Änderungen in Lebensstil und Nährstoffangebot als treibende Kräfte der Genomevolution in Escherichia coli und Salmonella

Inaugural - Dissertation

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

vorgelegt von

Thomas Laubach
aus Düsseldorf

Düsseldorf, Juli 2014

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

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

Referent: Univ. Prof. Dr. Martin J. Lercher
Korreferent: Univ. Prof. Dr. William Martin

Tag der mündlichen Prüfung: 11. Dezember 2014

Inhaltsverzeichnis

Publikationen

Kohl, J., I. Paulsen, T. Laubach, A. Radke, A. von Haeseler. **HvrBase++: a phylogenetic database for primate species.** *Nucleic Acids Res.* 34 (2006):D700-4.

Laubach, T. und A. von Haeseler. **TreeSnatcher: Coding Trees from Images.** *Bioinformatics* 23.24 (2007):3384-385.

Laubach, T., A. von Haeseler und M. J. Lercher. **TreeSnatcher Plus: Capturing Phylogenetic Trees from Images.** *BMC Bioinformatics* 13.1 (2012):110.

1 Zusammenfassung

Die Erhellung der Ursachen der Pathogenität vieler Stämme von *Escherichia coli* und aller *Salmonellen*, die nahe Verwandte von *E. coli* sind, ist von großer Bedeutung für die medizinische Forschung. Bakterien evolvieren fortwährend, was die Behandlung der von ihnen hervorgerufenen Krankheitsbilder erschwert. Sie tauschen Resistenzen gegen Antibiotika über lateralen Gentransfer (LGT) aus, der in der Bakterienevolution eine große Rolle spielt, die Ableitung von Phylogenien aber erschwert. Können wir verstehen, was die Evolution von *E. coli* durch LGT antreibt? Wie wird LGT von der Lebensumgebung der Bakterien beeinflußt?

Für Anschlußuntersuchungen sind zunächst statistisch robuste Phylogenien notwendig, da auf ihnen später die anzestralen Verteilungen von Gen-Anwesenheit und -abwesenheit mittels Maximum Parsimony geschätzt werden. Mittels *MMN*, einem für diese Arbeit entwickelten Perl-Programm, wurden zunächst Maximum-Likelihood-Phylogenien auf Grundlage konkatenierter universeller Genfamilien berechnet. Die *E. coli*-Phylogenie enthält 61 der 63 vollständig sequenzierten Genome, die *Salmonellen*-Phylogenie alle 25 (Stand 9/2013). Sie wurden mit Phylogenien verglichen, die mit zwei anderen Methoden inferiert worden sind, welche darauf abzielen, den Effekt von LGT auf die Ableitung der Phylogenie zu vermindern. Eine Methode („iTol“) basiert auf ausgewählten COGs, die andere („Maslov“) benutzt als Distanzmaß die Zahl rekombinanter Regionen zwischen zwei Genomen. Beide lieferten – entgegen der Erwartung - statistisch schlecht unterstützte Phylogenien, die sich deutlich von den ML-Phylogenien unterscheiden.

Zwei Phylogenien, die subjektiv als ähnlich empfunden werden, werden von den populären Distanzverfahren, insbesondere der *Split-Distanz*, nicht als ähnlich ausgewiesen. Das Distanzmaß *ssRF* wurde konzipiert, das diesen Nachteil beseitigt und nur als signifikant empfundene Splits bei der Distanzberechnung berücksichtigt. Es wurde für die Phylogenievergleiche in dieser Arbeit verwendet.

Für eine der untersuchten Phylogenien war keine maschinenlesbare Datenquelle mehr vorhanden. Für einen solchen Fall wurde vom Verfasser das Java-Programm *TreeSnatcher Plus* entwickelt. Damit wurde die Abbildung der Phylogenie digitalisiert und in einen Newick-Ausdruck umgewandelt.

Um sich den eingangs gestellten Fragen zu nähern, wurde die Verteilung der anzestralen Gengehalte der Verteilung der rekonstruierten essentiellen Nährstoffe auf den Phylogenien von *E. coli* und *Salmonellen* gegenübergestellt. Die These war, daß LGT auf dem Stammbaum nicht mit gleichmäßiger Rate geschehen ist, sondern daß die betrachteten Bakterien immer dann die Fähigkeit erworben (verloren) haben, bestimmte essentielle Nährstoffe aus der Umwelt metabolisch zu nutzen, wenn sie gleichzeitig Gene hinzugewonnen (abgegeben) haben. Das für diese Untersuchung konzipierte Verfahren lieferte mathematisch assoziierte Gen/Nährstoff-Paarungen, die auch biologisch sinnvoll sind. Neben untersuchenswerten Paaren mit hypothetischen Proteinen befanden sich in der Ergebnisliste insbesondere auch Paare aus Nährstoff und dem Gen für den entsprechenden Transporter - eine Information, die so nicht explizit in den Eingangsdaten vorhanden gewesen ist. Das Verfahren lieferte also eine Mehrzahl von Ergebnissen, die die o. g. These unterstützen.

Die Evolution von Bakterien wird möglicherweise auch von Änderungen in ihrem Lebensstil mitgestaltet. So ist vorstellbar, daß sich apathogene E. coli an einen Lebensstil als Pathogene anpaßten, indem sie verstärkt Gene aufnahmen oder abgaben. In einer zweiten Analyse wurde diese Hypothese untersucht. Die Ergebnisse deuten an, daß Lebensstil-Änderungen bei *E. coli* tatsächlich von einer erhöhten Rate von Gen-Änderungen begleitet waren.

2 Abstract

The elucidation of the causes of the pathogenicity of many strains of *Escherichia coli* and of all *Salmonella* who are close relatives of *E. coli*, is of great importance for medical research. Bacteria evolve continually, which complicates the treatment of the diseases caused by them. They exchange resistances against antibiotics by means of lateral gene transfer (LGT), which plays a major role in bacterial evolution. However, it makes the derivation of phylogenies difficult. Can we understand what drives the evolution of E. coli by LGT? How is LGT influenced by the living environment of the bacteria?

First of all, statistically robust phylogenies are required as on them ancestral gene presence/absence patterns will be estimated using Maximum Parsimony. By means of *MMN*, a Perl program developed for this thesis, Maximum Likelihood phylogenies based on concatenated universal gene families were inferred. The *E. coli* phylogeny contains 61 of the 63 fully sequenced genomes, the *Salmonella* phylogeny all 25 (as of 9/2013). They were compared with phylogenies that have been inferred with two other methods, which aim to reduce the effect of LGT on phylogeny inferences. The first method ("iTol") is based on selected COGs, the other ("Maslov") uses as distance measure the number of recombinant regions between two genomes. Both delivered - contrary to expectation - statistically insufficiently supported phylogenies that differ significantly from the phylogenies obtained with Maximum Likelihood.

Two lineage trees that are subjectively perceived as similar, are not recognized by the popular distance methods, in particular the split distance, as similar. The distance measure *ssRF* has been designed which eliminates this drawback. It only considers as significant perceived splits in its distance calculation. In this work, it was used for the comparisons of the phylogenies.

For one of the studied phylogenies no machine-readable data source was present. For such a case, the author had developed the Java program *TreeSnatcher Plus*. It was used to digitize the phylogeny that was contained in a picture and to convert it into a Newick expression.

To approach the questions asked at the outset, the distribution of ancestral gene content was compared to the distribution of the reconstructed essential nutrients on the phylogenies of *E. coli* and *Salmonella*. The hypothesis was that LGT has not occurred on the family tree at an even rate, but that the bacteria always acquired (lost) the ability to use certain essential nutrients from the environment metabolically whenever they simultaneously gained (lost) genes. The procedure that were designed for this study yielded mathematically associated gene/nutrient pairings that are biologically meaningful. In addition to pairs with hypothetical proteins worth investigating there were also pairs of nutrient and the gene for the corresponding transporter - information that has not explicitly been present in the input data. Thus, the method provided a plurality of results that support the above hypothesis.

The evolution of bacteria might be shaped also by changes in their lifestyle. Thus, it is conceivable that non-pathogenic *E. coli* adapted to a lifestyle as pathogens by accepting or donating genes with an increased rate. In a second analysis of this hypothesis has been investigated. The results indicate that lifestyle changes in *E. coli* were indeed accompanied by an increased rate of gene changes.

3 Einführung

3.1 Escherichia coli und Salmonellen sind nach wie vor eine globale Bedrohung für den Menschen

Wie ist das Bakterium *Escherichia coli* zu dem geworden, was es ist - ein unverzichtbarer Bestandteil der menschlichen Darmflora, ein wichtiges Objekt der Biotechnologie und ein Modellorganismus der Mikrobiologie einerseits, und ein gefürchtetes Pathogen von Warmblütern, verantwortlich für mehr als 160 Millionen Durchfallerkrankungen und einer Million Todesfälle pro Jahr (Hahn et al., 2008) andererseits?

Mit dieser ambivalenten Fragestellung haben sich schon einige Generationen von Wissenschaftlern beschäftigt, seit der Kinderarzt Theodor Escherich das „Kolibakterium“ im Jahre 1885 erstmalig beschrieben hat (Escherich, 1885). Die Ursachen der Pathogenität von *Escherichia coli* (kurz „*E. coli*“) liegen aber nach wie vor im Dunkeln. Die Erhellung der Ursachen besitzt nicht nur eine abstrakte, lediglich akademische Komponente, sondern ist für die medizinische Forschung von entscheidendem Interesse. Die Erwartung ist nämlich, daß sich die über *E. coli* gewonnenen Erkenntnisse auch auf andere Bakterien übertragen lassen, und daß man später in der Lage sein wird, durch Bakterien hervorgerufene Epidemien vorherzusagen und zu verhindern. Traditionell werden die als *Shigellen* bezeichneten Bakterien als Stämme von *E. coli* gewertet (Pupo et al., 2000): sagt man *E. coli*, meint man in der Regel auch *Shigellen*. Neuerdings gibt es aber wieder Kritik an dieser Einordnung (Zuo et al., 2013).

Die mit *E. coli* nahe verwandten Bakterien der Gattung *Salmonella* wurden im Jahre 1900 von Joseph Lignieres nach dem US-amerikanischen Tierarzt Daniel Elmer Salmon benannt. Die *Salmonellen* gehören ebenso wie *E. coli/Shigellen* zu der Familie der Enterobacteriaceae und sind, ebenso wie viele Stämme von *E. coli*, Pathogene von Säugetieren. Beide Organismen kommen weltweit innerhalb und außerhalb von Mensch und Tier vor. *Salmonellen* lösen beim Menschen die meldepflichtige Salmonellose aus, die sich durch meist spontan ausheilende Durchfallerkrankungen äußert. Für sehr junge und alte Menschen sowie immungeschwächte Patienten können Salmonellosen allerdings sehr gefährlich werden. Im Jahr 2011 trat ein bis dahin unbekannter enterohämorrhagischer *E. coli*-Stamm erstmals in Erscheinung, als er vor allem in Norddeutschland eine Epidemie auslöste, in deren Verlauf 53 Menschen starben. Fast 3000 Personen erkrankten an Gastroenteritis und 855 am hämolytisch-urämischen Syndrom (HUS). Nachdenklich macht dabei die Tatsache, daß serologische Untersuchungen den verantwortlichen Stamm zwar als O104:H4 identifizierten, man jedoch nach seiner Sequenzierung zu dem Ergebnis kam, daß er eine Mischform aus EHEC (enterohämorrhagische *E. coli*) und EAEC (enteroaggregative *E. coli*) darstellt, der bis auf weiteres als HUSEC („HUS-assoziierte *E. coli*“) bzw. STEC („Shiga-Toxin produzierende *E. coli*“) bezeichnet wird.

Wie bereits angedeutet, entwickeln Bakterien sich stetig weiter und erwerben dabei Resistenzen, die sie mitunter gegen Antibiotika immun machen. Das ist ein ernstes Problem im Zusammenhang mit der Behandlung bakterieller Infektionen, mit dem die Medizin sich zunehmend konfrontiert sieht. Schon heute (2014) sterben Menschen an den Folgen eigentlich heilbarer Bakterieninfektionen, weil sie auf kein Antibiotikum mehr ansprechen.

3.2 Die Vererbung bei Prokaryoten geschieht nicht nur vertikal

Da Bakterien sich durch Teilung vermehren, kann ihre Genealogie prinzipiell als Baum dargestellt werden. Nach dieser Metapher werden Gene von einem Bakterium entlang der Ahnenreihe an seine zwei Nachkommen weiter vererbt. Der Weg jedes Gens läßt sich dann leicht verfolgen. Bakterien tauschen genetisches Material sehr wahrscheinlich aber auch untereinander während ihres Lebens aus, und das sogar über Artgrenzen hinweg. Man glaubt, daß dieser als „lateraler“ oder „horizontaler Gentransfer“ („LGT/HGT“) bezeichnete Prozeß für die Vermittlung von Antibiotikaresistenzen zwischen verschiedenen Bakterien verantwortlich ist und ein nahezu gewöhnlicher Bestandteil der Prokaryotenevolution ist.

Es gibt diverse Ansätze, den Einfluß von HGT auf Genome zu modellieren und seinen Effekt auf Stammbaumrekonstruktionen vorherzusagen (Pal et al., 2005; Abby et al., 2011; Schönknecht et al., 2013). Die molekularen Ursachen sind aber noch nicht vollständig verstanden, einige Überlegungen dazu werden in Kapitel 9 vorgestellt. Es sind mehrere mutmaßliche Mechanismen bekannt, durch die genetisches Material zwischen – nicht notwendigerweise nah verwandten - Prokaryoten ausgetauscht wird. Hauptsächlich sind dies Konjugation (Thomas und Nielsen, 2005; Norman et al., 2009), Transduktion (Frost et al., 2005) und Transformation. Nicht nur chromosomale Gene sind von LGT betroffen, auch Gene auf Plasmiden. Plasmide gelten als Träger von Antibiotika-Resistenzen und werden vererbt.

Akzeptiert man horizontalen Gentransfer als einen dem vertikalen Gentransfer gleichgestellten adaptiven Prozeß, läßt sich prokaryotische Evolution nicht mehr als Baum darstellen, eher als Netz mit unterschiedlich gewichteten Kanten (Kunin et al., 2005; Huson et al., 2006). Autoren zukünftiger Inferenz-Programme sollten versuchen, den Einfluß lateralen Gentransfers auf Genome zu modellieren und ihn damit explizit in ihre Modelle zu integrieren. Der Verfasser ist allerdings der Ansicht, daß die Vererbung bei Prokaryoten maßgeblich von ihren vertikalen und weniger von ihren horizontalen Anteilen geprägt ist. Es ist unstrittig, daß Prokaryoten ihr genetisches Material verdoppeln und es weitgehend unverändert an zwei Nachkommen vererben. Dieser „vertikale“ Prozeß läßt sich durch einen bifurzierenden Baum darstellen. Aus den Ergebnissen ihrer Studien folgern Touchon und Kollegen (Touchon et al., 2009), daß während der Evolution von *E. coli* innerhalb ihrer Klade, also in flacher phylogenetischer Tiefe, LGT keine allzu große Rolle gespielt hat. Ihre Erkenntnisse ermuntern ebenfalls dazu, die Verwandtschaftsverhältnisse zumindest bei *E. coli* trotz der Anwesenheit von LGT in deren Stammesgeschichte als Baum darzustellen. Der Verfasser benutzte für diese Arbeit daher Programme zur Untersuchung der Verwandtschaftsverhältnisse von *E. coli* und *Salmonellen*, die entweder ausschließlich Bäume erzeugen oder auf solchen operieren.

Diese Ausführungen sollten klar gemacht haben, daß ein Verständnis der evolutionären Dynamik von *E. coli*, *Salmonellen* und anderer Pathogene von entscheidender Wichtigkeit ist, ist man doch bestrebt, die Ausbreitung der von ihnen hervorgerufenen Krankheitsbilder in Zukunft zu verhindern. Das allein wird aber nicht genügen. Es muß ebenso Klarheit darüber erlangt werden, welches die Faktoren sind, die diese Bakterien zu Pathogenen machen und in ferner Vergangenheit gemacht haben und wie die Pathogenitätsfaktoren an andere Bakterien weitergegeben werden. Diesbezügliche Forschungsvorhaben für *E. coli* und *Salmonellen* sind seit geraumer Zeit in vollem Gange (Johnson und Russo, 2002; Kaper et al., 2004; Kaur und Jain, 2012). Dieses Wissen muß in ein mathematisches Gewand gekleidet werden und in Form von Modellen[1] in Computerexperimente eingehen (McNeil und Aziz, 2009, Yang et al., 2010, Maslov et al., 2010). Der bakterielle

1 Modelle sind für die Wissenschaft äußerst nützlich, obwohl sie generell falsch sind. Die folgende populärwissenschaftliche Darstellung diskutiert dieses und verwandte Probleme von Modellen: http://www.utexas.edu/courses/bio301d/Topics/False.models/Text.html.

Metabolismus muß ebenso verstanden und hinreichend akkurat modelliert werden (Pal et al., 2006; Vieira et al., 2011; Closkey et al., 2013; Reed und Palsson, 2014), um zu verstehen, wie er durch die Umwelt beeinflußt wird. In-silico-Analysen sind aber kaum geeignet, derartige Fragen zu beantworten, wenn sie auf falschen Hypothesen beruhen. Die Verwandtschaftsverhältnisse der Bakterien sind eine solche Hypothese und müssen daher möglichst umfassend aufgeklärt werden.

3.3 Zielsetzung

Vor diesem Hintergrund war es zunächst Aufgabe dieser Arbeit, statistisch robuste Phylogenien für alle vollständig sequenzierten Genome von *E. coli* und *Salmonella* zu schätzen. Zwar wurden in der Vergangenheit diverse Phylogenien insbesondere für *E. coli* publiziert, jedoch umfaßten die Studien oft nur wenige Stämme (Sims und Kim, 2011; Chaudhuri et al., 2011), oder nicht alle Äste der Phylogenie waren durchgängig gut statistisch abgesichert (z. B. Reeves et al., 2011). Auch wurden viele Genome erst in den letzten Jahren vollständig sequenziert.

Das zentrale Anliegen dieser Arbeit war es jedoch, in zwei in-silico-Analysen die anzestralen Gengehalte, das Angebot essentieller Nährstoffe in der Umwelt (*E. coli* und *Salmonellen*) sowie Änderungen im Lebensstil (*E. coli*) auf den Phylogenien zu schätzen und zu ergründen, ob zwischen ihnen mathematische Abhängigkeiten existieren, die man biologisch erklären kann.

Für die Bauminferenz und andere Aufgaben wurden externe Programme und Web-Frontends benutzt. Sie werden in **Kapitel 4** beschrieben.

Um viele immer wiederkehrende handwerkliche Arbeitsschritte bei der Bauminferenz zu automatisieren, entwickelte der Verfasser das Programm *MMN*, das in **Kapitel 5** vorgestellt wird. Es schätzt einen ML-Speziesbaum für eine vorher festgelegte Menge von *E. coli*- und *Salmonellen*-Stämmen und ist vielfältig parametrisierbar. *MMN* arbeitet grundsätzlich nach der Konkatenationsmethode. Dabei wird ein multiples Sequenzalignment aus Genfamilien zusammengesetzt. *MMN* setzt diverse externe Programme ein.

Für den Vergleich phylogenetischer Bäume wurde das Programm *computeSSRF* realisiert, das ein neuartiges Distanzmaß für Phylogenien benutzt, welches dem menschlichen Denken eher entspricht als herkömmliche Distanzmaße. Das Maß, seine Motivation sowie die Realisation des entsprechenden Programms sind Thema von **Kapitel 6**.

Darstellungen von Phylogenien, für die keine maschinenlesbare Datenquelle (mehr) vorhanden ist, müssen manuell in das Newick-Format konvertiert werden, hat man den Wunsch, sie in einem Forschungsvorhaben weiter zu verwerten. Der Verfasser hat im Rahmen seiner Masterarbeit das Programm *TreeSnatcher Plus* verfaßt, das erstmals eine halbautomatische Erkennung beliebiger phylogenetischer Bäume in Pixelbildern und deren Umwandlung in das gängige Newick-Format ermöglichte. Im Rahmen dieser Arbeit wurden Fehler bereinigt, neue Funktionen hinzugefügt und eine Anleitung verfaßt, um das Programm online anzubieten. Für einen Beitrag in einer Fachzeitschrift (Laubach et al., 2012) wurde die Leistung des Programms auf einem Benchmark-Datensatz des Programms *TreeRipper* (Hughes, 2011) gemessen. Ferner hat der Verfasser das Programm um eine Texterkennung (OCR) erweitert. Die Benchmarks und Verbesserungsmöglichkeiten für *TreeSnatcher Plus* werden in **Kapitel 7** thematisiert.

Die verschiedenen Verfahren, mit denen für diese Arbeit Phylogenien unterschiedlicher Zusammensetzung für 61 *Escherichia coli* und 25 *Salmonellen*-Stämme (siehe **Tab. 8.1**) geschätzt wurden, sind Gegenstand von **Kapitel 8**. Drei der fünf verwendeten Techniken (Maximum Likelihood-Paradigma/universelle Genfamilien, Bayesianisches Paradigma/universelle Genfami-

lien, Maximum-Likelihood-Paradigma/COGs) schätzen einen Baum auf einem multiplen Sequenzalignment aus konkatenierten Genfamilien, das mit *MMN* hergestellt worden ist. Die vierte Technik (Neighbor-Joining, Distanzmatrix) gilt als weniger leistungsfähig als diese und wurde zu Vergleichszwecken mit der fünften Methode in die Arbeit aufgenommen. Letztere verwendet ebenfalls Neighbor-Joining, benutzt aber ein neuartiges Distanzmaß, das bisher noch nicht Gegenstand einer Veröffentlichung gewesen ist. Nach erfolgter Beschreibung der Techniken zur Bauminferenz werden die Bäume untereinander und mit von anderen Arbeitsgruppen veröffentlichten Bäumen verglichen. Grundlage dafür sind die Distanzen, die vorher mit *computeSSRF* errechnet worden sind. Der Versuch einer qualitativen Bewertung der Unterschiede wird unternommen. Abschließend erfolgt eine Diskussion der verwendeten Methoden und der Ergebnisse.

Geschieht lateraler Gentransfer auf dem Stammbaum mit gleichmäßiger Rate, oder haben die Vorfahren von *E. coli* und *Salmonellen* während der Evolution immer dann die Fähigkeit erworben oder verloren, bestimmte essentielle Nährstoffe aus der Umwelt metabolisch zu nutzen, wenn sie gleichzeitig Gene hinzugewonnen oder abgegeben haben? Um diese Frage geht es in **Kapitel 9**. Die Verteilungen anzestraler Gengehalte auf den *E. coli*- und *Salmonellen*-Phylogenien werden mit einem Maximum-Parsimony-Ansatz geschätzt. Die Rekonstruktion der Verteilung der essentiellen Nährstoffe auf der Phylogenie, die *E. coli* metabolisieren muß, um zu überleben, wird mit einem Verfahren von Elhanan Borenstein und Kollegen (Borenstein et al., 2008) bewerkstelligt. Abschließend wird betrachtet, inwieweit die mathematisch gewonnenen signifikanten Gen/Nährstoff-Paare auch biologisch sinnvoll sind. Eine Beurteilung der Methodik schließt das Kapitel ab.

In **Kapitel 10** wird die These aufgestellt, daß Änderungen im Lebensstil bei *E. coli* mit verstärktem lateralen Gentransfer einhergingen. Dazu werden zwölf Lebensstil-Merkmale untersucht, die über die von *E. coli* bekannten Pathotypen definiert worden sind. Die Verteilung der anzestralen Gengehalte auf der *E. coli*-Phylogenie wird aus **Kapitel 9** übernommen und mit den anzestralen Verteilungen der individuellen und der kumulierten Lebensstil-Änderungen verglichen. Letztere Verteilungen wurden ebenfalls mit einem Parsimonie-Ansatz geschätzt. Es erfolgt eine Schlußbetrachtung, in der die eingesetzte Strategie mit Blick auf die Ergebnisse hinterfragt wird.

Kapitel 11 beschließt die Arbeit mit einem Ausblick. Es wird darum gehen, ob die für diese Arbeit gesteckten Ziele erreicht worden sind und inwieweit die hier charakterisierten Methoden und Ergebnisse für künftige Studien nützlich sein können.

4 Externe Software und Web-Frontends

Für die im Rahmen dieser Arbeit durchgeführten Projekte wurde neben den Programmen des Verfassers eine Vielzahl fremder Programme und Web-Frontends verwendet, die in diesem Kapitel beschrieben werden. Für nahezu alle Programme und Frontends wurden Aufsätze in wissenschaftlichen Fachzeitschriften veröffentlicht. Die vom Verfasser selbst erstellten Programme werden jeweils in den betreffenden Kapiteln dargestellt.

BEAST und Tracer

Mit *BEAST* haben Alexei Drummond und Andrew Rambaut (Drummond und Rambaut, 2007) eine Applikation für die MCMC-basierte Analyse („Markov-Chain-Monte-Carlo“) molekularer Sequenzen vorgelegt. Nach Aussage der Autoren eignet sich *BEAST* besonders zur Analyse von Datensätzen, die die Anwesenheit einer molekularen Uhr fordern. *Trace*r von Andrew Rambaut, Marc Suchard und Alexei Drummond dient der Analyse der Ausgabedateien von MCMC-Programmen wie *BEAST*, *MrBayes* und *LAMARC* (Kuhner, 2009). In dieser Arbeit wurde die bis September 2013 entwickelte Version von *Tracer* genutzt. Mittlerweile existiert die Software *BEAST2* (Drummond et al. 2014). Für weitere Informationen zu *Tracer* siehe auch **Abschnitt 8.2.2.2**.

