9
110640219 117622299 26245924  -         -         -         16127999 -        91209060  5
110640220 117622300 26245927  157155229 15799686 157159473 16128000 15829260 91209062  9
110640221 117622301 26245928  157154855 15799687 157159474 16128001 15829261 91209063  9
110640222 162317588 161486342 157158697 15799688 157159475 16128002 15829262 162138515 9
110640223 117622303 26245930  157158038 15799689 157159476 16128003 15829263 91209065  9
```

**Abbildung 6.6:** Beispiel: Ausgabe von `synti.pl` im Standardformat. Abbildung a) zeigt die ausführliche Variante („long“). Hierbei werden neben den Namen der Organismen, die Namen der Gene, verschiedene Genidentifikatoren und die normalisierte Identität zu der Referenz (in diesem Beispiel K12) dargestellt. Die Kurzfassung die in Abbildung b) gezeigt ist, enthält nur die GI für jedes Protein und nur eine Zeile pro Familie. Die Reihenfolge der Genome ist alphabetisch wie bei der ausführlichen Variante.

Im Laufe der Entwicklung des Algorithmus entstanden zahlreiche Varianten, von denen die Wichtigsten in Tabelle 6.2 aufgeführt sind. Die Vorbereitung ist bei allen Methoden gleich. Nur `BLAST`-Ergebnisse, die den Anforderungen genügen, werden berücksichtigt. Alle beschriebenen Methoden können wahlweise nur reziproke

BLAST-Treffer benutzen (engl. *bidirectional blast hit*). Optional können weitere Ausgaben gewählt werden, um zusätzliche Informationen zu erhalten.

Das Perlskript bietet die Wahl zwischen 5 verschiedenen Ausgaben, die unterschiedliche Anwendungsmöglichkeiten haben.

- .long – Eine ausführliche Ausgabe (Abb. 6.6 a) erlaubt die detaillierte Analyse der Ergebnisse. Jede Proteinfamilie wird durch einen Block beschrieben, der eine Zeile für jedes im Datensatz vorhandene Genom enthält. Für jeden Organismus steht in dieser Liste, ob ein Gen an der Proteinfamilie beteiligt ist, welche Funktion es hat und an welcher Position es zu finden ist. Abschließend folgt eine Zeile, in der die Anzahl der an der Familie beteiligten Gene aufgeführt ist, gefolgt von einer Leerzeile zur Trennung der Einträge.
- Drei Varianten der kurzen Ausgabe (Abb. 6.6 b). Hier wird jede Familie durch eine Zeile dargestellt. Die am Datensatz beteiligten Genome werden jeweils in den Spalten dargestellt. Die Reihenfolge entspricht alphabetisch den Bezeichnungen der Genome und ist in einer .log Datei abgespeichert. Ein „-“ steht dafür, daß in dem entsprechenden Genom kein Homolog zu der Proteinfamilie gefunden wurde. Die letzte Spalte enthält die Anzahl der Mitglieder der Proteinfamilie. Die in dieser tabellarischen Ausgabe enthaltenen Werte unterscheiden sich je nach Variante:

  .sid — Die von dem Programm vergebenen Positions-IDs werden ausgegeben. Das ermöglicht die schnelle Identifikation von Sprüngen innerhalb eines Genoms oder das Erkennen von Inversionen, bei denen absteigende statt aufsteigender Positions-IDs gefunden werden.

  .synti — Diese Dateien enthalten die Gen-Identifikationsnummer (GI) wie sie beim NCBI verwendet werden.

  .gid — Dateien mit dieser Endung enthalten als Schlüssel die Kurzbezeichnungen der zugehörigen Gene (zum Beispiel „lacZ“). Diese Darstellung erlaubt eine einfache optische Überprüfung der Ergebnisse, weil für vollständig annotierte Genome, Gene mit einer identischen Funktion in der Regel identische Kürzel erhalten.

**Tabelle 6.2:** Varianten des Algorithmus zur Bestimmung von Positionsorthologen. X bezeichnet die Anzahl von Treffern, die in der richtigen Reihenfolge vorliegen müssen, um im Gesamtergebnis berücksichtigt zu werden. Y steht für die Anzahl der Lücken, die zwischen zwei Treffern erlaubt sind. Ein * hinter diesen Werten signalisiert, daß es sich um abänderbare Parameter handelt.

| Variante | Funktionsweise | X | Y | Ergebnis |
|---|---|---|---|---|
| 2 Punkt (engl. *2-point*) | Werden zwei Treffer in zwei Genomen gefunden, werden sie berücksichtigt. Durch eine quadratische Lückenbestrafung werden möglichst nahe Treffer bevorzugt | 2 | | Sehr viele Gene lassen sich in Gruppen einordnen, allerdings sind auch viele nicht konservierte Gruppen dabei. |
| 3 Punkt (engl. *3-point*) | Gleicher Ansatz wie bei der 2 Punkt Methode, allerdings werden mindestens 3 Treffer gefordert | 3 | | Immer noch einige falsch positive Treffer, aber auch fehlende Gruppen aufgrund des strikten Kriteriums |
| 5 Punkt lange Läufe (engl. *5-point-long-run*) | Hierbei wird von einem Startpunkt aus solange weitergegangen bis kein weiterer Treffer mehr erzielt werden kann. Ist die Anzahl der Treffer kleiner x, so werden die eingetragenen Treffer wieder aus den Ergebnissen entfernt. | 5* | 5* | Beste Ergebnisse unter den getesten Methoden. Sehr wenige falsch Positive, kaum falsch Negative. Durch die Verwendung von Parametern läßt sich die Methode den Bedürfnissen anpassen. |
| ... reziprok (engl. *reciprocal*) | Alle Methoden existieren auch in einer reziproken Variante. Dabei werden nur Treffer berücksichtigt, bei denen das getroffene Gen auch die Suchsequenz findet. | - | - | Die Ergebnisse unterscheiden sich nur marginal von denen ohne reziproke Suche. Da es jedoch das stringentere Kriterium ist, wurde es als Standard verwendet. |

- .matlab – Bei diesen Dateien handelt es sich um Matrizen im sogenannten „sparse“-Format, das heißt nur von 0 abweichende Werte werden eingetragen. Es handelt sich hierbei um Dotplots, in denen die Position von Genen in den Datenbankgenomen gegen die Position in der Referenz aufgetragen ist. Die eingetragenen Werte entsprechen der laufenden Nummer des Genoms. Die Matrix läßt sich zum Beispiel mit dem Programm `MATLAB` einlesen und darstellen, und vermittelt einen Eindruck über die Konservierung innerhalb der betrachteten Genome.

Im folgenden Kapitel wird die Methode zur Erstellung von Positionsorthologen auf einen Datensatz von *E. coli* angewendet und mit Daten eines anderen Verfahrens zur Erstellung von Genfamilien verglichen.

# 7 Der Einfluß der Genomgröße auf die Ableitung von Genaustauschen bei *E. coli*

## 7.1 Genverteilungmuster

Im Laufe der Evolution sind prokaryotische Genome dem Prozeß des lateralen Gentransfers (LGT siehe Seite 62) unterworfen, also dem Austausch von Genen zwischen Organismen, die keine direkte Verwandtschaft miteinander aufweisen. Eine der drei gängigen Methoden zur Ableitung von lateralem Gnetransfer ist der Vergleich von Genbäumen mit Phylogenien der Organismen. Die Häufigkeit von LGT-Ereignissen beeinflußt die Größe von Genomen und die Zusammensetzung derselben.

Darausfolgend wurde eine Häufigkeit für LGT in Prokaryoten von einem Ereignis pro Proteinfamilie und Lebenszeit abgeleitet (Dagan und Martin, 2007). Diese Analysen zeigen wie die Berechnungen aus Kapitel 5 welchen Einfluß lateraler Gentransfer zwischen verschiedenen Arten auf die Genomevolution von Prokaryoten hat. Im folgenden Kapitel soll untersucht werden, inwieweit auch Gentransfer zwischen verschiedenen Stämmen einer Art an der Genomevolution beteiligt ist.

Für die Analysen in diesem Kapitel wurden mithilfe zweier Methoden aus neun *Escherichia coli* Genomen Genfamilien errechnet. Die Genome werden im Folgenden mit den Bezeichnungen der Stämme benannt, das vorangestellte „*E. coli*“ wird weggelassen. Die Liste mit den vollständigen Namen findet sich im Anhang auf Seite 110.

Der erste Datensatz enthält mit `mcl` (siehe Kapitel 4.2.8) erstellte Genfamilen. Diese werden im Folgenden als COGs (engl. *Clusters of orthologues Genes*) bezeichnet.

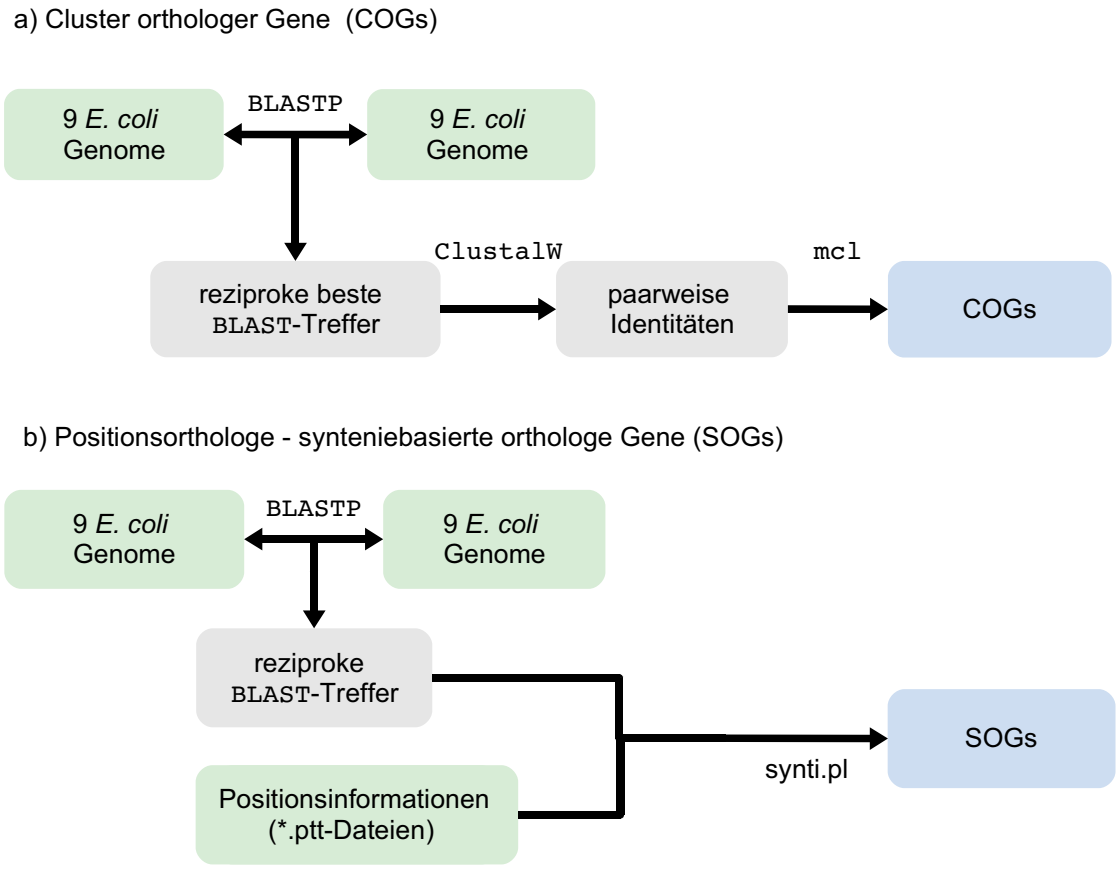



**Abbildung 7.1:** Vergleich der Arbeitsabläufe für die Erstellung von Clustern über Sequenzhomologie (COG) und über die Position auf dem Chromosom (SOG)

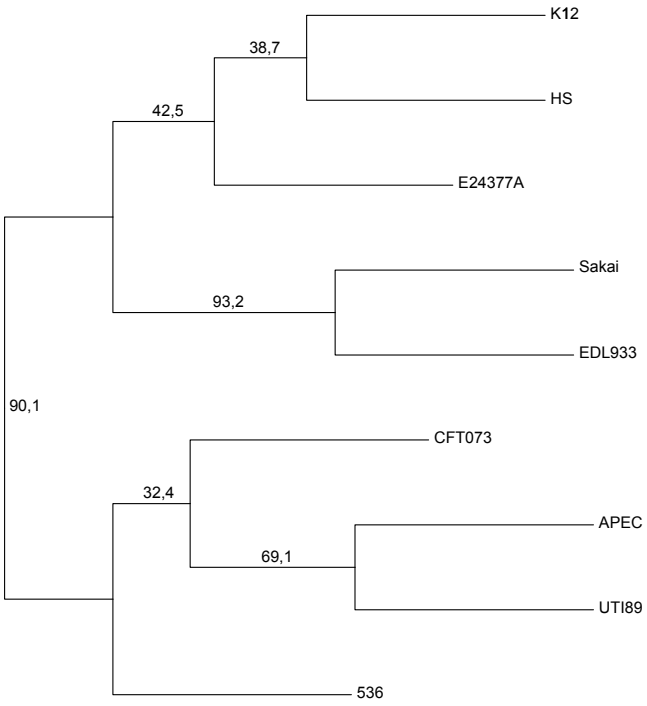



**Abbildung 7.2:** Konsensusbaum erstellt aus *neighbor-joining*-Bäumen, für die Nukleotidsequenzen von 2.584 universellen Genfamilien von *Escherichia coli*. Dieser Baum wurde für die Berechnung der Netzwerke der Gentransfers als Referenztopologie benutzt. Die Zahlen geben den Anteil der Einzelbäume in Prozent an, in denen der betreffende Ast enthalten war.

Da bei dieser Methode mehrere Proteine je Genom in eine Genfamilie eingeordnet werden können (Paraloge), wurde hier ein sehr stringenter Schwellenwert benutzt. Erst bei 90 % Sequenzidentität wurden Treffer für die Clustererstellung berücksichtigt. Bei diesem Schwellenwert traten in den universellen Clustern weniger als 1 % Paraloge auf. Bei niedrigeren Schwellenwerten wiesen bis zu 6,1 % der universellen Proteinfamilien Paraloge auf.

Der zweite Datensatz wurde synteniebasiert mit Hilfe der vorher beschriebenen Methode (Kapitel 6.3) erstellt. Da bei dieser Methode grundsätzlich keine Paralogen berücksichtigt werden, konnte hier ein vergleichsweise niedriger Schwellenwert gesetzt werden. Da die Einordnung jedoch nicht aufgrund der Sequenzidentität sondern aufgrund der Position getroffen wurde, spielt dieser Wert nur eine untergeordnete Rolle. Die Proteinfamilien aus diesem Datensatz wurden analog zum Begriff COG als synteniebasierte orthologe Gene (kurz SOGs) bezeichnet.

Für die Genverteilungsmuster wurde zunächst eine Anwesenheitsverteilung aller Proteinfamilien durchgeführt (engl. *presence-absence-pattern* – PAP). Dabei handelt es sich um eine Matrix, in der jede Zeile einer Genfamilie entspricht und jede Spalte einem Genom. Enthält die aktuelle Genfamilie mindestens ein Gen aus dem betrachteten Genom, so wird an der entsprechenden Stelle eine 1 gesetzt ansonsten eine 0. Diese werden absteigend entsprechend der Anzahl der vorhandenen Genome und anschließend aufsteigend numerisch sortiert. Proteinfamilien die mehrfach auftreten, werden durch die Sortierung hintereinander aufgeführt, so daß die Höhe eines Musters proportional zu seiner Häufigkeit ist.

## 7.2 Netzwerke der Rekombination

Die Methode zur Erstellung von Netzwerken lateralen Gentransfers (Dagan et al., 2008), die hier Netzwerke der Rekombination genannt werden, da es sich bei den analysierten Gentransfers um Übertragungen innerhalb einer Art handelt, erfordert eine Referenztopologie. Für diese wurde aus 2.584 *neighbor-joining*-Bäumen, von den zu universellen Genfamilien gehörenden Nukleotidsequenzen, ein Konsensus-Baum berechnet (Abb. 7.2 auf der vorherigen Seite). Dazu wurden die Nukleotidsequenzen mit `ClustalW` aligniert und mit den `PHYLIP`-Programmen `dnadist` und `neighbor` *neighbor-joining*-Bäume berechnet. Dann wurde aus diesen Bäumen mit dem Programm `consense` ein Konsensusbaum erstellt.

Im ersten Schritt wurde überprüft, welches Modell auf die verwendeten Daten zutrifft:

loss-only Jede Genfamilie ist bereits im letzten gemeinsamen Vorfahren vorhanden, besitzt also einen festgelegten Genursprung. Abweichende Genverteilungen werden immer durch den Verlust von Genen erklärt.

ori1 Es ist ebenfalls nur ein Genursprung vorhanden, dieser ist aber nicht zwangsläufig identisch mit der Wurzel, sondern kann frei auf der Referenztopologie festgelegt werden. Des Weiteren sind nur Genverluste möglich.

oriX Die Erweiterungen des Modells ori1 erlauben X (= 2, 4, 8 oder 16) Ursprünge die auf dem Baum verteilt werden können. Wird für eine Genfamilie mehr als ein Ursprung ermittelt, so handelt es sich dabei um abgeleiteten Gentransfer.

Die Abschätzung welches Modell zutrifft, erfolgte über eine Berechnung der Vorläufergenomgröße (engl. *ancestral genome size*). Abhängig von den abgeleiteten Ursprungsorten und Genverlusten wurde für jeden hypothetischen Knoten (engl. *hypothetical taxonomic unit* – HTU) in der Referenztopologie und für jede Genfamilie ermittelt, ob ein Vertreter im entsprechenden (hypothetischen) Organismus vorhanden war oder nicht. Daraus ließ sich für jeden HTU eine Genomgröße berechnen.

Die HTU-Genomgrößen wurden für die verschiedenen Modelle mit den tatsächlichen Genomgrößen verglichen. Wurde kein LGT zugelassen, so war die Größe der hypothetischen Genome wesentlich größer als die der tatsächlichen, da jedes gegenwärtig beobachtete Gen im Vorläufer enthalten sein mußte. Wurde dagegen unbegrenzt LGT erlaubt, so sank die Vorläufergenomgröße sehr stark ab. Daher wurde als das zu den verwendeten Daten passende Modell dasjenige ausgewählt, das für die HTU Genomgrößen die zu den gegenwärtigen Genomgrößen ähnlichste Verteilung produzierte.

Für das ausgewählte Modell wurden die abgeleiteten Transferereignisse auf entsprechende Kanten verteilt (engl. *inferred LGT*). Mit Hilfe eines Matlab Skriptes wird zunächst die Referenztopologie in Schwarzschattierungen (Farbkodierung entsprechend der Genomgröße) wiedergegeben. Die Transfers wurden in farbigen Linien (Farbkodierung entsprechend der Menge des lateralen Gentransfers zwischen den beteiligten Knoten) auf die Referenztopologie projiziert.

Gene, die keiner Proteinfamilie zugeordnet werden können, werden als *Singletons* bezeichnet. Für diese wurde eine Analyse nach den nächsten Nachbarn entsprechen des Kriteriums „bester Treffer in einer `BLAST`-Suche“ (BBH) entsprechend der Methode in Kapitel 5.3 auf Seite 34 durchgeführt.

Für den Datensatz bestehend aus neun *E. coli* Stämmen wurde sowohl für die mit `mcl` generierten (COG) Daten als auch für die auf Positionsorthologen basierenden (SOG) Daten das LGT1 Modell als das Wahrscheinlichste ausgewählt.

Die Datensätze für die synteniebasierten Proteinfamilien unterscheiden sich methodenbedingt von denen der Cluster orthologer Gene (COG). Da bei dem COG-Ansatz die Position im Genom keine Rolle spielt, ist es möglich, daß mehrere Gene pro Genom in eine Proteinfamilie eingehen. Die SOG-Methode ist referenzbasiert und enthält daher nur Gene, die Homologe zu Sequenzen in dem Referenzgenom haben. Es ist zwar möglich einen Konsensus zu berechnen, um zu einem vollständigen Datensatz zu kommen, dafür sind jedoch sehr strikte Vorgaben nötig, die einen

**Tabelle 7.1:** Überblick über die Datensätze der Proteinfamilien: Es werden in den folgenden Analysen nur COG* und SOG verwendet. Bei COG handelt es sich um Cluster, die mithilfe des Programmes `mcl` aus dem Originaldatensatz berechnet wurden. Der COG* Datensatz wurde so modifiziert, daß er mit dem referenzbasierten Ansatz der synteniebasierten orthologen Gene vergleichbar ist.