Bioperl

Bioperl von Stajich und Kollegen (Stajich et al., 2002) ist ein Gemeinschaftsprojekt, das sich zum Ziel gesetzt hat, Perl-Programme zu schreiben, die für biologische Fragestellungen nützlich sind. *Bioperl* bietet objektorientierte Methoden für eine Vielzahl von biologischen Problemen. Der Verfasser dieser Arbeit hat die Funktionalität von *Bioperl* zum Lesen und Schreiben verschiedener Sequenzformate während der Entwicklung des *MMN*-Programms (siehe Kapitel 4) eingesetzt.

BLAST

Das „Basic Local Alignment And Search Tool“, *BLAST* (Altschul, 1997), oft der Einfachheit halber „Blast“ geschrieben, implementiert eine schnelle heuristische Suche ähnlicher Nukleotid- oder Proteinsequenzen in Datenbanken. Die Version *blastp* für Proteinsequenzen ist in dem vom *NCBI* angebotenen Paket *Legacy BLAST* (http://blast.ncbi.nlm.nih.gov), enthalten und stellt eines der Grundmodule des Perl-Programms *MMN* des Verfassers dar (**vgl. Kapitel 4**). Mittlerweile rät das *NCBI* dazu, anstelle von *Legacy BLAST* ihr moderneres Paket *BLAST+* zu benutzen.

DendroPy/SumTree

SumTrees (Sukumaran und Holder, 2010) gehört zur *DendroPy*-Programmsammlung und ist ein Programm von Jeet Sukumaran und Mark T. Holder, mit dem man Support-Zahlen aus einem nichtparametrischen Bootstrap oder Bayesian Posterior Probabilities für Splits oder Claden eines phylogenetischen Baumes zusammenfassen kann.

In dieser Arbeit wurde *SumTree* dazu benutzt, die Split-Supports der Genbäume den Splits eines Speziesbaumes aufzuprägen.

Figtree

Die in Java geschriebene Software *FigTree* von Andrew Rambaut (Rambaut, 2012) dient der grafischen Darstellung phylogenetischer Bäume und der Produktion entsprechender

publikationsreifer Abbildungen. Der Verfasser hat die meisten der in dieser Arbeit dargestellten Phylogenien mit *FigTree* produziert.

Gblocks

Das Programm *Gblocks* (Castresana, 2000) löscht schlecht alignierte Positionen und divergente Regionen in einem Alignment aus Nukleotid- oder Proteinsequenzen. Solche Positionen sind entweder nicht homolog oder mit mehreren Substitutionen angereichert. Laut Castresana wählt *Gblocks* die Blöcke auf eine ähnliche Weise aus wie der Mensch. Jeder Block muß bestimmten Kriterien genügen: er enthält keine langen Segmente zusammenhängender, nichtkonservierter Positionen, er hat keine Gaps, und die flankierenden Positionen sind hochkonserviert. *Gblocks* ist sehr schnell und ermöglicht die Automatisierung einer phylogenetischen Analyse großer Datenmengen, da es die Notwendigkeit, multiple Alignments von Hand zu editieren, weitgehend beseitigt. Im Jahr 2007 konnten Gerard Talavera und Jose Castresana zeigen (Talavera und Castresana, 2007), daß die Entfernung von Blöcken mittels *Gblocks* für die meisten Alignments, die allerdings nicht zu kurz sein dürfen, zu besseren Bäumen führt, obwohl das Verfahren mit einem Informationsverlust einhergeht. Sie stellten darüberhinaus fest, daß bereinigte Alignments zwar bessere Topologien produzieren, diese sich aber paradoxerweise durch eine schlechtere Bootstrap-Unterstützung auszeichnen.

GLOOME

GLOOME (Cohen und Pupko, 2010), die „Gain Loss Mapping Engine“, ist ein System aus Web-Frontend und dahinterliegendem Server. Für ein vom Benutzer hochzuladendes FASTA-Alignment von 0/1-Sequenzen, sog. „phyletischer Muster“, und bei Bedarf eine Phylogenie im Newick-Format schätzt *GLOOME* ast- und positionsspezifische Gewinn- und Verlust-Ereignisse (engl. „gain and loss events“). Ist kein Baum vorhanden, schätzt *GLOOME* ihn aus den phyletischen Mustern. Dazu setzt es eine modellbasierte Distanzberechnung und Neighbor Joining ein.

GLOOME stellt für die Detektion der Ereignisse die Techniken *Stochastic Mapping* und *Parsimony* zur Verfügung. Entscheidet sich der Benutzer für das *Parsimonie*-Kriterium, kann er die relativen Kosten von Gewinn- und Verlust-Ereignissen angeben. *Stochastic Mapping* (Nielsen, 2002; Minin und Suchard, 2008) vollzieht sich in einem probabilistischen Bezugsrahmen. Für ein gegebenes evolutionäres Modell werden Erwartungswert und Wahrscheinlichkeit für die Ereignisse pro Ast und jeden der vier möglichen Konfigurationen der Zeichen 0 und 1 am Anfang und Ende eines Asts berechnet (siehe dazu http://gloome.tau.ac.il/overview.php). Der totale Erwartungswert für einen Ast ergibt sich dann über die gewichtete Summe aus den vier Konfigurationen.

Die Webanwendung stellt eine Auswahl unterschiedlich komplexer evolutionärer Modelle bereit, mit der die Dynamik von Gewinn- und Verlust-Ereignissen modelliert wird.

Obwohl der Fokus von *GLOOME* auf *Stochastic Mapping* liegt, nutzte der Verfasser dieser Arbeit ausschließlich das *Parsimonie*-Kriterium, um anzestrale Gewinn- und Verlust-Ereignisse für verschiedene *Escherichia coli*- und *Salmonellen*-Phylogenien zu schätzen (siehe **Abschnitt 10.4**).

iTol: Interactive Tree Of Life

Das Web-Frontend *iTol* (http://itol.embl.de) der Autoren Letunic und Bork (Letunic und Bork, 2006; Letunic und Bork, 2011) ist ein frei verfügbares Werkzeug zur Anzeige, interaktiven Manipulation und Annotation phylogenetischer Bäume. Es zeichnet sich besonders durch seine Fähigkeit aus, diverse Pixel- und Vektorgrafik-Formate exportieren zu können. Der vom Frontend in der Grundeinstellung angezeigte Baum wurde mit Methoden generiert, die die Gruppe um Francesca D. Ciccarelli im Fachblatt *Science* publiziert hat (Ciccarelli et al., 2006). Ihr Verfahren nutzt

hochkonservierte Gene. Für diese Arbeit wurden Verwandtschaftsbäume für *Escherichia coli* und *Salmonellen* mit einer Methode bestimmt, die der von Ciccarelli et al. ähnelt.

JModeltest/JModeltest2

Die in Java geschriebene GUI-Applikation J*Modeltest/JModeltest2* (Posada et al., 2012) wählt mit statistischen Methoden für eine gegebene Nukleotidsequenz das Modell aus, das die Substitution der Nukleotide am besten beschreibt. Dazu setzt die Applikation fünf Auswahlverfahren ein: hierarchischer und dynamischer Likelihood-Ratio-Test (hLRT, dLRT), Akaike- (AIC) und Bayesianisches Informationskriterium (BIC) sowie ein auf der Entscheidungstheorie fußendes Kriterium (DT). Bei Bedarf schätzt das Programm die Unsicherheit bei der Modellauswahl, die Wichtigkeit von Parametern oder über alle Modelle gemittelte Parameter und Phylogenien. Das Programm J*Modeltest/JModeltest2* bedient sich intern *PhyML* (Guindon und Gascuel, 2003).

KEGG

KEGG, die „Kyoto Encyclopedia of Genes and Genomes“, ist eine Datenbank-Resource der Kanehisa Laboratories, die der Organisation NPO Bioinformatics Japan angehören. Sie hat sich dem Ziel verpflichtet, zum Verständnis der Funktionen der biologischen Systeme Zelle, Organismus und Ökosystem beizutragen. Dazu integriert sie molekulare Daten aus Genomsequenzierungsprojekten und Hochdurchsatzexperimenten. KEGG stellt eine Programmierschnittstelle zum Download verschiedener Daten bereit, darunter biochemische Stoffwechselwege, Informationen zu Genen, Enzymen, Reaktionen und Liganden.

Für diese Arbeit wurde die Programmierschnittstelle im Zusammenhang mit der Rekonstruktion essentieller Nährstoffe in der anzestralen Umwelt von Escherichia coli benutzt.

MCL

Der *MCL*-Algorithmus („Markov Cluster Algorithm“) ist ein schneller und skalierbarer, unüberwachter Clustering-Algorithmus für Graphen, welcher stochastische Flüsse in Graphen simuliert. Sein Entdecker, Stijn van Dongen, hat ihn in seiner Dissertation (van Dongen, 2000) beschrieben und seine mathematischen Grundlagen untersucht. Die Anwendung des *MCL*-Algorithmus im Bereich der Detektion von Proteinfamilien wurde etwas später publiziert (Enright et al., 2002) und mit der griffigen Bezeichnung „*TRIBE-MCL*“ versehen.

Die vorliegende Arbeit benutzt zur Bestimmung orthologer Gruppen sowohl *MCL* als auch die von Christian Eßer für seine Dissertation (Eßer, 2010) entwickelte Methode, die auf Sequenzidentität und Syntenie beruht (vgl. Abschnitt *MpBatch/Alignomat*).

MAFFT

MAFFT ist ein von Kazutaka Katoh und Kollegen (Katoh et al., 2002; Katoh et al., 2005) konzipiertes Programm zur Erstellung eines multiplen Sequenzalignments (MSA). Es bedient sich zweier Techniken, die in vergleichbaren Programmen bisher nicht zum Einsatz gekommen sind. Zum einen identifiziert es homologe Regionen mittels schneller Fouriertransformation (FFT), wobei eine Aminosäuresequenz in eine Folge von Mengen- und Polaritätswerten pro Aminosäurerest überführt wird. Zum anderen benutzt *MAFFT* ein vereinfachtes Scoring-System, das die benötigte CPU-Zeit verkürzt und die Genauigkeit der Alignments verbessert, und das sowohl bei Sequenzen mit großen Insertionen und Anhängen als auch bei wenig verwandten Sequenzen ähnlicher Länge. *MAFFT* implementiert als Heuristiken eine progressive und eine iterative Methode.

In dieser Arbeit kamen die Versionen v6.902b vom 15.05.2012 und v6.864b vom 10.11.2011 zum

Einsatz, die nach Setzen des *auto*-Parameters selbständig die geeignete von *MAFFT* angebotene Bearbeitungsstrategie auswählen.

Mesquite

Das Open-Source-Projekt *Mesquit*e ist eine Software für Evolutionsbiologie, die Wissenschaftlern behilflich ist, vergleichende Daten über Organismen zu analysieren. Der Schwerpunkt liegt auf der phylogenetischen Analyse, aber es sind auch Module für Populationsgenetik und andere Bereiche vorhanden, die bei Bedarf hinzugefügt werden können. In dieser Arbeit wurde *Mesquite* verwendet, um Äste aus phylogenetischen Bäumen zu entfernen und Multifurkationen einzuführen. Desweiteren wurden die Abbildungen einiger Bäume mit Mesquite angefertigt.

Wayne und David Maddison sind die Initiatoren des mit der Programmiersprache Java erstellten Projektes (Maddison und Maddison, 2011).

MpBatch/Alignomat

Die von Christian Eßer für seine Dissertation (Eßer, 2010) entwickelten Perl-Skripten *synti und alignomat (*in seiner Dissertation, Kapitel 6, als *synti* bezeichnet*)* identifizieren Genfamilien aufgrund ihrer Position im Genom. Dazu werden zunächst die paarweisen Sequenzähnlichkeiten zwischen allen Proteinen aus den verglichenen Organismen mittels *blastp (*siehe *BLAST)* berechnet. *MpBatch* realisiert einen Algorithmus, der versucht, orthologe Gruppen derart anzulegen, daß möglichst viele Proteine in identischer Reihenfolge im als Referenz dienenden Genom und dem Zielgenom vorliegen. Zweck ist es, solche Proteine in einer Familie zu vereinen, die bei einer ausreichenden Sequenzidentität eine sehr ähnliche Position haben. *Alignomat* obliegt die Aufgabe, die von *MpBatch* erzeugten Orthogruppen in einzelne FASTA-Dateien (Lipman und Pearson, 1985) zu schreiben.

Die Methode liest zunächst aus den PTT-Dateien die Positionsinformationen für alle Gene ein, dann ordnet sie jedem eine eindeutige ID zu. Im Anschluß werden die BLAST-Ergebnisse verarbeitet, wobei nichtreziproke und solche Treffer ausgeschlossen werden, deren *normalisierte Identität* (Zahl identischer Positionen/Länge des kürzeren der beiden verglichenen Proteine) kleiner ist als ein über den Parameter `-n` spezifizierter Schwellenwert. Für *E. coli* wird 70 % als Standard-Schwellenwert verwendet. Der chromosomalen Reihenfolge des Referenzgenoms folgend, versucht der Algorithmus, für jedes Gen in jedem anderen Genom parallele Abschnitte zu finden. Dabei werden alle Treffer aus der BLAST-Suche verwendet. Gelingt es, mehrere aufeinander folgende Sequenzen mit ihren Homologen aneinander auszurichten, wird diese Reihe solange weiter verfolgt, bis sich kein weiterer Treffer mehr finden läßt. Für das betrachtete Paar von Genen werden alle untersuchten Gene markiert, um ihre Mehrfachverwendung auszuschließen. Lassen sich weniger als fünf Gene konsekutiv aneinander ausrichten, werden alle Treffer verworfen, und die Suche wird mit dem nächsten Gen fortgesetzt. Das Skript gibt eine Tabelle aus, die für jedes Genom eine Spalte enthält und zusätzliche eine mit der Größe der Proteinfamilien. Jede Zeile repräsentiert eine Proteinfamilie. Wurde nur ein Genom als Referenz ausgewählt, werden die Proteinfamilien an dieser ausgerichtet. Die Einträge der Tabelle sind die GI-Nummern der entsprechenden Proteine.

MrBayes

Die Autoren John Huelsenbeck, Bret Larget, Paul van der Mark, Fredrik Ronquist, Donald Simon und Maxim Teslenko haben mit *MrBayes* ein Programm für die Bayesianische Inferenz phylogenetischer Bäume und die Wahl evolutionärer Modelle geschaffen. *MrBayes* nutzt eine auf Markov-Chain-Monte-Carlo (MCMC) basierende Methodik, um die posteriore Verteilung von

Modellparametern zu schätzen. Im Gegensatz zu Programmen wie *PhyML* und *RAxML* liefert *MrBayes* konzeptbedingt nicht einen optimalen Baum, sondern eine Schar sehr guter Bäume.

Weitere Informationen zu *MrBayes* folgen in **Abschnitt 8.2.2.2**.

NCBI FTP Server

Alle in dieser Arbeit verwendeten 86 Bakteriengenome wurden zu verschiedenen Zeitpunkten in den Jahren 2010 bis 2013, spätestens Ende September 2013, vom FTP Server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) des *National Center for Biotechnology Information*, *NCBI*, heruntergeladen.

PHYLIP

PHYLIP ist der Name eines von Joseph Felsenstein (Felsenstein, 1981) entwickelten Software-Pakets zur Erstellung phylogenetischer Bäume („phylogeny inference package“) unter einem Maximum-Likelihood-Framework und gleichzeitig des Sequenzformats, das durch dieses Paket populär gemacht worden ist. Die Webseite zu *PHYLIP* wird von der Universität Washington betrieben (http://evolution.genetics.washington.edu/PHYLIP/general.html). Sequenzen im *PHYLIP*-Format sind entweder sequentiell oder verschachtelt angeordnet. Viele Phylogenieprogramme, unter ihnen *PhyML* und *RAxML* (Stamatakis, 2006), akzeptieren ausschließlich Alignments im *PHYLIP*-Format.

Für diese Arbeit wurde eine Auswahl der Programme in *PHYLIP* verwendet: CONSENSE berechnet Consensus-Bäume primär anhand des Majority-Rule-Kriteriums. Der Benutzer kann bei Bedarf aber auch Strict-Consensus oder Varianten des Majority Rule-Consensus benutzen. NEIGHBOR erzeugt Bäume nach der Neighbor-Joining-Methode von Saitou und Nei (Saitou und Nei, 1987) oder nach der UPGMA-Methode (Sokal und Michener, 1958). SEQBOOT erzeugt aus einem gegebenen Datensatz eine Vielzahl neuer Datensätze unter Verwendung von Bootstrap-Resampling. TREEDIST berechnet die Branch-Score-Distanz zwischen zwei Bäumen und berücksichtigt dabei sowohl Topologie als auch Astlängen. Zusätzlich berechnet das Programm den Robinson-Foulds-Abstand („RF-Distanz“), der lediglich Unterschiede zwischen den Topologien, nicht aber Astängen, berücksichtigt. Die verwendete Programmversion stammt vom April 2001.

PhyML

Bei *PhyML* der Autoren Stephane Guindon und Olivier Gascuel (Guindon und Gascuel, 2003) handelt es sich um ein Programm zur Ableitung von Phylogenien, das auf dem Maximum-Likelihood-Prinzip beruht. Im Kern benutzt *PhyML* einen Hill-Climbing-Ansatz, der gleichzeitig sowohl Baumtopologie als auch Astlängen modifiziert. Ihr Algorithmus beginnt bei einem Anfangsbaum, der vorher von einer schnellen distanzbasierten Methode erzeugt worden ist, und modifiziert ihn iterativ, um seine Likelihood zu verbessern. Da in jeder Iteration Topologie und Astlängen gleichzeitig geändert werden, nähert sich das Verfahren schnell einem Optimum an.

Die in dieser Arbeit eingesetzte Version 3 dieser Software (Guindon und Gascuel, 2010) mit der Kennung 20121204 bietet neben anderen Verbesserungen erweiterte Möglichkeiten zur Erzeugung eines Startbaumes unter einem Maximum-Parsimony-Kriterium an. Auf einem Rechencluster wurde ferner *PhyML-MPI*, die für den parallelen Betrieb auf Rechenclustern compilierte Version des Programms, eingesetzt.

Proteinortho

Die Software P*roteinortho* dient der Detektion von Orthologen/Koorthologen im Rahmen

großangelegter phylogenetischer Analysen. Marcus Lechner und Mitautoren (Lechner et al., 2011) haben das Programm für den Betrieb auf Rechenclustern optimiert. Für einen Satz paarweiser Sequenzähnlichkeiten, die vorher beispielsweise mit *Blast* (Altschul, 1997) generiert wurden, bestimmt es Orthologe mit einer auf reziproken Best-Hits basierenden Heuristik. Die Software zeichnet sich dadurch aus, daß sie in den meisten Fällen schneller als vergleichbare Programme arbeitet.

RAxML

Neben *PhyML* wurde das verwandte Programm *RAxML* (Stamatakis, 2014) zur Inferenz großer Phylogenien eingesetzt. *RAxML* bietet eine parallele Abarbeitung von Bootstrap-Suchen auf *Linux*-Clustern sowie über POSIX-Threads an. Leider offeriert es im Gegensatz zu PhyML nur wenige Nukleotidsubstitutionsmodelle, unter ihnen GTR. *RAxML* bietet in neueren Versionen neben dem ursprünglichen einen sehr viel schnelleren Algorithmus zur Baumsuche. Der als „Rapid Hill-climbing“ bezeichnete Ansatz verzichtet darauf, Topologien weiter zu verfeinern, die wahrscheinlich keine bessere Likelihood erzielen, und liefert dabei Bäume, die nur unwesentlich schlechter als die sind, die der ursprüngliche Suchalgorithmus liefert.

In dieser Arbeit wurde überwiegend Version 7.4.2 vom 23. November 2012 mit „Rapid Hill Climbing“ eingesetzt.

TOPD/FMTS

Mit dem Perl-Programm *TOPD/FMTS* von Autor Pere Puigbo (Puigbo, 2007) lassen sich Unterschiede zwischen Bäumen evaluieren. Neben bereits veröffentlichten Methoden zum Vergleich von Phylogenien bietet es eigene an. Es kombiniert das Programm *TOPD* („Topological Distance“), das zwei Bäume mit gleichen Taxa oder beschnittene Bäume vergleicht, mit dem Programm *FMTS* („From Multiple To Single“), das Bäume auf Grundlage multipler Genfamilien in Bäume auf Grundlage von Einzelgenen konvertiert.

5 Halbautomatische Erkennung von Phylogenien in Rastergrafiken

Auszüge dieses Kapitels wurden bereits veröffentlicht in:

Laubach, T., A. von Haeseler und M. J. Lercher. **TreeSnatcher Plus: Capturing Phylogenetic Trees from Images**. *BMC Bioinformatics* 13.1 (2012): 110.

5.1 Einleitung

Jedes wissenschaftliche Feld, das Informationen mithilfe von Computern verarbeitet, muß technische Darstellungen zur späteren Wiederverwendung in einem maschinenlesbaren Format vorhalten und archivieren. Dazu muß die Struktur einer Darstellung in seine geometrischen Primitive zerlegt werden. Je mehr Vorwissen zum Bildinhalt für den Computer verfügbar ist, desto weniger Fehler macht dieser während der Analyse. Heutzutage digitalisieren und archivieren Computerprogramme jahrzehntealte architektonische Zeichungen, für die die verwendeten Symbole, Piktogramme und Schrifttypen in der Regel genormt sind. Dahingegen gibt es kein Computerprogramm, das automatisch beliebig geformte phylogenetische Bäume aus einer Abbildung in einen maschinenlesbaren Ausdruck, z. B. das bis heute nicht veröffentlichte Newick-Format, konvertieren kann. Die Stile, in denen phylogenetische Bäume bisher veröffentlicht worden sind, sind genauso vielfältig wie die Softwarepakete, die zur Herstellung der Bäume und Bilder verwendet werden können. Eine umfassende Link-Sammlung zu solchen Programmen hat Joseph Felsenstein (Felsenstein, 2008) online gestellt (siehe http://evolution.genetics.washington.edu/phylip/doc/sequence.html). Da es keine festen Regeln für das Design solcher Bäume gibt, darf ein Programm, das auf die Erkennung prinzipiell beliebiger phylogenetischer Bäume abzielt, keinerlei Vorwissen voraussetzen, von der bloßen Existenz des dargestellten Baumes einmal abgesehen.

In diesem Kapitel wird die vom Verfasser realisierte Java-Applikation *TreeSnatcher Plus („TSP“)* beschrieben, die mit Unterstützung des Benutzers eine Digitalisierung phylogenetischer Bäume in Bildern erlaubt. *TSP* wurde in dieser Arbeit praktisch eingesetzt, um eine von Reeves und Kollegen (Reeves et al., 2011) veröffentlichte *E. coli*-Phylogenie maschinenlesbar zu machen (siehe Abschnitt **8.3.3**). Vor *TSP* gab es andere Programme, deren Ziel es war, phylogenetische Darstellungen digital zu erfassen. *TreeThief* von Autor Andrew Rambaut (Rambaut, 2000) für *Mac OS 9* war die erste Applikation, die das Bild eines Baumes in seine computerlesbare Repräsentation übersetzt. Das Programm erlaubte es dem Benutzer, einen Baum zu erfassen, indem er schrittweise jeden Knoten mit der Maus anklickt. *TreeRipper* von Joseph Hughes (Hughes, 2011) konvertiert Darstellungen rechteckiger Phylogenien automatisch in das Newick-Format, sofern sie mehreren eng definierten Kriterien genügen. Er bietet ein Web-Frontend an, das es Benutzern gestattet, Abbildungen zur automatischen Bearbeitung hochzuladen. Hughes hat zusammen mit *TreeRipper* außerdem einen Satz von Darstellungen phylogenetischer Bäume veröffentlicht, der für andere Programme als Benchmark dienen kann.

Obwohl Wissenschaftler die Ergebnisse ihrer Studien zur Stammbaumforschung gewöhnlich durch Abbildungen phylogenetischer Bäume illustrieren, haben sich die Herausgeber der einschlägigen

wissenschaftlichen Fachzeitschriften bisher noch nicht auf Standards geeinigt. Ein solcher Standard könnte beispielsweise fordern, daß als Voraussetzung für eine Veröffentlichung phylogenetische Bäume in einem maschinenlesbaren Format in einer der einschlägigen Online-Datenbanken, beispielsweise *TreeBASE* (www.treebase.org/treebase-web/home.html), *MorphoBank* (www.morphobank.org) oder *Dryad* (www.datadryad.org), hinterlegt werden müssen. Das Fehlen solcher Standards hat dazu geführt, daß bis heute Phylogenien oft nur in bildlicher Form veröffentlicht werden, was sowohl ihre Weiterverwendung in anderen Studien als auch die Archivierung der Forschungsergebnisse erschwert. Leebens-Mack und Kollegen schlagen einen Fahrplan für die Entwicklung minimaler Standards für das Berichtswesen im Zusammenhang mit Daten aus phylogenetischen Analysen, MIAPA (engl. „Minimal Information about a Phylogenetic Analysis"), vor (Leebens-Mack et al., 2006). Sie betreiben eine Webseite, auf welcher sie mögliche Barrieren für die Verwertung von Daten aus wissenschaftlichen Analysen diskutieren (siehe http://www.evoio.org/wiki/BarriersToReUse). Prinzipiell sollte es möglich sein, Daten in elektronischer Form von den Autoren wissenschaftlicher Publikationen zu erhalten. Dieser offensichtliche Ansatz erscheint jedoch nicht praktikabel. In einem Beispiel aus einer anderen wissenschaftlichen Disziplin verweigerten 73 % der befragten Autoren die Herausgabe ihrer Daten (Wicherts et al., 2006). Möchte man trotzdem die meisten veröffentlichen Bäume wiederverwenden, scheint deren Digitalisierung die zurzeit einzig realistische Option zu sein.