| | COGs | COGs* | SOGs |
|---|---|---|---|
| Gene | 44.355 | 30.851 | 32.852 |
| Universelle Familien | 2.537 | 2.537 | 2.922 |
| Gene ohne BBH | 2.288 | 269 | 140 |
| Proteinfamilien | 6.688 | 3.843 | 3.988 |

Vegleich unterschiedlicher Methoden schwierig machen. Es ist jedoch sehr einfach möglich aus den COG-Daten einen referenzbasierten Datensatz zu erstellen, in dem nur Proteinfamilien betrachtet werden, die mindestens eine Sequenz des Referenzgenoms enthalten. In den hier gezeigten Ergebnissen diente *E. coli K12* als Referenzgenom. Der so erhaltene Datensatz wird als COG* bezeichnet. Tabelle 7.1 zeigt eine Übersicht über die Eigenschaften der Datensätze. Insgesamt waren im COG*-Datensatz 30.851 Gene vertreten, im SOG Datensatz 1.999 mehr. Bei den Positionsorthologen wurden 3,8 % mehr Proteinfamilien gebildet als mit `mcl`.

Die Verteilung der Familiengröße (Abb. 7.3) zeigt die Anzahl der Genome, die in jeder Proteinfamilie enthalten waren. Der Anteil der universellen Proteinfamilien war im SOG Datensatz 7 % höher als bei COG*. Dementsprechend ist der Anteil der anderen Familien mit zwei bis acht Mitgliedern im COG* Datensatz höher.

Die Auftrittsverteilungen (Abb. 7.4) sollen verdeutlichen wie sich die Proteinfamilien über die betrachteten Spezies verteilten. Die Abbildungen zeigen nur die Proteinfamilien, die nicht universell waren, und die Sequenzen aus mindestens zwei Genomen enthielten. Die Höhe der Muster ist proportional zu ihrer Häufigkeit. Den größten Anteil machten Proteinfamilien aus, die nahezu universell waren, aber in einem Genom fehlten. Die anderen Familien traten zumeist nur sehr selten auf, mit Ausnahme einiger Verteilungsmuster die *Escherichia coli* K12 mit den relativ nah verwandten Stämmen HS, E24377A sowie den O157 Stämmen O157_Sakai und O157_EDL933 zusammenfaßen. Ebenso fiel in beiden Datensätzen ein Block von 85 Clustern im COG* Datensatz und 61 im SOG-Datensatz auf, der Proteinfamilien kennzeichnet, die in allen untersuchten *E. coli* Stämmen vorkamen, aber nicht in den beiden *E. coli* O157 Stämmen EDL933 und Sakai.

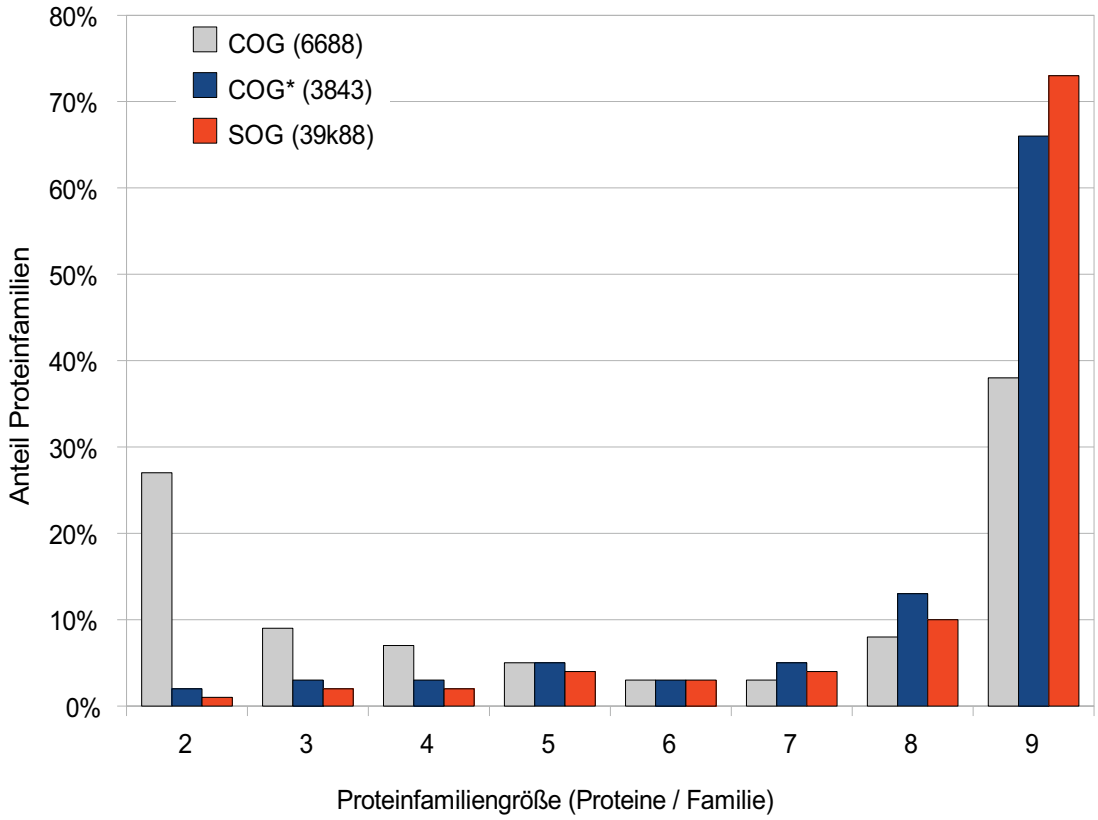



**Abbildung 7.3:** Verteilung der Familiengröße in den Datensätzen COG, COG* und SOG (Beschreibung siehe Tabelle7.1). Die Familiengröße ist dabei die Anzahl der in einer Familie vertretenen Genome. Bei den Clustern orthologer Gene kann jede Familie auch mehrere Homologe für ein Genom enthalten. Die Zahl der im Datensatz enthaltenen Proteinfamilien ist in der Legende in Klammern angegeben.

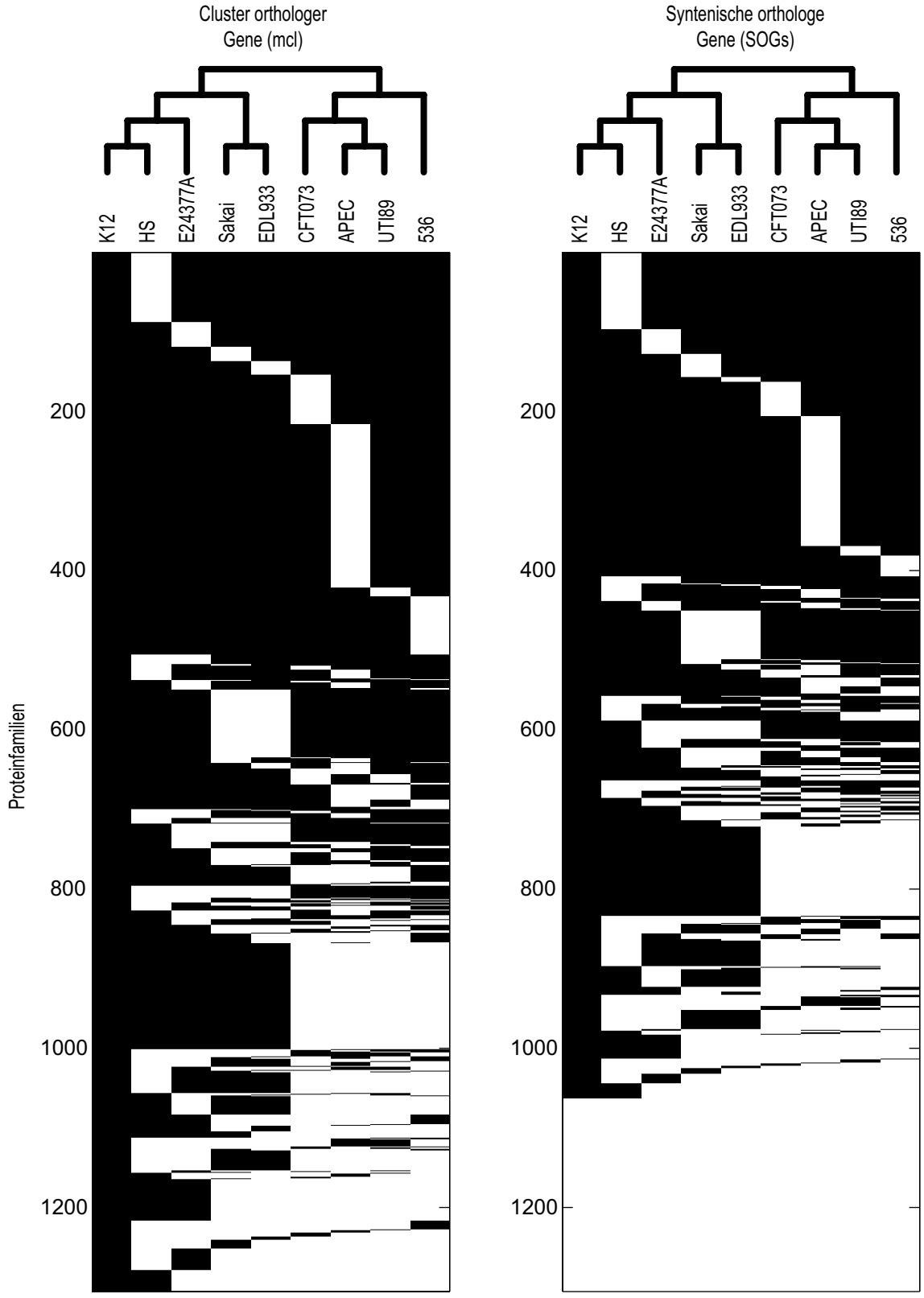



**Abbildung 7.4:** Verteilung der verschiedenen Auftrittsverteilungen (engl. *presence-absence-patterns*) von Proteinfamilien für neun Stämme von *E. coli*. Die `mcl`-Cluster sind auf der linken Seite dargestellt und die Syntenieorthologen auf der rechten. Auf der Y-Achse sind alle nicht universellen Proteinfamilien mit mindestens zwei enthaltenen Taxa aufgetragen. Die Referenztopologie aus Abb. 7.2 ist oberhalb der Abbildung wiedergegeben um die Verwandtschaftsverhältnisse zu verdeutlichen.

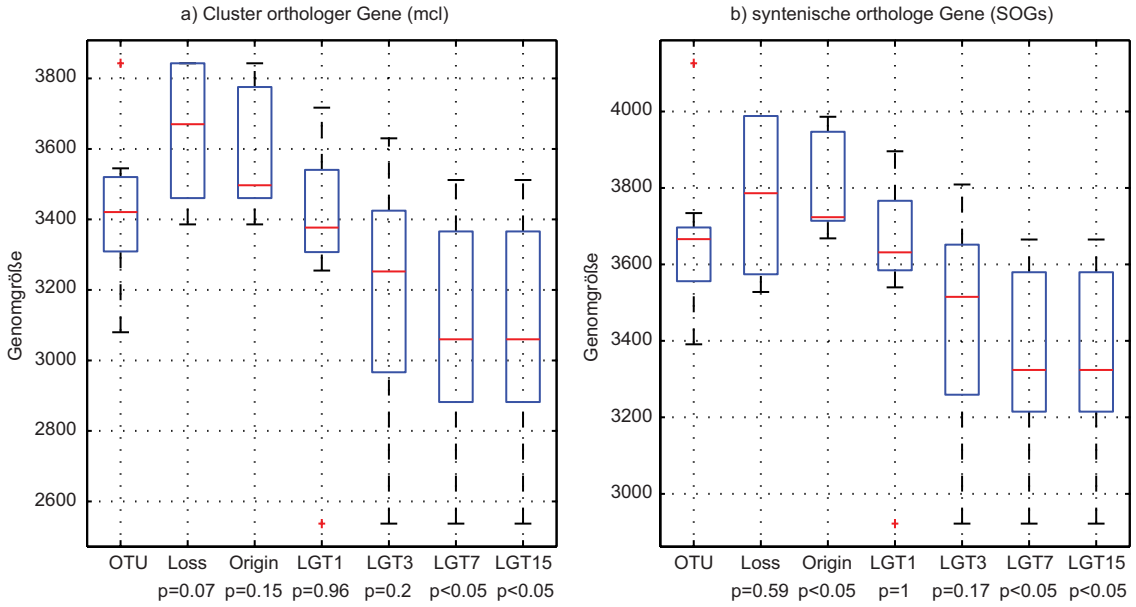



**Abbildung 7.5:** Abschätzung der Genomgröße für die hypothetischen Vorläufergenome für neun Stämme von *E. coli.* Abbildung a) zeigt die aufgrund einer Auswahl von `mcl` Clustern berechneten Genomgrößen. Abbildung b) beruht auf der Verwendung von Syntenieorthologen. Die zweite Zeile unter den Diagrammen gibt den p-Wert (*engl.* p-value) für die Hypothese wieder, daß die Verteilung der Vorläufergenomgrößen ähnlich ist mit der der heutigen Genome, berechnet mithilfe des Rangsummentests. Für beide Methoden wird das Modell LGT1 für die weitere Bearbeitung ausgewählt. OTU bezeichnet die Genömgrößen der verwendeten Genome. Die anderen Balken stellen die hypothetischen Genomgrößen für die im Text beschriebenen Modelle dar.

Für die Erstellung von LGT-Netzwerken wird zunächst eine Abschätzung der Genomgrößen der Vorläufergenome durchgeführt (Abb. 7.5). Dabei werden die Proteinfamilien mit einer Referenztopologie verglichen und entsprechend unterschiedlichen evolutionären Modellen wird versucht die Proteinverteilungen mit der Topologie in Deckung zu bringen. Die Größenverteilung der hypothetischen Genome wird mit der der aktuellen Genome verglichen. Mithilfe eines statistischen Tests wird ermittelt, ob sich die Größe der hypothetischen Genome signifikant von denen der aktuellen Genome unterscheidet. Für beide untersuchten Datensätze gibt das „LGT1“-Modell, das für jede Proteinfamilie maximal einen lateralen Gentransfer erlaubt, am besten die beobachteten Daten wieder und wurde daher für die weiteren Untersuchungen ausgewählt.

Die Berechnung der Netzwerke minimalen Gentransfers (Abb. 7.6) erfolgt, indem die abgeleiteten Genursprünge für das Verteilungsmuster so auf dem Baum verteilt

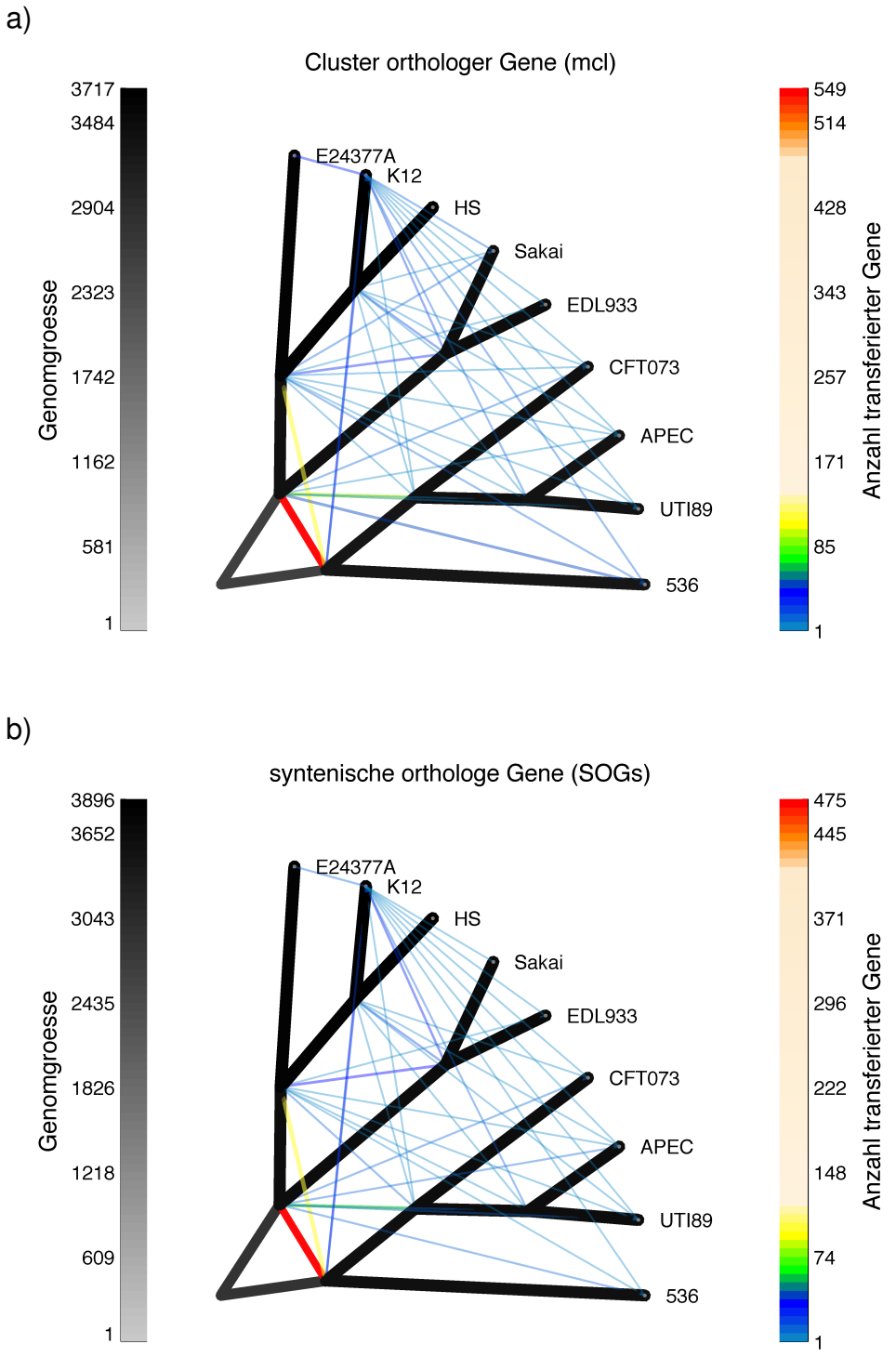



**Abbildung 7.6:** Darstellungen der Rekombinationsereignisse für neun Stämme von *E. coli*. Teil a) wurde aus den `mcl`-Clustern des COG*-Datensatzes berechnet. Teil b) beruht auf der Verwendung von Syntenieorthologen. Die Referenztopologie aus Abb. 7.2 wurde hierfür mithilfe von *Salmonella* als Außengruppe (engl. *outgroup*) gewurzelt.

werden, daß der Baum das Verteilungsmuster erklärt. Das heißt nicht, daß die Gentransfers genau an diesen Stellen stattgefunden haben, sondern daß zwischen den beiden Teilbäumen, die sich unterhalb der abgeleiteten Ursprünge befinden, ein Transfer stattgefunden haben muß. Die Richtung des Transfers wird dabei nicht berücksichtigt.

Im Fall von universellen Proteinfamilien liegt der einzige Ursprung immer in der Wurzel. Sollte ein zweiter Ursprung notwendig sein, das heißt es wird ein lateraler Gentransfer abgeleitet, so wird dies durch eine Linie zwischen den beiden ermittelten Ursprungsorten angedeutet. Entsprechend der Häufigkeit des abgeleiteten Gentransfers zwischen zwei Knoten innerhalb der Referenztopologie werden diese Linien farbkodiert.

Gentransfers wurden zwischen den meisten Taxa für beide Datensätze in geringer Menge abgeleitet. Lediglich drei Kanten traten deutlich hervor. Die deutlichste mit 475-549 transferierten Genen war diejenige, die die beiden Tochterknoten der Wurzel miteinander verbindet. Diese Linie repräsentiert Gene, die nicht universell sind und bei denen ein Gentransfer zwischen den beiden Teilbäumen stattgefunden hat. Hier wird deutlich, daß die Linien wie oben beschrieben keine exakten Karten sind, die genau berechnete Punkte verbinden, sondern Veranschaulichungen über die Menge des zwischen den Teilbäumen nötigen Transfers.

Der zweithäufigste Gentransfer verbindet die Teilbäume der CFT073, APEC, UTI89 und 536 Gruppe und der HS, K12 und E24377A Gruppe. Der Dritte verbindet den Teilbaum mit den beiden O157 Stämmen, K12, HS und E24377A mit dem Teilbaum, der APEC, CFT073 und UTI89 enthält. Insgesamt wurden für den COG* Datensatz 1.093 LGT-Ereignisse in 3.843 Clustern abgeleitet, das entspricht 28,4 %. Für den SOG Datensatz wurden 904 LGTs in 3.988 Proteinfamilien abgeleitet, was bedeutet, daß in 22,7 % aller Proteinfamilien ein Gentransfer stattgefunden hat.

Beide untersuchten Datensätze enthalten für das Referenzgenom *E. coli* K12 Gene, die keiner Proteinfamilie zugeordnet werden können. Für die 269 Gene aus dem COG*-Datensatz und die 140 aus dem SOG-Datensatz (vgl. Tabelle 7.1), für die das der Fall war wurde untersucht ob sich Orthologe in den Prokaryoten finden ließen (Abb. 7.7). Die dafür verwendete Methode entspricht der auf Seite 34 beschriebenen BBH Methode.

Für 49 der Gene, die im SOG-Datensatz keiner Familie zugeordnet werden konnten, und 67 der Gene aus dem COG*-Datensatz, ließen sich auch mit dieser Methode

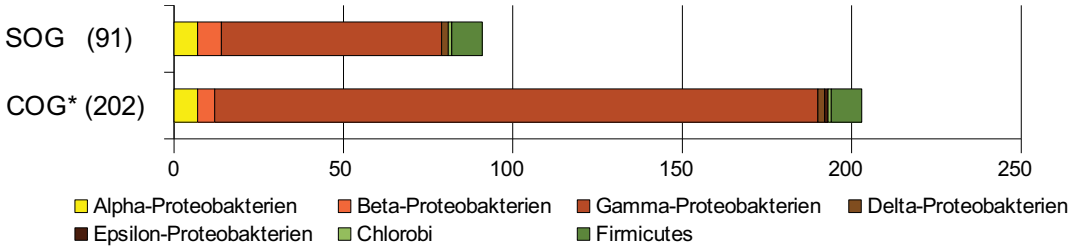



**Abbildung 7.7:** Taxonomische Verteilung der nächsten Nachbarn für die Gene die keiner Genfamilie zugeordnet werden konnten. Das Kriterium für den nächsten Nachbarn war hier der beste Treffer in einer BLAST Datenbanksuche (BBH)

keine Homologen identifizieren. Für die übrigen ließen sich nächste Nachbarn innerhalb der Bakterien identifizieren. Der weitaus größte Anteil findet sich unter den Proteobakterien, davon die meisten in der Gruppe der $\gamma$-Proteobakterien. Wenige Proteine finden ihren nächsten Nachbarn innerhalb der Firmicutes einzelne auch innerhalb der Chlorobi.

# 8 Ableitung von Rekombination aus Positionsorthologen

Rekombination ist ein Prozeß, der bei Prokaryoten einen ähnlichen Einfluß hat wie die sexuelle Fortpflanzung bei Eukaryoten. Durch den Austausch mehrerer homologer Bereiche beispielsweise zwischen Chromosomen und Plasmiden ist eine Anhäufung von Mutationen möglich.

Die Anzahl der vollständig sequenzierten *E. coli* Genome stieg in den letzten Jahren stetig an. Damit gibt es im Moment keine andere Spezies, von der mehr vollständig sequenzierte Genome verfügbar sind. Deswegen eignet sich *E. coli* sehr gut für die Überprüfung und Bewertung der in Kapitel 6 beschriebenen Syntenie-Methode. Unter diesen Genomen gibt es eine Vielfalt von Datensätzen, die eine offensichtliche Konservierung aufweisen (Abb.6.1).

Die direkt messbare Größe beim Vergleich zweier Sequenzen ist die Anzahl der Austausche beziehungsweise die Sequenzidentität. Der Vergleich vieler homologer Sequenzen gibt zusätzlich die Möglichkeit, zwischen unterschiedlichen Arten von Austauschen zu unterscheiden. Zu den Substitutionen werden nur diejenigen Veränderungen gezählt, die durch den Austausch einer Base hervorgerufen wurden.

Als Rekombinationen werden Austausche bezeichnet, die vermutlich durch Rekombinationsereignisse herbeigeführt wurden. Hierbei lagern sich zwei homologe DNA Stränge aneinander an und es findet ein Austausch statt. In der Regel wird dabei ein homologer Abschnitt zwischen zwei DNA Elementen (beispielsweise einem Chromosom und einem Plasmid) durchgeführt. Die mittlere Größe von per Rekombination ausgetauschten Sequenzabschnitten wird mit etwa 1000 bp beziffert (Milkman und Bridges, 1990), daher kann der Austausch eines Sequenzabschnitts zu mehreren Basenaustauschen führen.

In der Literatur wurde oft der Einfluß von Rekombination auf Genomsequenzen diskutiert. Die Definitionen sind dabei jedoch sehr unterschiedlich. Brisse et al.

(2009) definierten unabhängig von der Verteilung in unterschiedlichen Genomen eine Akkumulation mehrerer Austausche in einem abgegrenzten Sequenzabschnitt als Rekombination, während einzelne Austausche als Substitution gewertet wurden. Guttman und Dykhuizen (1994) untersuchten in *E. coli* vier Loci nach Hinweisen auf Rekombination und Substitutionen und fanden seit dem letzten gemeinsamen Vorfahren drei Rekombinationsereignisse und keine Substitution.

Während Punktmutationen in der Regel durch Lesefehler entstehen und somit als zufälliger Prozeß anzusehen sind, findet Rekombination zwischen homologen Sequenzen statt. Das heißt die Wahrscheinlichkeit, daß die auf diese Weise abgeänderte Sequenz funktionell ist, ist wesentlich höher als bei einem vollständig zufälligen Prozeß. Außerdem können so verschiedene Mutationen auf einmal in ein Gen eingefügt werden, was ansonsten sehr viele unwahrscheinlichere Einzelereignisse erfordern würde. Daher wird vermutet, daß der Einfluß von Rekombination auf das Genom eines Organismus einen höheren Einfluß hat als zufällige (Punkt-) Mutationen (Milkman und Bridges, 1990).

## 8.1 Die Verteilung kompatibler und inkompatibler Splits im Genom von *E. coli*

Für jede der 2.665 universellen synteniebasierten Proteinfamilien wurde mithilfe von `ClustalW` (Thompson et al., 1994) ein Alignment erstellt, aus dem mit dem `PHYLIP`-Paket *neighbor-joining*-Bäume erstellt wurden. Zusätzlich wurden mit `PhyML` Wahrscheinlichkeitsbäume berechnet. Mithilfe eines Programms, das von David Bryant zur Verfügung gestellt wurde, wurde die Kompatibilität der Alignments zu unterschiedlichen Referenzbäumen ermittelt. Für jede Alignmentposition wird dabei beurteilt, mit welchem Split sie kompatibel ist. Referenztopologien werden in der gleichen Weise aufgetragen wie die Splits, um einen Vergleich zu ermöglichen.

Im unteren Bereich des ersten Teils von Abbildung 8.1 (hellgrau, dunkelgrau) werden von links nach rechts alle möglichen Splits dargestellt. Es sind alle Taxa, die im Baum auftreten aufgelistet. Alle Taxa, die in einer Spalte in der gleichen Farbe eingefärbt sind, gehören dabei zu einem Split. Die Sortierung der Splits ist so gewählt, daß zunächst die internen Splits absteigend entsprechend ihrer Häufigkeit eingezeichnet werden. Anschließend folgen die externen Splits (die Blätter) in der gleichen Weise

sortiert. Die oberen Zeilen (hellgrau, rosa) repräsentieren die Referenztopologien. Alle hervorgehobenen Splits sind in dem entsprechenden Baum enthalten. Diese Abbildung dient als Legende für die Kompatibilitätsplots und das Histogramm. Die Reihenfolge der Splits ist in allen Teilabbildungen identisch. Im mittleren Bereich ist die Kompatibilität der Genfamilien aufgetragen. Die Genfamilien sind entsprechend der Reihenfolge auf dem Referenzgenom sortiert. Die Farbkodierung entspricht dabei der Häufigkeit des entsprechenden Splits im betrachteten Gen.

Die Farbkodierung und die Reihenfolge der Splits in Abbildung 8.2 ist identisch mit derjenigen in Abbildung 8.1, allerdings sind die Proteinfamilien hier nach der Auftrittsverteilungen sortiert. Während sich in Abbildung 8.1 erkennen lässt, ob benachbarte Gene ähnliche Splitverteilungen aufweisen, lassen sich in Abbildung 8.2 Auftrittsverteilungen für Splits erkennen, die häufiger sind als andere. Im oberen Bereich ist ein Histogramm wiedergegeben, das den Anteil der betreffenden Splits an der Gesamtzahl der Splits wiedergibt.

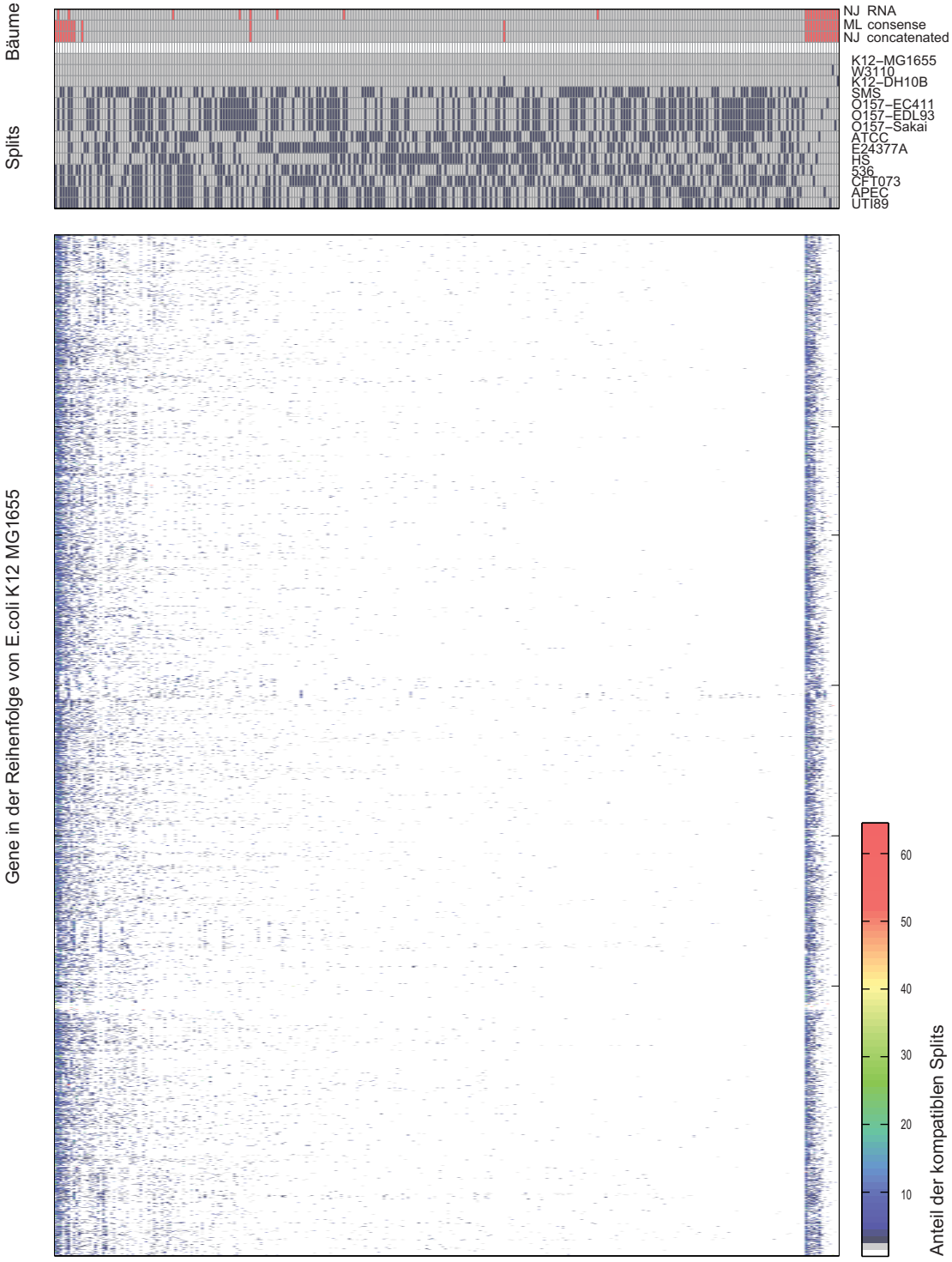



**Abbildung 8.1:** Der obere Teil der Abbildung beschreibt die in der Abbildung enthaltenen Splits. Dunkelgrau markierte Taxa sind in dem betreffenden Split enthalten. Die oberen drei Zeilen beschreiben unterschiedliche Bäume. Die magenta-eingefärbten Splits sind in den betreffenden Bäumen enthalten. Der untere Teil der Abbildung stellt die Proteine in der Reihenfolge des Referenzgenoms und die Splits in dem zugehörigen Alignment dar.

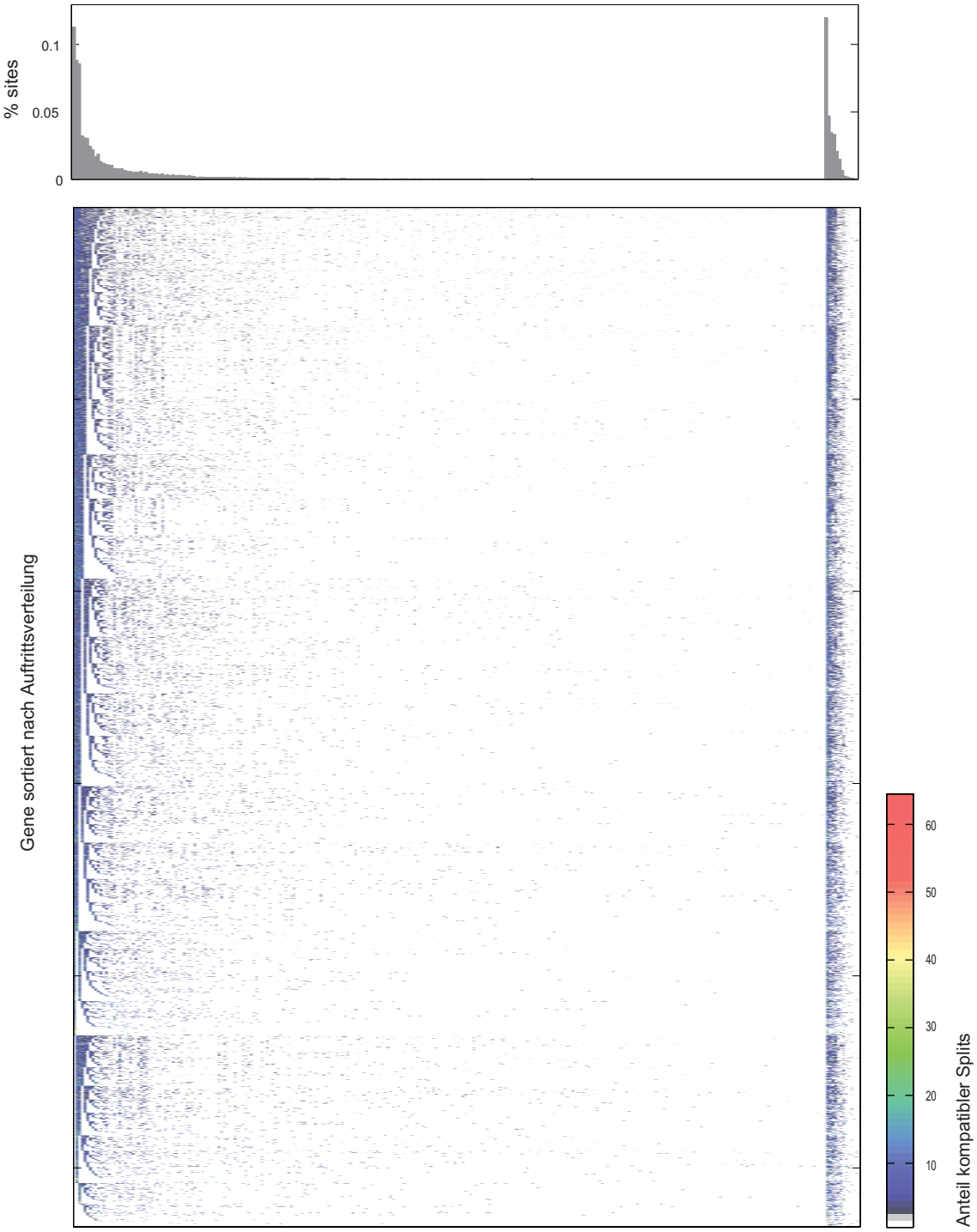



**Abbildung 8.2:** Die obere Abbildung zeigt die Häufigkeit der in den Daten enthaltenen Splits. Die Reihenfolge der Splits ist identisch mit der in Abb. 8.1. Der untere Teil der Abbildung stellt die Proteine sortiert nach ihrem Auftrittsmuster (engl. *presence-absence-pattern*) dar.

```
536             000000000010000000000001
APEC            000000001110000000000010
ATCC            000011110000000000000100
CFT073          000000000110000000001000
E24377A         000000110000000000010000
HS              000011110000000000100000
K12_DH10B       000110110000000001000000
K12_MG1655      001110110000000010000000
O157_EC4115     010000100000000100000000
O157_EDL933     110000100000001000000000
O157_Sakai      110000100000010000000000
SMS             000000000000100000000000
UTI89           000000001110100000000000
W3110           001110110001000000000000
```

**Abbildung 8.3:** Beispiel eines Baums in Split-Annotation. Jede Zeile steht dabei für einen OTU. Jede Spalte steht für einen Split; durch die Ziffern 0 und 1 wird gekennzeichnet auf welcher Seite des Splits sich der OTU befindet.

## 8.2 Berechnung von Rekombinations- und Substitutionsraten

Im Abschnitt 8.1 wurde für alle Alignmentpositionen der universellen Genfamilien von 14 *E. coli*-Stämmen überprüft mit welchen Splits sie kompatibel sind. Im nächsten Schritt soll nun eine sehr ähnliche Analyse zu einer neuen Darstellungsform beitragen. Die Datenbasis stellen wie im vorigen Abschnitt, die gleichen 2.665 Alignments mit insgesamt 2.734.536 Positionen dar. Die Kompatibilität wird hier allerdings nur für die in den Referenztopologien enthaltenen Splits überprüft.

Alignments und Bäume sollen nach Anhaltspunkten für Rekombinationsereignisse und Substitutionen untersucht werden. Dafür werden alle Positionen aller Alignments eines Datensatzes mit einer Referenztopologie verglichen. Basenverteilungen, die sich durch die Topologie erklären lassen, werden als Substitutionen gewertet. Verteilungen, für die das nicht gilt, werden als Hinweise auf Rekombination gewertet. Durch die Betrachtung aller Kanten eines phylogenetischen Baumes ist es möglich einen Ort (Zeitpunkt) für Substitutionen und Rekombinationen zu ermitteln.

Diese Analyse erfolgte mithilfe des Programms `recombinator.pl`. Das Programm liest einen Baum in Split-Annotation (Abb. 8.3) ein und erstellt für jeden Split eine Liste mit Nachbarsplits (Abb. 8.4). Für jedes Alignment werden alle Positionen auf vorher definierte Kriterien überprüft. Die Anzahl der an einer Position vorkommenden Nukleotide bildet das erste Kriterium. Nur für Positionen mit zwei Basen kann eine Aussage getroffen werden. Invariante Positionen sind uninformativ und für Positionen mit mehr als zwei Nukleotiden läßt sich die Kompatibilität nicht in einem vergleichbar

einfachen Prozeß ermitteln. Des Weiteren werden Positionen ausgeschlossen, wenn sie Ambiguitäten aufweisen, das heißt mehrere Nukleotide können in einem Genom an dieser Position auftreten.

Daraus folgt für den verwendeten Datensatz, daß sich nur ein kleiner Anteil von 5,06 % der Alignmentpositionen für die Ratenbestimmung eignet (Abb. 8.6). 91,03 % aller Alignmentpositionen waren invariant und daher für diese Fragestellung uninformativ. 0,16 % der Positionen enthielten mehr als zwei Nukleotide und wurden deshalb aus der Datenmenge entfernt. Die größte Menge der Daten wurde nicht berücksichtigt, weil sie Lücken enthielten (3,7 %). Des Weiteren wurden 0,04 % der Alignmentpositionen ausgeschlossen, da sie Ambiguitäten enthielten, bei denen eine eindeutige Zuordnung nicht möglich gewesen wäre. Insgesamt verblieben 138.263 Alignmentpositionen, in denen zwei verschiedene Nukleotide auftraten.

Eine Position wird als Substitution gewertet, wenn sie sich exakt auf einen Split projizieren lässt. Das heißt die Basenverteilung an der entsprechenden Position deckt sich mit der Topologie des Baumes (siehe Abb. 8.5). Als rekombinant werden Positionen gewertet, bei denen das nicht der Fall ist. Hierbei werden Splits als Orte der Rekombination gewertet, wenn beide Teilbäume auf einer Seite des Splits das gleiche Nukleotid aufweisen, und auf der anderen Seite ein Teilbaum nur das andere Nukleotid enthält und der andere beide.