5.2 Material und Methoden

Die Dateien zu diesem Kapitel befinden sich innerhalb der Ordnerstruktur im Projektordner *Kapitel_TSP*. Kopien des Quellcodes zu *TreeSnatcher Plus*, die PDF-Tutorien zum Programm sowie die in TSP verarbeiteten Benchmark-Bilder von Joseph Hughes sind ebenfalls dort zu finden.

Neue Funktionalitäten in TreeSnatcher Plus

Der Verfasser hat im Rahmen seiner Masterarbeit die Java-Applikation *TreeSnatcher Plus* (Laubach et al., 2012) entwickelt, die es erstmals ermöglichte, mit Benutzerassistenz phylogenetische Bäume in Pixelbildern mit Methoden aus dem Bereich Bildverarbeitung zu analysieren, zu vermessen und schließlich in das Newick-Format zu überführen. Letzteres ist ein gängiges, aber nicht offiziell standardisiertes Datenaustauschformat für Bäume. Während die unter Umständen notwendige Vorverarbeitung eines Eingabebildes bei dem Vorgänger von *TSP*, *TreeSnatcher* (Laubach und von Haeseler, 2007) noch außerhalb des Programms erfolgen mußte, kann ein zu verarbeitendes Pixelbild nunmehr innerhalb von *TSP* mit diversen Techniken vorbearbeitet werden (vgl. **Abb. 5.1**).

Die grafische Benutzeroberfläche (engl. „Graphical User Interface", abgekürzt „GUI") von *TSP* basiert auf dem Java Swing API. Das Programm bietet neben einer Undo-Funktionalität die Möglichkeit, den aktuellen Bearbeitungszustand zu speichern und wieder herzustellen („Snapshot"). Dieser kann auch in Form eines Bildes gespeichert werden, das auf Wunsch verschiedene Schichten mit visuellen Informationen enthält. *TSP* erkennt zuverlässig Knoten und Kanten der dargestellten Phylogenie und baut daraus eine programminterne Repräsentation in Form eines mathematischen Graphen. Es berechnet die Astlängen für beliebig geformte Äste und erkennt dabei die horizontal verlaufenden Astanteile in rechteckigen Phylogenien, selbst wenn die Darstellung geringfügig „gekippt" ist. Vom Programm berechnete und vom Benutzer eingegebene Astlängen dürfen gleichzeitig verwendet werden. Desweiteren kann der Benutzer einen existierenden Baum modifizieren oder einen völlig neuen konstruieren. Der Verfasser hat die *Linux*- und *Windows*-

Versionen von *TSP* mit einer einfachen OCR-Funktionalität ausgestattet. Dabei kommt die von Jörg Schulenburg programmierte OCR-Engine *GOCR* in der Version 0.49 (Schulenburg, 2013) zum Einsatz. Der Benutzer von *TSP* hat die Möglichkeit, horizontal angeordneten Text mit einer Auswahlbox zu markieren. Das Programm versucht dann eine Texterkennung und weist den erkannten Text dem am nächsten stehenden Taxon zu. Die Erkennungsrate schwankt von akzeptabel bis sehr schlecht, abhängig von Schriftart und Auflösung.

Bearbeitungswerkzeuge

TreeSnatcher Plus akzeptiert Bilddateien in den Formaten PNG, JPG/JPEG und GIF. Das Format PDF wird derzeit nicht unterstützt, jedoch sind Werkzeuge für die Extraktion von Bildern aus PDFs anderweitig verfügbar, für *Linux* beispielsweise die *Xpdf-Suite* (www.foolabs.com/xpdf). *TreeSnatcher Plus* bietet die folgenden, aus Standard-Algorithmen (Burger und Burge, 2005) modifizierten Vorverarbeitungs-Werkzeuge: Stift, Radiergummi, Linie, Flächenfüller, Stempel, Histogrammspreizung, Farbreduktion, Graustufenkonvertierung, Binarisierung über lokales und globales Thresholding, Farbbereichsmanipulation, Inversion, Median- und Minimumfilter, Verwaschen, Schärfen, Aufhellen und Abdunkeln, Ausdünnen/Skelettieren. Vor der automatischen Plazierung der Knoten muß der Benutzer das Bild geeignet vorbereiten. Insbesondere muß das Bild in eine Liniengrafik umgewandelt werden, in der sich Text und Grafikelemente nicht mit dem Baum überschneiden dürfen. Sollte das Bild diesen Anforderungen nicht genügen, wird das Programm die Baumtopologie nicht fehlerfrei identifizieren können.

Arbeitsablauf

Die Arbeit mit *TreeSnatcher Plus* vollzieht sich entlang einer generellen Abfolge globaler Schritte, von denen jeder mindestens einmal ausgeführt wird, und das entweder für einen Teil des Bildes oder für das Gesamtbild. Dem Benutzer obliegt die Gesamtaufsicht über den Prozeß. Er nimmt geeignete Korrekturen vor und wiederholt Teilschritte, falls nötig. Der Arbeitsablauf sieht wie folgt aus:

1. **Eingabebild lesen:** Das Programm liest das Eingabebild und wandelt es in ein Graustufenbild um. Der Benutzer beschneidet das Bild nach Belieben. Auf diese Weise löst er Teilbäume aus dem Gesamtbaum oder wählt eine Untermenge der Taxa.
2. **Vorverarbeitung**: Der Benutzer bereitet das Bild mit den Vorverarbeitungswerkzeugen geeignet vor.
3. **Binarisierung**: Der Benutzer gibt einen Schwellenwert an. Die Graustufen unterhalb des Schwellenwerts oder gleich dem Schwellenwert werden schwarz, alle anderen werden weiß. Ziel dieses Schrittes ist es, Vordergrund und Hintergrund voneinander zu trennen.
4. **Skelettieren**: Der Benutzer dünnt halbautomatisch den Vordergrund in dem Teil des Bildes aus, der den Baum enthält. Dieser Schritt ist notwendig, damit das Programm die Pfade zwischen den Linienkreuzungen finden kann (siehe Schritt 8).
5. **Fluten des Vordergrundes**: Der Benutzer markiert eine Position auf einem Ast des Baumes. Das Programm färbt den gesamten von dort erreichbaren, zusammenhängenden Vordergrund ein. Darauf folgende Schritte behandeln den so eingefärbten Bereich als „den Baum“. Alle anderen Vordergrundbestandteile werden ignoriert.
6. **Plazierung der inneren und äußeren Knoten**: Das Programm schlägt Positionen für Linienkreuzungen und Linienenden vor. Diese repräsentieren Verzweigungen und Blätter des Baumes. Jeder Position wird ein logischer Baumknoten zugeordnet. Der Benutzer kann

Knoten hinzufügen, löschen und verschieben. Im skelettierten Bild werden schwarze Pixel, die Nachbar genau eines anderen schwarzen Pixels sind, zu einer Blattposition. Schwarze Pixel, die Nachbar von mindestens drei schwarzen Pixeln sind, werden Kandidaten für eine Verzweigungsposition. Falls es mehrere Kandidatenpositionen für eine Verzweigung gibt, wird diese aus dem Durchschnitt dieser Positionen gebildet.

7. **Wahl des Baum-Typs**: Das Programm kann zwischen beliebig geformten (aber nicht zirkulären und polaren) Bäumen sowie rechteckigen Bäumen unterscheiden. Die Wahl des Typs beeinflußt, wie das Programm Astlängen behandelt. Der Baum-Typ muß gewählt werden, bevor Schritt 8 aufgerufen wird.
8. **Erkennung der Äste**: Das Programm verfolgt lückenlose Vordergrundpfade zwischen jedem Paar von Knoten, um die Äste des Baumes zu identifizieren. Gibt es mehrere mögliche Pfade, entscheidet es sich für den kürzesten. Sollte das Programm einen Ast nicht erkannt oder falsch gesetzt haben, kann der Benutzer entweder das Bild mithilfe der Malwerkzeuge modifizieren und Schritt 8 wiederholen oder einen neuen Ast zwischen zwei Knoten manuell hinzufügen und dessen Länge manuell setzen.
9. **Bestimmung der Astlängen**: Die Genauigkeit hängt davon ab, wie gut die ausgedünnte Baumstruktur, die Plazierung der Knoten und der Originalbaum übereinstimmen. Für beliebig geformte Bäume basiert die Astlänge in Pixeln auf der gesamten Länge des Vordergrundpfades zwischen den beiden definierenden Knoten. Für rechteckige Bäume ist die Astlänge jeweils die Summe der Längen der horizontalen Pfadsegmente. Der Benutzer darf eigenständig definierte Astlängen eintippen und sie mit den vom Programm berechneten Astlängen mischen. Indem der Benutzer mit der Maus eine Maßstabsleiste im Bild markiert, kann er den Baum an dieser definierten Länge skalieren.
10. **Zuweisen der Taxon-Namen**: Der Benutzer klickt hintereinander mit der rechten Maustaste auf jeden Blattknoten und gibt dann in einem Textfeld den entsprechenden Taxon-Namen ein. Alternativ kann der Benutzer horizontal geschriebene Taxon-Namen markieren, woraufhin das Programm eine Texterkennung versucht und den erkannten Text dem am nächsten gelegenen Blattknoten zuweist (nur für rechteckige Topologien).
11. **Wahl der Baumwurzel**: Das Programm wählt, ggf. benutzerassistiert, den inneren Knoten, relativ zu welchem der gewurzelte Newick-Ausdruck für den Baum berechnet wird.
12. **Konstruktion des Newick-Ausdrucks**: Das Programm berechnet den Newick-Ausdruck für den dargestellten Baum und zeigt ihn an. Der Benutzer kann ihn entweder in der Zwischenablage speichern oder ihn in eine Textdatei exportieren.

5.3 Ergebnisse

5.3.1 Einfluß verschiedener Faktoren auf den Digitalisierungsvorgang

Ein Bild, das einen einheitlich schwarzen, rechteckigen phylogenetischen Baum auf einem gleichmäßig hellen Hintergrund in genügend hoher Auflösung anzeigt, und in welchem keine Vordergrundelemente, z. B. Text, mit der Baumstruktur überlappen, bedarf nur wenig bis keiner Vorverarbeitung. Genügen dem Benutzer Zahlen als Taxon-Namen, kann so der gesamte Erkennungsprozeß der Baumtopologie innerhalb von Minuten vonstatten gehen. Im Allgemeinen

gibt es jedoch Bereiche in einem Eingabebild, die einer manuellen Vorverarbeitung bedürfen. Dies können Äste sein, die nicht vollständig von anderen Vordergrundelementen getrennt sind, beispielsweise Zahlen und Buchstaben. Für solche Fälle stellt *TSP* ein besonderes Werkzeug zur Verfügung, das vom Benutzer gemalten Vordergrund mit einem weißen "Rahmen" ausstattet.

Um Astlängen bestimmen zu können, muß *TSP* die Pfadlänge in Pixeln zwischen Verzweigungen auf dem skelettierten Vordergrund feststellen. Zusätzlich muß das Programm für Rechtecktopologien zuverlässig Knicke in den Ästen und die horizontalen Astanteile erkennen. Das funktioniert besser bei höher aufgelösten Bildstrukturen. Je besser sich die Äste im Originalbild und die im skelettierten Bild überdecken, umso genauer können die Astlängen berechnet werden.

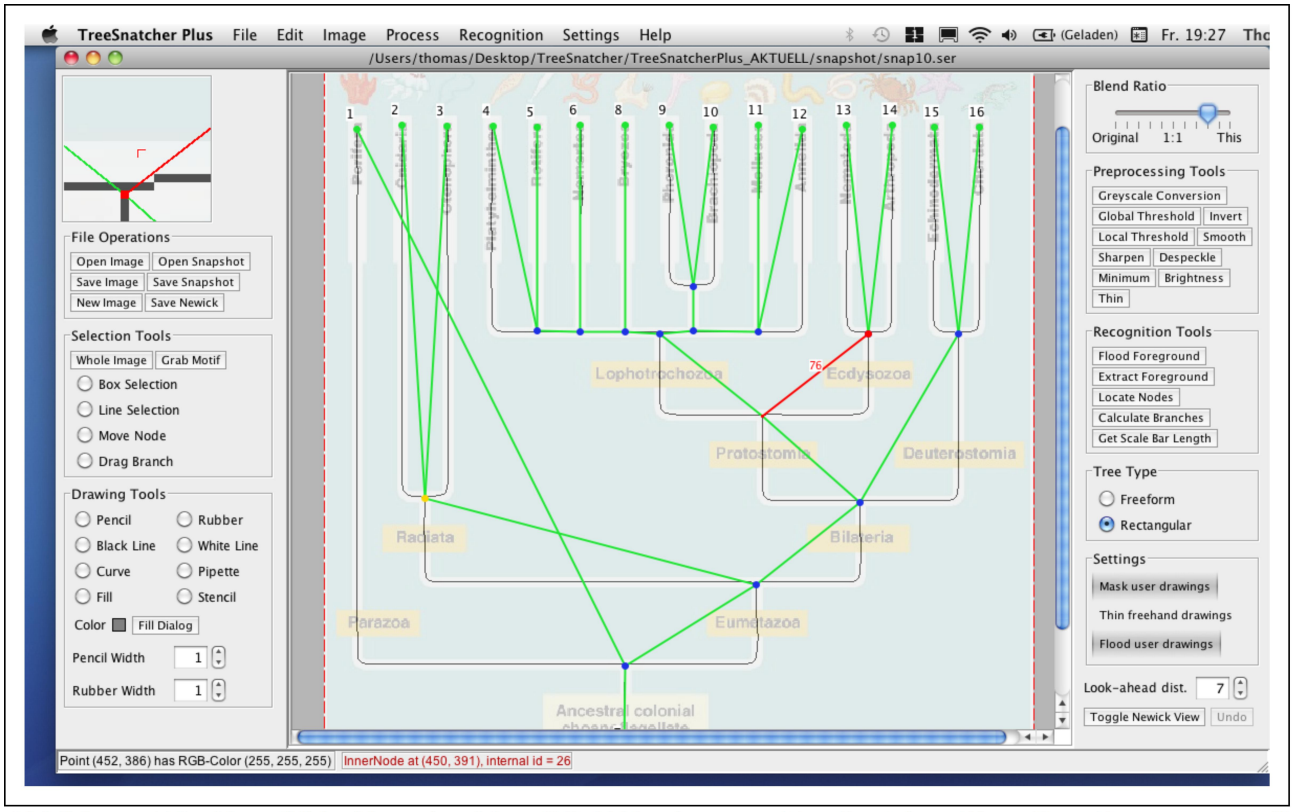


Abbildung 5.1: *Hauptbildschirm der Mac OS X-Version von TreeSnatcher Plus.* Das Programm stellt Originalbild und aktuellen Bearbeitungszustand überlagert dar. Nach Graustufenumwandlung, Binärisierung und Skelettierung wurde der Vordergrund an Stellen, wo Ast und Schrift überlappen, manuell korrigiert. Das Programm erkennt danach selbstständig Knoten und Astlängen und kann den zur Baumtopologie äquivalenten Newick-Ausdruck generieren.

Die für die gesamte Digitalisierung des dargestellten Baumes benötigte Zeit hängt von mehreren Faktoren ab, so beispielsweise Bildgröße, Bildqualität, Größe des Baumes und Komplexität der Topologie, aber auch davon, ob Astlängen berechnet werden sollen und Taxon-Namen eingetippt werden müssen, ob eine Vorverarbeitung notwendig und wie erfahren der Benutzer im Gebrauch von *TSP* ist.

5.3.2 Benchmarks

Um die Leistungsfähigkeit von *TreeSnatcher Plus* zu testen, wurden die 100 Bäume aus dem Benchmark-Set für *TreeRipper (*siehe Dateien im Ordner *Benchmarks)* von Joseph Hughes digitalisiert, zusätzlich neun andere Bilder mit nichtrechteckiger Topologie sowie ein weiteres Bild aus dem Satz von Benchmark-Bildern. Die Ergebnisse sind in der *Microsoft Excel*-Tabelle

12859_2011_5272_MOESM1_ESM.xls aus dem Ordner *Papers/BMCBioinformatics* hinterlegt. Die Taxon-Namen wurden in allen Fällen manuell eingegeben. Da *TSP* nicht vollautomatisch arbeitet, obliegt es dem Benutzer, die jeweiligen Vorverarbeitungs- und Analyseschritte zu initiieren. Abhängig von seinen Fähigkeiten schwankt daher die Zeit, die er zur Bearbeitung verschiedener Aufgaben benötigt. Für die Benchmark-Bäume lag die Bearbeitungszeit zwischen 30 und 1.800 Sekunden mit einem Durchschnitt von 165 Sekunden auf einem PC, der mit einem Core i7-960 Prozessor ausgestattet ist. Das Eintippen der Speziesnamen über die Tastatur dauerte durchschnittlich 4,4 Minuten und maximal 35 Minuten. Die kürzeste Baumdigitalisierung erfolgte innerhalb einer Minute, die längste dauerte 45 Minuten. Die Erkennung der Topologie eines Baumes mit mehr als 100 Taxa inkl. Eingabe der Speziesnamen (Ordner *1471-2148-6-93* im Benchmark-Set) nahm 17 Minuten in Anspruch. Die gleiche Topologie und die Speziesnamen lieferte das *TreeRipper*-Frontend innerhalb von fünf Minuten, allerdings mußten alle Namen manuell nachbearbeitet werden.

Zusätzlich wurden neun Abbildungen von Bäumen mit nichtrechteckiger Topologie (Verzeichnis *OtherTrees* im Ordner *Benchmarks*) analysiert. Wie zuvor wurden Topologie und Astlängen erkannt. Da *TSP* Astlängen in beliebig geformten Bäumen, jedoch nicht speziell in zirkulären oder polaren Bäumen messen kann, wurden die horizontalen Anteile der Äste in den Bäumen *„bustard“*, *„TreeofLife“* und *„Phylogenetic_Tree_of_Life“* mit dem Werkzeug „Line Selection Tool“ mit geringerer Genauigkeit vermessen. Im Fall der Abbildung *„vert_tree“* mußte der Baum manuell nachgezeichnet werden. Mit *TSP* können auch gekippte Darstellungen rechteckiger Bäume zuverlässig analysiert werden. Um zu testen, wie gut das gelingt, wurde *Bild 1471-2148-6-99-1* aus dem Benchmark-Set um 6° nach rechts rotiert (*„RotatedTree“* im Ordner *Benchmarks/OtherTrees*). Die Vorverarbeitungsschritte waren für das Originalbild und das rotierte Bild sehr ähnlich. Im rotierten Bild erkannte *TSP* 10 von 81 horizontalen Astanteilen nicht gegenüber 3 im Originalbild.

Die Korrektur überlappender Vordergrundstrukturen, die Eingabe von Astlängen und die Eingabe von Taxon-Namen stellen generell die zeitaufwendigsten Schritte dar. Für bestimmte Bäume werden ggf. mehrere Anläufe benötigt, bis eine Abfolge von Verarbeitungsschritten gefunden wird, die zu einer zuverlässigen Erkennung der Baumtopologie führt.

5.4 Diskussion

Mit Blick auf die Leistungsfähigkeit der heute verwendeten Bildanalyse- und OCR-Methoden sind für das gesetzte Ziel, in Bildern dargestellte phylogenetische Bäume zu digitalisieren und zu archivieren, zwei verschiedene Strategien denkbar: entweder ein vollautomatischer Ansatz, der auf einem fest definierten Baumtyp arbeitet und einen festgelegten Illustrationsstil fordert, oder ein halbautomatischer Ansatz, der für eine Vielzahl von Bäumen geeignet ist.

TreeRipper strebt eine vollautomatische Verarbeitung von Baumtopologien an. Um das zu erreichen, beschränkt sich das Programm auf Darstellungen rechteckiger Topologien, welche bestimmte eng umgrenzte Kriterien erfüllen müssen. Unter den diese Kriterien erfüllenden Abbildungen in seinem Benchmark-Set erkennt *TreeRipper* die Verzweigungsmuster von lediglich 32 % der Bäume. Die Abbildungen mußten zuvor allerdings keiner Vorverarbeitung unterzogen werden. Die Leistungsfähigkeit der vom Programm angebotenen Texterkennung ist vom Schrifttyp und der Auflösung der zu verarbeitenden Bilder abhängig. Der Verfasser hält eine vollautomatische Stapelverarbeitung (engl. „batch processing“) rechteckiger Baumtopologien angesichts dieser Ergebnisse bis auf weiteres für unrealistisch, obwohl Joseph Hughes angibt, sein Programm ausdrücklich für diesen Zweck geschaffen zu haben. Die Benutzung von *TreeRipper* empfiehlt sich

daher für die Digitalisierung von Darstellungen phylogenetischer Bäume, die den strikten Anforderungen des Programms genügen.

TreeSnatcher Plus kann jede Baumtopologie erkennen und damit archivieren, nimmt aber die Notwendigkeit einer manuellen Vorverarbeitung der Bilder in Kauf. Abhängig von der Komplexität der Abbildung erfordert es eine unter Umständen beträchtliche Benutzerinteraktion, kann dafür aber prinzipiell beliebige Bäume erfassen. Durchschnittlich benötigt *TSP* für die Digitalisierung eines Baumes unter drei Minuten. Die Leistung der verwendeten Schrifterkennung ist noch nicht gut, insbesondere ist sie auf horizontal angeordneten Text beschränkt. Die Taxon-Namen müssen daher in vielen Fällen nach wie vor manuell eingegeben bzw. nachkorrigiert werden. Obwohl *TreeSnatcher Plus* nicht vollkommen automatisch arbeitet, ist es geeignet, prinzipiell jede beliebige Darstellung einer Phylogenie zu digitalisieren und für eine zukünftige wissenschaftliche Anwendung zu erhalten.

Verbesserungspotential für *TreeSnatcher Plus* gibt es an mehreren Stellen. Hier wird nur auf die prominenteste hingewiesen. In der Literatur werden am häufigsten rechteckige Baumtopologien angetroffen. Es lohnt sich daher, das Programm speziell für diesen Baumtyp zu optimieren. Zwar kann der Benutzer dem Programm signalisieren, daß der zu erkennende Baum eine Rechtecktopologie besitzt, jedoch wird dieses Wissen bisher nur zur Bestimmung der Astlängen eingesetzt und trägt nicht zur Erkennung der Baumtopologie bei. Die nötige Erweiterung des Programms ist mit verhältnismäßig geringem Aufwand möglich. Danach könnte *TSP* überlappende Vordergrundregionen bei rechteckigen Baumtopologien weitgehend autonom korrigieren, da es die Ausrichtungen der Äste kennen würde. Ultrametrische Bäume würde es ebenso zuverlässig bearbeiten. Die Leistung der Schrifterkennung für rechteckige Bäume könnte stark verbessert werden, würde Vorwissen zu der zu verarbeitenden Phylogenie genutzt. Dem Programm wären nämlich in diesem Fall die wahrscheinlichen Positionen und die Ausrichtung der Taxon-Namen im Bild bekannt. Dieses Wissen würde *TSP* in die Lage versetzen, die Taxon-Namen weitgehend automatisch zu finden und diese den passenden Blattknoten zuzuweisen.

TreeSnatcher Plus wurde eigens zu dem Zweck entwickelt, Abbildungen phylogenetischer Bäume zu digitalisieren, deren Datengrundlage nicht mehr zu beschaffen ist. Eine höhere Akzeptanz verbesserter Minimalstandards für das Berichtswesen in der phylogenetischen Forschung würde jedoch garantieren, daß phylogenetische Daten in zukünftigen Veröffentlichungen maschinenlesbar sind und damit auch anderen Forschern zur Verfügung stehen. Es ist seit Jahren verbindlich, daß DNA-Sequenzen auch elektronisch in maschinenlesbarer Form in einem der akzeptierten Formate veröffentlicht werden. Daher erscheint es wünschenswert, daß sich die Herausgeber wissenschaftlicher Fachzeitschriften auf einen vergleichbaren Standard für Phylogenien einigen.

6 Ein erweiterter Robinson-Foulds-Abstand für ausschließlich signifikante Splits

6.1 Einleitung

Einer der entscheidenden Schritte in der Stammbaumforschung ist der Vergleich von phylogenetischen Bäumen. Bei entsprechenden Forschungsvorhaben können Bäume, die nicht deckungsgleich sind und mit verschiedenen Methoden hergestellt wurden, im Vordergrund stehen, aber auch die Frage, wie leistungsfähig verschiedene Methoden der Stammbaumrekonstruktion sind. Generell ist zum Vergleich zweier Bäume die Definition eines Abstands zwischen ihnen erforderlich. Die Distanz zwischen zwei Phylogenien manifestiert sich in der Zahl der Unterschiede zwischen ihnen. Zwei intrinsische Eigenschaften von Bäumen können verglichen werden: Topologie und Astlängen. Einige der am häufigsten verwendeten Distanzmaße verwenden ausschließlich topologische Informationen, so z. B. die *Robinson-Foulds-Metrik* (Robinson und Foulds, 1978; Robinson und Foulds, 1981), *Nearest-Neighbor-Interchange* (Waterman und Smith, 1978), *Quartet Distance* (Estabrook et al., 1985) und *Subtree Pruning and Regrafting* (Hein, 1990). Im Jahr 2007 haben Pattengale und Kollegen eine randomisierte Approximation der ursprünglichen Robinson-Foulds-Metrik („RF-Distanz“) veröffentlicht (Pattengale et al., 2007), die schneller ist als der ältere Ansatz. Eine auf der Topologie basierende Distanzmethode vermag es, in zwei gegebenen Bäumen die Bereiche zu quantifizieren, die einander widersprechen. Der *gewichtete Robinson-Foulds-Abstand* („weighted Robinson-Foulds“, RF_W), die *geodäsische Distanz* (Billera et al., 2001; Kupczok et al., 2008) und die *Branch-Score-Distanz* (Kuhner und Felsenstein, 1994) benutzen neben der Topologie zusätzlich noch die Astlängen. Ein Distanzmaß, das Astlängen berücksichtigt, ist besonders nützlich, wenn man Bäume mit vielen unterschiedlichen Astlängen vergleicht, sofern sich die evolutionären Raten in den Ästen nicht zu sehr unterscheiden.

Phylogenien können nicht nur anhand der intrinsischen Eigenschaften Topologie und Astlängen miteinander verglichen werden, sondern auch anhand der extrinsischen Eigenschaft *Bootstrap-Support-Wert* (ab jetzt mit BSW abgekürzt) an ihren Ästen. Ein BSW an einem Ast in einer rekonstruierten Phylogenie drückt, umgangssprachlich formuliert, aus, wie sehr man darauf vertrauen kann, daß dieser sich hinsichtlich seiner Position in der wahren Phylogenie an der richtigen Stelle befindet. Man könnte einen Ast mit einem niedrigen BSW, der als einziger zwischen zwei zu vergleichenden Phylogenien unterscheidet, als nicht-signifikant werten und ihn daher ignorieren. Konsequenz daraus wäre dann, dass man die Phylogenien als praktisch gleich ansieht.

Der Begriff des BSW hat allerdings durchaus ein mathematisches Fundament. 1985 schlug Joseph Felsenstein (Felsenstein, 1985) vor, die von Efron eingeführte Technik namens „Bootstrap“ (Efron, 1979) auf das Baumproblem zu übertragen. Seitdem wurden viele Anstrengungen unternommen, die daraus resultierenden Anwendungen sowie verschiedenste statistische Eigenschaften zu untersuchen. Susko und Mitautoren (Susko et al., 2006) verwendeten BSW, um die Frage zu beantworten, inwieweit ein Split in einem Baum die Taxa korrekt an beiden Seiten des Splits plaziert hat. Berry und Gascuel beschäftigten sich mit der allgemeinen Frage, wie man einen Bootstrap-Baum (Berry und Gascuel, 1996) richtig interpretiert. Sie untersuchten in diesem Zusammenhang Fehlermaße, die auf eine Generalisierung der *RF-Distanz* zurückgehen. Ihr Ziel war es, die Divergenz zwischen wahrer und geschätzter Phylogenie zu quantifizieren. Efron et al. konnten zeigen, daß das Bootstrap-Verfahren auf Bäumen zwar nicht statistisch verzerrt (engl. „biased“) ist (Efron et al., 1996), es aber unter den Aspekten Hypothesentesten und

Konfidenzniveaus die damals geltenden Standards noch nicht erreichte. Cardona und Kollegen haben eine *RF-Distanz* zwischen phylogenetischen Netzwerken erarbeitet (Cardona et al. 2009; Asano et al., 2012). Die ursprüngliche *RF-Distanz* kann keine ungewurzelten Phylogenien mit gewurzelten vergleichen. Um diese Einschränkung aufzuheben, haben Gorecki und Eulenstein eine weitere Variante der *RF-Distanz*, *urRF*, entwickelt (Gorecki et al., 2012).

Verschiedene Programmpakete wurden veröffentlicht, die mehrere gängige Methoden für den Phylogenievergleich bündeln (Bogdanovicz et al., 2012; Puigbo et al., 2007). In ihrer Dissertation (Kupczok, 2010) erörtert Anne Kupczok die Eigenschaften verschiedener Distanzmaße und stellt Methoden zur Vereinigung von Phylogenien vor. Joseph Felsenstein bietet in seinem Buch „Inferring Phylogenies“ (Felsenstein, 2004) einen historischen Überblick über die Vergleichsmetriken für Phylogenien.

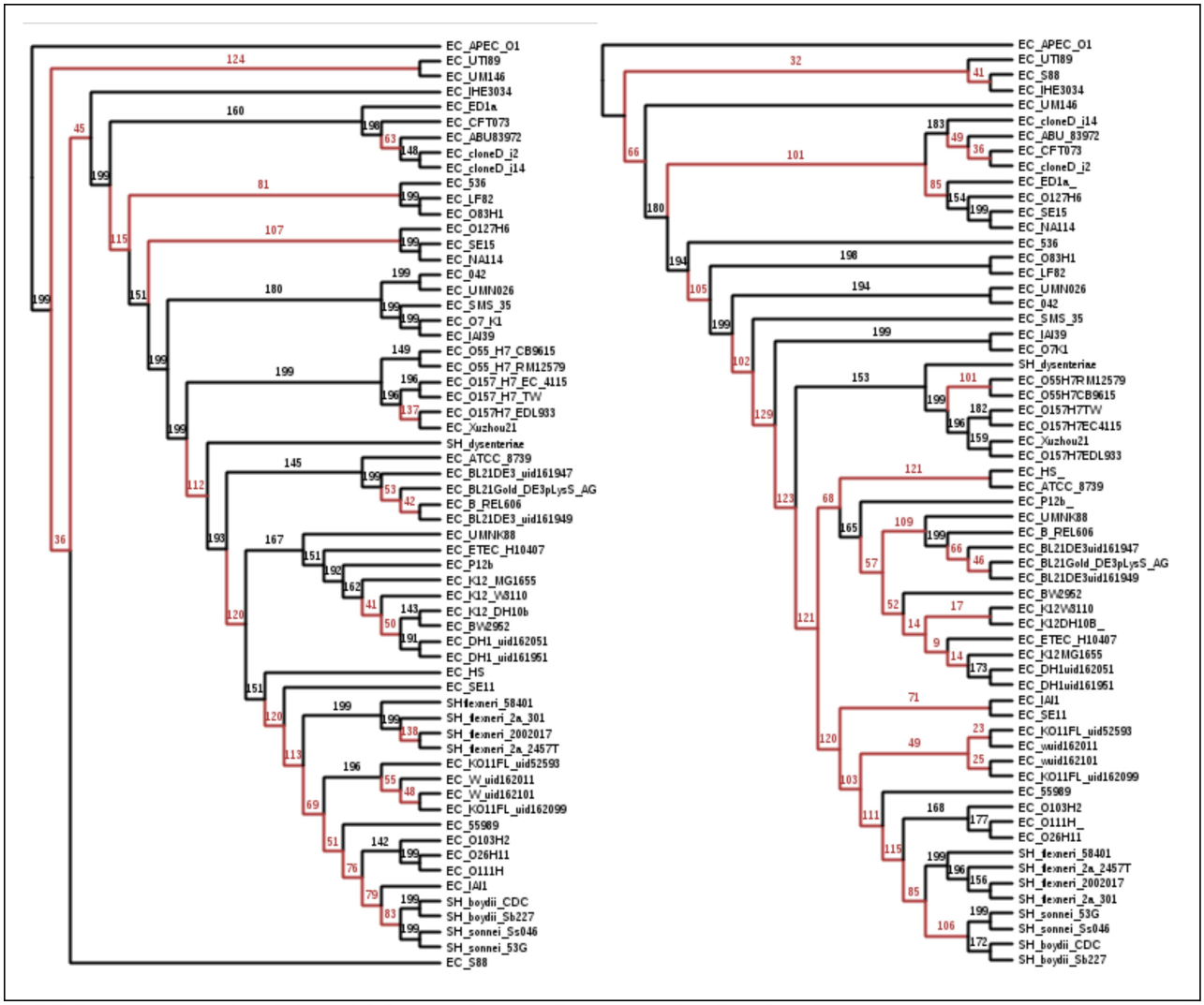


Abbildung 6.1*:* Zwei auf den gleichen 29 konservierten Genfamilien beruhende ungewurzelte, bifurzierende *E. coli*-Phylogenien mit Bootstrap-Werten. Die farblich hervorgehobenen Splits können aufgrund ihrer Bootstrap-Werte unterhalb 70 % als lediglich beschränkt vertrauenswürdig angesehen werden. Zur Berechnung der *ssRF-Distanz* werden solche Äste, abhängig von der Höhe der Bootstrap-Schwelle, aus dem Baum entfernt, wodurch die an ihnen befindlichen Teilbäume zu Multifurkationen zusammengezogen werden (vgl. **Abb. 6.3**).

Im Rahmen dieses Kapitels wird eine vom Verfasser realisierte Erweiterung der *RF-Distanz* vorgestellt, *ssRF*, die zwischen signifikanten und nichtsignifikanten Ästen unterscheidet. Der Algorithmus benutzt entweder den BSW oder alternativ eine Bayesian Posterior Probability (BPP) als Maß für die Signifikanz eines Asts. Er reduziert zunächst zwei ungewurzelte, bifurzierende Bäume mit BSW („Konfidenzwerten“) an den Ästen auf ihre gemeinsamen Taxa. Dann werden alle Splits unterhalb der initial vom Benutzer (ggf. auch subjektiv) festgelegten Signifikanzschwelle aus der Menge aller Splits entfernt. Für die Newick-Repräsentation eines Baumes bedeutet das, daß Teilbäume in Multifurkationen zusammengezogen werden. Die *ssRF-Distanz* zwischen den Bäumen ist die Zahl der Splits im ersten Baum, die mit dem zweiten Baum inkompatibel sind, zuzüglich der Zahl der Splits im zweiten Baum, die mit dem ersten Baum inkompatibel sind. Ein Split *S* teilt die Taxa im Baum in zwei disjunkte Mengen S_1 und S_2 auf. Zwei Splits *A*, *B* sind kompatibel, wenn für sie gilt: $(A_1 \cap B_1 = \emptyset) \vee (A_1 \cap B_2 = \emptyset) \vee (A_2 \cap B_1 = \emptyset) \vee (A_2 \cap B_2 = \emptyset)$.

6.2 Material und Methoden

Die für das in diesem Kapitel vorgestellte Projekt benötigten Dateien befinden sich innerhalb der Ordnerstruktur im Projektordner *Distanzmaß*.

6.2.1 Herleitung des erweiterten Distanzmaßes

Die vom Verfasser als *ssRF* bezeichnete Distanz stellt eine Erweiterung des *Robinson-Foulds-Abstands* (Robinson und Foulds, 1979; Robinson und Foulds, 1981) dar. 1981 konnten Robinson und Foulds beweisen, daß ihr Distanzmaß eine Metrik[2] ist.

Robinson-Foulds-Abstand

Die *RF-Distanz* ist die symmetrische Differenz zwischen zwei Phylogenien, betrachtet aus einer mengentheoretischen Perspektive. Dabei handelt es sich um die Summation der Zahl der Splits in jeweils einem der Bäume, die aber nicht im jeweils anderen Baum vorkommen. Die Distanz zwischen zwei Bäumen T_1 und T_2 ist dann wie folgt definiert:

$$RF(T_1, T_2) := |(T_1 \cup T_2)| - |(T_1 \cap T_2)|$$

Gewichteter Robinson-Foulds-Abstand

1979 veröffentlichten Robinson und Foulds eine verbesserte Variante ihres Distanzmaßes, RF_W, die als *gewichteter RF-Abstand* („weighted Robinson-Foulds“) bezeichnet wird (Robinson und Foulds, 1979). Werden Astlängen als Gewichte benutzt, kann RF_W als die Summe der absoluten Differenzen der Astlängen über alle Äste der Topologie ausgedrückt werden.

2 Ob ssRF ebenfalls eine Metrik ist, wurde im Rahmen dieser Arbeit nicht untersucht.

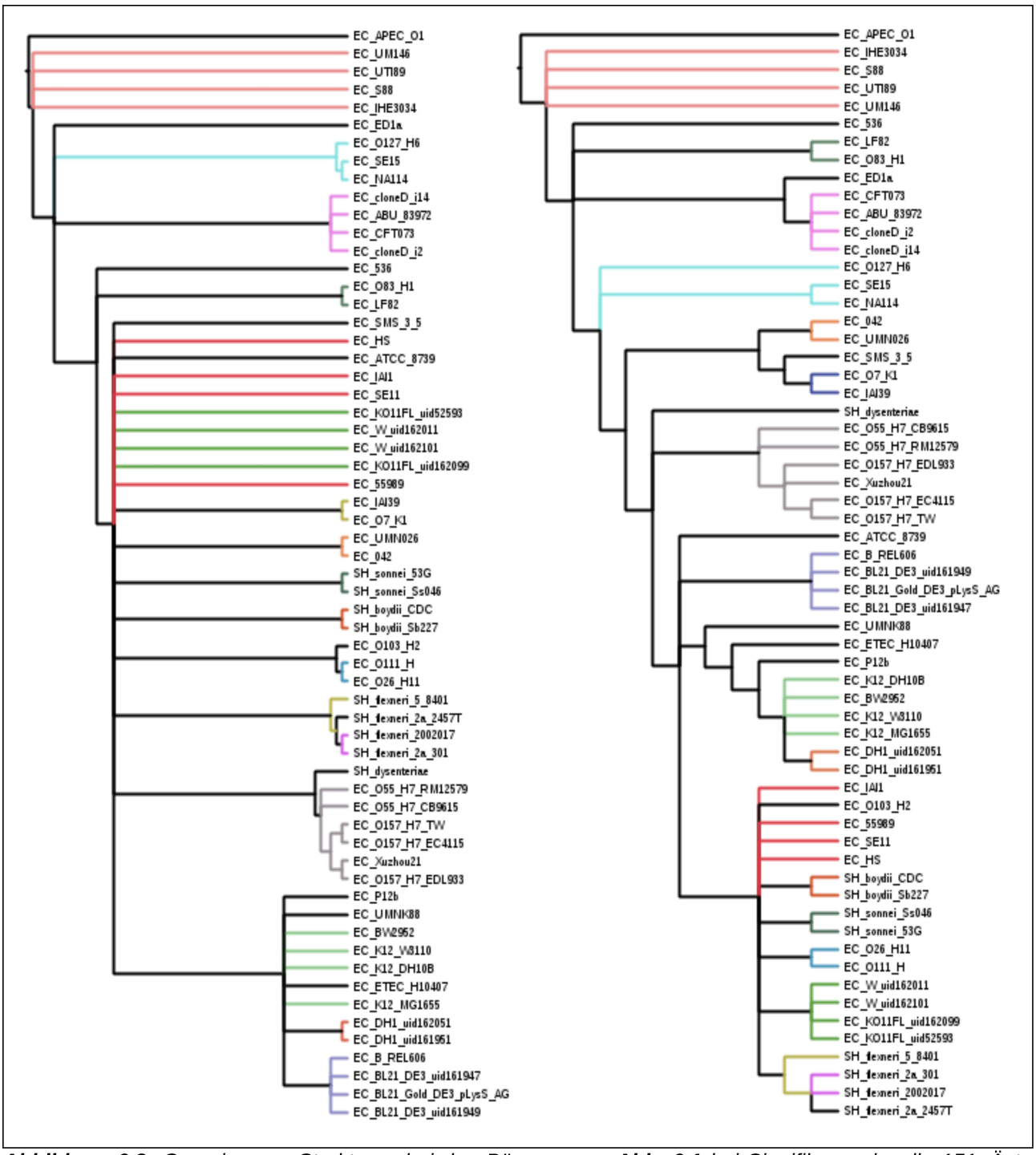


Abbildung 6.2: *Gemeinsame Strukturen bei den Bäumen aus* ***Abb. 6.1*** *bei Signifikanzschwelle 151.* Äste mit geringerem Bootstrap-Support wurden aus beiden Bäumen entfernt, und betroffene Teilbäume wurden in Multifurkationen zusammengezogen, wodurch die statistische Unsicherheit hinsichtlich des Verzweigungsmusters abgebildet wird. Das Bild zeigt unmittelbar sichtbare gleiche Regionen in den Bäumen. Der *RF-Abstand* zwischen den Bäumen beträgt 19, der *ssRF-Abstand* 2. Sie unterscheiden sich durch zwei signifikante Splits und erscheinen subjektiv ähnlicher, als der *RF-Abstand* es suggeriert.

Für die Bäume T_1, T_2, die Menge w_i aller positiven Gewichte an den Ästen von T_i, die Menge $\sum(T_i)$ aller Splits in T_i und das Gewicht $w_i(A|B)$ des Asts, der mit dem Split $A|B$, $i = 1, 2$ korrespondiert, ist die Metrik definiert über

$$RF_w(T_1, T_2) := \sum_{A|B \in \sum(T_1) \cup \sum(T_2)} |w_1(A|B) - w_2(A|B)|$$

Die gewöhnliche *RF-Distanz* tritt bei dieser Definition als Spezialfall auf, wenn alle Äste das gleiche Gewicht besitzen. Es ist möglich, eine *RF-Distanz* zu definieren, die Bootstrap-Werte anstelle von Astlängen als Gewichte benutzt.

Robinson-Foulds-Abstand für ausschließlich signifikante Splits

Die *ssRF-Distanz* zwischen zwei Bäumen ist die Anzahl Splits des ersten Baumes, die mit allen Splits im zweiten Baum inkompatibel sind, zuzüglich der Anzahl Splits des zweiten Baumes, die mit allen Splits im ersten Baum inkompatibel sind. Für die Bäume T_1 und T_2, die Splits S_1 von T_1 und S_2 von T_2 sowie $X_1 :=$ $S_1 \cup$ *{ Split s | s ist kompatibel mit* S_1 *}* und S_1 := *{ Splits in* T_1 *}* und X_2 analog definiert, kann die Distanz über die Gleichung

$ssRF(T_1, T_2) := |S_1 - X_2| \cup |S_2 - X_1|$

ausgedrückt werden. Die *ssRF-Distanz* berücksichtigt sowohl die Topologie als auch die Signifikanz der Äste eines Baumes. Dabei benutzt sie den Bootstrap-Support als Maß für das Vertrauen in die Äste: solche oberhalb des gesetzten Schwellenwerts werden als signifikant bewertet, wohingegen solche unter dem Schwellenwert in einem Vergleich zweier Bäume ignoriert werden. Für die Newick-Repräsentation eines Baumes bedeutet dies, daß von einem Ast ausgehende Teilbäume, welcher einen zu niedrigen Bootstrap-Support besitzt, in Multifurkationen zusammengezogen werden.

6.2.2 Vergleich zweier E. coli-Phylogenien auf konservierten Genen

Um die Praxistauglichkeit der entwickelten Distanzmethode zu testen, wurden zwei Phylogenien mit jeweils 61 *Escherichia coli*-Stämmen miteinander verglichen, die zuvor mit einem Maximum-Likelihood-Ansatz auf einem Alignment aus 29 konkatenierten konservierten Genfamilien (COGs) inferiert worden waren. Der in **Abbildung 6.1** links dargestellte Baum zeigt den *E. coli*-Teil einer Phylogenie mit ursprünglich 61 *E. coli*- und 25 *Salmonellen*-Stämmen (Datei *outtree_cor* im Verzeichnis *Kapitel_Phylogenien/iTol/iTol_29COGs/EC_iTol_29COGs/MMN/data/alignments/gblocks_reverse_concatenated*), der nach der in **Kapitel 8, Abschnitt 2.2.4** dargestellten Methode konstruiert wurde. Der rechte Baum basiert auf COGs mit *E. coli*- und *Salmonellen*-Genen (Datei *outtree_cor* im Ordner *Kapitel_Phylogenien/iTol/iTol_29COGs/ECSA_iTol_29COGs/MMN/data/alignments/gblocks_reverse/concatenated)*. Seine Konstruktion wird an der gleichen Stelle beschrieben. Die Beschreibungen beziehen sich allerdings auf die dort abgedruckten Bäume, die auf 26 COGs basieren.

6.2.3 Algorithmus zur Berechnung des ssRF-Abstands

Das Perl-Programm `computeSSRF.pl` (Verzeichnis *Distanzmaß*) wurde entwickelt, um den *ssRF-Abstand* zwischen zwei ungewurzelten, bifurzierenden phylogenetischen Bäumen mit BSW an allen Ästen zu berechnen. Es setzt als Signifikanzschwellen alle unterschiedlichen BSW ein, die in den beiden zu vergleichenden Bäumen existieren, und liefert eine Tabelle der *ssRF*-Distanzen zwischen ihnen pro Signifikanzschwelle zurück.

Beim Aufruf von einer *Linux*-Shell nimmt das Programm zwei separate Dateien als Eingabe entgegen (Schalter `-i` und `-j`), von der jede genau einen Newick-Ausdruck enthalten muß. Dieser wird abgewiesen, und das Programm wird vorzeitig beendet, falls er nicht geeignet ist, d. h. er entweder keine BSW enthält oder nicht wohldefiniert ist. Im nächsten Schritt werden die Taxa aus

dem ersten Baum, T_1, entfernt, die nicht auch im zweiten Baum, T_2, zu finden sind. Das kann bedenkenlos gemacht werden, da Taxa, welche nur in einem Baum vorhanden sind, nicht zur *ssRF-Distanz* beitragen. Zur Ermittlung der Menge aller Signifikanzschwellen wird eine Liste aller BSW in beiden Bäumen erstellt. Danach werden für jede Schwelle *b* die Split-Zerlegungen für T_1 und T_2 bestimmt[3]. Aus dieser Menge werden die Splits kleiner *b* entfernt. Die verbleibenden Splits in T_1 werden als S_1 bezeichnet, die verbleibenden Splits in T_2 als S_2. Dann wird die Menge C_1 der Splits in T_1, die kompatibel mit T_2 sind, berechnet und ebenso die Menge C_2 der Splits in T_2, die kompatibel mit T_1 sind. Als letzter Schritt wird der Abstand ss$RF(T_1,T_2)$ über $|S_1| - |C_1| + |S_2| - |C_2|$ berechnet. Abschließend wird die Ergebnistabelle ausgegeben.

Das Programm gibt neben der *ssRF-Distanz* auch die *RF-Distanz (RF)*, die *gewichtete RF-Distanz* (RF_W), welche BSW als Gewichte benutzt, die Zahl der Splits sowie die Zahl kompatibler und gemeinsamer Splits aus (vgl. **Tabelle 6.1**).

Bei der Berechnung der RF_W-Distanz folgt *computeSSRF* standardmäßig der Definition von Shi und Kollegen, wie diese sie in ihrer Arbeit (Shi et al., 2010) verwendet haben. Sie entspricht in zwei Punkten nicht der ursprünglichen Definition von Robinson und Foulds (Robinson und Foulds, 1979). Robinson und Foulds addieren zusätzlich noch einen weiteren Term, nämlich die Summe der BSW für die Splits, die nur in einem der Bäume vorhanden sind, aber nicht im anderen, und andersherum. Für das vorliegende Programm ist diese ursprüngliche Definition jedoch nicht sinnvoll, denn es soll $RF_W = 0$ liefern, wenn zwei identische, bifurzierende Phylogenien sich nur in den (signifikanten) Gewichten an den Ästen unterscheiden.

Algorithmus 6.1: Berechnung des ssRF-Abstands mit dem Programm *computeSSRF*

Eingabe: Zwei Dateien mit jeweils einem Newick-Ausdruck inkl. Bootstrap-Werten
Ausgabe: Tabelle mit ssRF-Abständen für alle Signifikanzschwellen

Lies die Bäume T_1 und T_2 aus den Newick-Ausdrücken
Entferne aus T_1 die Taxa, die nicht in T_2 sind
Entferne aus T_2 die Taxa, die nicht in T_1 sind
Bestimme die Menge *B* aller individuellen Bootstrap-Werte in T_1 und T_2

Foreach *b* in *B* /* *alle Bootstrap-Werte bzw. Signifikanzschwellen* */
Berechne die Splits-Zerlegung $\overline{S}_1$ *für* T_1
Berechne die Splits-Zerlegung $\overline{S}_2$ für T_2
Entferne aus $\overline{S}_1$ die Splits mit einem Bootstrap-Wert kleiner *b*
Entferne aus $\overline{S}_2$ die Splits mit einem Bootstrap-Wert kleiner *b*
S_1 := Menge der verbliebenen Splits in T_1
S_2 := Menge der verbliebenen Splits in T_2

Berechne die Menge C_1 der Splits in T_1, die mit (allen Splits in) T_2 kompatibel sind
Berechne die Menge C_2 der Splits in T_2, die mit (allen Splits in) T_1 kompatibel sind

Distanz := $ssRF(T_1, T_2) = |S_1| - |C_1| + |S_2| - |C_2|$
End
Gib die Tabelle der Distanzen pro Signifikanzschwelle aus

Algorithmus 6.1: *Berechnung des ssRF-Abstands mit computeSSRF*

Sollte die Addition des Terms ausdrücklich erwünscht sein, kann das durch Hinzufügen des Schalters `-a` erzwungen werden. Ferner nehmen Shi und Kollegen im Gegensatz zu Robinson und

3 Das Programm für die Zerlegung der Splits wurde von Christian Eßer für seine Dissertation (Eßer, 2010) entwickelt und vom Verfasser für die Benutzung im Programm *computeSSRF* modifiziert (siehe Kapitel 4).

Foulds eine Normierung vor, indem sie RF_W noch durch die Summe der Zahl der Splits ausschließlich im ersten Baum zuzüglich der Zahl der Splits im zweiten Baum teilen (falls `-a` gesetzt ist, wird zu dieser Summe noch die Zahl der Splits, die die Bäume gemeinsam haben, addiert). Mit dem Schalter `-n` kann diese Normierung unterbunden werden.

Wurde der Schalter `-w` gesetzt und dabei ein Verzeichnis angegeben, speichert das Programm dort zusätzliche Daten für jede Bootstrap-Schwelle, die während des Programmlaufs generiert werden: die Splits pro Baum, die Splits in Baum 1, die zu Baum 2 kompatibel sind und andersherum, inkompatible Splits sowie die Splits, die beide Bäume gemeinsam haben. Der Name der Dateien beginnt immer mit einem der Präfixes *„tree1_“* oder *„tree2_“*, dann folgt eine Zeichenkette, die die Art der enthaltenen Daten angibt, gefolgt von dem Suffix *„_bs_*140.*txt“*, falls 140 die Bootstrap-Schwelle ist, für die diese Daten erzeugt wurden. **Algorithmus 6.1** faßt den Ablauf zusammen.