```
------------------------------------------------------------------------
     00000000011000       0       12       2
              |           O157_EDL933
               |          O157_Sakai
14       0       a                00000000001000
15       0       a                00000000010000
1        0       b       *        11111111000111
16       0       b                00000000100000
                                  11111111111111

------------------------------------------------------------------------
     00000000111000       1       11       3
             |            O157_EC4115
              |           O157_EDL933
               |          O157_Sakai
0        1       a                00000000011000
16       1       a                00000000100000
6        1       b       *        11010000000110
7        1       b                00101111000001
                                  11111111111111

------------------------------------------------------------------------
     00000001000001       2       12       2
            |             K12_MG1655
                 |        W3110
11       2       a                00000000000001
17       2       a                00000001000000
18       2       b                00000010000000
3        2       b       *        11111100111110
                                  11111111111111

------------------------------------------------------------------------
     00000011000001       3       11       3
           |              K12_DH10B
            |             K12_MG1655
                 |        W3110
18       3       a                00000010000000
2        3       a                00000001000001
4        3       b       *        11011000111110
5        3       b                00100100000000
                                  11111111111111

------------------------------------------------------------------------
```

**Abbildung 8.4:** Für jeden Split des Baumes wird ausgegeben welche Astnummer er hat (zweite Zahl jedes Blocks) und welche Taxa auf einer Seite des Splits liegen (alle anderen liegen auf der anderen Seite). Außerdem sind alle benachbarten Splits angegeben. Die mit a gekennzeichneten liegen auf der Seite der im Muster mit „1“ angegebenen Taxa. Ein Sternchen in der 4. Spalte signalisiert, daß der Split gegenüber dem eingegebenen Muster invertiert wurde. Dies verändert die Daten nicht, setzt sie jedoch so um, daß deutlich wird, daß die Summe aller Nachbarsplits immer ein Muster ergibt, daß nur aus Einsen besteht, in dem also alle Taxa vorkommen.

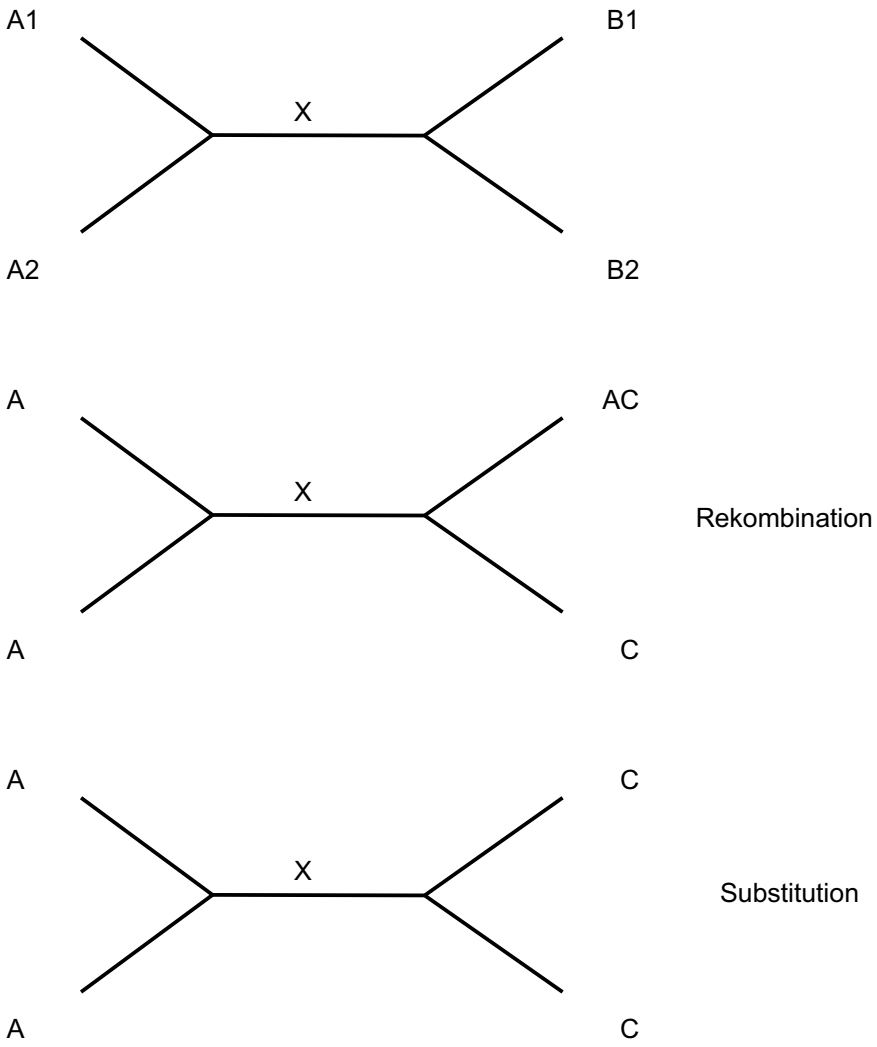



**Abbildung 8.5:** Beispiel zur Ermittlung von rekombinanter Splits. Betrachtet werden die Nachbarn des zu untersuchenden Splits X. Sind die Nachbarn auf beiden Seiten homogen verteilt (nur eine Base ist vorhanden) also A1 = A2 und B1 = B2, aber A1 =! B1, so liegt ein Substitution vor, da sich die Verteilung der Basen auf die Topologie des Baumes übertragen läßt. Enthält jedoch genau ein Nachbar mehrere Basen (zum Beispiel B1) und der andere Nachbar derselben Seite (B2) ist homogen verteilt und unterscheidet sich von den Nachbarn auf der anderen Seite (A1, A2) so wird der Split X als Rekombinationsort gewertet.

```
2665 files

Gruppe                 Anzahl    Anteil

0-all                2734536  100.0000 %
A-1-Character(s)     2489363   91.0342 %
A-2-Character(s)      138263    5.0562 %
A-3-Character(s)        4497    0.1645 %
A-4-Character(s)         100    0.0037 %
A-ambiguous             1090    0.0399 %
A-gapped              101266    3.7032 %
B-compatible         2489363   91.0342 %
B-out                 102356    3.7431 %
C0-1-ambi                  6    0.0002 %
C0-2-ambi                979    0.0358 %
C0-3-ambi                101    0.0037 %
C0-4-ambi                  4    0.0001 %
C6-1-gapped            97622    3.5700 %
C6-2-gapped             3528    0.1290 %
C6-3-gapped              115    0.0042 %
C6-4-gapped                1    0.0000 %
```

**Abbildung 8.6:** Erster Teil der Ausgabe des Programms für 2.665 `ClustalW` Alignments verglichen mit dem oben dargestellten Konsensus-Baum der aus den NJ-Bäumen der einzelnen Alignments berechnet wurde. Hier wird die grundsätzliche Zusammensetzung der Daten verdeutlicht: Der 0'er Block enthält alle Positionen der Alignments. In Block A werden die Positionen nach Anteil der auftretenden Nukleotide aufgelistet. Positionen mit Ambiguitäten oder Lückenpositionen werden dabei nicht berücksichtigt und in eigene Rubriken eingeordnet. Nur für die Positionen, die in der Gruppe A-2 eingeordnet werden, können verlässliche Aussagen über Rekombinations- und Substitutionsereignisse gemacht werden.
In Block C werden die nicht berücksichtigten Gruppen genauer charakterisiert, die zweite Ziffer bezeichnet dabei die Anzahl der an der Position vorhandenen Nukleotide.

| Branch | Subst | | Recomb | | S:R |
|---|---|---|---|---|---|
| All | 99448 | 719.27 | 38815 | 280.73 | 3:1 |
| 0 | 91 | 0.0658 | 5 | 0.0036 | 18:1 |
| 1 | 12278 | 8.8802 | 10693 | 7.7338 | 1:1 |
| 2 | 12 | 0.0087 | 10 | 0.0072 | 1:1 |
| 3 | 4211 | 3.0456 | 5403 | 3.9078 | 1:1 |
| 4 | 3403 | 2.4613 | 1260 | 0.9113 | 3:1 |
| 5 | 1841 | 1.3315 | 4853 | 3.5100 | 1:3 |
| 6 | 11903 | 8.6090 | 2772 | 2.0049 | 4:1 |
| 7 | 4475 | 3.2366 | 2937 | 2.1242 | 2:1 |
| 8 | 4244 | 3.0695 | 5474 | 3.9591 | 1:1 |
| 9 | 3061 | 2.2139 | 1676 | 1.2122 | 2:1 |
| 10 | 15680 | 11.3407 | 4019 | 2.9068 | 4:1 |
| 11 | 78 | 0.0564 | 25 | 0.0181 | 3:1 |
| 12 | 134 | 0.0969 | 133 | 0.0962 | 1:1 |
| 13 | 16632 | 12.0292 | 14814 | 10.7144 | 1:1 |
| 14 | 34 | 0.0246 | 2 | 0.0014 | 17:1 |
| 15 | 37 | 0.0268 | 6 | 0.0043 | 6:1 |
| 16 | 220 | 0.1591 | 22 | 0.0159 | 10:1 |
| 17 | 1 | 0.0007 | 2 | 0.0014 | 1:2 |
| 18 | 18 | 0.0130 | 27 | 0.0195 | 1:2 |
| 19 | 2863 | 2.0707 | 5749 | 4.1580 | 1:2 |
| 20 | 6514 | 4.7113 | 15169 | 10.9711 | 1:2 |
| 21 | 4558 | 3.2966 | 6432 | 4.6520 | 1:1 |
| 22 | 2052 | 1.4841 | 5012 | 3.6250 | 1:2 |
| 23 | 307 | 0.2220 | 109 | 0.0788 | 3:1 |
| 24 | 4801 | 3.4724 | 10455 | 7.5617 | 1:2 |

**Abbildung 8.7:** Zweiter Teil der Ausgabe des Programms für 2.665 `ClustalW` Alignments verglichen mit einem Konsensusbaum, der aus den ML-Bäumen der einzelnen Alignments berechnet wurde. Die erste Spalte bezeichnet den Ast, für den die Berechnung gilt (siehe oben). Die zweite und vierte Spalte enthalten die Anzahl der ermittelten Substitutions- und Rekombinationsereignisse jeweils gefolgt von ihrem Anteil (in Promille), bezogen auf die Positionen, die zwei Nukleotide aufweisen, das heißt bezogen auf die Menge von Positionen, für die Aussagen getroffen werden können.
Die letzte Spalte enthält das auf ganze Zahlen gerundete Verhältnis zwischen Substitutionen und Rekombinationen an dem entsprechenden Split. Für jede Alignmentposition können mehrere Ereignisse abgeleitet werden, in der Summe wird aber nur ein Ereignis pro Alignmentposition berücksichtigt.

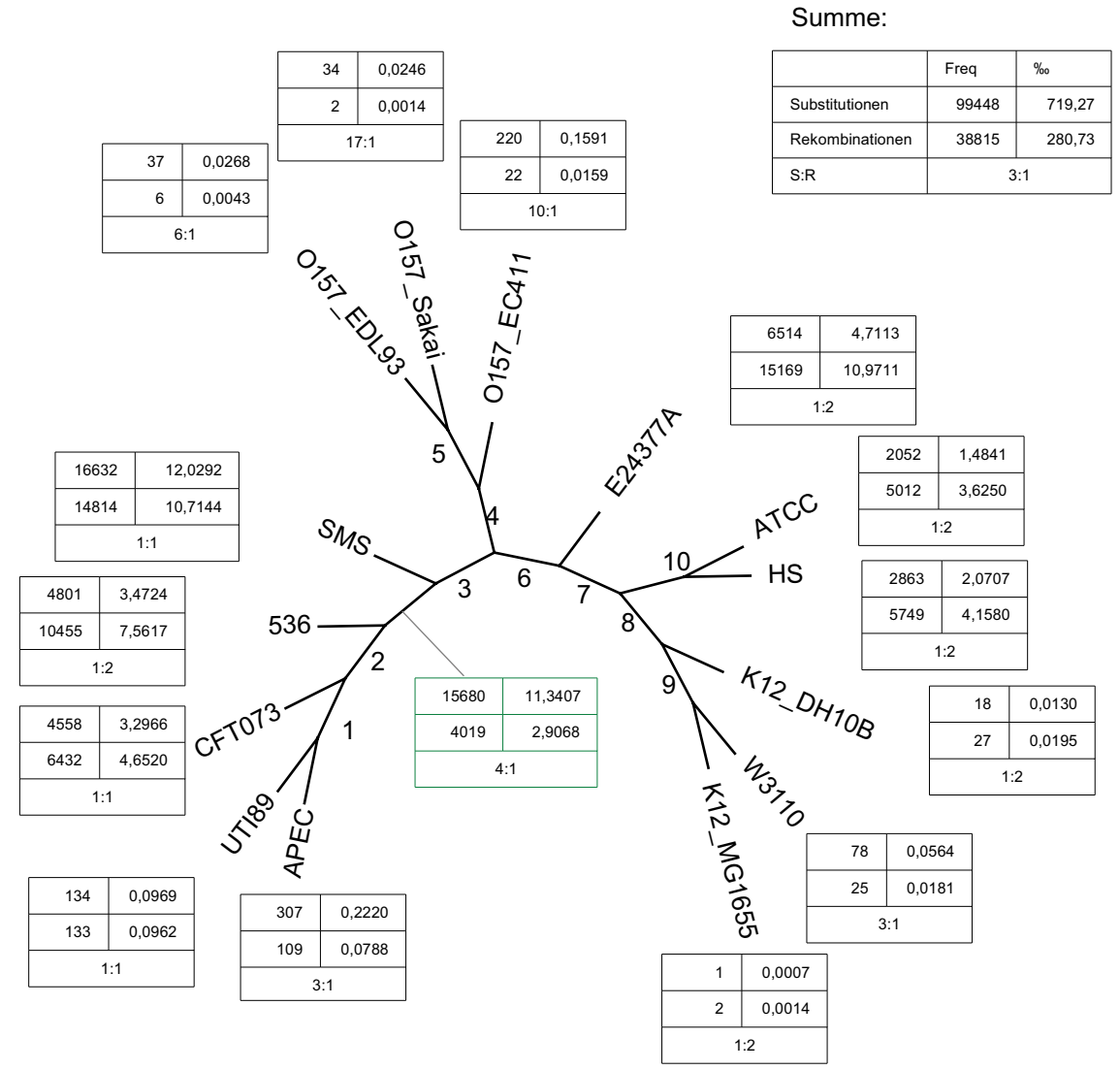
Der beschriebene Datensatz wurde mit drei unterschiedlichen Referenztopologien verglichen.

1. Konsensus-Baum über 2.665 ML-Bäume / NJ-Baum über 2.665 konkatenierte Alignments (gleiche Topologie, wird im folgenden als „Konsensus-Baum“ bezeichnet.
2. *neighbor-joining*-Baum für rRNA Sequenzen
3. zufällige Topologie

In Abbildung 8.8 sind die Ergebnisse für den Vergleich mit dem Konsensusbaum wiedergegeben. Die Topologie des Baumes ist ohne Berücksichtigung der tatsächlichen Astlängen wiedergegeben. In den schwarz umrandeten Kästen finden sich die Ergebnisse für externe Kanten, in grün diejenigen für die internen. Aus Gründen der Übersichtlichkeit wurden die internen Kanten hier mit Nummern beschriftet, die dazugehörigen Werte finden sich unterhalb der Abbildung. Der Anteil bezogen auf die Zahl der untersuchten Positionen ist in Promille angegeben. Im Schnitt werden 71,9 % aller Alignmentpositionen durch Substitutionen erklärt. Einige Äste weisen aber auch einen erhöhten Anteil von Rekombinationen auf, beispielsweise der Teilbaum, in dem sich die beiden Stämme *E. coli* HS und *E. coli* ATTC wiederfinden.

Die Ergebnisse für den rRNA-Baum finden sich in Abbildung 8.9. In diesem Baum finden sich einige farblich hervorgehobene interne Kanten für die keine Substitutionen abgeleitet wurden. Für die Kante, die den Split darstellt, der die beiden Stämme *E.coli* K12_MG1655 und *E. coli* K12_DH10B von den restlichen Stämmen trennt wurden weder Rekombinations- noch Substitutionsraten bestimmt. Insgesamt ist der Anteil von Rekombination höher als bei dem Konsensus-Baum und beträgt 3:2 verglichen mit den ermittelten Substitutionsraten.

Für den zufälligen Baum lassen sich für keine interne Kante Substitutionsraten bestimmen. Nur für wenige wurden Rekombinationen abgeleitet. Insgesamt werden 72,3 % der untersuchten Alignmentpositionen durch Rekombination erklärt.



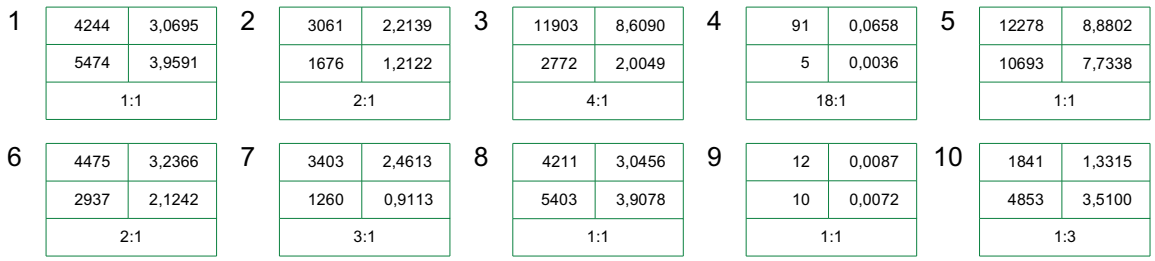




**Abbildung 8.8:** Substitutions- und Rekombinationsraten für *E. coli*, ermittelt mit einer Referenztopologie, die sowohl bei der Erstellung eines *neighbor-joining*-Baumes mit einem konkatenierten Alignment aus 2.665 Proteinfamilien entstand als auch bei einem Konsensus-Baum erstellt aus 2.665 *maximum-likelihood*-Bäumen. Die Raten für die mit 1-10 beschrifteten internen Kanten sind im unteren Bereich der Abbildung aufgeführt. Die Raten für interne Kanten stehen in grünen Kästen, diejenigen für externe in schwarzen.

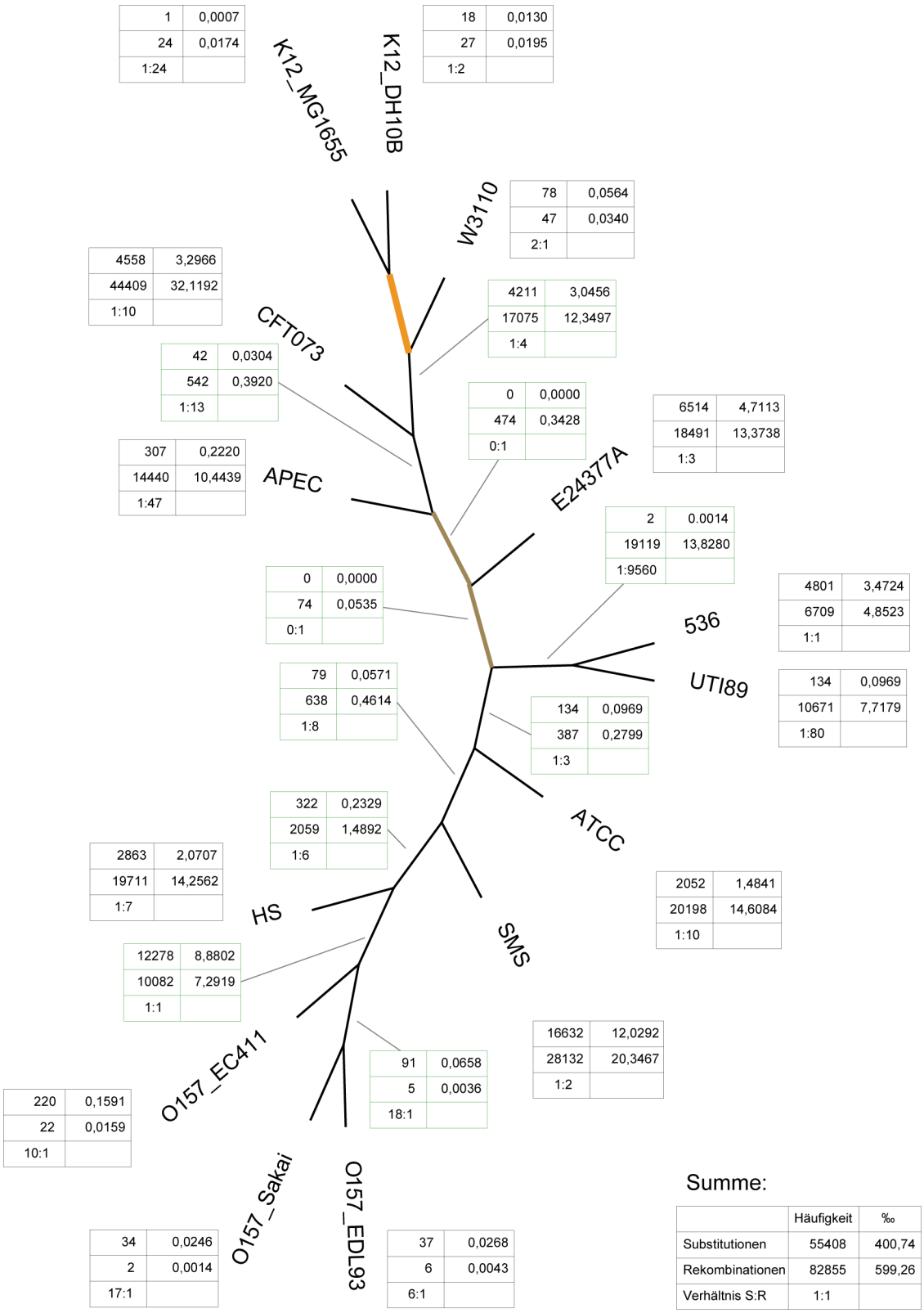
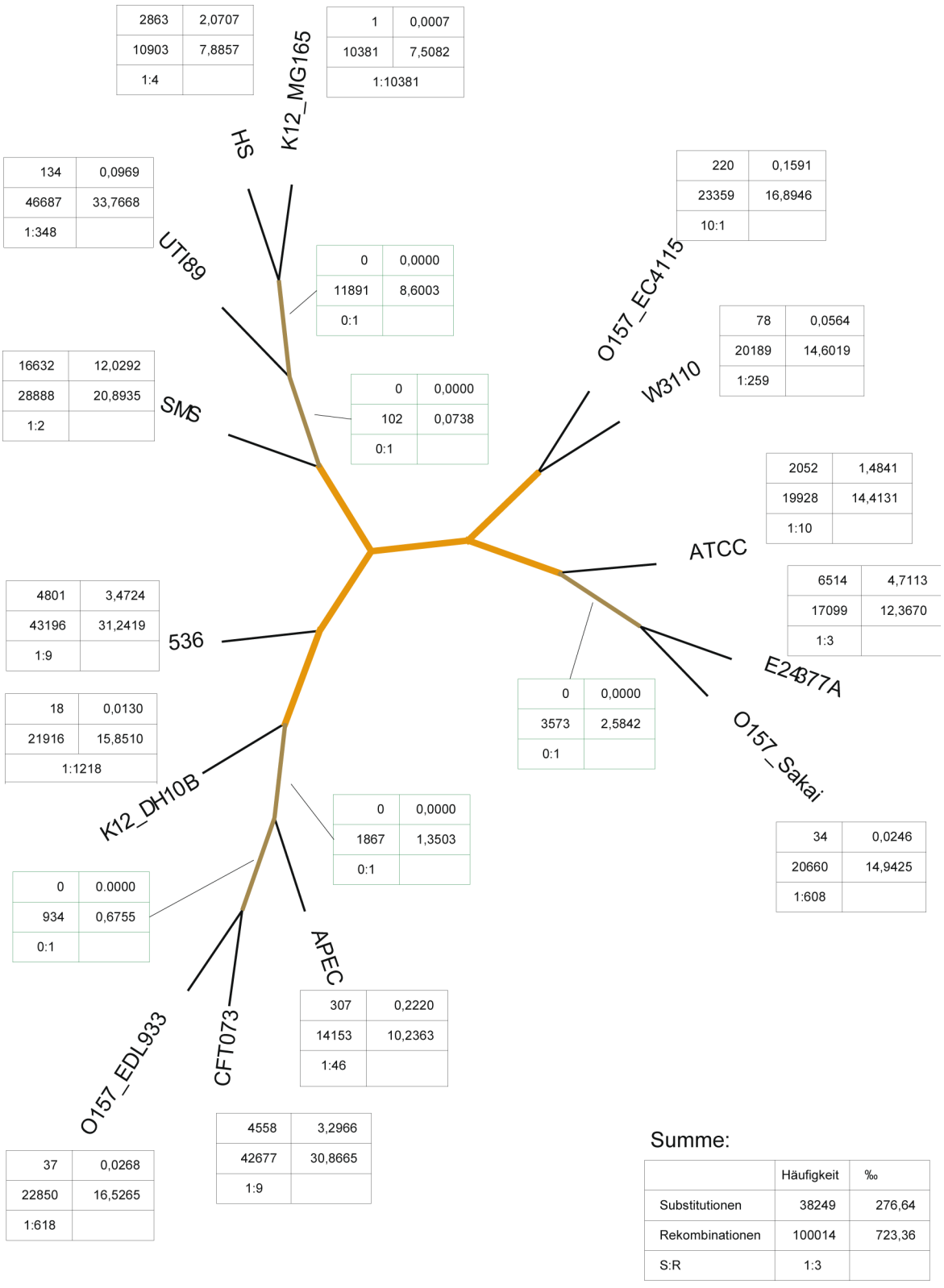



**Abbildung 8.9:** Substitutions- und Rekombinationsraten für 14 Stämme von *E. coli*, ermittelt mit einem *neighbor-joining*-Baum erstellt aus rRNA Sequenzen als Referenzbaum. Für die hervorgehobenen internen Kanten (orange) werden weder Substitutionen noch Rekombination abgeleitet. Für die braunen Kanten wurde Rekombination abgeleitet, aber keine Substitution. Die Raten für interne Kanten stehen in grünen Kästen, diejenigen für externe in schwarzen.




Summe:

| | Häufigkeit | ‰ |
|---|---|---|
| Substitutionen | 38249 | 276,64 |
| Rekombinationen | 100014 | 723,36 |
| S:R | 1:3 | |

**Abbildung 8.10:** Substitutions- und Rekombinationsraten für 14 Stämme von *E. coli*, ermittelt mit einer Zufallstopologie als Referenzbaum. Für die hervorgehobenen internen Kanten (orange) werden weder Substitutionen noch Rekombinationen abgeleitet. Für die braunen Kanten wurde Rekombination abgeleitet, aber keine Substitution. Die Raten für interne Kanten stehen in grünen Kästen, diejenigen für externe in schwarzen.

# 9 Diskussion

## 9.1 Chimäre Bakterienchromosomen

Im Kapitel 5 wurde die Frage nach der Zusammensetzung $\alpha$-proteobakterieller Genome in Bezug auf die nächsten Nachbarn der einzelnen Gene untersucht. Der nächste Nachbar der Gene wurde dabei in einer prokaryotischen Datenbank zunächst als der beste BLAST-Treffer außerhalb der eigenen Gattung bestimmt, mit dem Ergebnis, daß bis zu 36 % der Gene einen nächsten Nachbarn außerhalb der $\alpha$-Proteobakterien aufwiesen. Da bekannt ist, daß BLAST-Ergebnisse nicht immer mit phylogenetischen Ergebnissen übereinstimmen (Koski und Golding, 2001), wurden auch phylogenetische Methoden verwendet, um die Ergebnisse zu überprüfen. Hierbei wurden unterschiedliche Fehlerquellen auf ihre Wirkung überprüft. Die Zusammensetzung der Datenbank wurde in unterschiedlicher Weise abgewandelt (Abb. 5.11, 5.12). Außerdem wurden die Ergebnisse auf solche beschränkt in denen eine große Datenbasis zur Verfügung gestellt werden konnte (Bäume mit mindestens 15 Taxa, Abb. 5.10). Durch diese Methoden wurde immer eine geringere Menge von Ergebnissen ermittelt, da eine Filterung stattfand. Der Ausschluß von Genomen aus der Ergebnismenge der $\alpha$-Proteobakterien zur Überprüfung, ob die Datenbankzusammensetzung einen Einfluß auf die Ergebnisse hat, führte erwartungsgemäß zu einer Verringerung des Anteils nächster Nachbarn innerhalb der $\alpha$-Proteobakterien. Diese Reduktion lag jedoch weit unterhalb dessen was zu erwarten gewesen wäre, wenn die Ergebnisse proportional zur Datenbankzusammensetzung abgenommen hätten.

Die Verwendung strikterer Methoden, beispielsweise unterschiedlicher Methoden zur Rekonstruktion von phylogenetischen Bäumen und die Beschränkung auf große Datensätze, führen erwartungsgemäß zu einer Erhöhung des Anteils der Gene mit einem $\alpha$-Proteobakterium als nächstem Nachbarn. Allein die Verwendung von Phylogenien reduziert bereits die Anzahl der Gene mit nächsten Nachbarn, da

hier nur noch Sequenzen mit mehreren Treffern in der Datenbanksuche verwendet werden können. Die Beschränkung auf Datensätze mit 90 % Bootstrap-Unterstützung (Abb. 5.8) oder auf Bäume mit mindestens 15 Taxa (Abb. 5.10) sorgen dafür, daß die Datenmenge reduziert wird und hauptsächlich diejenigen nächsten Nachbarn übrig bleiben, die ein starkes phylogenetisches Signal aufweisen. Das stärkste Signal war bereits bei den auf Sequenzähnlichkeit basierenden nächsten Nachbarn (5.6), beruhend auf den Ergebnisssen der BLAST-Datenbanksuche das der $\alpha$-Proteobakterien. Alle Überprüfungen der Daten haben zu Veränderungen in der Zusammensetzung der Ergebnisse geführt. Allerdings blieb in allen Datensätzen dasselbe Bild erhalten, so daß die Ergebnisse gegenüber methodischen Einflüssen weitgehend robust waren.

Mit Hilfe von Sequenzähnlichkeit wurden 630 eukaryotische Gene identifiziert die auf $\alpha$-Proteobakterien hindeuten (Gabaldón und Huynen, 2003). Allerdings gibt es tausende von Genen, die auf einen eubakteriellen Ursprung hindeuten, aber nicht eindeutig den $\alpha$-Proteobakterien zugeordnet werden können (Esser et al., 2004; Rivera und Lake, 2004; Embley und Martin, 2006). Einen Nachbarn in solchen Gruppen zu finden, wird oft als Hinweis auf einen Ursprung des Genes in der betreffenden Gruppe gewertet (Baughn und Malamy, 2002). Dabei werden jedoch die methodischen Probleme bei der Erstellung von Phylogenien deren Aufzweigung sehr weit zurückreicht oft vernachlässigt (Susko et al., 2006).

Wenn wir uns aber von der Vorstellung eines Bakterienchromosom trennen, das über lange Perioden statischen Charakter hat, und stattdessen ein dynamisches Modell heranziehen in dem die Übergänge fließend sind (engl. *fluid chromosome model*), so würden wir nach einem gemeinsamen Vorfahren für alle eukaryotischen Sequenzen suchen. Dieser müsste sich aber nicht unbedingt innerhalb der $\alpha$-Proteobakterien finden, da nicht bekannt ist welche Ansammlung von Genen der Vorfahre der Mitochondrien besaß. Frühere Analysen der Genomevolution von $\alpha$-Proteobakterien gingen auf die Genomgrößen, und die funktionellen Gruppen der Proteine ein, ließen aber die Sequenzähnlichkeit außer acht (Boussau et al., 2004).

Daher ist die Frage interessant wieviele $\alpha$-proteobakterielle Gene auch unter Verwendung des *nearest-neighbor*-Kriteriums $\alpha$-proteobakterieller Abstammung zu sein scheinen. Die Antwort liegt bei den untersuchten Genomen zwischen 33 % (bei *Magnetococcus*) und 97 % bei *Sinorhizobium*. Die mitochondrialen Genome wiesen

keine Abweichungen in ihrer Zusammensetzung der nächsten Nachbarn im Vergleich zu den $\alpha$-Proteobakterien auf (Abb. 5.6).

Prokaryotische Genome werden nicht nur durch Vererbung in ihrem Gengehalt beeinflusst, sondern auch durch den Verlust von Genen und lateralem Gentransfer (Kunin et al., 2005; Lerat et al., 2005). Diese Tatsache wird jedoch nur selten zu Fragestellungen zum Ursprung des Mitochondriums und dem von eukaryotischen Genen in Bezug gesetzt (Esser et al., 2004). Die Ergebnisse deuten darauf hin, daß rezente $\alpha$-proteobakterielle Genome eine dynamische Sammlung von Genen aufweisen, deren Ursprünge in vielen Verwandtschaftsbereichen liegen. Daraus läßt sich ableiten, daß auch der Vorfahr der Mitochondrien vermutlich über ein mosaikartiges Genom verfügt hat. Daher ist das Kriterium für ein kernkodiertes Gen, durch endosymbiontischen Gentransfer in das Genom gelangt zu sein, nämlich die gemeinsame Verzweigung mit einem $\alpha$-proteobakteriellen Gen in einem phylogenetischen Baum (Kurland und Andersson, 2000), zu strikt, da es ein konserviertes Bakteriengenom vorraussetzt, in dem LGT und Genverlust keine Rolle spielen.

Wird die Endosymbiontentheorie mit der Annahme eines dynamischen Bakterienchromosoms kombiniert, kann diese Annahme dahingehend abgeändert werden, daß kernkodierte Gene mitochondrialen Ursprungs eine gemeinsame Abstammung zu den Eubakterien zeigen – vorrausgesetzt, daß die Methoden der molekularen Phylogenie in der Lage sind die Ereignisse, die vor 1,5 Milliarden Jahren stattgefunden haben, korrekt zu rekonstruieren (Embley und Martin, 2006). Sie müssen aber nicht notwendigerweise zu einem Satz von $\alpha$-proteobakteriellen Genen moderner Genome gehören.

## 9.2 Syntenie: Nukleotid- vs. Proteinsequenzen

Die meisten Methoden, die testen, ob ein Gen seinen Ursprung in einem lateralen Gentransfer hat, beruhen auf Gen- oder Proteinfamilien (Gogarten et al., 2008; McInerney et al., 2008). Daher wurde in Kapitel 6 eine Methode entwickelt, die Genfamilien auf Basis konservierter syntenischer Abschnitte erstellen sollte. Zunächst wurde überprüft, inwieweit sich die verfügbaren Daten für solche Analysen eignen. Dafür wurde zunächst die Genomsequenz von *Salmonella typhii* in kurze Fragmente von 500 bp unterteilt, mit denen jeweils eine `BLAST`-Datenbanksuche durchgeführt wurde. Die Ergebnisse der positionsbezogenen Sequenzähnlichkeit

zeigen, daß sich mit dem einfachen `blastn` nur für sehr eng begrenzte Bereiche eine Menge von Treffern erzielen lässt, die eine Suche nach Positionsorthologen erlauben. Werden die Nukleotidsequenzen in Aminosäuren übersetzt, zeigt sich eine starke Erhöhung der Trefferdichte. Allerdings hat die Anwendung von `tblastx` auch sehr große Nachteile, da der Algorithmus aufwendiger ist und sich die Laufzeiten stark verlängern. Außerdem werden auch nicht proteinkodierende Bereiche in Aminosäuresequenzen übersetzt. Hier besteht die umgekehrte Problematik wie bei kodierenden Sequenzen, eine Änderung die beispielsweise auf der RNA-Ebene nur kleine Auswirkungen hat (zum Beispiel die Insertion eines Nukleotids) würde bei der Übersetzung für eine Leserasterverschiebung sorgen.

Daher bietet es sich für die Entwicklung eines Algorithmus zur Ermittlung von Positionsorthologen an, statt von der DNA-Ebene von der Protein-Ebene auszugehen. Das erlaubt zum einen die Verwendung schnellerer `BLAST`-Varianten für das Auffinden homologer Sequenzen (`blastp`). Zum Anderen ermöglicht es eine Vereinfachung der Positionsangaben, was zu wesentlich intuitiveren Kriterien für den Schwellenwert führt, der bestimmt, bei welchem Abstand zwei Treffer noch als Treffer bezeichnet werden. Wie in den Methoden beschrieben, wurde jedem Gen aufgrund seiner Startposition in Nukleotiden eine Positions-ID in Proteinen zugeordnet. Proteine mit aufeinanderfolgenden Positions-IDs lagen also benachbart auf dem Chromosom. Auf der DNA-Ebene müssten hier viele Sonderfälle berücksichtigt werden, inklusive überlappender Genbereiche und langer Abschnitte mit nicht-proteinkodierenden Sequenzen. Die mit der Methode berechneten Familien von positionsorthologen Genen werden im folgenden als synteniebasierte orthologe Gene oder kurz SOGs bezeichnet.

Für die Daten berechnet aus neun Stämmen (siehe Kapitel 7) von *Escherichia coli* lassen sich Proteinfamilien bestimmen, die in guter Deckung mit den mit dem Markov-Cluster-Algorithmus (`mcl`) ermittelten Proteinfamilien sind. Unterschiede ergeben sich zwangsläufig dadurch, daß die SOG Methode nur ein Gen pro Organismus erlaubt, so daß sich Proteine, die sich bei `mcl` in einer Proteinfamilie befinden, sich bei SOG in mehrere Proteinfamilien aufspalten können (Tab. 7.1). Die ausführliche Beschreibung eines mit der SOG-Methode erstellten Datensatzes und der Vergleich mit `mcl` basierten Daten findet sich im folgenden Abschnitt.

Bei dem in Abb. 6.5 gezeigten Beispiel für eine Synteniekarte handelt es sich um einen manuell aus den SOG-Daten erstellten Ausschnitt. Ähnliche Formen

der Veranschaulichung in ausführlicherer Form existieren beispielsweise in Form der Datenbank YGOB (engl. *yeast gene order browser*) (Byrne und Wolfe, 2005). Dort wurden die Gene verschiedener Hefestämme einander gegenüber gestellt. Eine solche Darstellungsform wäre in begrenztem Maß auch für die in dieser Arbeit vorgestellten Methode zur Erstellung von orthologen Genfamilien möglich. Allerdings sollte das Augenmerk auf den referenzbasierten Datensätzen liegen, da sich die Anordnung ansonsten schwierig gestalten würde. Eingeschobene Gene in den Genomen der Vergleichsgenome könnten dabei durchaus in die Darstellung einbezogen werden, solange es sich um kurze Abschnitte handelt. Hier wäre ähnlich wie im Beispiel eine Darstellungsform nötig, die deutlich macht, daß es sich in solchen Bereichen um keine berechnete Syntenie handelt, sondern um ein Hinzufügen externer Daten, um Lücken zu füllen. Beispielsweise durch das Einfärben der entsprechenden Abschnitte.

Die Erstellung von phylogenetischen Analysen erfordert eine erhöhte Aufmerksamkeit bei der Auswahl der Sequenzen. Die Berechnungen reduzieren die Sequenzen auf mathematische Eigenschaften, und jede Eingabe erzeugt in der Regel auch eine Ausgabe. Statistische Angaben können helfen herauszufinden, ob es sich bei den verwendeten Daten um brauchbares Ausgangsmaterial handelt. Gerade wenn phylogenetische Analysen im großen Maßstab durchgeführt werden sollen, ist es sinnvoll durch die Verwendung von vorher definierten Proteinfamilien eine größere Sicherheit in die Datensätze zu bringen, indem gleiche Kriterien für alle Daten angelegt werden.

Synteniebasierte Orthologe unterliegen hierbei einer Einschränkung, die aber auch ein Vorteil ist. Die Datensätze enthalten keine Paraloge. Für jedes verwendete Genom wird exakt ein Protein für jede Proteinfamilie ausgewählt. Bei hierarchischen Clusterverfahren ist es möglich, daß je nachdem wie die Parameter gewählt werden, ähnliche Proteinfamilien zu einem großen Cluster vereinigt werden. Das kann zu schwer interpretierbaren Ergebnissen führen. Da der Algorithmus bei synteniebasierten Orthologen aufgrund der Position innerhalb des Genoms eindeutige Zuordnungen trifft, entfällt diese Problematik. Das kann jedoch dazu führen, daß bei duplizierten Genen die Einordnung aufgrund der Position erfolgt, obwohl die Sequenzähnlichkeit zu einer anderen Kopie wesentlich höher ist.

Das Ausgangsmaterial für die Suche ist eine `BLAST`-Datenbanksuche, die eigentliche Rechenzeit des Algorithmus ist sehr kurz, die Hauptprobleme bei der Effizienz sind

das Einlesen der BLAST-Ergebnisse und die Auslastung des Arbeitsspeichers. Da die Qualität der Ergebnisse nur einen Vorteil bringen kann, wenn die Konservierung der Sequenzreihenfolge zwischen den betrachteten Spezies relativ hoch ist, ist der nützliche Einsatzbereich der Vergleich näher verwandter Spezies, so daß die Menge der zu bearbeitenden Daten (beispielsweise etwa 300 Genome auf einem Rechner mit 32 GB Arbeitsspeicher) ausreichend ist. Bei einem breit angelegten Datensatz, der alle Verwandtschaftsbereiche der Prokaryoten abdeckte, wurden auch in weiter entfernten Verwandtschaftsbereichen syntenische Abschnitte gefunden. Hierbei handelte es sich allerdings in der Regel um einzelne stark konservierte Gengruppen. Das deckt sich mit Ergebnissen die mit zunehmender Distanz zwischen Organismen eine sehr starke Abnahme der Konservierung der Genreihenfolge bemerkten, aber für wenige Bereiche auch zwischen Archaebakterien und Eubakterien noch Abschnite mit konservierter Genreihenfolge fanden (Tamames, 2001). Diese Konservierung ist dabei oft nicht exakt, geht aber über die Operon-Ebene hinaus (Lathe et al., 2000).

## 9.3 Kartierung von Rekombination in *E. coli*

Nachdem in Kapitel 5 untersucht wurde wie groß der Einfluß von lateralem Gentransfer in prokaryotischen Genomen ist, sollte in Kapitel 7 der Einfluß von Rekombination innerhalb von neun Stämmen von *E. coli* untersucht werden. Dazu wurden Genfamilien basiererend auf einem hierarchischem Clusteralgorithmus benutzt (Cluster orthologer Gene – COGs), und solche basierend auf den Positionsorthologen wie sie im vorigen Abschnitt beschrieben wurden (syntenische orthologe Gene – SOGs). Da die Methoden sich grundsätzlich unterscheiden, weil der SOG Datensatz auf *E. coli* K12 als Referenz beruht, während die COGs ohne Referenz berechnet wurden wurde der Datensatz COG* eingeführt. Er enthält genau die Genfamilien, die einen Vertreter des Referenzgenoms aufweisen und ist damit direkt mit den Ergebnissen der SOGs vergleichbar.