6.3 Ergebnisse

Für die beiden in **Abbildung 6.1** gezeigten *Escherichia coli*-Phylogenien wurden die *ssRF-Distanzen* mit dem Programm *computeSSRF* mit den Standardeinstellungen berechnet. Der Aufruf erfolgte über `perl computeSSRF.pl -i ec_iTol_alt.tre -j ec_iTol_neu.tre -w Lauf` innerhalb des Ordners *Distanzmaß*.

Abbildung 6.2 visualisiert die Verzweigungsmuster der beiden Phylogenien bei Bootstrap-Schwelle 151. Die Phylogenien weisen an mehreren Stellen unmittelbar sichtbare strukturelle Gemeinsamkeiten auf, die in der Abbildung jeweils in der gleichen Farbe hervorgehoben sind. Die in **Abbildung 6.1** rot eingefärbten Äste und weitere Äste mit einem niedrigeren BSW als 151 sind in den Bäumen nicht mehr enthalten. Die betroffenen Teilbäume wurden an Multifurkationen zusammengezogen. Für den initialen Zustand, d. h., alle Splits sind vorhanden, ist der *RF-Abstand* zwischen den Bäumen 62, der *ssRF-Abstand* ebenfalls. Mit steigender Signifikanzschwelle nimmt die Zahl der Splits in beiden Bäumen ab, ebenso die Zahl gemeinsamer Splits. Aus diesen Größen ergibt sich direkt die *RF-Distanz*, die entsprechend sinkt. Die Zahl kompatibler Splits ist von der Zahl der Multifurkationen in den Bäumen abhängig. Da letztere Zahl sowohl zwischen den Bäumen als auch bei unterschiedlichen Schwellen schwankt, verhält sie sich monoton, jedoch nicht streng monoton[4]. Insgesamt sinkt die Zahl kompatibler Splits jedoch, bis sie am Ende der noch in den Bäumen vorhandenen Zahl der Splits gleicht. Die *gewichtete RF-Distanz* steigt im Beispiel, von Ausnahmen abgesehen, insgesamt an, da mit höherer Bootstrap-Schwelle auch immer höhere BSW summiert werden, die aber an einer abnehmenden Zahl von Splits normiert werden. Ein Vergleich mit den anderen Distanzmaßen bietet sich nicht an. Wie RF_W in einem konkreten Anwendungsfall berechnet wird, muß anhand der Anforderungen entschieden werden.

Im Projektordner befindet sich noch der Unterordner *OhneNormierung*, der die Ergebnisse für einen Programmlauf auf den gleichen Bäumen ohne Normierung von RF_W enthält. Es fällt auf, daß der *gewichtete RF-Abstand* in diesem Fall - wie zu erwarten – fällt.

Die *ssRF-Distanz* zwischen den Bäumen sinkt mit steigender Schwelle monoton und schneller als die *RF-Distanz*, da sie zusätzlich von der Entfernung nichtsignifikanter Splits beeinflußt ist (siehe **Abb. 6.3**). Bei Bootstrap-Schwelle 139 (ca. 70 %) ist die *ssRF-Distanz* zwischen den Bäumen 4. Mit anderen Worten, die Bäume unterscheiden sich durch vier signifikante Splits, wenn man alle Äste wegläßt, „denen man nicht vertraut“[5]. Es bietet sich dabei die Interpretation an, daß die Bäume

4 Diese Vermutung wurde nicht mathematisch bewiesen.
5 Ein Ast mit einem BSW von mindestens 70 % wird in phylogenetischen Studien gewöhnlich als signifikant angesehen.

zwar grundsätzlich verschieden sind, jedoch in geringerem Maße, als es die *symmetrische Differenz/RF-Distanz* suggeriert. Bei Schwelle 151 (vgl. **Tab. 6.1**) unterscheiden zwei Äste zwischen den Bäumen, und ab der Schwelle 153 gibt es keinen Split mehr, der zwischen beiden Bäumen unterscheidet.

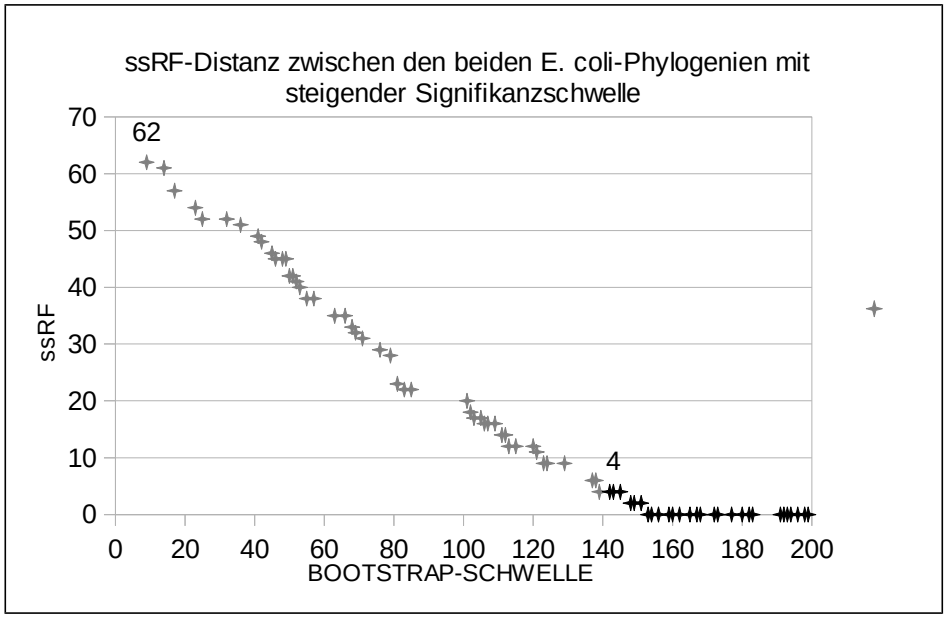


Abbildung 6.3: *ssRF-Abstände pro Signifikanzschwelle zwischen den verglichenen Phylogenien.* Ab Schwelle 151 sind alle Splits (im Diagramm schwarz) signifikant. Der *ssRF-Abstand* fällt mit steigender Schwelle monoton.

6.4 Diskussion

In seinem Buch „Inferring Phylogenies“ (Felsenstein, 2004, S. 534) führt Joseph Felsenstein aus, daß es nur beschränkt sinnvoll ist, nach signifikanter Ähnlichkeit zweier Bäume oder nach signifikanten Unterschieden zu fragen, wenn dabei die zugrunde liegenden Daten ignoriert werden. Noch weniger sinnvoll sei die bloße Frage nach Unterschieden, denn sobald es einen Unterschied gibt, sind die Bäume nicht gleich. Auch ist es zwar möglich, für zwei Bäume festzustellen, daß sie einander stärker ähneln, als bloßer Zufall es erklären könnte, aber dieses Wissen ist selten hilfreich. Es können nämlich gemeinsame Merkmale in den Bäumen dafür verantwortlich sein, die in biologischer Hinsicht irrelevant sind. Felsenstein schränkt allerdings ein, daß es trotzdem sinnvoll sein kann, explizit nach den signifikanten Unterschieden zwischen zwei Bäumen zu fragen, wenn das Hauptaugenmerk bei einer phylogenetischen Analyse auf den zugrunde liegenden Daten liegt. Bisher wurde aber kein Distanzmaß vorgestellt, das einen statistischen Zusammenhang zwischen Merkmalen in einem Baum und den Daten, die ihn erzeugt haben, herzustellen vermag. Ebensowenig werden persönliche Präferenzen unterstützt, wie etwa, ob der Mensch eine bestimmte Eigenheit eines Baumes als wichtig ansieht.

Das in diesem Kapitel vorgestellte Distanzmaß *ssRF* füllt dieses Lücke ansatzweise, steuert es doch bei der Distanzberechnung über eine Signifikanzschwelle für BSW oder BPP, welche Äste berücksichtigt werden und welche nicht. Zwar beruhen Bootstrap-Werte auf einem soliden statistischen Unterbau, die Wahl der Schwelle vollzieht sich dennoch subjektiv und unterliegt damit der Präferenz des Benutzers.

Im Fall der hier untersuchten *E. coli*-Phylogenien wäre es verlockend, die Bäume bei Signifikanzschwelle 153 als „praktisch gleich“ oder „signifikant ähnlich“ zu beschreiben, da es keinen Split mehr gibt, der zwischen ihnen unterscheidet. Allerdings ließe man dabei die zu diesem Zeitpunkt eingeführten Multifurkationen außer acht, die darauf hinweisen, daß sich die Verzweigungsmuster der wahren Bäume möglicherweise beträchtlich voneinander unterscheiden. Dieser vermeintliche Nachteil des Verfahrens ist allerdings tatsächlich ein Vorteil: bei der Konstruktion der Phylogenie werden keine signifikanten Daten unterdrückt, und die statistischen Unsicherheiten in ihnen werden über das veränderte Verzweigungsmuster unmittelbar auf die Phylogenie abgebildet.

Für welchen Anwendungsfall ergibt sich durch die Benutzung des Distanzmaßes *ssRF* ein Vorteil? In diesem Projekt wurde *ssRF* eingesetzt, um *Escherichia coli*- oder *Salmonellen*-Phylogenien miteinander zu vergleichen, die mit einem Maximum-Likelihood-, einem Bayesianischen Ansatz oder Neighbor-Joining konstruiert wurden und entweder auf ausgewählten konservierten Genen oder auf universellen orthologen Genfamilien basieren (vgl. diesbezüglich **Kapitel 7** sowie **Kapitel 8**). Mehrere Phylogenien wiesen hohe strukturelle Gemeinsamkeiten auf. Die Vermutung lag nahe, daß sich die Bäume ab der als signifikant empfundenen Schwelle von 70 % nicht mehr voneinander unterscheiden, was durch ihren *ssRF-Abstand* untermauert werden sollte. Zwar deckte sich dessen Höhe mit dem subjektiven Empfinden der Ähnlichkeit der Bäume, Unterschiede waren dennoch in jedem Fall vorhanden. Obwohl im Rahmen dieser Arbeit nicht gänzlich zufriedenstellend, ist die Verwendung der *ssRF-Distanz* in vergleichbaren phylogenetischen Studien trotzdem empfehlenswert, bietet sie doch den Vorteil, als unbedeutend empfundene Äste ausblenden[6] und verfolgen zu können, wie robust eine Phylogenie bei steigender Bootstrap-Schwelle ist.

Gegenüber der symmetrischen Differenz, welche für zwei Phylogenien, die viele Strukturen, aber keine Partitionen gemeinsam haben, maximal groß werden kann (Felsenstein, 2004, S. 529), entspricht die *ssRF-Distanz* eher dem subjektiven Ähnlichkeitsempfinden des Menschen. Sofern die zu vergleichenden Phylogenien sehr unterschiedlich lange Äste besitzen, sollte zusätzlich der *gewichtete RF-Abstand*, der Astlängen als Gewichte benutzt, berechnet werden. Für andere Anwendungsszenarien sollte eines der in der Einleitung angesprochenen Distanzmaße ausgewählt werden.

6 Man könnte ssRF dergestalt benutzen, daß man nicht ausschließlich (automatisch generierte) Bootstrap-Gewichte verwendet, sondern die Gewichte für Äste, von denen man überzeugt ist, manuell sehr hoch setzt. So könnte man einen Teilbaum, dessen Verzweigungsmuster als gesichert gilt, konservieren. Diese Praxis müßte selbstverständlich begründbar sein.

Bootstrap-Schwelle	Zahl der Splits in Baum 1	Zahl der Splits in Baum 2	Zahl gemeinsamer Splits	Zahl kompatibler Splits in Baum 1	Zahl kompatibler Splits in Baum 2	RF-Distanz	Gewichtete RF-Distanz	ssRF-Distanz
9	58	58	27	27	27	62	45	62
14	57	58	27	27	27	61	46	61
17	55	58	27	27	29	59	47	57
23	54	58	27	27	31	58	48	54
25	53	58	27	27	32	57	49	52
32	52	58	26	26	32	58	48	52
36	51	58	26	26	32	57	49	51
41	50	57	26	26	32	55	50	49
42	49	56	26	26	31	53	51	48
45	49	55	26	27	31	52	52	46
46	49	54	26	27	31	51	53	45
48	48	54	26	26	31	50	53	45
49	48	53	26	26	30	49	54	45
50	46	53	25	25	32	49	56	42
51	46	52	25	25	31	48	56	42
52	46	51	25	25	31	47	57	41
53	45	51	25	25	31	46	58	40
55	45	50	25	26	31	45	59	38
57	45	49	25	26	30	44	59	38
63	44	49	25	26	32	43	60	35
66	44	48	25	26	31	42	61	35
68	42	48	25	25	32	40	62	33
69	41	48	25	25	32	39	63	32
71	41	47	25	25	32	38	64	31
76	40	47	25	25	33	37	65	29
79	40	46	25	25	33	36	66	28
81	40	45	25	29	33	35	67	23
83	40	44	25	29	33	34	68	22
85	40	43	24	29	32	35	67	22
101	38	43	24	28	33	33	69	20
102	36	43	23	27	34	33	69	18
103	35	43	23	27	34	32	70	17
105	34	43	23	26	34	31	71	17
106	33	43	23	26	34	30	72	16
107	32	43	23	25	34	29	72	16
109	32	42	22	25	33	30	72	16
111	31	42	22	25	34	29	73	14
112	30	42	22	24	34	28	74	14
113	30	41	22	25	34	27	75	12
115	30	40	22	25	33	26	75	12
120	29	39	22	24	32	24	77	12
121	28	37	21	23	31	23	78	11
123	26	37	20	22	32	23	79	9
124	26	37	20	22	32	23	80	9
129	26	36	20	22	31	22	81	9
137	25	36	20	22	33	21	81	6
138	25	35	19	22	32	22	81	6
139	25	34	19	23	32	21	82	4
142	24	34	18	22	32	22	82	4
143	24	33	17	22	31	23	83	4
145	24	32	17	22	30	22	83	4
148	24	31	17	23	30	21	84	2
149	24	30	17	23	29	20	84	2
151	24	29	17	23	28	19	85	2
153	24	26	17	24	26	16	87	0
154	23	26	17	23	26	15	88	0
156	22	26	17	22	26	14	88	0
159	21	26	17	21	26	13	89	0
160	20	26	17	20	26	12	90	0
162	20	25	17	20	25	11	91	0
165	20	24	17	20	24	10	92	0
167	19	24	17	19	24	9	93	0
168	19	23	17	19	23	8	95	0
172	18	23	17	18	23	7	96	0
173	17	23	16	17	23	8	97	0
177	16	23	15	16	23	9	97	0
180	15	23	14	15	23	10	97	0
182	14	22	13	14	22	10	98	0
183	13	22	12	13	22	11	98	0
191	12	22	11	12	22	12	98	0
192	12	21	11	12	21	11	98	0
193	12	20	11	12	20	10	99	0
194	12	19	11	12	19	9	99	0
196	10	19	10	10	19	9	99	0
198	8	16	8	8	16	8	99	0
199	7	15	7	7	15	8	100	0

***Tabelle 6.1**: Ergebnistabelle des computeSSRF-Laufs.* Die Tabelle entspricht dem Inhalt der Datei *ssrfDistances.txt* im Ordner *Distanzmaß/Lauf.* Alle Splits ab Bootstrap-Schwelle 139 (entspricht ca. 70 % des höchsten Bootstrap-Wertes 199) sind signifikant. Ab Schwelle 153 gibt es keinen signifikanten Split mehr, der zwischen den Phylogenien unterscheidet.

7 Automatisierte Ableitung einer Phylogenie aus einem Alignment konkatenierter Genfamilien von E. coli und Salmonellen

Ein Anliegen der vorliegenden Arbeit bestand darin, mit verschiedenen Methoden Phylogenien aller sequenzierten *Escherichia coli*- und *Salmonellen*-Stämme abzuschätzen. Die dazu notwendigen Arbeitsschritte lassen sich sehr gut automatisieren und somit auch protokollieren. Zu diesem Zweck wurde das Programm *MMN* in der Programmiersprache Perl entwickelt, das in diesem Kapitel vorgestellt wird.

7.1 Funktionalität des Programms MMN

MMN berechnet für einen gegebenen Datensatz molekularer Sequenzen von *Escherichia coli*-/*Shigellen*- und/oder *Salmonellen*-Stämmen halbautomatisch eine Phylogenie. Die Inferenz wird mit einem Maximum-Likelihood-Framework auf Grundlage des multiplen Sequenzalignments der konkatenierten (ggf. universellen) Gene jedes Bakteriums durchgeführt. Das Verfahren benutzt eine definierte Abfolge von Schritten, von denen jeder über vom Benutzer im Vorfeld einzustellende Parameter gesteuert werden kann. Ein Schritt kann entweder individuell ausgeführt werden, sofern alle Daten aus den vorigen Schritten an vorgesehener Stelle in der Ordnerstruktur von *MMN* hinterlegt sind, oder es werden alle Schritte hintereinander ausgeführt. Das Programm wird in der *Linux*-Shell mittels `perl MMNmain.pl` aus dem Verzeichnis *MMN* aufgerufen.

Die Gesamtfunktionalität des Programms wurde auf mehrere Perl-Dateien („packages“) aufgeteilt. Zum Lesen von Sequenzen aus Dateien in den Formaten FASTA und PHYLIP ist die Installation des *BioPerl*-Core-Pakets (Stajich und Birney, 2000) erforderlich. *MMN* verwendet die aus **Tabelle 7.2** ersichtliche Dateistruktur. Das Programm behandelt sämtliche Verzeichnisse als relativ zum übergeordneten Verzeichnis *MMN*. Das ermöglicht es, für ein neues Projekt den ganzen Ordner *MMN* einfach an eine andere Stelle zu kopieren und alle nicht mehr benötigten Textdateien im Ordner *tables* und ggf. nicht mehr notwendige Dateien im Ordner *data* zu löschen. Weitere Informationen können dem ausführlich kommentierten Quellcode, insb. der Datei *Config.pm* im Ordner *Core,* entnommen werden.

Es folgt eine Erklärung der Programmschritte (siehe **Abb. 7.1**):

Download der Sequenzen

Das Programm lädt für jedes in der Datei *bacteriaTable.txt* markierte Genom die chromosomalen und plasmidischen Sequenzdaten (Suffixes **.faa, *.fna, *.ppt* und **.fna*, siehe auch **Kapitel 8**, **Material und Methoden**) in das Verzeichnis *downloads_NCBI* herunter. Zum jetzigen Zeitpunkt können nur *Escherichia coli/Shigellen*- und *Salmonellen*-Genome benutzt werden.

Zusammenfassen chromosomaler und plasmidischer Proteinsequenzen

Chromosomale und plasmidische Proteinsequenzen werden in neu generierten FASTA-Dateien zusammengefaßt und im Ordner *extendedSeqs* gesammelt, ebenso die entsprechenden

Informationen zur Genabfolge in den PTT-Dateien. Sowohl chromosomale als auch plasmidische Sequenzdaten enthalten phylogenetische Signale, die den Gesamtaufbau der Phylogenie bestimmen.

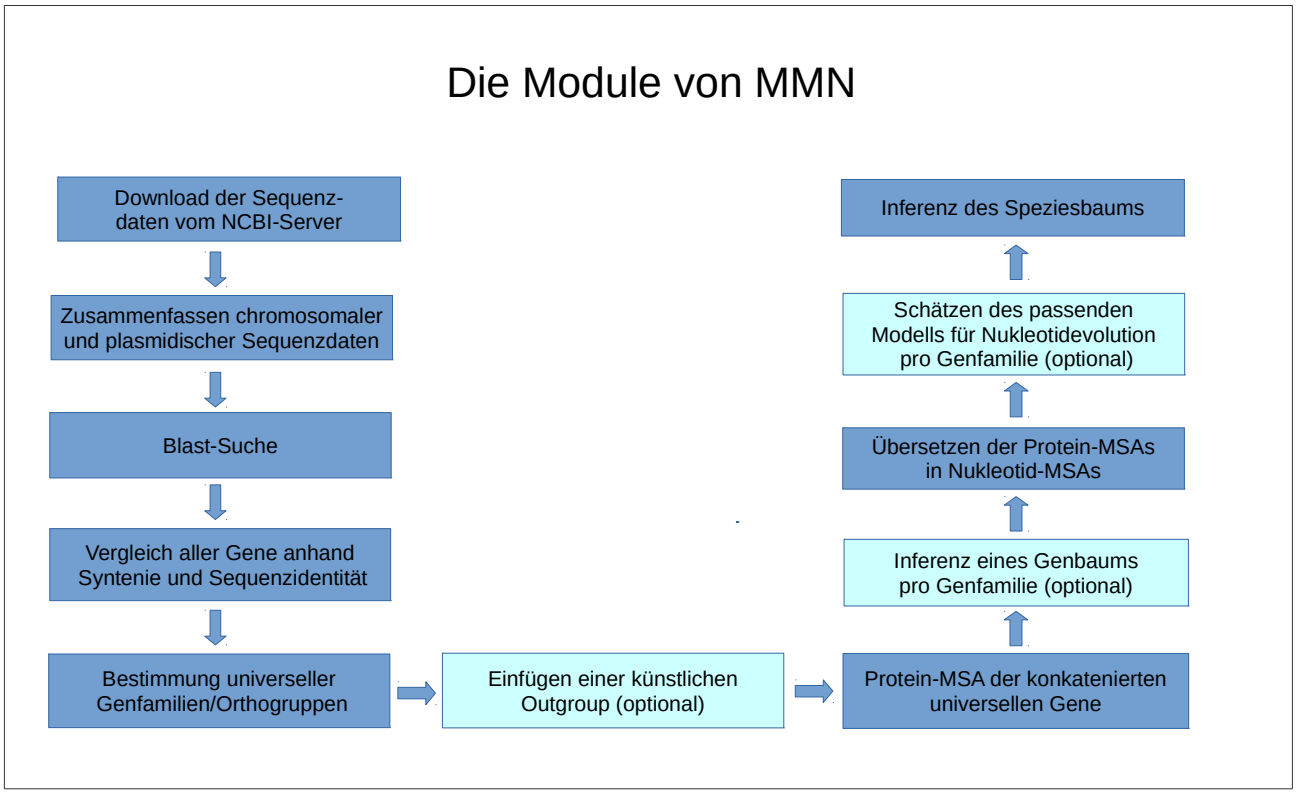


Abbildung 7.1*: Die Module von MMN*

MMN erzeugt für jeden Bakterienstamm ein neue FASTA-Datei, indem es die plasmidischen Proteinsequenzen hinter die chromosomalen schreibt. Das Gleiche geschieht für die PTT-Dateien. Da alle Gen-Positionsinformationen in aufsteigender Reihenfolge angeordnet sein müssen, werden die Positionsinformationen für die gerade hinzugefügten plasmidischen Gene neu durchnummeriert. Die Gen-Start- und Stop-Daten in den FASTA-Dateien werden ebenfalls entsprechend modifiziert. Das Programm *mpBatch*, welches die orthologen Gruppen generiert, wäre jetzt in der Lage, hintereinanderstehende chromosomale und plasmidische Proteinsequenzen in den neu erzeugten FASTA-Dateien (Ordner *extendedSeqs*) als einen zusammenhängenden Block zu interpretieren. Das muß unterbunden werden, da das biologisch sinnlos ist. Dazu wird in die PTT-Dateien vor die Positionsinformationen für ein plasmidisches Gen immer ein künstliches Gen einer bestimmten Länge eingefügt. Die Länge des künstlichen Gens ist stets höher als der Parameter *g* von *mpBatch*, der den maximalen Abstand zwischen zwei zu vergleichenden Sequenzfragmenten angibt.

Blast-Suche

Das Programm *blastAll* aus dem *Basic Local Alignment and Search Tool*, *Blast* (Altschul et al., 1990), wird eingesetzt, um paarweise Sequenzähnlichkeit, Sequenzidentität und den Erwartungswert[7] (*e-value*) für alle Genpaare und Bakterien zu berechnen. Dazu wird eine *Blast*-Datenbank im Ordner *data/blast/db* angelegt. Die Ergebnisse der *Blast*-Suche werden im Ordner *data/blast/result* deponiert. Die *Blast*-Suche kann auch parallel auf einem Rechencluster erfolgen. In diesem Fall befinden sich alle Dateien im Ordner *data/blast/parallel*.

7 BLAST normalisiert den *e-value* an der Sequenzlänge. Dieser drückt aus, wie hoch die Wahrscheinlichkeit dafür ist, daß die in der vorliegenden BLAST-Datenbank gefundene Sequenzidentität nur durch Zufall so niedrig ist. Bei Experimenten dient der *e-value* gewöhnlich als Schwelle: Genpaare mit einer paarweisen Identität oberhalb dieser Schwelle werden verworfen.

Konstruktion orthologer Gruppen bzw. Genfamilien

MMN berechnet orthologe Gruppen (ab jetzt mit „OG“ abgekürzt) für eine vom Benutzer bestimmbare Untermenge der Bakteriengenome (Datei *organisms_synteny.txt* im Ordner *tables*), die sich im Verzeichnis *extendedSeqs* befinden. Mit der Wahl einer Untermenge legt der Benutzer gleichzeitig fest, daß alle anderen Genome in den Folgeschritten ignoriert werden. *MMN* benutzt die von Christian Eßer im Rahmen seiner Doktorarbeit entwickelten Perl-Programme *mpbatch* und a*lignomat* (Eßer, 2010), deren Funktionsweise in **Kapitel 4** dargestellt worden ist. Bei der Erstellung der Genfamilien bzw. OG berücksichtigt sein Algorithmus Sequenzähnlichkeit und Syntenie der Gene. Dabei werden aber nur Genkopien berücksichtigt, die in mindestens 50 % aller Organismen vorhanden sind. Das Hilfsprogramm a*lignomat* schreibt die resultierenden Genfamilien (Ordner *data/synteny/mpBatch_result*) in Sequenzdateien im FASTA-Format und legt diese im Ordner *data/synteny/alignomat_result* ab.