Die Verteilung der Familiengrößen (Abb. 7.3) zeigt einen erhöhten Anteil an universellen Clustern bei Verwendung von SOGs. Das ist unter anderem auf die strikteren Schwellenwerte zurückzuführen, die nötig waren, um im COG*-Datensatz den Anteil der Proteinfamilien, die mehrere Paraloge enthielten, zu reduzieren. Allerdings ist die Anzahl der universellen Proteinfamilien bei Verwendung des gleichen Schwellenwerts von 90 % Identität zwischen Referenz und Treffer mit 2.802 immer

noch um 10,5 % höher als beim COG* Datensatz. Die Menge der Proteine, die nicht in eine Proteinfamilie eingeordnet werden können, ist mit 245 um 8,9 % niedriger.

Die Abbildung der verschiedenen Auftrittsverteilungen (Abb. 7.4) stellt die Häufigkeit und die Verteilung von verschiedenen Kombinationen von Organismen dar. Durch die Sortierung fällt auf, daß es zwar eine große Anzahl unterschiedlicher Verteilungen gibt, es aber auch eine Anzahl von Verteilungen gibt, die nicht auftreten. Diese würden zumeist eins von mehreren eng verwandten Genomen enthalten, z.B. *E. coli* O157_Sakai, aber nicht *E. coli* O157_EDL933. Da die Ähnlichkeiten zwischen diesen Genomen sehr hoch sind, ist die Wahrscheinlichkeit, daß ein Gen aus dem einen Genom in eine Genfamilie eingeordnet werden kann, aber keins aus dem anderen sehr gering.

Für die oben genannte Paarung *E. coli* O157_Sakai und *E. coli* O157_EDL933 lassen sich in den COG*-Daten 33 Verteilungen finden, die nur eins der beiden Genome enthalten, davon sechs in größerer Häufigkeit. In den SOG-Daten ist dies nur für 29 der Fall, von denen auch nur zwei in größeren Häufigkeiten auftreten (Abb.7.4). Bei den Rekombinationsnetzwerken unterscheiden sich die beiden Datensätze nur marginal. Das spricht dafür, daß die beiden Methoden zur Erstellung von Proteinfamilien Ergebnisse liefern, die sich systembedingt unterscheiden, nicht jedoch zu komplett unterschiedlichen Ergebnissen führen. Die Untersuchung der nächsten Nachbarn der *singletons* ergab realtiv große Unterschiede zwischen den untersuchten Datensätzen, da sich für den COG*-Datensatz nahezu doppelt so viele Proteine nicht in Gruppen einordnen lassen. Da sich die Menge der nächsten Nachbarn hauptsächlich in der Gruppe der Proteine mit nächsten Nachbarn in den $\gamma$-Proteobakterien unterscheidet, spricht vieles dafür, daß es sich hierbei um Proteine handelt die aufgrund unterschiedlicher Schwellenwerte im `mcl` Datensatz nicht berücksichtigt wurden.

Bei der Bestimmung der Gesamtrate für Rekombination wurden bei den Positionsorthologen weniger Genfamilien mit Transfers vorhergesagt. Das spricht dafür, daß sich die so gewonnenen Daten besser durch die Referenztopologie erklären lassen, als die durch den hierarchischen Clusteralgorithmus berechneten Daten. Gentransfer spielen nicht nur zwischen sehr weit entfernt verwandten Prokaryoten eine Rolle, sondern auch zwischen verschiedenen Stämmen derselben Spezies gibt es eine große Menge Gentransfers in Form von Rekombination.

In dieser Arbeit wurden im COG* Datensatz für 27,1 % aller Genfamilien und im SOG Datensatz für 22,7 % aller Genfamilien ein LGT ermittelt. Ähnliche Ergebnisse ergab auch eine Studie mit Hilfe eines referenzbasierten Verfahrens (Kunin und Ouzounis, 2003*b*), in der für 51 Genome mit unterschiedlichen Referenztopologien laterale Gentransfers abgeleitet wurden. Dort wurden mit Hilfe eines 16S rRNA-Baumes und mit unterschiedlichen Evolutionsmodellen für 18 - 41 % aller untersuchten Genfamilien Gentransfers ermittelt. Diese Rate lag bei der Vergleichstopologie die auf Genverteilungen beruhte leicht höher, und bei einer zufälligen Topologie bei 80 - 95 % (Kunin und Ouzounis, 2003*a*).

Gene, die sich nicht in synteniebasierte orthologe Gene (SOGs) einordnen lassen (engl. *singletons* oder *orphans* – einelementige Menge bzw. Waise), haben ihre nächsten Nachbarn hauptsächlich innerhalb der $\gamma$-Proteobakterien. Insgesamt können aber mit Hilfe der SOG-Methode mehr Proteine einer Familie zugeordnet werden, als mit den Clustern orthologer Gene (COGs), die mithilfe der `mcl` Methode erstellt wurden. Die Methode der Erstellung von Proteinfamilien hat jedoch keinen großen Einfluß auf die Ergebnisse der Analysen in der vorliegenden Arbeit, die auf diesen beruhen. Allerdings spricht eine niedrigere Rate von Transfer-Ereignissen pro Proteinfamilie bei den Positionsorthologen dafür, daß sich die berechneten Daten besser durch die benutzte Referenztopologie erklären lassen.

## 9.4 Rekombinationsraten

Für die Proteinfamilien auf Synteniebasis wurde für 22,1 % der Familien Rekombination abgeleitet. In Kapitel 8 wurde eine weitere Methode benutzt um den Einfluß von Rekombination und Substitution auf die Genomevolution von *E. coli* zu untersuchen. Hierbei wurde die Rate nicht in Einheiten von Proteinfamilien sondern in Sequenzpositionen ermittelt, die auf Rekombination zurückzuführen sind.

Für jede der 2.665 universellen synteniebasierten Proteinfamilien wurde ein *neighbor-joining*-Baum erstellt, aus denen ein Konsensusbaum berechnet wurde. Zusätzlich wurde ein Wahrscheinlichkeitsbaum aufgrund von konkatenierten Alignments berechnet. Da beide Methoden zur gleichen Topologie führten, wurde diese als erste Referentopologie für die weiteren Analysen verwendet. Anschließend wurde die Kompatibilität der Alignments zu den unterschiedlichen Splits untersucht. Zusätzlich wurden die Splits, die in den Referenztopologien enthalten sind, über der Abbildung

wiedergegeben, um einen Vergleich zu ermöglichen. Als kompatibel gilt eine Alignmentposition, wenn sie sich auf einen Split projizieren lässt.

Die Abbildungen 8.1 und 8.2 zeigen im Hauptteil den Anteil kompatibler Splits für die auf der Y-Achse aufgetragenen Genfamilien an. Auf der X-Achse sind alle auftretenden Splits aufgetragen, deren Sortierung durch den oberen Teil von Abb. 8.1 vorgegeben wird. In diesem Bereich ist in Abb 8.2 ein Histogramm zu sehen, das die Häufigkeit der Splits bezogen auf die Alignmentpositionen der Genfamilie wiederspiegelt. Der Unterschied zwischen den beiden Abbildungen liegt in der Sortierung der Genfamilien. Während sie in Abb. 8.1 nach der Reihenfolge der Gene im Referenzgenom sortiert sind, stehen sie in Abbildung 8.2 aufgelistet nach Auftrittsmustern.

Ein großer Anteil der Splits, der durch die Alignments unterstützt wurde liegt in den terminalen Ästen – den Blättern. Da die Sortierung der Splits ansonsten nach absteigender Häufigkeit der Unterstützung erfolgt, lässt sich im oberen Teil von Abb. 8.1 leicht die Unterstützung für die Referenztopologien ablesen. Während bei dem ML-Baum und dem NJ-Konsensusbaum neun der elf internen Kanten zu denen mit der stärksten Unterstützung gehören, ist das bei dem Baum der auf rRNA Sequenzen beruht, nur für zwei Kanten der Fall.

Des Weiteren lässt sich auch ablesen welche Splits zwar eine hohe Unterstützung in den zugrunde liegenden Daten haben, aber trotzdem keinen Eingang in die Phylogenie gefunden haben. Einer der beiden würde *E. coli* 536 und *E. coli* CFT073 zu einer monophyletischen Gruppe machen, der andere die Stämme *E. coli* APEC, *E. coli* UTI89 und *E. coli* 536. Trotz der hohen Unterstützung stehen diese beiden Splits im Widerspruch zu häufigeren Splits und wurden daher in den erstellten Phylogenien nicht berücksichtigt. Solche Verwandtschaftsbeziehungen könnten lediglich mit Netzwerkmethoden korrekt wiedergegeben werden (Bryant und Moulton, 2004; Huson und Bryant, 2006).

Die Splits des rRNA Baumes finden sich fast ausschließlich unter denjenigen deren Häufigkeit in dem Histogramm im oberen Abschnitt von Abb. 8.2 kaum noch abgelesen werden kann. Das macht deutlich, das sich die phylogenetische Geschichte einiger weniger Sequenzen – auch wenn es sich dabei um so wichtige und essentielle handelt wie das bei rRNA Sequenzen der Fall ist – nicht mit der Geschichte des gesamten Genoms decken muß (Esser et al., 2004; Dagan und Martin, 2006).

Wenn die nach den Auftrittsverteilungen sortierte Kompatibilitätsverteilung in Abbildung 8.2 betrachtet wird, so fällt ins Auge, daß es keine Proteinfamilie zu geben scheint, die eine identische Zusammensetzung von Splits aufweist, was bei der Anzahl der betrachteten möglichen Kombination von Splits nicht verwundert. Einige wenige Muster treten jedoch hervor. In der nach Genreihenfolge sortierten Abbildung 8.1 treten einige Muster deutlicher hervor, in denen einige aufeinanderfolgende Genfamilien sehr ähnliche Verteilungsmuster aufweisen. Insbesondere gibt es einige Abschnitte die auffällig abweichende Verteilungen aufweisen und sich stark von dem recht gleichmäßigen Grundmuster abheben. Das deutet daraufhin, das es auch hier Sequenzeigenschaften gibt die sich über konservierte Bereiche fortsetzen.

Da die meisten Methoden zur Ableitung von Rekombinationsereignissen auf dem Vergleich mit einer Referenzphylogenie basieren ist die Auswahl der Referenzphylogenie ein entscheidender Schritt, da jeder Unterschied in der Topologie unterschiedliche Ergebnisse zur Folge hat (Kunin und Ouzounis, 2003*a*). Die Phylogenien aus dem vorherigen Abschnitt (`neighbor-joining` rRNA, ML-Konsensusbaum und NJ konkateniert) wurden als Referenzphylogenien für die Berechnung von Substitutions- und Rekombinationsraten bezogen auf Alignmentpositionen benutzt. Wie im vorigen Abschnitt sollte dabei zwischen kompatiblen und unkompatiblen Splits unterschieden werden. Allerdings sollte anstatt einer Lokalisation verschiedener Muster auf dem Chromosom hier eine projektion der abgeleiteten Ereignisse auf einer Referenztopologie stattfinden. Zusätzlich zu den bereits erwähnten Topologien wurde hier zusätzlich ein zufälliger Baum generiert und als Referenz gewählt.

Für jeden Ast in jedem Baum wurde sowohl die Rekombinationsrate als auch die Substitutionsrate bestimmt. Zusätzlich wurde falls möglich das Verhältnis von Substitutions- zu Rekombinationsrate als ganzzahliges Verhältnis berechnet. Für einige Splits, die keinerlei Unterstützung in den Daten hatten, für die also entweder keine Substitutionsereignisse und/oder keine Rekombinationsereignisse abgeleitet werden konnten, war das nicht möglich. Solche Splits sind in den Abbildungen 8.8-8.10 farblich hervorgehoben.

Bei dem ML-Konsensusbaum ist dies jedoch nicht der Fall. Hier lassen sich für alle Kanten alle Raten bestimmen. Im Schnitt lassen sich 71,9 % aller betrachteten Positionen durch Substitutionen erklären. Bei den internen Kanten ist die Substitutionsrate durchweg ausgeglichen oder höher als die Rekombinationsrate, mit Ausnahme des Splits, der die Stämme ATCC und HS miteinander verbindet, hier wurden mit

4.853 Positionen für fast dreimal soviele Positionen Rekombination abgeleitet wie für Substitutionen. Bei den externen Kanten werden in der Regel mehr Rekombinationen abgeleitet, was vermutlich an den kurzen terminalen Ästen liegt auf denen sich nur wenige Substitutionen ereignet haben.

Wird der rRNA Baum als Referenz ausgewählt, so fällt auf, daß die Daten für einige Kanten keinerlei Unterstützung enthalten. Für drei Kanten, die in Abbildung 8.9 farblich hervorgehoben sind, können keine Substitutionsraten bestimmt werden. Da alle Daten, die nicht durch Substitution (also durch die Topologie des Baums) erklärt werden können, durch Rekombination erklärt werden müssen, ist das Verhältnis von Substitution zu Rekombination im Vergleich zu dem ML-Baum in Richtung der Rekombination verschoben. 60 % der Daten werden nun durch Rekombination erklärt. Werden Extremfälle nicht berücksichtigt, in denen sehr hohe Ratenverhältnisse durch wenige einzelne Substitutionen hervorgerufen werden (zum Beispiel ein Verhältnis von 1:9560 bei nur zwei Substitutionen) so treten hier auch Äste auf, bei denen 80mal mehr Daten durch Rekombination erklärt werden als durch Substitution. Insgesamt treten bei neun von 14 externen Kanten teilweise deutlich höherere Rekombinationsraten auf.

Bei dem zufällig erstellten Baum lassen sich noch weniger Daten durch Substitutionen erklären. Hier konnte für keine interne Kante eine Substitution abgeleitet werden. Für fünf von elf Kanten konnten Rekombinationsraten bestimmt werden. Insgesamt werden fast dreimal soviele Alignmentpositionen durch Rekombinationen erklärt, als durch Substitutionen. Alle Substitutionen werden an externen Kanten abgeleitet, diese Werte sind daher identisch mit allen anderen Referenzbäumen, da sich hier nur die Menge der rekombinanten Stellen ändert.

Bei dem Vergleich der Abbildungen 8.1 und 8.8 fällt auf, daß sich die höheren Rekombinationsraten die sich für einige Äste des Baumes durch die Unterstützung für die zugehörigen Splits erklären lassen. So weisst der Split der *E. coli* HS und *E.coli* ATCC zusammenfasst obwohl er im Konsensusbaum enthalten ist nur eine relativ geringe Unterstützung in den Alignments auf (Abb. 8.2). Umgekehrt verhält es sich mit dem Split der die beiden Stämme *E. coli* 536 und *E. coli* CFT073 verbindet. Obwohl er die zehnthöchste Unterstützung unter den möglichen internen Splits aufweist ist er nicht im Konsensusbaum enthalten. Das wiederum führt zu erhöhten Rekombinationsraten bei den terminalen Splits der beiden Stämme.

Daraus folgt, daß die Auswahl der Referenztopologien einen enormen Einfluß auf die Abschätzung der Substitutions- und Rekombinationsraten hat. Eine Topologie, die möglichst viele Splits mit einer hohen Unterstützung durch die Alignments enthält, führt zu einer niedrigeren Abschätzung von Rekombination. Dagegen führt die Auswahl eines fehlerhaften Referenzbaums zu einer Überschätzung des Einflusses der Rekombination. Die Verteilung der abgeleiteten Ereignisse auf der Topologie kann auf der einen Seite zwar auf eine Vorliebe der Reaktion zwischen bestimmten Arten oder hier Stämmen hinweisen. Auf der anderen Seite kann eine ehöhte Rekombinationsrate aber auch auf eine fehlerhafte Topologie hinweisen, die in einigen Fällen durch die Berechnung der Raten für alle Äste relativ genau eingeordnet werden kann. So ist im RNA Baum die Gruppe APEC, UTI89 und CFT073 im Gegensatz zum Konsensusbaum nicht monophyletisch. Alle internen und externen Kanten, die im RNA-Baum zwischen diesen drei Taxa liegen, weisen stark erhöhte Rekombinationsraten auf. Drei von vier internen Kanten werden durch die Daten nur sehr wenig unterstüzt.

## 9.5 Schlußfolgerung und Ausblick

In dieser Arbeit wurden unterschiedliche Ansätze präsentiert, um lateralen Gentransfer (LGT) in Prokaryoten zu untersuchen: Die Analyse nächster Nachbarn, die Abschätzung über die Genomgrößen gemeinsamer Vorfahren sowie Karten für die Ableitung von Rekombination. Diese bieten in dieser Reihenfolge ansteigende Qualitäten für die Ableitung von LGT.

Die Analyse nächster Nachbarn der Gene von $\alpha$-Proteobakterien kann als Abschätzung für rezente Genome auf der taxonomischen Ebene der Klasse gesehen werden. Bei den Auftrittswahrscheinlichkeiten handelt es sich um ein vereinfachtes Modell. Hier gehen Sequenzähnlichkeiten nur indirekt in die Ergebnisse ein. Allerdings können, in diesem auf Genfamilien basierenden Ansatz, paraloge Gene in die Berechnung eingehen und die Ergebnisse beeinflussen. Die Ableitungen von Rekombinations- und Substitutionsraten berücksichtigen Sequenzeigenschaften und basieren auf den syntheniebasierten Genfamilien, die zur Entdeckung wahrer Orthologer gedacht sind.

Alle benutzten Methoden zeigen das gleiche Ergebnis. Lateraler Gentransfer ist kein seltenes Ereignis, sondern spielt eine wichtige Rolle in der Evolution prokaryotischer

Genome, sowohl innerhalb einer Art als auch in größer werdenden taxonomischen Einheiten. In der Nachbarschaftsanalyse der $\alpha$-Proteobakterien zeigte sich, daß bis zu 36,1 % der Genome nächste Nachbarn außerhalb der eigenen Klasse aufwiesen und damit einen Hinweis auf möglichen lateralen Gentransfer liefern. Verschiedene Untersuchungen auf systematische Abweichungen mithilfe unterschiedlicher phylogenetischer Methoden bestätigten diese Ergebnisse.

Es wurde bereits gezeigt, daß es auch in unterschiedlichen Stämmen einer Art einen hohen Anteil von Variation gibt (van Passel et al., 2008). Daher wurde zusätzlich eine Analyse durchgeführt bei der durch die Verwendung von Genanwesenheitsverteilungen (engl. *presence-absence-patterns*) der Einfluß von lateralem Gentransfer zwischen neun *Escherichia coli* Stämmen bestimmt werden sollte. Dabei zeigte sich, daß lateraler Gentransfer nicht nur zwischen unterschiedlichen Klassen und Gattungen stattfindet, sondern auch zwischen verschiedenen Stämmen einer Art. Hier ließ sich in 27,1 % aller Genfamilien ein lateraler Gentransfer ableiten. Die Familien wurden dabei zunächst mit einem hierarchischen Clusterverfahren abgeleitet. Zusätzlich wurde eine Methode entwickelt, mit der es möglich ist Genfamilien aus wahren Orthologen zu erstellen, indem Positionsinformationen in die Berechnung eingehen. Damit kann verhindert werden, daß Paraloge die Ergebnisse der Analysen verfälschen. Mit den so erstellten synteniebasierten orthologen Genen wurde für 22,7 % der Genfamilien ein Gentransfer abgeleitet, was sich mit den Ergebnisen deckt bei denen unter verschiedenen Bedingungen für 18 - 41 %aller Genfamilien LGTabgeleitet wurde (Kunin und Ouzounis, 2003*a*).

Alle Ergebnisse deuten darauf hin, daß lateraler Gentransfer auf den verschiedenen taxonomischen Ebenen eine große Rolle spielt. Daher handelt es sich bei den berechneten Werten eher um eine untere Schranke für die Abschätzung. Das führt dazu, daß der bifurzierende phylogenetische Baum keine geeignete Darstellungsform für die Verwandtschaft prokaryotischer Organismen ist. Hier sollte dazu übergegangen werden, Netzwerke für die Darstellung phylogenetischer Verwandtschaft zu benutzen, mit denen die Darstellung widersprüchlicher Signale und ein realistischeres Abbild der prokaryotischen Evolution möglich ist. Damit können entweder widersprüchliche Signale gleichzeitig (Bryant und Moulton, 2004; Huson und Bryant, 2006) oder verschiedene Informationen getrennt voneinander in mehreren Ebenen dargestellt werden, beispielsweise eine Referenztopologie mit überlagerter Darstellung von abgeleiteten Gentransfers (Dagan und Martin, 2009).

Das hat nicht nur einen Einfluß auf Verwandtschaftsverhältnisse von Prokaryoten, sondern geht weit darüber hinaus. Phylogenetische Analysen beruhen auf gegenwärtig beobachtbaren Ausgangsdaten und machen Abschätzungen über den Zustand dieser Daten zu sehr weit zurückliegenden Zeitpunkten. Allerdings haben wir nur ein wages Bild davon wie beispielsweise die Zusammensetzung der Genome der Vorfahren der Eukaryoten aussah.

Die Mutationsrate für *E. coli* wurde mit $4 \times 10^{-15}$ Änderungen pro Generation und Nukleotid beziffert (Milkman und Stoltzfus, 1988). Der Einfluß von Rekombination wurde wenige Jahre später 50 mal höher eingeschätzt, als der von Substitutionen (Milkman und Bridges, 1990). In dieser Arbeit wurden Rekombinations- und Substitutionsraten durch den Vergleich von Sequenzalignments mit einer Referenztopologie abgeschätzt. Für die auf Basis eines Konsensusbaumes, erstellt aus *maximum-likelihood*-Bäumen für 2.665 universelle Gene, berechneten Raten lag das Verhältnis der Sequenzveränderungen durch Substitutionen zu Rekombinationen bei 3:1. Um die Beständigkeit der Methode zu überprüfen, wurden auch ein rRNA-Baum und ein zufälliger Baum als Referenztopologie herangezogen. Beim rRNA-Baum wurde ein Verhältnis von Substitutions- zu Rekombinationsrate von 2:3 ermittelt, für den zufälligen Baum lag der Wert bei 1:3. Dadurch konnte gezeigt werden, daß die Abschätzung sehr stark von der Auswahl einer Referenztopologie abhängt und daß die Rekombinationsrate umso höher wird, je schlechter die zugrunde liegende Referenztopologie ist.

Da Proteinfamilien auf Synteniebasis eher in der Lage sind, evolutionäre Prozesse korrekt wiederzugeben als Methoden, die nur auf Sequenzähnlichkeiten beruhen, könnte die Anwendung dieser Methode in vielen Bereichen, in denen Proteinfamilien die Grundlagen von Berechnungen darstellen, zu einer Verbesserung der Ergebnisse beitragen. Grundsätzlich ist es auch möglich, die Methode zur Rekonstruktion von Positionsorthologen auf größere Datensätze anzuwenden. Die erreichbaren Datenmengen sind zwar durch die geringe Menge der konservierten Bereiche bei nur entfernt verwandten Arten beschränkt, ermöglichen jedoch eine Auswahl von Genen, bei denen von einer gemeinsamen Vergangenheit ausgegangen werden kann. Der Übergang von referenzbasierten Berechnungen zu Konsensusfamilien ermöglicht vermutlich, in großen Datensätzen eine angemessene Abdeckung der Daten durch Genfamilien zu gewährleisten.

# A Anhang

## Schwellenwerte für die Cluster orthologer Gene (mcl)

Für die Clusterberechnung mithilfe von mcl wurde ein sehr strikter Schwellenwert von 90% angelegt. Damit wurde erreicht, das die Datensätze eine geringe Menge von Paralogen enthielten. Da bei den syntenibasierten Clustern ein Gen je Genom Bestandteil eines Clusters werden kann, konnten so vergleichbare Datensätze erzielt werden.

| Schwellenwert für die die `mcl` Daten | Einträge | | Universelle Proteinfamilien | |
|---|---|---|---|---|
| | gesamt | mit Paralogen | gesamt | mit Paralogen |
| 30 | 39.558 | 3,24 % | 2.915 | 6,07 % |
| 40 | 39.733 | 2,46 % | 2.904 | 3,99 % |
| 50 | 39.733 | 1,96 % | 2.886 | 2,53 % |
| 60 | 39.735 | 1,59 % | 2.851 | 1,40 % |
| 70 | 39.608 | 1,29 % | 2.811 | 0,85 % |
| 80 | 39.314 | 1,04 % | 2.719 | 0,44 % |
| 90 | 38.712 | 0,82 % | 2.537 | 0,20 % |
| 95 | 37.953 | 0,60 % | 2.224 | 0,09 % |

# Liste der verwendeten Genome

Übersicht über die verwendeten Genomdaten. Alle verwendeten Genome wurden vom FTP Server des „National Center for Biotechnology Information“ (NCBI) [1] heruntergeladen. Eine Übersicht findet sich in Tabelle 4.1 auf Seite 14

## Wie $\alpha$-proteobakteriell sind $\alpha$-Proteobakterien

### Referenzgenome

6 Mitochondrien Genome

- Homo sapiens (NC_001807)
- Malawimonas jakobiformis (NC_002553)
- Marchantia polymorpha (NC_001660)
- Reclinomonas americana (NC_001823)
- Spizellomyces punctatus (NC_003052)
- Triticum aestivum (NC_007579)

18 $\alpha$-Proteobakterien

- Agrobacterium tumefaciens str. C58
- Bartonella henselae str. Houston-1
- Bradyrhizobium japonicum USDA 110
- Brucella melitensis 16M
- Brucella suis 1330
- Caulobacter crescentus CB15
- Magnetococcus sp. MC-1
- Magnetospirillum magnetotacticum MS-1
- Mesorhizobium loti MAFF303099
- Novosphingobium aromaticivorans DSM124
- Rhodobacter sphaeroides 2.4.1
- Rhodopseudomonas palustris CGA009
- Rhodospirillum rubrum
- Rickettsia conorii str. Malish 7
- Rickettsia prowazekii
- Sinorhizobium melioti
- Wolbachia sp. (Endosymbiont of Drosophila melanogaster)
- Zymomonas mobilis subsp. mobilis ZM4

[1] `ftp://ftp.ncbi.nih.gov/Genoms/Bacteria`

## Datenbank

231 Genome vollständig sequenzierter Prokaryoten

| Species | Gruppen-Kürzel | Sequenzen |
|---|---|---|
| Acinetobacter sp. ADP1 | 0315 | 3325 |
| Aeropyrum pernix K1 | 5100 | 2694 |
| Agrobacterium tumefaciens str. C58 | 0311 | 4556 |
| Anaplasma marginale str. St. Maries | 0311 | 949 |
| Aquifex aeolicus VF5 | 3210 | 1522 |
| Archaeoglobus fulgidus DSM 4304 | 5110 | 2421 |
| Azoarcus sp. EbN1 | 0312 | 4598 |
| Bacillus anthracis str. 'Ames Ancestor' | 3280 | 5309 |
| Bacillus anthracis str. Ames | 3280 | 5126 |
| Bacillus anthracis str. Sterne | 3280 | 5287 |
| Bacillus cereus ATCC 10987 | 3280 | 5844 |
| Bacillus cereus ATCC 14579 | 3280 | 5124 |
| Bacillus cereus ZK | 3280 | 5134 |
| Bacillus clausii KSM-K16 | 3280 | 4108 |
| Bacillus halodurans C-125 | 3280 | 4066 |
| Bacillus licheniformis ATCC 14580 | 3280 | 4011 |
| Bacillus licheniformis DSM 13 | 3280 | 4196 |
| Bacillus subtilis subsp. subtilis str. 168 | 3280 | 4106 |
| Bacillus thuringiensis serovar konkukian str. 97-27 | 3280 | 5197 |
| Bacteroides fragilis NCTC 9343 | 3220 | 4236 |
| Bacteroides fragilis YCH46 | 3220 | 4578 |
| Bacteroides thetaiotaomicron VPI-5482 | 3220 | 4778 |
| Bartonella henselae str. Houston-1 | 0311 | 1488 |
| Bartonella quintana str. Toulouse | 0311 | 1142 |
| Bdellovibrio bacteriovorus HD100 | 0313 | 3583 |
| Bifidobacterium longum NCC2705 | 3200 | 1727 |
| Bordetella bronchiseptica RB50 | 0312 | 4994 |
| Bordetella parapertussis | 0312 | 4185 |
| Bordetella pertussis Tohama I | 0312 | 3447 |
| Borrelia burgdorferi B31 | 3320 | 850 |
| Borrelia garinii PBi | 3320 | 832 |
| Bradyrhizobium japonicum USDA 110 | 0311 | 8317 |
| Brucella abortus biovar 1 str. 9-941 | 0311 | 3085 |
| Brucella melitensis 16M | 0311 | 3198 |
| Brucella suis 1330 | 0311 | 3273 |
| Buchnera aphidicola str. APS (Acyrthosiphon pisum) | 0315 | 564 |
| Buchnera aphidicola str. Bp (Baizongia pistaciae) | 0315 | 504 |
| Buchnera aphidicola str. Sg (Schizaphis graminum) | 0315 | 545 |
| Burkholderia mallei ATCC 23344 | 0312 | 4764 |
| Burkholderia pseudomallei K96243 | 0312 | 5729 |
| Campylobacter jejuni RM1221 | 0314 | 1838 |
| Campylobacter jejuni subsp. jejuni NCTC 11168 | 0314 | 1634 |
| Candidatus Blochmannia floridanus | 0315 | 583 |
| Caulobacter crescentus CB15 | 0311 | 3737 |
| Chlamydia muridarum Nigg | 3230 | 904 |
| Chlamydia trachomatis D/UW-3/CX | 3230 | 897 |
| Chlamydophila abortus S26/3 | 3230 | 932 |

| Species | Gruppen-Kürzel | Sequenzen |
|---|---|---|
| Chlamydophila caviae GPIC | 3230 | 998 |
| Chlamydophila pneumoniae AR39 | 3230 | 1110 |
| Chlamydophila pneumoniae CWL029 | 3230 | 1052 |
| Chlamydophila pneumoniae J138 | 3230 | 1069 |
| Chlamydophila pneumoniae TW-183 | 3230 | 1113 |
| Chlorobium tepidum TLS | 3240 | 2252 |
| Chromobacterium violaceum ATCC 12472 | 0312 | 4407 |
| Clostridium acetobutylicum ATCC 824 | 3280 | 3672 |
| Clostridium perfringens str. 13 | 3280 | 2660 |
| Clostridium tetani E88 | 3280 | 2373 |
| Corynebacterium diphtheriae | 3200 | 2272 |
| Corynebacterium efficiens YS-314 | 3200 | 2942 |
| Corynebacterium glutamicum ATCC 13032 | 3200 | 3099 |
| Coxiella burnetii RSA 493 | 0315 | 2046 |
| Dehalococcoides ethenogenes 195 | 3250 | 1580 |
| Deinococcus radiodurans | 3270 | 2936 |
| Desulfotalea psychrophila LSv54 | 0313 | 3118 |
| Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough | 0313 | 3379 |
| Ehrlichia ruminantium str. Gardel | 0311 | 950 |
| Ehrlichia ruminantium str. Welgevonden | 0311 | 1846 |
| Enterococcus faecalis V583 | 3280 | 3265 |
| Erwinia carotovora subsp. atroseptica SCRI1043 | 0315 | 4472 |
| Escherichia coli CFT073 | 0315 | 5379 |
| Escherichia coli K12 | 0315 | 4242 |
| Escherichia coli O157:H7 | 0315 | 5361 |
| Escherichia coli O157:H7 EDL933 | 0315 | 5349 |
| Francisella tularensis subsp. tularensis SCHU S4 | 0315 | 1603 |
| Fusobacterium nucleatum subsp. nucleatum ATCC 25586 | 3290 | 2077 |
| Geobacillus kaustophilus HTA426 | 3280 | 3540 |
| Geobacter sulfurreducens PCA | 0313 | 3447 |
| Gloeobacter violaceus PCC 7421 | 3260 | 4430 |
| Gluconobacter oxydans 621H | 0311 | 2664 |
| Haemophilus ducreyi 35000HP | 0315 | 1717 |
| Haemophilus influenzae Rd KW20 | 0315 | 1709 |
| Haloarcula marismortui ATCC 43049 | 5110 | 3959 |
| Halobacterium sp. NRC-1 | 5110 | 2058 |
| Helicobacter hepaticus ATCC 51449 | 0314 | 1875 |
| Helicobacter pylori 26695 | 0314 | 1553 |
| Helicobacter pylori J99 | 0314 | 1491 |
| Idiomarina loihiensis L2TR | 0315 | 2628 |
| Lactobacillus acidophilus NCFM | 3280 | 1864 |
| Lactobacillus johnsonii NCC 533 | 3280 | 1772 |
| Lactobacillus plantarum WCFS1 | 3280 | 3059 |
| Lactococcus lactis subsp. lactis Il1403 | 3280 | 2266 |
| Legionella pneumophila str. Lens | 0315 | 2878 |
| Legionella pneumophila str. Paris | 0315 | 3027 |
| Legionella pneumophila subsp. pneumophila str. Philadelphia 1 | 0315 | 2942 |
| Leifsonia xyli subsp. xyli str. CTCB07 | 3200 | 2030 |
| Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 | 3320 | 3658 |
| Leptospira interrogans serovar lai str. 56601 | 3320 | 4725 |

| Species | Gruppen-Kürzel | Sequenzen |
|---|---|---|
| Listeria innocua | 3280 | 2968 |
| Listeria monocytogenes | 3280 | 2846 |
| Listeria monocytogenes str. 4b F2365 | 3280 | 2821 |
| Magnetococcus sp. MC-1 | 0310 | 3540 |
| Magnetospirillum magnetotacticum MS-1 | 0311 | 10108 |
| Mannheimia succiniciproducens MBEL55E | 0315 | 2384 |
| Mesoplasma florum L1 | 3280 | 683 |
| Mesorhizobium loti MAFF303099 | 0311 | 6752 |
| Methanocaldococcus jannaschii DSM 2661 | 5110 | 1770 |
| Methanococcus maripaludis S2 | 5110 | 1722 |
| Methanopyrus kandleri AV19 | 5110 | 1687 |
| Methanosarcina acetivorans str. C2A | 5110 | 4540 |
| Methanosarcina mazei Goe1 | 5110 | 3371 |
| Methanothermobacter thermautotrophicus str. Delta H | 5110 | 1869 |
| Methylococcus capsulatus str. Bath | 0315 | 2959 |
| Mycobacterium avium subsp. paratuberculosis str. k10 | 3200 | 4350 |
| Mycobacterium bovis AF2122/97 | 3200 | 3920 |
| Mycobacterium leprae | 3200 | 1605 |
| Mycobacterium tuberculosis CDC1551 | 3200 | 4189 |
| Mycobacterium tuberculosis H37Rv | 3200 | 3991 |
| Mycoplasma gallisepticum R | 3280 | 726 |
| Mycoplasma genitalium G-37 | 3280 | 480 |
| Mycoplasma hyopneumoniae 232 | 3280 | 691 |
| Mycoplasma mobile 163K | 3280 | 635 |
| Mycoplasma mycoides subsp. mycoides SC | 3280 | 1016 |
| Mycoplasma penetrans HF-2 | 3280 | 1037 |
| Mycoplasma pneumoniae M129 | 3280 | 688 |
| Mycoplasma pulmonis | 3280 | 772 |
| Nanoarchaeum equitans Kin4-M | 5120 | 536 |
| Neisseria gonorrhoeae FA 1090 | 0312 | 2002 |
| Neisseria meningitidis MC58 | 0312 | 2025 |
| Neisseria meningitidis Z2491 | 0312 | 2065 |
| Nitrosomonas europaea ATCC 19718 | 0312 | 2461 |
| Nocardia farcinica IFM 10152 | 3200 | 5683 |
| Nostoc sp. PCC 7120 | 3260 | 5368 |
| Novosphingobium aromaticivorans DSM 12444 | 0311 | 3771 |
| Oceanobacillus iheyensis HTE831 | 3280 | 3496 |
| Onion yellows phytoplasma OY-M | 3280 | 754 |
| Parachlamydia sp. UWE25 | 3230 | 2031 |
| Pasteurella multocida subsp. multocida str. Pm70 | 0315 | 2014 |
| Photobacterium profundum | 0315 | 5413 |
| Photorhabdus luminescens subsp. laumondii TTO1 | 0312 | 4683 |
| Picrophilus torridus DSM 9790 | 5110 | 1535 |
| Pirellula sp. | 3300 | 7325 |
| Porphyromonas gingivalis W83 | 3220 | 1909 |
| Prochlorococcus marinus str. MIT 9313 | 3260 | 2265 |
| Prochlorococcus marinus subsp. marinus str. CCMP1375 | 3260 | 1882 |
| Prochlorococcus marinus subsp. pastoris str. CCMP1986 | 3260 | 1712 |
| Propionibacterium acnes KPA171202 | 3200 | 2297 |
| Pseudomonas aeruginosa PAO1 | 0315 | 5566 |

| Species | Gruppen-Kürzel | Sequenzen |
|---|---|---|
| Pseudomonas putida KT2440 | 0315 | 5350 |
| Pseudomonas syringae pv. tomato str. DC3000 | 0315 | 5607 |
| Pyrobaculum aerophilum str. IM2 | 5100 | 2605 |
| Pyrococcus abyssi | 5110 | 1784 |
| Pyrococcus furiosus DSM 3638 | 5110 | 2065 |
| Pyrococcus horikoshii OT3 | 5110 | 2061 |
| Ralstonia solanacearum | 0312 | 3440 |
| Rhodobacter sphaeroides 2.4.1 | 0311 | 4103 |
| Rhodopseudomonas palustris CGA009 | 0311 | 4815 |
| Rhodospirillum rubrum | 0311 | 3791 |
| Rickettsia conorii str. Malish 7 | 0311 | 1374 |
| Rickettsia prowazekii | 0311 | 834 |
| Rickettsia typhi str. Wilmington | 0311 | 838 |
| Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67 | 0315 | 4359 |
| Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 | 0315 | 4093 |
| Salmonella enterica subsp. enterica serovar Typhi | 0315 | 4395 |
| Salmonella enterica subsp. enterica serovar Typhi Ty2 | 0315 | 4323 |
| Salmonella typhimurium LT2 | 0315 | 4368 |
| Shewanella oneidensis MR-1 | 0315 | 4630 |
| Shigella flexneri 2a str. 2457T | 0315 | 4068 |
| Shigella flexneri 2a str. 301 | 0315 | 4443 |
| Silicibacter pomeroyi DSS-3 | 0311 | 4252 |
| Sinorhizobium meliloti | 0311 | 3341 |
| Staphylococcus aureus subsp. aureus COL | 3280 | 2618 |
| Staphylococcus aureus subsp. aureus MRSA252 | 3280 | 2656 |
| Staphylococcus aureus subsp. aureus MSSA476 | 3280 | 2598 |
| Staphylococcus aureus subsp. aureus Mu50 | 3280 | 2699 |
| Staphylococcus aureus subsp. aureus MW2 | 3280 | 2632 |
| Staphylococcus aureus subsp. aureus N315 | 3280 | 2593 |
| Staphylococcus epidermidis ATCC 12228 | 3280 | 2485 |
| Staphylococcus epidermidis RP62A | 3280 | 2487 |
| Streptococcus agalactiae 2603V/R | 3280 | 2124 |
| Streptococcus agalactiae NEM316 | 3280 | 2094 |
| Streptococcus mutans UA159 | 3280 | 1960 |
| Streptococcus pneumoniae R6 | 3280 | 2043 |
| Streptococcus pneumoniae TIGR4 | 3280 | 2094 |
| Streptococcus pyogenes M1 GAS | 3280 | 1696 |
| Streptococcus pyogenes MGAS10394 | 3280 | 1886 |
| Streptococcus pyogenes MGAS315 | 3280 | 1562 |
| Streptococcus pyogenes MGAS8232 | 3280 | 1845 |
| Streptococcus pyogenes SSI-1 | 3280 | 1861 |
| Streptococcus thermophilus CNRZ1066 | 3280 | 1915 |
| Streptococcus thermophilus LMG 18311 | 3280 | 1889 |
| Streptomyces avermitilis MA-4680 | 3200 | 7577 |
| Streptomyces coelicolor A3(2) | 3200 | 7769 |
| Sulfolobus solfataricus P2 | 5100 | 3001 |
| Sulfolobus tokodaii str. 7 | 5100 | 2826 |
| Symbiobacterium thermophilum IAM 14863 | 3200 | 3337 |
| Synechococcus elongatus PCC 6301 | 3260 | 2525 |
| Synechococcus sp. WH 8102 | 3260 | 2517 |

| Species | Gruppen-Kürzel | Sequenzen |
|---|---|---|
| Synechocystis sp. | 3260 | 3169 |
| Synechocystis sp. PCC 6803 | 3260 | 3167 |
| Thermoanaerobacter tengcongensis MB4 | 3280 | 2588 |
| Thermococcus kodakaraensis KOD1 | 5110 | 2306 |
| Thermoplasma acidophilum | 5110 | 1478 |
| Thermoplasma volcanium GSS1 | 5110 | 1526 |
| Thermosynechococcus elongatus BP-1 | 3260 | 2475 |
| Thermotoga maritima MSB8 | 3330 | 1846 |
| Thermus thermophilus HB27 | 3270 | 1982 |
| Thermus thermophilus HB8 | 3270 | 2238 |
| Treponema denticola ATCC 35405 | 3320 | 2767 |
| Treponema pallidum subsp. pallidum str. Nichols | 3320 | 1031 |
| Tropheryma whipplei str. Twist | 3200 | 808 |
| Tropheryma whipplei TW08/27 | 3200 | 783 |
| Ureaplasma parvum serovar 3 str. ATCC 700970 | 3280 | 611 |
| Vibrio cholerae O1 biovar eltor str. N16961 | 0315 | 3828 |
| Vibrio fischeri ES114 | 0315 | 3802 |
| Vibrio parahaemolyticus RIMD 2210633 | 0315 | 4832 |
| Vibrio vulnificus CMCP6 | 0315 | 4537 |
| Vibrio vulnificus YJ016 | 0315 | 5028 |
| Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis | 0315 | 611 |
| Wolbachia endosymbiont of Drosophila melanogaster | 0311 | 1195 |
| Wolbachia endosymbiont strain TRS of Brugia malayi | 0311 | 805 |
| Wolinella succinogenes | 0314 | 2044 |
| Xanthomonas axonopodis pv. citri str. 306 | 0315 | 4427 |
| Xanthomonas campestris pv. campestris str. ATCC 33913 | 0315 | 4181 |
| Xanthomonas oryzae pv. oryzae KACC10331 | 0315 | 4637 |
| Xylella fastidiosa 9a5c | 0315 | 2766 |
| Xylella fastidiosa Temecula1 | 0315 | 2034 |
| Yersinia pestis biovar Medievalis str. 91001 | 0315 | 3895 |
| Yersinia pestis CO92 | 0315 | 3885 |
| Yersinia pestis KIM | 0315 | 4090 |
| Yersinia pseudotuberculosis IP 32953 | 0315 | 3901 |
| Zymomonas mobilis subsp. mobilis ZM4 | 0311 | 1998 |

**Taxonomische Gruppen** Die Perlskripte verwenden für die Einordnung der verschiedenen taxonomischen Gruppen Taxonomiedateien. In diesen wird jeder Art ein Gruppenkürzel zugeordnet (beispielsweise 3280 für die Vertreter der Firmicutes). Zusätzliche Kürzel dienen für die Zählung von Sequenzen mit bestimmten Eigenschaften. Diese internen Kürzel sind in der folgenden Tabelle kursiv aufgeführt.