Bei der Rekonstruktion der Phylogenie sollen nur universelle Gene berücksichtigt werden. Das Programm speichert daher Genfamilien, für die in jedem Organismus genau eine Kopie vorhanden ist, im Ordner *fullGroups ab* und alle anderen in *otherGroups*. Die Familien in *otherGroups* werden - außer bei der Berechnung von Genbäumen - nicht mehr gebraucht.

Anfügen einer künstlichen Outgroup (optional)

Sollte es für ein Projekt, in dem ein Speziesbaum für *E. coli*-Genome geschätzt wird, sinnvoll erscheinen, kann *MMN* einen künstlichen Bakterienstamm aus Genen von *Salmonellen-S*tämmen konstruieren. Zweck ist es, mit diesem Stamm eine Outgroup für das *E. coli*-Phylum im Speziesbaum zu erhalten. Dazu wird jeder universellen Genfamilie von *E. coli* ein *Salmonellen*-Gen hinzugefügt. *MMN* inspiziert alle paarweisen *Blast*-Vergleiche für die Gene innerhalb der Genfamilie. Dann fügt es das *Salmonellen*-Gen, das zu allen Genen in der Familie eine maximale Ähnlichkeit aufweist, der Familie hinzu, sofern der *e-value* des *Salmonellen*-Gens nicht höher als ein vom Benutzer festzulegender Schwellenwert ist. Tritt Letzteres ein, wird die gesamte Familie von der weiteren Benutzung ausgeschlossen.

Oben dargestelltes Prozedere kann durch Setzen entsprechender Parameter dergestalt abgewandelt werden, daß einem Speziesbaum für *Salmonellen*-Stämme eine Outgroup aus *E. coli*-Genen hinzugefügt wird. Diese Funktionalität ist im Programm vorhanden, wurde aber in jüngerer Zeit nicht mehr benutzt. Sie diente ursprünglich dem Zweck, eine Wurzelposition für die ungewurzelte *E. coli*- bzw. *Salmonellen*-Phylogenie vorzuschlagen und so die Phylogenien zu wurzeln, ohne daß sich dadurch der statistische Support der Äste deutlich verschlechtert[8]. Für die konkrete Anwendung war das aber nicht der Fall.

Multiples Sequenzalignment der Proteine in den universellen Gruppen

Ab diesem Schritt benutzt *MMN* ausschließlich die universellen Genfamilien, entweder mit oder ohne künstliche Outgroup. Für jede universelle Familie aus Proteinsequenzen wird ein multiples Sequenzalignment (MSA) mit *MAFFT* (Katoh et al., 2002) im Ordner *data/alignments/mafft* erzeugt. *MAFFT* benutzt dazu eine FFT-Methode („Schnelle Fouriertransformation“), die automatisch die beste ihm zur Verfügung stehende Strategie für die zu alignierenden Sequenzen wählt. Aus den MSAs werden dann mittels *Gblocks* (Castresana, 2000) schlecht alignierte und divergente Regionen entfernt. Die modifizierten Alignments werden im Ordner *data/alignments/gblocks* abgelegt.

8 Das Wurzeln der Phylogenien gelang mit diesem Vorgehen, aber die Bootstrap-Unterstützung verschlechterte sich deutlich. Es deutet einiges darauf hin, daß bei der Wahl der an die Orthogruppen anzufügenden Gene ausgerechnet solche ausgewählt wurden, die wahrscheinlich lateral transferiert worden sind. Dadurch kam es zu widersprüchlichen phylogenetischen Signalen.

Inferenz von Genbäumen (optional)

MMN kann Genbäume für die universellen Familien mit der Applikation *phyML* (Guindon und Gascuel, 2003) für die Phylogenieinferenz mittels eines Maximum-Likelihood-Ansatzes schätzen. Vorher müssen die FASTA-Alignments allerdings in das PHYLIP-Format konvertiert werden. Die Genbäume werden im Ordner *data/phyml_GeneTrees* gespeichert.

Rückübersetzung der Proteinsequenzen in Nukleotidsequenzen

Da bei der Proteinbiosynthese die Basentripletts in nur eine Aminosäure übersetzt werden, unterstellt man Nukleotidsequenzen, daß sie ein reicheres phylogenetisches Signal als Proteinsequenzen tragen. *MMN* übersetzt daher die vorliegenden multiplen Proteinalignments zurück in Nukleotidsequenzen, wobei es die von *MAFFT* eingeführten Gaps konserviert. Da dies ohne das Wissen, welche Basentripletts ursprünglich in die hier und jetzt sichtbaren Aminosäuren übersetzt worden sind, nicht ohne Informationsverlust machbar ist, werden die passenden Nukleotidsequenzen aus den FNA-Dateien als Schablone benutzt.

Der Algorithmus iteriert jeweils über alle Proteinsequenzen in einer ALN-Datei (enthält ein MSA) im Ordner *data/alignments/mafft* und übersetzt sie hintereinander zurück in Nukleotidsequenzen. Die erfolgreich übersetzten MSAs werden im Ordner *data/alignments/mafft_reverse* gespeichert.

Auffinden der als Schablone dienenden Nukleotidsequenz:

Für eine gegebene Proteinsequenz wird zunächst das Präfix der FASTA-Datei, aus der die Sequenz ursprünglich stammt, ermittelt. Über das Präfix, eine NC-Nummer, wird die GI-Nummer für das betreffende Gen der Tabelle *synti.consense* entnommen, die mit einem Eintrag in der PTT-Datei mit dem gleichen Präfix korrespondiert. Die gesuchte Nukleotidsequenz wird dann der FFN-Datei mit der passenden NC-Nummer entnommen.

Rückübersetzung:

Der Algorithmus bewegt sich an der Proteinsequenz (PS) entlang und inspiziert schrittweise jede Aminosäure. Gleichzeitig bewegt er sich an der Schablone (codierende Region, NS) entlang und inspiziert schrittweise jedes Basentriplett. Dabei wird das rückübersetzte Nukleotidalignment (NA) inkrementell erzeugt. Wird ein Basentriplett in der NS aufgefunden, das für die Aminosäure in der PS codiert, wird das Triplett in das NA aufgenommen. Aus einem Gap in der PS werden drei Gaps im NA. Ein „X“ in der PS bedeutet, daß die Aminosäure nicht identifiziert werden konnte. Daher wird das Triplett an der entsprechenden Position in der NS in das NA übernommen. Steht in der PS ein „J“, gemäß *IUPAC-Code für Nukleotide* entweder Leucin oder Isoleucin, wird das entsprechende Triplett ebenfalls in das NA übernommen. Für die restlichen Fälle wird zur Übersetzung der Basentripletts in Aminosäuren die Übersetzungstabelle für Prokaryoten genutzt, wie das NCBI sie vorschlägt (Elzanowski und Ostell, 2013).

Sofern es gelungen ist, alle Proteinsequenzen im MSA aus der Quelldatei in Nukleotidsequenzen zurück zu übersetzen, wird eine Zieldatei mit einem MSA aus den übersetzten Sequenzen angelegt. Schlug die Rückübersetzung mindestens einer Sequenz fehl, wird das MSA als unbrauchbar markiert. Das bedeutet, daß keine Sequenz in dieser Datei bei der Berechnung des Speziesbaumes berücksichtigt wird.

Während der Rückübersetzung können zwei Fehler geschehen:

- Der Algorithmus trifft in der Mitte der PS auf eine Aminosäure, die nicht zum vorgefundenen Basentriplett paßt. Dieser Fehler ist fatal und führt dazu, daß das gesamte Ursprungsalignment - die Genfamilie - zukünftig ignoriert wird.

- Eine NS beginnt mit einem anderen Startcodon als „ATG“[9]. Der Algorithmus toleriert dies zunächst und schreibt das Codon in das NA. Trifft er danach allerdings nochmals auf ein Basentriplett, das er nicht in die angetroffene Aminosäure übersetzen kann, bricht er die Rückübersetzung ab. Auch dieser Fehler ist fatal.

Gblocks (Castresana, 2000) wird nochmals eingesetzt, um aus den resultierenden NAs im Ordner *data/alignments/mafft_reverse s*chlecht alignierte oder divergente Regionen zu entfernen. Die modifizierten Alignments befinden sich im Ordner *data/alignments/gblocks_reverse*. Diese werden dann mittels *BioPerl* (Stajich und Birney, 2000) in das *PHYLIP*-Format[10] übersetzt und in *data/alignments/phylip* gespeichert.

Auswahl eines Modells für Nukleotidevolution je universeller Genfamilie (optional)

Für jedes Nukleotidalignment kann das Java-basierte Werkzeug *JModeltest2* das auf die Sequenzen am besten passende Modell für Nukleotidevolution schätzen. Das Modell für ein bestimmtes Alignment kommt dann bei der Ableitung des entsprechenden Genaumes zum Einsatz, indem es als Parameter für *phyML* spezifiziert wird. Bei der Ableitung des Speziesbaumes wird standardmäßig das Modell für Nukleotidevolution verwendet, das am häufigsten von J*Modeltest2* vorgeschlagen worden ist.

Für die Auswahl eines Modells bedient sich *JModeltest2* des *Akaike Information Criterions, AIC*. David Posada und Thomas R. Buckley (Posada und Buckley, 2004) und auch Diego Pol (Pol, 2004) heben hervor, daß eine bayesianische, insbesondere eine auf dem *AIC* beruhende Strategie für die Modellauswahl mehrere Vorteile gegenüber einer solchen besitzt, die den oft benutzten *Likelihood Ratio Test, LRT* (Felsenstein, 1973; Felsenstein, 1981), benutzt. Einer der Vorteile besteht darin, daß eine *AIC*-basierte Strategie zur gleichen Zeit verschachtelte und nicht-verschachtelte Modelle vergleichen kann. *MMN* führt für jedes Nukleotid-MSA im Ordner *data/alignments/phylip* einen Test durch und schreibt die Ergebnisse in den Ordner *data/modeltests.*

Inferenz des Speziesbaumes

Im letzten Schritt schätzt *MMN* einen Speziesbaum für alle Bakterien. Es gibt zwei grundsätzlich verschiedene Strategien zur Abschätzung einer Phylogenie auf Grundlage von Genen bzw. Genfamilien. Die eine besteht darin, einen Consensus-Baum aus den individuell geschätzten Genbäumen zu berechnen, die andere, Gensequenzen in einem universellen Alignment zusammenzufassen und für dieses einen Speziesbaum zu schätzen. Gadagkar und Mitautoren (Gadagkar et al., 2005) berichten, daß letzterer Ansatz präzisere Bäume hervorbringt[11].

MMN erzeugt ein globales Nukleotid-MSA aus den Nukleotid-MSAs der universellen Genfamilien. Jede Zeile in diesem MSA enthält die betreffenden Gensequenzen eines Bakteriums hintereinandergeschrieben. Die Reihenfolge der Gene ist in jeder Zeile identisch. Das globale MSA wird dann ins *PHYLIP*-Format übersetzt.

Die Bauminferenz geschieht mittels *PhyML* der Autoren Stephane Gascuel und Olivier Guindon (Guindon und Gascuel, 2003). *PhyML* erwartet als Argumente ein MSA und ein Modell für Nukleotidevolution. Als Modell sollte das gewählt werden, das J*Modeltest2* am häufigsten als das

9 Die Translation der DNS bei Bakterien und Archaeen kann auch über die Codons „GTG“ oder „TTG“ bzw. über „AUG“, „GUG“ und „UUG“ bei mRNA initiiert werden (Elzanowski und Ostell, 2013), aber der Algorithmus sucht nicht gezielt nach diesen Codons.
10 Gelegentlich erfolgte die Konvertierung extern mit dem Programm `sortal.pl`. Dieses von einem unbekannten Autor verfaßte Programm wurde von Gabriel Gelius-Dietrich, Heinrich-Heine-Universität Düsseldorf, für seine Doktorarbeit (Gelius-Dietrich, 2008) modifiziert.
11 Eine schon vor der Jahrtausendwende entbrannte Kontroverse („total evidence debate“, nachzulesen z. B. bei Felsenstein, 2004, S. 536) zwischen Befürwortern und Gegnern von Konkatenationsmethoden ist zum jetzigen Zeitpunkt (Frühjahr 2014) noch nicht beigelegt.

zu den universellen Genfamilien passende bestimmt hat[12]. Zurzeit muß das Modell noch manuell in der Datei *BuildSpeciesTree.pm* in Zeile 571 festgelegt werden.

MMN fordert, daß in einem vorhergehenden Programmschritt das MSA mit dem festen Namen *fgRevNucGenome_ALL.faa.phy* im Ordner *data/alignments/gblocks_reverse/concatenated* angelegt worden ist.

7.2 Ablaufsteuerung der Module von MMN

Es folgt an dieser Stelle eine Kurzerklärung der Parameter in der Datei *Config.pm*. Genauere Informationen zum Gebrauch der Parameter sind in der Datei selbst und in **Tabelle 7.3** zu finden, insbesondere die Speicherorte der jeweiligen Datentypen relativ zum Ordner *MMN*. **Tabelle 7.1** gibt über den Aufbau von *Config.pm* Auskunft.

Der Benutzer legt mit den in **Tabelle 7.3** beschriebenen Schaltern fest, welche Module von *MMN* ausgeführt werden. Gültige Belegungen für jeden Schalter sind 0 oder 1. Für die Mehrzahl der Schalter ist nur eine Stellung dargestellt. Es handelt sich dabei um Programmfunktionalitäten, die aus bereits vorhandenen Daten neue Daten erzeugen. Wird der Schalter so eingestellt, daß keine neuen Daten erzeugt werden, erwartet das Programm, daß die Daten bereits an der richtigen Stelle innerhalb des Ordners *data* vorhanden sind. Sofern neue Daten erzeugt werden, werden die Daten, die sich bereits im Zielverzeichnis befinden, in eines der Backup-Verzeichnisse verschoben (vgl. Datei *Config.pm,* Zeilen 268 - 356) und dabei mit einem Zeitstempel versehen.

Programminterne Parameter (brauchen normalerweise nicht verändert werden)	
Zeilen 106 - 214	Von *MMN* benutzte Verzeichnisnamen relativ zum Ordner *MMN*
Zeilen 268 - 356	Backup-Verzeichnisse
Zeilen 357 - 433	Programminterne Listen, Hashes und Skalare
Zeilen 757 - 768	Beschaffung des Zeitstempels
Vom Benutzer ggf. zu ändernde Parameter	
Zeilen 215 – 267	Verzeichnisse mit den Executables der externen Programme *Blast*, *MAFFT*, *Gblocks*, *PhyML*, *PHYLIP*, *Java*, *JModelTest2*, *PAUP* innerhalb des Ordners *exec* Hinweis: Es kann vorkommen, daß auf einem individuellen System ein externes Programm aufgrund fehlender Benutzerrechte nicht innerhalb des Ordners *exec* installiert werden darf.
Zeilen 434 - 481	Parameter für die Programme *Blast, mpBatch, MAFFT, jModeltTest2, phyML*
Zeilen 487 - 750	Schalter für die Ablaufsteuerung von *MMN*

Tabelle 7.1*: Aufbau der Konfigurationsdatei Config.pm*

Die meisten zeitkritischen Module wurden sowohl für eine sequentielle als auch für eine parallele Abarbeitung auf Rechenclustern vorbereitet. Hier werden nur die Parameter für die stabile sequentielle Abarbeitung gezeigt. Erklärungen zu den Parametern für die parallelen Methoden können der Datei *Config.pm* sowie den verschiedenen Programmteilen entnommen werden.

12 Es ist biologisch unwahrscheinlich, daß durch die Benutzung lediglich eines einzelnen Modells die wahre Dynamik in der Evolution aller Nukleotide im gesamten MSA nachgebildet wird. Bis dato kennt die Forschung im Zusammenhang mit Konkatenationsmethoden allerdings keine bessere Möglichkeit.

Ordner in MMN	Datei	Erklärung
/tables	bacteriaTable.txt	Textdatei mit den Verzeichnisnamen der Bakteriengenome auf *NCBI*. Gewünschte Genome werden mit „i“ oder „o“ markiert. Die Datei wird auf Wunsch des Benutzers neu konstruiert.
	chromosomeFiles.txt	*NCBI*-Nummern aller Dateien mit chromosomalen Sequenzen
	filesPerStrain.txt	Verzeichnisnamen der Genome und zugehörige Dateien
	organismsSynteny.txt	Aufbau wie filesPerStrain.txt; Genome, die für die Konstruktion orthologer Gruppen benutzt werden sollen, werden mittels „x“ markiert.
	sequenceType.txt	Ordnet jeder *NCBI*-Nummer eines der Schlagworte „chromosomal“ oder „plasmidisch“ zu.
/exec		Verzeichnisse der *Linux*-Executables *Blast, MAFFT, Gblocks* etc.
/Core	Config.pm	Sämtliche Einstellungen werden in diesem Modul vorgenommen.
	AddSingleBestHitToOGs.pm	Fügt jeder orthologen Gruppe von *E. coli*-Genen das am besten passende Salmonellen-Gen hinzu. In der letzten *MMN*-Version wurde diese Funktionalität nicht mehr gebraucht.
	alignomat.pl	Setzt das Ergebnis von *mp_batch_013.pl* zu orthologen Gruppen zusammen (Autor: Christian Eßer)
	BuildBlastDB.pm	Erstellt eine *Blast*-Datenbank
	BuildGeneTrees.pm	Erlaubt die Konstruktion von Genbäumen mit verschiedenen Methoden
	CalculateSynteny.pm	Integriert die von Christian Eßer realisierten Funktionalitäten in *MMN*
	InitProject.pm	Erzeugt alle notwendigen Verzeichnisse, nimmt Prüfarbeiten vor.
	MakeAlignments.pm	Erzeugt multiple Alignments mit *MAFFT* und entfernt uninformative Regionen
	MakeOrthologs.pm	Separiert universelle Genfamilien von den berechneten Orthogruppen.
	mp_batch_013.pl	Berechnet den Abstand zwischen Genen basierend auf Ähnlichkeit und Synthenie
	PerformBlast.pm	Führt eine *Blast*-Suche auf den Sequenzen im Datensatz durch.
	ReconstructAncestralStates.pm	Inferiert den anzestralen Gengehalt für die Spezies im Datensatz unter Verwendung von PAUP, wird in der neuesten *MMN*-Version nicht mehr verwendet.
/IO	AppendChromosomesPlasmids.pm	Faßt chromosomale und plasmidische Sequenzen in neuen Dateien zusammen, modifiziert ebenfalls die PTT-Dateien.
	AssignNames.pm	Ordnet den heruntergeladenen Dateien die im Projekt verwendeten Kürzel zu.
	GetSequences.pm	Lädt die ausgewählten Dateien vom *NCBI*-Server herunter.
	RevTransSeq.pm	Übersetzt ein Alignment von Proteinsequenzen in Nukleotid-sequenzen und verwendet die Original-Nukleotidsequenzen als Schablone.
/data	alignments/	Multiple Protein- und Nukleotidalignments im FASTA- und PHYLIP-Format
	blast/	*Blast*-Datenbank und Ergebnisse der sequentiellen oder parallelen *Blast*-Suche
	downloads_NCBI/	Speicherort der vom *NCBI*-Server heruntergeladenen Dateien
	extendedSeqs/	Speicherort der aus chromosomalen und plasmidischen Sequenzen zusammengesetzten FASTA- und PTT-Dateien
	modeltests/	Ergebnisse von *JModeltest* pro Genfamilie/Orthogruppe
	phyml_GeneTrees/	Genbäume für jede Genfamilie/Orthogruppe
	phyml_SpeciesTrees/	Speziesbaum auf Basis des Alignments der konkatenierten Gene in den Orthogruppen
	synteny/	Ausgabedateien von *mpBatch* und *Alignomat*: alle auf Grundlage von Sequenzähnlichkeit und Syntenie ermittelten Genfamilien, universelle Genfamilien („full groups/“), alle anderen Genfamilien („otherGroups/“)
MMNMain.pl	Hauptprogramm	Arbeitet sequentiell alle Programmschritte ab. Durch Setzen der Parameter in der Datei Config.pm werden Schritte ein- oder ausgeschaltet.

Tabelle 7.2*: Verzeichnisstruktur von MMN*

Schalter für die Ablaufsteuerung (Datei *Config.pm*, Zeilen 487 - 750)

Herstellung des Datensatzes	
rebuildBacteriaList	1 - Erzeugt eine aktuelle Liste der E. coli/Salmonellen-Genome auf dem FTP-Server
downloadDataset	0 - Lädt die in *bacteriaTable.txt* ausgewählten Sequenzen herunter,
checkFiles	0 - Gleicht herunterzuladende und vorhandene Dateien ab
Blast-Suche	
buildBLASTDB	0 - Erzeugt eine neue *Blast*-Datenbank
runBLASTALL	0 - Startet eine neue *Blast*-Suche
backupIntermediateBLASTResults	0 - Sichert ein vorhandenes *Blast*-Ergebnis vor einer neuen *Blast*-Suche, 1 - sichert keine vorhandenen *Blast*-Ergebnisse
Erzeugung universeller Proteinfamilien	
useExistingBLASTResultFile	0 - Kopiert alle *Blast*-Ergebnisse in eine neue Datei, sehr aufwendig
calculateSynteny	0 - Berechnet Orthogruppen für die in *organisms_synteny.txt* markierten Organismen
generateOrthologousGroups	1 - Schreibt die Orthogruppen in FASTA-Dateien und teilt sie in universelle und übrige Genfamilien auf
appendOrthologousGroups	0 - Fügt den universellen E. coli-Familien ein Salmonellen-Gen an, oder vice versa
tryToAddProteinToAllGroups	Falls *appendOrthologousGroups* = 0: 0 - Versucht, allen universellen Genfamilien ein Gen anzufügen (siehe Datei *Config.pm*), 1 - Ignoriert solche Familien, bei denen das vorher nicht gelungen ist (Suffix *.bad)
readBLASTResultInMemory	Falls *appendOrthologousGroups* = 0: 0 - Liest das vollständige *Blast*-Ergebnis in den Speicher (schneller), 1 - Liest das *Blast*-Ergebnis zeilenweise (langsamer, spart Speicherplatz)
addSingleBestEC	Falls *appendOrthologousGroups* = 0: 0 - *MMN* soll einen E. coli-Baum mit einem künstlichen Salmonellen-Stamm als Outgroup, 1 - einen Salmonellen-Baum mit einem künstlichen E. coli-Stamm als Outgroup berechnen
Globales Alignment von Proteinsequenzen	
buildMultipleAlignment	0 - Berechnet neue MAFFT-Alignments
removeDivergentRegions_AA	0 - Enfernt nichtinformative Positionen aus den Alignments
unifyPHYLIP	0 - kopiert die PHYLIP-Alignments in eine Datei
Rückübersetzung der Protein-MSAs in Nukleotid-MSAs	
reverseTranslateFASTA	0 - übersetzt Proteinalignments zurück in Nukleotidalignments
removeDivergentRegions_NT	0 - Enfernt nichtinformative Positionen aus den Alignments
convertFASTAtoPHYLIP	0 - konvertiert die zurückübersetzten Alignments vom FASTA- ins PHYLIP-Format
concatenateRevFASTA	0 - konkateniert die zurückübersetzten Alignments und schreibt sie in eine einzige Datei
convertRevFASTAtoPHYLIP	0 - übersetzt die Datei mit allen FASTA-Alignments in das PHYLIP-Format
Auswahl der Modelle für die Nukleotidevolution	
runModeltests	0 - Sucht für jede universelle Genfamilie das passende Modell für Nukleotidevolution
Ableitung der Genbäume	
computeGeneTrees	0 - Berechnet Genbäume für alle universellen Genfamilien
Ableitung des Speziesbaums	
computeConcatSpeciesTree	0 - Berechnet einen Speziesbaum auf Basis der konkatenierten Gene der universellen Genfamilien (Voreinstellung)
computeSchemeSpeciesTree	0 - Berechnet einen Speziesbaum auf Grundlage der konkatenierten Gene ausgewählter universellen Genfamilien. Es handelt sich dabei um die Familien, für die JModelTest2 in das gleiche Schema (nicht Modell!) für Nukleotidevolution ausgewählt hat
*useGuideTree***	1 - PhyML verfeinert einen Guide Tree, der vom Benutzer im Ordner *data/alignments/gblocks_reverse/concatenated* abgelegt wurde 0 - PhyML schätzt einen Speziesbaum auf Grundlage des MSAs der konkatenierten Gene aus den universellen Genfamilien
computeConsenseSpeciesTree	0 - Berechnet einen Consensus-Baum für die Äste in den Genbäumen, die in mindestens 50% der Bäume vorkommen („Majority Rule Consensus“)
computeModelSpeciesTree	0 - Berechnet verschiedene Spezies-Bäume. Falls J*ModelTest2* für mindestens 20 Proteine das gleiche Modell für Nukleotidevolution vorgeschlagen hat, wird für diese jeweils ein Baum konstruiert

Tabelle 7.3*: Schalter für die Ablaufsteuerung von MMN*

8 Inferenz robuster Phylogenien für Escherichia coli und Salmonellen mit verschiedenen Methoden und Vergleich mit Literaturbäumen

8.1 Einleitung

Für die Rekonstruktion des anzestralen Lebensstils von Bakterien oder des Nährstoffangebots in den anzestralen Umgebungen ihrer Vorfahren ist eine statistisch belastbare Rekonstruktion ihrer phylogenetischen Beziehungen von außerordentlicher Bedeutung. In den letzten beiden Kapiteln dieser Arbeit werden zwei Projekte vorgestellt, die diese Thematik im Zusammenhang mit lateralem Gentransfer untersuchen. Dieses Kapitel beschreibt, wie robuste Phylogenien für die 61 *Escherichia coli*-Stämme und 25 *Salmonellen*-Serotypen aus **Tabelle 8.1** geschätzt wurden.