| Gruppen-Kürzel | taxonomische Gruppe |
|---|---|
| 0310 | Proteobacteria |
| 0311 | Alphaproteobacteria |
| 0312 | Betaproteobacteria |
| 0313 | Deltaproteobacteria |
| 0314 | Epsilonproteobacteria |
| 0315 | Gammaproteobacteria |
| 0316 | Magnetotactic Cocci |
| 3200 | Actinobacteria |
| 3210 | Aquificae |
| 3220 | Bacteroidetes |
| 3230 | Chlamydiae |
| 3240 | Chlorobi |
| 3250 | Chloroflexi |
| 3260 | Cyanobacteria |
| 3270 | Deinococcus-Thermus |

| Gruppen-Kürzel | taxonomische Gruppe |
|---|---|
| 3280 | Firmicutes |
| 3290 | Fusobacteria |
| 3300 | Planctomycetes |
| 3320 | Spirochaetes |
| 3330 | Thermotogae |
| 5100 | Crenarchaeota |
| 5110 | Euryarchaeota |
| 5120 | Nanoarchaeota |
| 8000 | Eukaryota |
| 9000 | unidentified |
| *0001* | Genduplikation |
| *9991* | Keine Treffer |
| *9992* | Zu Kurz |
| *9993* | Fehler |
| *9999* | Proteine (insgesamt) |

# Stämme von *Escherichia coli*

Die Namen der Stämme von *E. coli* wurden im Text aufgrund der besseren Lesbarkeit mit abgekürzten Stammbezeichnungen benannt. In der folgenden Tabelle sind diese Bezeichnungen für die beiden *E. coli* Datensätze aufgeführt.

| Genom | NC | Bezeichnung in den Datensätzen | |
|---|---|---|---|
| | | 9 Genome | 14 Genome |
| *Escherichia coli* 536 | NC_008253 | 536 | 536 |
| *Escherichia coli* APEC | NC_008563 | APEC | APEC |
| *Escherichia coli* ATCC | NC_010468 | | ATCC |
| *Escherichia coli* CFT073 | NC_004431 | CFT073 | CFT073 |
| *Escherichia coli* E24377A | NC_009801 | E24377A | E24377A |
| *Escherichia coli* HS | NC_009800 | HS | HS |
| *Escherichia coli* K12 DH10B | NC_010473 | | K12_DH10B |
| *Escherichia coli* K12 MG1655 | NC_000913 | K12 | K12_MG1655 |
| *Escherichia coli* O157 EC4115 | NC_011353 | | O157_EC4115 |
| *Escherichia coli* O157 EDL933 | NC_002655 | EDL933 | O157_EDL933 |
| *Escherichia coli* O157 Sakai | NC_002695 | Sakai | O157_Sakai |
| *Escherichia coli* SMS | NC_010498 | | SMS |
| *Escherichia coli* UTI89 | NC_007946 | UTI89 | UTI89 |
| *Escherichia coli* W3110 | AC_000091 | | W3110 |

# Literaturverzeichnis


**Adams III, E.** Consensus techniques and the comparison of taxonomic trees. *Syst Zool*, 1972. **21**(4):390–397.

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

**Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, A. K., Eriksson, A. S., Winkler, H. H. und Kurland, C. G.** The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 1998. **396**:133–140.

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

**Atkinson, H. J., Morris, J. H., Ferrin, T. E. und Babbitt, P. C.** Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, 2009. **4**.

**Azad, R. K. und Lawrence, J. G.** Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput Biol*, 2005. **1**:461–473.

**Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F.-J., Dupré, J., Dagan, T., Boucher, Y. und Martin, W.** Prokaryotic evolution and the tree of life are two different things. *Biol Direct*, 2009. **4**:34–53.

**Baughn, A. D. und Malamy, M. H.** A mitochondrial-like aconitase in the bacterium bacteroides fragilis: implications for the evolution of the mitochondrial krebs cycle. *Proc Natl Acad Sci USA*, 2002. **99**:4662–4667.

**Beiko, R. G., Harlow, T. J. und Ragan, M. A.** Highways of gene sharing in prokaryotes. *PNAS*, 2005. **102**:14332–14337.

**Blattner, F. R., Plunkett, G. r., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. und Shao, Y.** The complete genome sequence of *Escherichia coli* K-12. *Science*, 1997. **277**:1453–1474.

**Boussau, B., Karlberg, E., Frank, A., Legault, B. und Andersson, S.** Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc Natl Acad Sci USA*, 2004. **101**:9722–9727.

**Brisse, S., Fevre, C., Passet, V., Issenhuth-Jeanjean, S., Tournebize, R., Diancourt, L. und Grimont, P.** Virulent clones of *Klebsiella pneumoniae*: identification and evolutionary scenario based on genomic and phenotypic characterization. *PLoS One*, 2009. **4**:e4982.

**Brown, J. R.** Ancient horizontal gene transfer. *Nat Rev Genet*, 2003. **4**:121–132.

**Bryant, D. und Moulton, V.** Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*, 2004. **21**:255–265.

**Byrne, K. P. und Wolfe, K. H.** The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, 2005. **15**:1456–1461.

**Cavalier-Smith, T.** The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*, 2002. **52**:297–354.

**Cavalier-Smith, T.** Predation and eukaryote cell origins: a coevolutionary perspective. *Int J Biochem Cell Biol*, 2009. **41**:307–322.

**Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T. und Pfanner, N.** Importing mitochondrial proteins: machineries and mechanisms. *Cell*, 2009. **138**:628–644.

**Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. und Bork, P.** Toward automatic reconstruction of a highly resolved tree of life. *Science*, 2006. **311**:1283–1287.

**Dagan, T. und Martin, W.** The tree of one percent. *Genome Biol*, 2006. **7**:118.

**Dagan, T. und Martin, W.** Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA*, 2007. **104**:870–875.

**Dagan, T. und Martin, W.** Getting a better picture of microbial evolution en route to a network of genomes. *Phil Trans Roy Soc Lond B*, 2009. **364**:2187–2196.

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

**Darwin, C.** On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *New York: D. Appleton*, 1859.

**Dayhoff, M., Schwartz, R. und Orcutt, B.** A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 1978. **5**:345–352.

**Delcher, A. L., Salzberg, S. L. und Phillippy, A. M.** Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics*, 2003. **Chapter 10**.

**Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W. und Dagan, T.** Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol*, 2008. **25**:748–761.

**Dolezal, P., Likic, V., Tachezy, J. und Lithgow, T.** Evolution of the molecular machines for protein import into mitochondria. *Science*, 2006. **313**:314–318.

**van Dongen, S.** *Graph clustering by flow simulation*. Dissertation, Universiteit Utrecht, 2000.

**Doolittle, W. F.** A paradigm gets shifty. *Nature*, 1998. **392**:15–16.

**Doolittle, W. F.** Phylogenetic classification and the universal tree. *Science*, 1999. **284**:2124–2128.

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

**Emelyanov, V. V.** Common evolutionary origin of mitochondrial and rickettsial respiratory chains. *Arch Biochem Biophys*, 2003. **420**:130–141.

**Enright, A. J., Van Dongen, S. und Ouzounis, C. A.** An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 2002. **30**:1575–1584.

**Esser, C. und Martin, W.** Supertrees and symbiosis in eukaryote genome evolution. *Trends Microbiol*, 2007. **15**:435–437.

**Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant,**

**D., Steel, M., Lockhart, P., Penny, D. und Martin, W.** A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*, 2004. **21**:1643–1660.

**Felsenstein, J.** Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, 1973. **25**:471–492.

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

**Felsenstein, J.** PHYLIP (Phylogeny Inference Package) version 3.6. Technischer Bericht, Department of Genome Sciences, University of Washington, Seattle, 2004.

**Frost, L. S., Leplae, R., Summers, A. O. und Toussaint, A.** Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*, 2005. **3**:722–732.

**Gabaldón, T. und Huynen, M. A.** Reconstruction of the proto-mitochondrial metabolism. *Science*, 2003. **301**:609–609.

**van der Giezen, M. und Tovar, J.** Degenerate mitochondria. *EMBO Rep*, 2005. **6**:525–530.

**Gogarten, J. P., Fournier, G. und Zhaxybayeva, O.** Gene transfer and the reconstruction of life's early history from genomic data. *Space Sci Rev*, 2008. **135**:115–131.

**Golubchik, T., Wise, M. J., Easteal, S. und Jermiin, L. S.** Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol*, 2007. **24**:2433–2442.

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

**Gray, M. W., Burger, G. und Lang, B. F.** Mitochondrial evolution. *Science*, 1999. **283**:1476–1481.

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

**Guttman, D. S. und Dykhuizen, D. E.** Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, 1994. **266**:1380–1383.

**Huang, J. und Gogarten, J. P.** Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet*, 2006. **22**:361–366.

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

**Ichioka, K., Yoshimura, K., Honda, T., Takahashi, A. und Terai, A.** Paracentric inversion of chromosome 7(q22-31) associated with nonobstructive azoospermia. *Fertil Steril*, 2005. **83**:455–456.

**John, P. und Whatley, F.** *Paracoccus denitrificans*: a present-day bacterium resembling the hypothetical free-living ancestor of the mitochondrion. In *Sym Soc Exp Biol*, 29. 39–40.

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

**Kimura, M.** Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research*, 1968. **11**:247–269.

**Kimura, M.** Recent development of the neutral theory viewed from the wrightian tradition of theoretical population genetics. *Proc Natl Acad Sci U S A*, 1991. **88**:5969–5973.

**King, J. L. und Jukes, T. H.** Non-darwininan evolution. *Science*, 1969. **164**:788–798.

**Kleine, T., Maier, U. G. und Leister, D.** DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol*, 2009. **60**:115–138.

**Koski, L. und Golding, G.** The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 2001. **52**:540–542.

**Koski, L. B., Morton, R. A. und Golding, G. B.** Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol*, 2001. **18**:404–412.

**Kunin, V. und Ouzounis, C. A.** The balance of driving forces during genome evolution in prokaryotes. *Genome Res*, 2003*a*. **13**:1589–1594.

**Kunin, V. und Ouzounis, C. A.** Gene trace-reconstruction of gene content of ancestral species. *Bioinformatics*, 2003*b*. **19**:1412–1416.

**Kunin, V., Goldovsky, L., Darzentas, N. und Ouzounis, C.** The net of life: reconstructing the microbial phylogenetic network. *Genome Res*, 2005. **15**:954–959.

**Kurland, C. G. und Andersson, S. G.** Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev*, 2000. **64**:786–820.

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

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

**Lang, B. F., Gray, M. W. und Burger, G.** Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet*, 1999. **33**:351–397.

**Lathe, W. C., 3rd, Snel, B. und Bork, P.** Gene context conservation of a higher order than operons. *Trends Biochem Sci*, 2000. **25**:474–479.

**Lawrence, J. G. und Ochman, H.** Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol*, 1997. **44**:383–397.

**Lawrence, J. G. und Ochman, H.** Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*, 1998. **95**:9413–9417.

**Lerat, E., Daubin, V., Ochman, H. und Moran, N.** Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*, 2005. **3**:e130.

**Lin, C. H., Bourque, G. und Tan, P.** A comparative synteny map of *Burkholderia* species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol Biol Evol*, 2008. **25**:549–558.

**Maddison, D. R., Swofford, D. L. und Maddison, W. P.** NEXUS: an extensible file format for systematic information. *Syst Biol*, 1997. **46**:590–621.

**Margulis, I.** Symbiosis in cell evolution. san francisco, ca: W. h, 1981.

**Margulis, L., Dolan, M. F. und Guerrero, R.** The chimeric eukaryote: Origin of the nucleus from the karyomastigont in amitochondriate protists. *PNAS*, 2000. **97**:6954–6959.

**Martin, W.** Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*, 1999. **21**:99–104.

**Martin, W.** Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci USA*, 2003. **100**:8612–8614.

**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**:1521–1539.

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

**Mayr, E.** Two empires or three? *PNAS*, 1998. **95**:9720–9723.

**McInerney, J. O., Cotton, J. A. und Pisani, D.** The prokaryotic tree of life: past, present... and future? *Trends Ecol Evol*, 2008. **23**:276–281.

**Milkman, R. und Bridges, M. M.** Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics*, 1990. **126**:505–517.

**Milkman, R. und Stoltzfus, A.** Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics*, 1988. **120**:359–366.

**Mirkin, B. G., Fenner, T. I., Galperin, M. Y. und Koonin, E. V.** Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 2003. **3**:2.

**Mokranjac, D. und Neupert, W.** Thirty years of protein translocation into mitochondria: unexpectedly complex and still puzzling. *Biochim Biophys Acta*, 2009. **1793**:33–41.

**Nakamura, Y., Itoh, T., Matsuda, H. und Gojobori, T.** Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*, 2004. **36**:760–766.

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

**Neupert, W. und Herrmann, J. M.** Translocation of proteins into mitochondria. *Annu Rev Biochem*, 2007. **76**:723–749.

**Ochman, H., Lawrence, J. G. und Groisman, E. A.** Lateral gene transfer and the nature of bacterial innovation. *Nature*, 2000. **405**:299–304.

**Page, R. D.** Introduction to comparing large sequence sets. *Curr Protoc Bioinformatics*, 2003. **Chapter 10**.

**van Passel, M. W. J., Marri, P. R. und Ochman, H.** The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLoS Comput Biol*, 2008. **4**:e1000059.

**Pisani, D., Cotton, J. A. und McInerney, J. O.** Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*, 2007. **24**:1752–1760.

**Reyes-Prieto, A., Weber, A. P. und Bhattacharya, D.** The origin and establishment of the plastid in algae and plants. *Annu Rev Genet*, 2007. **41**:147–168.

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

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

**Rodelsperger, C. und Dieterich, C.** Syntenator: Multiple gene order alignments with a gene-specific scoring function. *Algorithms Mol Biol*, 2008. **3**:14.

**Roettger, M., Martin, W. und Dagan, T.** A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Mol Biol Evol*, 2009. **26**:1931–1939.

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

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

**Sidow, A., Nguyen, T. und Speed, T.** Estimating the fraction of invariable codons with a capture-recapture method. *J Mol Evol*, 1992. **35**:253–260.

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

**Snel, B., Bork, P. und Huynen, M. A.** Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, 2002. **12**:17–25.

**Stackebrandt, E., Murray, R. G. E. und Trüper, H. G.** *Proteobacteria* classis nov. a name for the phylogenetic taxon that includes the “purple bacteria and their relatives“. *Int J Syst Bacteriol*, 1988. **38**:321–325.

**Steel, M., Huson, D. und Lockhart, P.** Invariable sites models and their use in phylogeny reconstruction. *Syst Biol*, 2000. **49**:225–232.

**Susko, E., Leigh, J., Doolittle, W. F. und Bapteste, E.** Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol*, 2006. **23**:1019–1030.

**Szpirer, C., Szpirer, J., Van Vooren, P., Tissir, F., Simon, J. S., Koike, G., Jacob, H. J., Lander, E. S., Helou, K., Klinga-Levan, K. und Levan, G.** Gene-based anchoring of the rat genetic linkage and cytogenetic mans: new regional localizations, orientation of the linkage groups, and insights into mammalian chromosome evolution. *Mammalian Genome*, 1998. **9**:721–734.

**Tamames, J.** Evolution of gene order conservation in prokaryotes. *Genome Biol*, 2001:RESEARCH0020.

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

**Thollesson, M.** LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. *Bioinformatics*, 2004. **20**:416–418.

**Thomas, C. M. und Nielsen, K. M.** Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol*, 2005. **3**:711–721.

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

**Thompson, J., Plewniak, F. und Poch, O.** BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 1999. **15**: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**:123–135.

**Welch, R. A., Burland, V., Plunkett, G. r., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S. und Blattner, F. R.** Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. *Proc Natl Acad Sci USA*, 2002. **99**:17020–17024.

**Woese, C., Kandler, O. und Wheelis, M.** Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, 1990. **87**:4576–4579.

**Wu, M., Sun, L. V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J. C., McGraw, E. A., Martin, W., Esser, C., Ahmadinejad, N., Wiegand, C., Madupu, R., Beanan, M. J., Brinkac, L. M., Daugherty, S. C., Durkin, A. S., Kolonay, J. F., Nelson, W. C., Mohamoud, Y., Lee, P., Berry, K., Young, M. B., Utterback, T., Weidman, J., Nierman, W. C., Paulsen, I. T., Nelson, K. E., Tettelin, H., O'Neill, S. L., Eisen, J. A., Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant, D., Steel, M. A., Lockhart, P. J., Penny, D. und Martin, W.** Phylogenomics of the reproductive parasite *Wolbachia pipientis* wmel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol*, 2004. **2**:E69.

**Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. und Woese, C. R.** Mitochondrial origins. *Proc Natl Acad Sci USA*, 1985. **82**:4443–4447.

**Yoon, H. S., Hackett, J. D., Van Dolah, F. M., Nosenko, T., Lidie, K. L. und Bhattacharya, D.** Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Mol Biol Evol*, 2005. **22**:1299–1308.

**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**:1099–1108.

# Danksagung

Ich danke Prof. Dr William Martin für das interessante Thema und für die Möglichkeit es in seinem Institut zu bearbeiten. Desweiteren für die Gelegenheit meine Ergebnisse auf internationalen Tagungen zu präsentieren und für viele wertvolle Tipps und Ratschläge, insbesondere bei der Fertigstellung dieser Arbeit.

Herrn Prof. Dr. Martin Lercher möchte ich für die Bereitschaft danken, das Zweitgutachten auch unter erschwerten Bedingungen zu übernehmen, und für das Interesse, das er meiner Arbeit entgegengebracht hat.

Frau Dr. Tal Dagan danke ich für die Betreuung während der vergangenen Jahre, die Diskussionsbereitschaft und die Einführung in die Geheimnisse von `MATLAB` und `mySQL`.

Allen gegenwärtigen und vergangenen Mitgliedern der Botanik III danke ich für das angenehme Betriebsklima in den vergangenen Jahren, nicht nur während der Kernarbeitszeit. Den wechselnden Bürokollegen (wenn ich richtig gezählt habe waren es acht) möchte ich für die Atmosphäre danken. Unseren Bioinformatikern danke ich für stetige Diskussionsbereitschaft und die Lösung größerer und kleinerer Probleme.

In alphabetischer Reihenfolge danke ich Oliver Deusch, Nicole Grünheit, Kathrin Hoffmann, Britta Pinzger und Verena Zimorski für den unermüdlichen Einsatz bei dem Versuch diese Arbeit von Rechtschreibfehlern zu befreien, sie mit einer angemessenen Anzahl von Kommas auszustatten und außerdem für viele Anregungen, die mir sehr weitergeholfen haben. Gabriel Gelius-Dietrich möchte ich für die Beantwortung aller Fragen zu LaTeXund Perl danken.

Meinen Eltern schließlich möchte ich einfach dafür danken, daß sie da sind.

# Erklärung

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

.....

(Christian Eßer)