In **Abschnitt 8.2** werden die verschiedenen Methoden beschrieben, die zur Rekonstruktion von Phylogenien für *E. coli* und *Salmonellen* eingesetzt wurden, und die rekonstruierten Phylogenien werden gezeigt. Zunächst wurden Bäume auf Grundlage eines Alignments der konkatenierten universellen Genfamilien erstellt. Bäume wurden zunächst unter einem Maximum-Likelihood-Ansatz (**Abschnitt 8.2.2.1**), danach mit einem Bayesianischen Ansatz (**Abschnitt 8.2.2.2**) und schließlich mit Neighbor-Joining (**Abschnitt 8.2.2.3**) inferiert. Mit einer Methode, die auf einem Alignment hochkonservierter Genfamilien (sog. COGs, „Clusters of Orthologous Groups“) arbeitet, wurden ebenfalls Phylogenien geschätzt (**Abschnitt 8.2.2.4**). Eine deutlich andere Strategie zur Bauminferenz wird in **Abschnitt 8.2.2.5** vorgestellt. Die entsprechende Methode definiert die Distanz zwischen zwei *E. coli*-Genomen über die Zahl rekombinanter Regionen zwischen ihnen und berechnet auf der Matrix aller paarweisen Distanzen einen Neighbor-Joining-Baum.

Abschnitt 8.3 vergleicht die rekonstruierten Phylogenien für *E. coli* untereinander, ebenso die für *Salmonella*. Grundlage dafür ist die statistische Unterstützung ihrer Splits durch die angewandten Baumrekonstruktionsmethoden. Die Distanzen zwischen den Bäumen werden mit dem vom Verfasser implementierten Programm *computeSSRF* (siehe **Kapitel 6**) berechnet, welches nur Äste, deren Bootstrap-Support mindestens so hoch wie eine gegebene Schwelle ist, berücksichtigt. Abschließend werden die *E. coli*- und *Salmonellen*-Phylogenien mit Bäumen verglichen, die von anderen Gruppen publiziert worden sind.

Eine Diskussion der Stärken und Schwächen (**Abschnitt 8.4**) der verwendeten Methoden sowie der Versuch einer Bewertung der geschätzten Phylogenien schließen das Kapitel ab.

In der Vergangenheit wurden andere Methoden angewandt, um Phylogenien für die Spezies *E. coli* und das Genus *Salmonella* zu schätzen, da Nukleotidsequenzdaten noch nicht zur Verfügung standen. Zunächst wurden die Antigene O, H und K zur Unterscheidung der Stämme benutzt (für ein Review siehe Orskov et al., 1977). Mit der Verfügbarkeit größerer Sammlungen von *E. coli*-Stämmen und mit Aufkommen des Verfahrens „Multi-locus enzyme elektrophoresis“ (MLEE) stellte sich allerdings heraus, daß die Klassifikation nach Serotypen kein geeignetes Werkzeug zur Unterscheidung der verschiedenen *E. coli*- und *Shigellen*-Stämme ist (Ochman und Selander, 1984a; Caugant et al., 1985). Ochman und Selander etablierten im Jahre 1984 ECOR, eine Referenzsammlung mit 72 *E. coli*-Stämmen (Ochman und Selander, 1984b). Die Stämme waren so handverlesen worden, daß sie eine maximale elektrophoretische Diversität und eine möglichst breite

E. coli-Genom	NCBI-Nummer
IAI1 uid59377	NC 011741
55989 uid59383	NC 011748
ATCC 8739 uid58783	NC 010468
HS	NC 009800
K-12 substr MG1655 uid57779	NC 000913
K-12 substr W3110 uid161931	NC 007779
UMN026 uid62981	NC 011751
APEC 01 uid58623	NC 008563
S88 uid62979	NC 011742
UTI89 uid58541	NC 007946
ED1a uid59379	NC 011745
536 uid58531	NC 008253
CFT073 uid57915	NC 004431
O127:H6 E2348 69 uid59343	NC 011601
IAI39 uid59381	NC 011750
SMS 3 5 uid58919	NC 010498
042 uid161985	NC 017626
ABU 83972 uid161975	NC 017631
BL21 DE3 uid161947	NC 012971
BL21 DE3 uid161949	NC 012892
BW2952 uid59391	NC 012759
B REL606 uid58803	NC 012967
DH1 uid161951	NC 017625
DH1 uid162051	NC 017638
H10407 uid161993	NC 017633
IHE3034 uid162007	NC 017628
KO11FL uid162099	NC 017660
KO11FL uid52593	NC 016902
K-12 substr DH10B uid58979	NC 010473
LF82 uid161965	NC 011993
NA114 uid162139	NC 017644
O103:H2 uid41013	NC 013353
O111:H 11128 uid41023	NC 013364
O157:H7 EC4115 uid59091	NC 011353
O157:H7 TW14359 uid59235	NC 013008
O26:H11 uid41021	NC 013361
O55:H7 CB9615 uid146655	NC 013941
O55:H7 RM12579 uid162153	NC 017656
O7:K1 CE10 uid162115	NC 017646
O83:H1 NRG 857C uid161987	NC 017634
P12b (O15:H17) uid162061	NC 017663
SE11 uid59425	NC 011415
SE15 uid161939	NC 013654
UM146 uid162043	NC 017632
UMNK88 uid161991	NC 017641
W uid162011	NC 017635
O157:H7 Xuzhou21 uid163995	NC 017906
BL21 Gold DE3 pLysS AG uid59245	NC 012947
D i14 uid162049	NC 017652
D i2 uid162047	NC 017651
O157:H7 EDL933 uid57831	NC 002655
W uid162101	NC 017664

Salmonella enterica serovar-Genom	NCBI-Nummer
Arizonae 62 z4 z23 RSK2980 uid58191	NC 010067
Agona SL483 uid59431	NC 011149
Choleraesuis SC B67 uid58017	NC 006905
Dublin CT 02021853 uid58917	NC 011204
Enteritidis P125109 uid59247	NC 011294
Gallinarum 287 91 uid59249	NC 011274
Gallinarum pullorum RKS5078 uid87035	NC 016831
Heidelberg B182 uid162195	NC 017623
Heidelberg SL476 uid58973	NC 011083
Newport SL254 uid58831	NC 011080
Paratyphi A AKU 12601 uid59269	NC 011147
Paratyphi A ATCC 9150 uid58201	NC 006511
Paratyphi B SPB7 uid59097	NC 010102
Paratyphi C RKS4594 uid59063	NC 012125
Schwarzengrund CVM19633 uid58915	NC 011094
Typhi CT18 uid57793	NC 003198
Typhi P stx 12 uid87001	NC 016832
Typhi Ty2 uid57973	NC 004631
Typhimurium 14028S uid86059	NC 016856
Typhimurium 798 uid158047	NC 017046
Typhimurium LT2 uid57799	NC 003197
Typhimurium SL1344 uid86645	NC 016810
Typhimurium ST4 74 uid84393	NC 016857
Typhimurium T000240 uid84397	NC 016860
Typhimurium UK 1 uid87049	NC 016863

Shigella-Genom	NCBI-Nummer
Sh. boydii Sb227 uid58215	NC 007613
Sh. boydii CDC3083 94 uid58415	NC 010658
Sh. flexneri 2a 301 uid62907	NC 004741
Sh. flexneri 2a 2457T uid57991	NC 004337
Sh. flexneri 5 8401 uid58583	NC 008258
Sh. dysenteriae Sd197 uid58213	NC 007606
Sh. flexneri 2002017 uid159233	NC 017328
Sh. Sonnei 53G uid84383	NC 016822
Sh. sonnei Ss046 uid58217	NC 007384

***Tabelle 8.1**: In dieser Arbeit verwendete Prokaryotengenome*

geografische Herkunft aufwiesen und von so vielen verschiedenen Wirten wie möglich stammten. Sie umfaßte pathogene und nichtpathogene Stämme, von ersteren allerdings lediglich UPEC.

Eine initiale, mit der UPGMA-Methode (Sneath und Sokal, 1974) erfolgte phylogenetische Analyse der Stämme mittels MLEE führte zur Definition von sechs großen phylogenetischen Gruppen, die als *A, B1, B2, C, D und E* bezeichnet wurden (Selander et al., 1987). Anschlußstudien, zunächst die von Herzer et al. (1990), die Neighbor-Joining benutzte, haben die Definition der großen phylogenetischen Gruppen für *Escherichia coli* bestätigt (Desjardins et al., 1995; Clermont et al., 2000). Die ECOR-Sammlung hat daher eine große Bedeutung für das Studium der Diversität von *E. coli* erlangt. Sie wurde später durch die als DEC-Sammlung bezeichnete Kollektion mit 78 Diarrhoe-auslösenden *E. coli* ergänzt (Whittam et al., 1993).

Als Alternative zu MLEE kam Ende der 90er Jahre „multilocus sequence typing“, MLST, auf. Eine der ersten Studien, die MLST benutzte, untersuchte die phylogenetischen Beziehungen zwischen pathogenen Stämmen aus der DEC-Sammlung und anderen pathogenen Stämmen sowie dem Laborstamm *K-12* (Reid et al., 2000). Dazu wurden die Sequenzen mehrerer Housekeeping-Gene konkateniert und mit der Methode „Split Decomposition“ von Daniel Huson (Huson, 1998) untersucht. Diese Methode fordert nicht, daß die evolutionären Beziehungen als baumartige Struktur abgebildet werden müssen. Stattdessen erlaubt sie es, einander widersprechende phylogenetische Informationen in Form eines Netzes darzustellen. Für die MLST-Daten von *E. coli* ergab sich eine überwiegend baumartige Struktur mit Ausnahme eines parallelen Pfades innerhalb der Gruppe *B2*. Das legte nahe, daß sich die Daten für eine phylogenetische Analyse eigneten. Die im Anschluß inferierte Phylogenie stimmt in wesentlichen Zügen mit früheren MLEE-Bäumen überein (Whittam et al., 1993).

Bis heute (2014) wird die Debatte geführt, ob insbesondere die Evolution der Prokaryoten besser als Baum (vgl. „Tree of Life“-Hypothese) oder als Netz dargestellt werden sollte. Befürworter von Netzdarstellungen argumentieren, daß Ereignisse lateralen Gentransfers unbedingt in die Visualisierung von Stammesgeschichten miteinfließen müssen, da sowohl laterale als auch vertikale Vererbung allgegenwärtige adaptive Prozesse in der Bakterienevolution sind (Dagan und Martin, 2006; Bapteste und Doolittle, 2007). Verfechter von Baumdarstellungen führen an, daß die vertikalen phylogenetischen Signale durch lateralen Gentransfer zwar verwässert werden, aber in der Regel dominant bleiben (Touchon et al., 2009). Der Verfasser hält die Darstellung der Prokaryotenevolution als Baum für angemessen, da Bakterien ihr genetisches Material verdoppeln und an jeweils zwei Nachkommen vererben. Dieser Prozeß hat binären Charakter und kann durch einen bifurzierenden Baum dargestellt werden.

8.2 Material und Methoden

Die für die in diesem Kapitel vorgestellten Analysen benötigten Dateien befinden sich innerhalb der Ordnerstruktur im Projektordner *Kapitel_Phylogenien*. Alle nicht vom Autor verfaßten Programme, die im Text erwähnt werden, wurden in **Kapitel 4** beschrieben.

8.2.1 Herkunft der verwendeten Genomdaten

Für die verschiedenen Experimente, die in diesem Kapitel vorgestellt werden, wurden chromosomale und plasmidische Sequenzdaten für 63 *Escherichia coli-*, *Shigellen-* und 25 *Salmonellen*-Stämme von der Genomdatenbank (ftp://ftp.ncbi.nlm.nih.gov/genomefs/) des *U. S.*

National Center for Biotechnology Information (*NCBI*) heruntergeladen. Da die Experimente sich über einen Zeitraum von ca. drei Jahren erstreckten, wurde der Datensatz sukzessive auf immer mehr Organismen ausgedehnt. Die hier vorgestellten Experimente beziehen sich ausschließlich auf den letzten Stand von September 2013.

Das *NCBI* hat ein allgemein akzeptiertes und benutztes Modell für biologische Sequenzdaten eingeführt (Ostell, 1995; Ostell, 1996), das es erlaubt, unkompliziert Sequenzen logisch miteinander zu verknüpfen und sie zu annotieren. Die sog. NC-Nummer identifiziert eindeutig genau ein Chromosom. Für jede NC-Nummer bietet der vom *NCBI* betriebene FTP-Server Dateien dreier Typen im FASTA-Format (Lipman und Pearson, 1985) sowie eines vierten, PTT, zum Herunterladen an:

PTT	Informationen zur Position von Genen innerhalb des Genoms
FAA	Aminosäuresequenzen, nur codierende Regionen
FNA	Nukleotidsequenzen, ganze Chromosomen oder Plasmide als ein Block
FFN	Nukleotidsequenzen, nur codierende Regionen

Für jeden Bakterienstamm wurden jeweils die Dateien dieser vier Formate heruntergeladen. In vielen Fällen wird neben einer Datei mit chromosomalen Sequenzdaten auch mindestens eine Datei mit Sequenzen plasmidischen Ursprungs angeboten. Ein Dateiname besteht aus einer NC-Nummer als Präfix und einem der vier Dateityp-Abkürzungen als Suffix.

Anmerkung: Das für diese Arbeit verfaßte Programm *MMN*, vorgestellt in **Kapitel 7**, lädt die Dateien für den Benutzer vom *NCBI* FTP-Server selbständig herunter. Es ist lediglich erforderlich, die gewünschten Genome in einer aktualisierbaren Liste zu markieren.

Tabelle 8.1 zeigt die in dieser Arbeit verwendeten Genome und ihre NC-Nummern.

8.2.2 Escherichia coli- und Salmonellen-Phylogenie auf Grundlage universeller Genfamilien

8.2.2.1 Inferenz in einem Maximum-Likelihood-Framework

In diesem Unterabschnitt wird dargestellt, wie mit einer Maximum-Likelihood-Methode Phylogenien für *E. coli* und *Salmonellen* und eine gemeinsame Phylogenie auf der Basis konkatenierter universeller Genfamilien erzeugt wurden, die dann als Referenzbäume *EC-Referenz*, *SA-Referenz* und *ECSA-Referenz* bei den sich anschließenden Vergleichen mit den mit anderen Methoden generierten Phylogenien dienen würden. Es wird auch gezeigt, wie äquivalente EC- und SA-Phylogenien mit einer Bayesianischen Methode geschätzt wurden.

Vorgehen bei der Berechnung der E. coli-Phylogenie

In *MMN*, dessen Arbeitsweise in **Kapitel 7** vorgestellt worden ist, wurde für die Berechnung der *E. coli*- und der *Salmonellen*-Phylogenien die in **Tabelle 8.2** gezeigte Parameterbelegung verwendet, alle anderen Parameter wurden in ihrer Standardeinstellung belassen. Die dargestellte Parameterbelegung wurde empirisch festgelegt. Der Erwartungswert 0,01 für *Blast* dient als obere Schranke für die paarweise Sequenzähnlichkeit der Proteine und filtert damit die Ergebnismenge von *Blast* aus dem Ordner *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/ blast/result*).

Blast		
e	0,01	In der Datenbank-Suche verwendeter Erwartungswert („expectation value“)
mpBatch		
g	4	Maximaler Abstand zwischen zwei Hits
m	7	Minimale Anzahl Hits
e	0,001	Erwartungswert, sollte höchstens so groß wie der für die *BLAST*-Suche verwendete Erwartungswert sein
n	75	Schwellenwert für die normalisierte Sequenzidentität
MAFFT		
auto		*MAFFT* bestimmt jeweils die für die Eingabe-Sequenzen am besten geeignete Alignment-Methode selbst
Gblocks		
t	p	Sequenztyp ist „Protein“
b5	a	Erlaubte Gap-Positionen: alle; Alignment-Positionen mit Gaps werden genauso wie andere behandelt. Die resultierenden Alignments enthalten u. U. Gaps.
JModeltest2		
s	11	Zahl der Substitutions-Schemata
f		Benutzt Modelle mit ungleichen Basen-Frequenzen
i		Benutzt Modelle mit einem Anteil invariabler Positionen
AICc		Berechnet das korrigierte Akaike-Informations-Kriterium
g	4	Benutzt Modelle mit variabler Substitutionsrate zwischen den Positionen, 4 Kategorien
p		Schätzt die Wichtigkeit der Parameter („Parameter Importance“)
v		Benutzt Model Averaging und Parameter Importance
w		Schreibt einen *PAUP*-Block (nicht benutzt)
a		Schätzt eine über alle Modelle gemittelte Phylogenie für jedes aktive Kriterium
PhyML		
f	e	Bestimmt die Equilibrium-Frequenzen der Nukleotide empirisch durch Zählen der verschiedenen Basen im Alignment
v	e	Schätzt den Anteil invarianter Positionen über Maximum-Likelihood
s	BEST	Sucht die beste Baumtopologie auf Grundlage von NNI („Nearest Neighbor Interchange“) und SPR („Subtree Pruning and Regrafting“)
o	tl	Optimiert Baumtopologie und Astlängen
n	1	Analysiert einen einzigen Datensatz
bootstrap	100	Berechnet 100 Bootstrap-Replikate
model	TN93	Benutzt als Modell für die Nukleotidsubstitution „Tamura-Nei 1993“ (TN93 wurde von *JModeltest2* vorher am häufigsten als das am besten geeignete Modell identifiziert. Diese Aussage bezieht sich auf die Alignments im Ordner *Referenzbaeume/MMN_ALL_EC_ohne_009801_002695/MMN/data/alignments/gblocks_reverse*.

Tabelle 8.2*: Für die Berechnung der Referenz-Phylogenien verwendete MMN-Parameter*

Das Programm *mpBatch* (Eßer, 2010) arbeitet auf dieser Ergebnismenge und filtert sie nochmals über einen eigenen Erwartungswert, hier 0,001. Es ermittelt orthologe Gruppen bzw. Genfamilien aufgrund von Sequenzähnlichkeit und Syntenie. Mit Blick auf die konkret verwendete Parameterbelegung sind beide Kriterien erfüllt, wenn das Programm beim Vergleich zweier Regionen 7 Gene in Folge findet, die maximal jeweils einen Abstand 4 voneinander haben und deren paarweise Sequenzidentität mindestens 75 % ist. Die Arbeitsweise des Programms wurde in **Kapitel 4** dargestellt. Sie folgt inhaltlich den Ausführungen in Eßer, 2010.

Die orthologen Gruppen befinden sich in den Ordnern *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/synteny/fullGroups* bzw. */otherGroups*. Multiple Alignments für die universellen Genfamilien wurden mittels *MAFFT* (Katoh et al., 2002) generiert, das die zu verwendenden Parameter selbst bestimmt. *MAFFT* wurde aufgrund seiner Schnelligkeit benutzt, insbesondere rechnet es parallel auf mehreren Threads. Die Alignments sind im Ordner *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/ alignments/mafft* abgelegt.

Gblocks (Castresana, 2000) entfernt schlecht alignierte und divergente Regionen aus den Alignments. Der verwendete Parameter legt fest, daß es dabei nicht alle Gaps bzw. einen Anteil von Gaps entfernt, sondern sie nicht anders behandelt als jede andere Position. Die dadurch

resultierenden Blocks dürfen also Gaps enthalten. Aus Zeitgründen wurde hier nicht mit unterschiedlichen Parametersetzungen experimentiert. Auch ist unklar, welche Parametersetzung der biologischen Wirklichkeit am besten entspricht. Die Alignments befinden sich im Ordner *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/alignments/gblocks*.

Die Proteinalignments wurden in *MMN* mit den ursprünglichen Nukleotidsequenzen als Schablone zurück in Nukleotidalignments übersetzt, wobei Gap-Positionen beibehalten wurden. Für jedes Nukleotidalignment schätzt *JModeltest2* (Darriba et al., 2012) das am besten zu den vorgefundenen Nukleotiden passende Modell für die Nukleotidsubstitution. Die verwendete Parameterbelegung stellt einen Kompromiß zwischen Genauigkeit und Rechenaufwand dar. Für die vorgelegten universellen Genfamilien von *E. coli* wählte *JModeltest2* am häufigsten das Modell TN93 von Tamura und Nei als passend aus. Die rückübersetzten Nukleotid-Alignments befinden sich im Ordner *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/alignments/mafft_reverse*.

Vor der Berechnung des Speziesbaumes konkateniert *MMN* die Genfamilien im Ordner *Referenzbaeume/MMN_ALL_EC_OKT12/MMN/data/alignments/gblocks_reverse* und konvertiert das resultierende multiple Sequenzalignment (MSA) in das PHYLIP-Format. Der *E. coli*-Speziesbaum für alle 63 vollständig sequenzierten Genome, hier nicht gezeigt, wurde mit *PhyML* (Guindon und Gascuel, 2003) aus diesem MSA der konkatenierten Genfamilien (Datei *fgRevNucGenome_ALL.faa.phy* im Ordner *MMN_ALL_EC_OKT12/MMN/data/alignments/gblocks_reverse/concatenated*) berechnet. Als Nukleotidsubstitutionsmodell wurde TN93 gewählt, das die Dynamik in der Evolution aller Nukleotide im MSA näherungsweise beschreibt[13].

Aus den 100 Bootstrap-Replikaten in der Datei *fgRevNucGenome_ALL.faa.phy_phyml_boot_trees.txt* wurde mit dem Programm *consense* aus der *PHYLIP*-Suite (Felsenstein et al., 1980) unter der erweiterten Majoritätsregel MRe ein Consensus-Baum (Datei *ec_outtree_phyml)* mit Bootstrap-Zahlen an den Ästen berechnet.

Anmerkung: *PHYLIP consense* liefert Newick-Ausdrücke, in denen allen Bootstrap-Zahlen immer ein Doppelpunkt vorausgeht. Die im Rahmen dieser Arbeit verwendeten Programme erwarten jedoch Bootstrap-Zahlen direkt nach einer schließenden Klammer. Daher wurden alle von *consense* produzierten Newick-Ausdrücke zunächst in dieses Format gebracht. Auf diese Tatsache wird ab jetzt nicht mehr gesondert hingewiesen.

E. coli-Baum mit 61 Stämmen

Der Baum für alle 63 vollständig sequenzierten *E. coli*- und *Shigellen*-Stämme enthält 54 *Escherichia coli*-Stämme und 9 *Shigellen*. Die Äste der Teilbäume, in denen sich diese Stämme befinden, weisen teilweise einen unzureichenden Bootstrap-Support von deutlich unter 70 % auf. Für die sich anschließenden Projekte, die in den folgenden Kapiteln dargestellt werden, werden allerdings statistisch robuste Phylogenien für *E. coli* und *Salmonellen* benötigt. Es wurden daher Experimente angestellt, in deren Verlauf die Taxa NC_011748, NC_009801, NC_002655, NC_002695, NC_013353 und NC_017906 (siehe **Tabelle 8.1** für die entsprechenden Klarnamen) einzeln oder in unterschiedlicher Kombination entfernt wurden, wobei sich teilweise der Bootstrap-Support der Äste im resultierenden Baum verbesserte. Das Optimum wurde mit der Entfernung der Taxa NC_009801 und NC_002695 erreicht[14].

13 Dieses Verfahren stellt mit Sicherheit lediglich eine grobe Näherung an die biologische Realität dar. Es ist unwahrscheinlich, daß die Dynamik der Nukleotidevolution für viele verschiedene Gene mit nur einem Modell hinreichend genau simuliert werden kann. Bis dato (2014) kennt man allerdings keine bessere Methode.

14 Die entfernten Stämme lieferten offenbar ein widersprüchliches phylogenetisches Signal. Im Interesse einer robusten Phylogenie erschien es dem Autor vertretbar, sie zu entfernen.

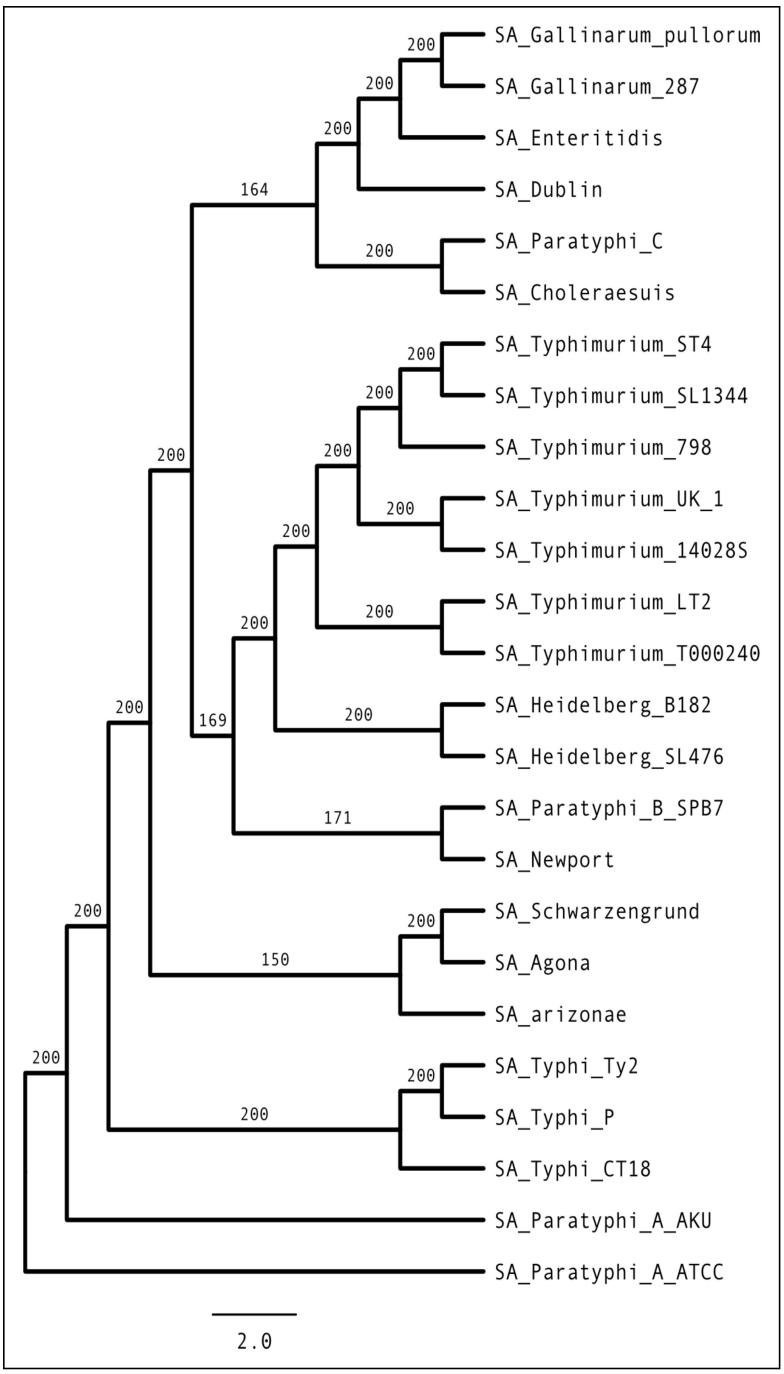


Abbildung 8.1*: Phylogenie aller 25 vollständig sequenzierten Salmonellen (Stand Oktober 2012) auf den konkatenierten universellen Genfamilien unter einem Maximum-Likelihood-Framework.* Es wurden 200 Bootstrap-Replikate berechnet. Die Astlängen sind proportional dargestellt. Alle Äste haben einen Bootstrap-Support von über 70 Prozent.

Taxa, deren Entfernung die Inferenz von Verwandtschaftsverhältnissen zwischen betrachteten Organismen stark beeinflußt, werden in der Literatur als „instabile Taxa“ (engl. „unstable taxa“) bezeichnet. Diego Pol und Ignacio H. Escapa haben ein Protokoll vorgestellt (Pol und Escapa, 2009), mit dem sich instabile Äste, entweder terminale Taxa oder Kladen, detektieren lassen. Ferner erlaubt es die Identifikation solcher Zeichen, die möglicherweise für Instabilität im Rahmen einer kladistischen Analyse verantwortlich sind.

Das *MMN*-Projekt zu diesem Abschnitt befindet sich in der Ordnerstruktur bei *Referenzbaeume/MMN_ALL_EC_ohne_009801_002695*. Die folgenden Ausführungen beziehen sich auf den Unterordner *MMN/data/alignments/gblocks_reverse/ concatenated*.

Mit *Phyml* wurde ein phylogenetischer Baum mit 200 Bootstrap-Wiederholungen auf der Grundlage des MSAs in *fgRevNucGenome_ALL.phy* geschätzt[15] und in der Datei *fgRevNucGenome_ALL.faa.phy_phyml_tree.txt* abgelegt. Bis auf einen Split mit 63,5 % weisen alle Splits eine statistische Unterstützung von mindestens 70 % auf.

MPI-basierte Berechnung einer Bootstrap-Phylogenie mittels PhyML-MPI

```
mpirun -n 20 ./phyml-mpi -i fgRevNucGenome_ALL.faa.phy --datatype nt -f e -v
e -s BEST -o tlr  --r_seed 11111 --use_median --bootstrap 200 --model TN93
```

Auf den Bootstrap-Bäumen wurde mit *PHYLIP* c*onsense* unter der *MRe* der in **Abbildung 8.2** dargestellte Consensus-Baum berechnet (Datei *outtree_phyml*).

Auf dem gleichen Alignment wurden mit RAxML (Stamatakis, 2014) unter Verwendung der Rapid-Bootstrapping-Methode ein Bootstrap-Baum (Datei *RaxML_bipartitions.raxml_out* im Ordner *raxml_200BootstrapReps*) mit 200 Wiederholungen sowie der Maximum-Likelihood-Baum mit dem höchsten Score bestimmt[16]. Als Nukleotidsubstitutionsmodell kam GTR („General Time Reversible“) zum Einsatz, das das speziellere Modell TN93 einschließt, welches *RAxML* nicht anbietet.

Parallele, auf POSIX-Threads basierende Berechnung eines Bootstrap-Baumes sowie des Maximum-Likelihood-Baumes mit dem höchsten Score mittels RAxML

```
raxmlHPC-PTHREADS -p 123614 -T 20 -s fgRevNucGenome_ALL.phy -m GTRGAMMA -n
raxml_output -x 12354 -f a -#200
```

Der entsprechende Consensus-Baum, erstellt wieder mit *PHYLIP consense* unter der MRe, liegt in Datei *raxml_consense.txt*.

Unter Verwendung der mit *JModeltest2* berechneten Nukleotidsubstitutionsmodelle für die Genfamilien wurden mit *PhyML* Genbäume generiert (Verzeichnis *Kapitel_Phylogenien/Referenzbaeume/MMN_ALL_SA_OKT12/MMN/data/phyml_GeneTrees*) und aus ihnen mit *PHYLIP consense* ein Consensus-Baum berechnet (Datei *geneTrees_consensus.phy*).

Die Unterstützung der Splits des Speziesbaumes durch die Genbäume, ausgedrückt in Prozent[17], wurde dem *E. coli*-Speziesbaum mit dem Python-Programm *SumTrees* aus dem *DendroPy*-Paket (Sukumaran und Holder, 2010) aufgeprägt. Die Bäume mit Split-Support-Werten sind in den Dateien *ec_phylipConsensus_mitSplitsSupport.tre* und *sa_phylipConsensus_mitSplitsSupport.tre* im Ordner *Referenzbaeume/Speziesbaum_und_Genbäume-Support* zu finden.

15 Für *PhyML* wurde der Parameter *use_median* gesetzt. Er legt fest, daß für die Bestimmung der Mitte jeder diskretisierten Klasse von Substitutionsraten nicht der Mittelwert, sondern der Median benutzt wird. Die Autoren weisen im Handbuch von *PhyML 3.0* darauf hin, daß durch diese Setzung gewöhnlich höhere Likelihood-Werte erzielt werden. Bei der Berechnung der Genbäume wurde allerdings der Mittelwert (Standardeinstellung) genommen. Das Gesagte gilt sowohl für die *E. coli*- als auch für die *Salmonellen*-Phylogenie. Es wurde nicht untersucht, inwieweit diese (irrtümlich vorgenommene) willkürliche Setzung den resultierenden Baum verändert.

16 Laut RAxML sind die Sequenzen von NC_017651 und NC_017652 (siehe Tabelle 1) identisch. Da unter *PhyML* mit beiden Sequenzen gearbeitet worden war, wurden beide Sequenzen im Datensatz belassen.

17 Die Prozentwerte wurden für die später vorzunehmenden Vergleiche in Zahlen zwischen 0 und 200 umgerechnet, da dies die Skalierung der Bootstrap-Zahlen in den übrigen Bäumen ist.

Berechnung der Split-Support-Werte (in %) der Genbäume auf dem E. coli-Speziesbaum mittels sumtrees.py

```
python sumtrees.py ec_ref_allGeneTrees.tre --unrooted -t
ec_phylipConsensus_mitGTS.tre --min-clade-freq=0.00 --decimals=0 >
ec_phylipConsensus_mitSplitsSupport.tre
```

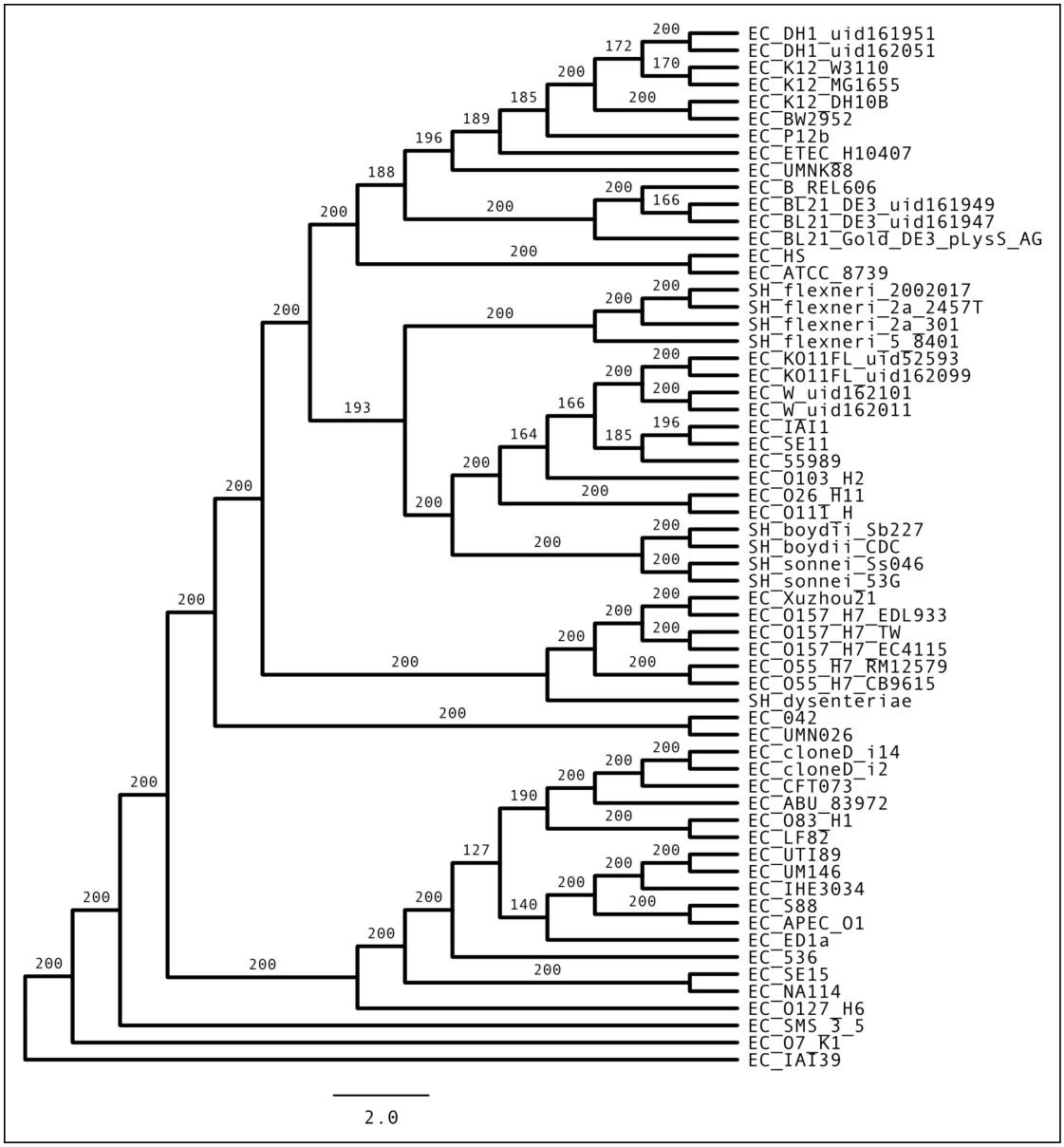


Abbildung 8.2*: E. coli-Phylogenie für 61 E. coli- und Shigella-Genome (Stand Oktober 2012) auf den konkatenierten Genfamilien der universellen Genfamilien unter einem Maximum-Likelihood-Framework.* Es wurden 200 Bootstrap-Replikate berechnet. Die Astlängen sind proportional dargestellt. Alle Äste bis auf einen mit einem Bootstrap Support von 127 haben einen Bootstrap-Support von mindestens 70 Prozent.

Salmonellen-Baum mit 25 Taxa

Die *Salmonellen*-Phylogenie, gezeigt in **Abb. 8.1**, (Datei *fgRevNucGenome_ALL.faa.phy_phyml_*

tree.txt) mit Bootstrap-Zahlen an den Ästen wurde mit *PhyML* aus dem MSA der konkatenierten Genfamilien (Datei *fgRevNucGenome_ALL.faa.phy* im Ordner *MMN_ALL_SA_OKT12/MMN/data/alignments/gblocks_reverse/concatenated)* für das Nukleotidsubstitutionsmodell TN93 berechnet. Zusätzlich wurde aus den Bootstrap-Replikaten in der Datei *fgRevNucGenome_ALL.faa.phy_phyml_boot_trees.txt* mit dem Programm *consense* aus der *PHYLIP*-Suite unter der erweiterten Majoritätsregel *MRe* ein Consensus-Baum (Datei sa_*outtree_phyml)* erstellt.

Auf den gleichen Eingabedaten wurde mit *RAxML* unter Verwendung der Rapid-Bootstrapping-Methode ein Bootstrap-Datensatz berechnet, aus dem anschließend mit *Consense* wieder ein Consensus-Baum (Datei *outtree* im Ordner *raxml_Consensus*) berechnet wurde. Alle Bäume basieren auf 200 Bootstrap-Replikaten. Alle Splits haben eine Unterstützung von über 70 %.

Für die universellen Genfamilien von *Salmonella* wurden unter dem Nukleotid-substitutionsmodellen TN93 und GTR Genbäume mit *PhyML* erzeugt[18] (Verzeichnis *Kapitel_Phylogenien/Referenzbaeume/MMN_ALL_EC_ohne_009801_002695/MMN/data/phyml_-GeneTrees*) und aus ihnen mit *PHYLIP consense* jeweils ein Consensus-Baum berechnet (Dateien sa_*geneTreesConsensus_GTR bzw. sa_geneTreesConsensus_TN93.txt*). Wie zuvor für den *E. coli*-Speziesbaum wurde auch dem *Salmonellen*-Speziesbaum die Unterstützung der Splits durch die Genbäume aufgeprägt. Die resultierenden Prozentwerte wurden wieder auf Zahlen zwischen 0 und 200 umgerechnet.

E. coli+Salmonellen-Baum mit 86 Taxa

Ebenfalls wurde eine Bootstrap-Phylogenie für die 61 *Escherichia coli*- und 25 *Salmonellen*-Genome (Datei *fgRevNucGenome_ALL.faa.phy_phyml_tree.txt* an der entsprechenden Position im Ordner *MMN_ECSA_OKT12*) auf den konkatenierten Genfamilien von *E. coli* und *Salmonellen* berechnet. Dabei wurden für *MMN* die gleichen Parameter wie zuvor benutzt.

Wurzeln der E. coli- und Salmonellen-Teilbäume

Für das Projekt, welches den Zusammenhang zwischen Nährstoff-Angebot in den anzestralen Umgebungen von *E. coli* und *Salmonellen* und lateralem Gentransfer auf den Phylogenien untersucht (vgl. **Kapitel 9**), war es notwendig, die *E. coli*- und *Salmonellen*-Bäume zu einem Baum zu verbinden. Da beide Teilbäume ungewurzelt sind, mußte dazu jeweils die Position in einem der beiden Bäume gefunden werden, an welcher der jeweils andere Baum angehängt würde. Implizit wird dabei festgelegt, an welchen Stellen die Teilbäume gewurzelt werden müssen. Das Programm *RAxML* bietet die Möglichkeit (Option *-r*), einen bifurzierenden Teilbaum einer zu schätzenden Phylogenie als unveränderlich festzulegen (engl. „Constraint Tree/Backbone Tree“). Die fehlenden Taxa werden sukzessive unter Benutzung eines Maximum-Parsimonie-Kriteriums hinzugefügt. Nachdem auf diese Weise ein Baum mit allen Taxa aufgebaut worden ist, optimiert *RAxML* ihn unter Maximum-Likelihood unter Einhaltung der durch den Constraint Tree auferlegten topologischen Einschränkungen.

Für den *E. coli*-Baum wurde über die beschriebene Methode die Wurzelposition dergestalt bestimmt, daß eine Phylogenie aller *E. coli*-Stämme und *Salmonellen* berechnet wurde, in der der Teilbaum *EC-Referenz* als Constraint Tree festgehalten wurde, dem schrittweise die *Salmonellen* angefügt wurden. Das als Datengrundlage benötigte MSA (Datei *fgRevNucGenome_ALL.faa.phy* im Ordner *ConstraintTrees/EC_mit_SA_ConstraintTree_ohne_009801_002695/MMN/data/alignments/gblocks_reverse/concatenated)* wurde mittels *MMN* mit der eingangs dargestellten Parameterbelegung für alle 86 *E. coli*- und *Salmonellen*-Genome berechnet. Es wurden jeweils 200

18 In den Experimenten des Autors waren der Consensus-Baum für die Salmonellen-Genbäume unter dem Modell TN93 und der Consensus-Baum unter dem Modell GTR hinsichtlich ihrer Topologie identisch (vgl. auch Fußnote 19).

Bootstrap-Wiederholungen berechnet. Aus Gründen der Schnelligkeit wurde das von *RAxML* angebotene Rapid-Bootstrapping-Verfahren benutzt (Option *-x*). Als Nukleotidsubstitutionsmodell kam *GTR* („General Time Reversible“) zum Einsatz, das das speziellere Modell *TN93* einschließt[19], welches *RAxML* nicht anbietet. Aus den Bootstrap-Replikaten in der Datei *RaxML_bootstrap.raxml_rapid* wurde mit dem Programm *Consense* unter der erweiterten Majoritätsregel ein Consensus-Baum (Datei *outtree* im Ordner *raxml_Consensus*) berechnet. Für den *Salmonellen-Baum* wurden die äquivalenten Schritte mit demselben Alignment durchgeführt.

Wurzelpositionen der E. coli- und Salmonellen-Teilbäume

Inferenz einer E. coli+Salmonellen-Phylogenie mit unveränderlichem Salmonellen-Teilbaum

```
raxmlHPC-PTHREADS -p 12345 -s fgRevNucGenome_ALL.faa.phy -m GTRGAMMA -r
sa_referenz.tre -n raxml_rapid -x 12354 -#200 -T 20
```

Inferenz eines E. coli+Salmonellen-Phylogenie mit unveränderlichem E. coli-Teilbaum

```
raxmlHPC-PTHREADS -p 12345 -s fgRevNucGenome_ALL.faa.phy -m GTRGAMMA -r
ec_referenz.tre -n raxml_rapid -x 12354 -#200 -T 20
```

Der Consensus-Baum (Datei *outtree*) befindet sich im Ordner *ConstraintTrees/SA_mit_EC_ ConstraintTree_ohne_009801_002695/MMN/data/alignments/gblocks_reverse/concatenated/raxml _Consensus*.

Konstruktion eines Gesamtbaumes aus den Teilbäumen

Die folgenden Erläuterungen beziehen sich auf die Dateien im Ordner *Gesamtbaum/ohne _009801_002695/ohneAstlängen.* Wie im vorigen Schritt erklärt, wurden geeignete Wurzelpositionen für den *E. coli*-Teilbaum und den *Salmonellen*-Teilbaum ermittelt. Der gewurzelte *Salmonellen*-Teilbaum wurde wie folgt gewonnen:

Der Newick-Ausdruck, der den Gesamtbaum mit fest belassenem *Salmonellen*-Teilbaum beschreibt (Datei *ec_mit_sa_constraintTree*), wurde unter Zuhilfenahme von *FigTree* und *Mesquite* so umgebaut, daß er an der Stelle, wo zuvor das *E. coli*-Phylum mit dem *Salmonellen*-Teilbaum verbunden war, gewurzelt ist (Datei *sa_mitWurzel.tre*). Der gewurzelte *E. coli*-Teilbaum wurde äquivalent hergestellt (Datei *ec_mitWurzel.tre*).

Beide Newick-Ausdrücke wurden manuell zu einem Ausdruck für den Gesamtbaum zusammengesetzt (siehe dazu Unterordner *ConstraintTrees/Wurzelpositionen_Teilbaeume*). Zwischenschritte sind in der Datei *combinedTree_Zwischenschritte.txt* gespeichert, der Gesamtbaum ist in der Datei *combinedTree.txt* zu finden.

Schätzen der Astlängen des Gesamtbaumes

Die folgenden Erläuterungen beziehen sich auf die Dateien im Ordner *Gesamtbaum/März2013_ FinalerBaum/ohne_009801_002695/mitAstlängen.* Der Gesamtbaum hat seine Wurzel an der Position, an welcher der *E. coli*- und *Salmonellen*-Teilbaum miteinander verbunden wurden. Da die dort zusammenlaufenden Äste manuell eingefügt worden sind, ist deren Länge unbekannt. Die Astlängen des Gesamtbaumes wurden mit *PhyML* wieder auf Grundlage des Alignments geschätzt, das bereits im Rahmen der Berechnungen für die Bestimmung der Wurzelpositionen der Teilbäume

19 Es wird in Wissenschaftskreisen angenommen, daß dies bei verschachtelten (engl. „nested“) Modellen zulässig ist. Das allgemeinere Modell GTR umschließt in diesem Fall das speziellere Modell TN93. Eine Quelle kann hier nicht benannt werden.

benutzt wurde. *PhyML* optimiert die Astlängen für eine Topologie, die beim Aufruf als Startbaum (Option *inputtree*) und gleichzeitig als Constraint-Baum (Option *constraint_file*) spezifiziert wird. Es sollen lediglich die Astlängen, nicht aber die Topologie, optimiert werden:

Optimieren der Astlängen der Phylogenie aus E. coli und Salmonellen mit PhyML

```
phyml -i fgRevNucGenome_ALL.faa.phy --datatype nt -f e -v e -s BEST -o l
--inputtree combinedTree.tre --constraint_file combinedTree.tre --r_seed
11223 --use_median --model TN93
```

Der Gesamtbaum, zu sehen als Kladogramm in **Abbildung 8.3**, wurde in der Datei namens *finalerBaum_mitAstlaengen.tre* („ECSA-Referenz“) abgelegt. Er wird mit unter Zuhilfenahme von *PhyML* geschätzten Astlängen in **Abschnitt 8.3.3, Abbildung 8.13** gezeigt.

EC-Referenz, SA-Referenz und ECSA-Referenz

Das *E. coli*-Monophylum im Gesamtbaum wird für den Rest dieses Kapitels als „EC-Referenz“, das *Salmonellen*-Monophylum als „SA-Referenz“ und der Gesamtbaum als „EC+SA-Referenz“ bezeichnet. Die aus dem Gesamtbaum herausgelösten Teilbäume sind in den Dateien *ec_referenz.tre* und *sa_referenz.tre* im Ordner *Gesamtbaum/März2013_FinalerBaum/ohne_009801_002695/mit Astlängen* abgelegt.

8.2.2.2 Inferenz in einem Bayesianischen Framework (MrBayes, BEAST)

Die Daten und Ergebnisse für diesen Abschnitt sind innerhalb des Projektordners *Kapitel_Phylogenien* im Ordner *MrBayes_Laeufe* gespeichert.

MrBayes benutzt den Bayesianischen Wahrscheinlichkeitsbegriff

Der Bayesianische Wahrscheinlichkeitsbegriff unterscheidet sich von dem herkömmlichen, „frequentistischen“ Wahrscheinlichkeitsbegriff in mehrfacher Hinsicht. Zum einen beruht er auf dem Theorem von Bayes, mit dem unbekannte Parameter geschätzt, ihre Konfidenz angegeben und Hypothesen für sie abgeleitet werden können. Zum anderen sind die unbekannten Parameter in der bayesianischen Statistik Zufallsvariablen. Schließlich erweitert sie den Wahrscheinlichkeitsbegriff dergestalt, als daß die Wahrscheinlichkeit einer Aussage ein Maß für ihre Plausibilität wird. Um statistische Probleme mithilfe dieses Wahrscheinlichkeitsbegriffes analysieren zu können, wurde das Konzept der „A-priori-Wahrscheinlichkeit“ eingeführt, welches vorsieht, daß etwaiges Vorwissen und subjektive Grundannahmen (engl. „priors“) zu dem zu untersuchenden Sachverhalt in einer Wahrscheinlichkeitsverteilung ausgedrückt werden.

Das Programm *MrBayes* (Ronquist und Hülsenbeck, 2003), das ein Bayesianisches Framework zur Inferenz einer Phylogenie für ein gegebenes Alignment nutzt, wurde verwendet, um eine *E. coli*- und eine *Salmonellen*-Phylogenie auf Grundlage der Alignments zu schätzen, die schon für die Berechnungen in **Abschnitt 8.2.2.1** benutzt wurden. *MrBayes* verwendet Priors, die auf den Bereich Bauminferenz zugeschnitten sind. Die A-posteriori-Wahrscheinlichkeit der phylogenetischen

Bäume und anderer Parameter des Substitutionsmodells können nicht analytisch bestimmt werden. Stattdessen wird die posteriore Wahrscheinlichkeitsverteilung der Bäume approximiert, indem abhängige Stichproben aus ihr gezogen werden und das verwendete Modell schrittweise angepaßt wird.

Im Gegensatz zu Programmen für die Bauminferenz unter Maximum-Likelihood (ML) wie *PhyML* und *RAxML* liefert *MrBayes* nicht einen Baum mit hoher Likelihood, sondern eine Menge von Bäumen mit möglichst hohen Likelihoods, die während des eingesetzten Markov Chain Monte Carlo-Prozesses („MCMC“) gefunden wurden. Weder die ML-basierten Verfahren noch die bayesianischen Verfahren, die im Kern ebenfalls mit ML arbeiten, finden garantiert den auf das gegebene Alignment am besten passenden Baum bzw. die am besten passenden Bäume, und erst recht nicht garantiert den „wahren“ Baum. Die in einem MCMC-Lauf gefundenen Bäume können von den Bäumen eines anderen Laufs, der auf der gleichen Datengrundlage beruht, abweichen, da das Verfahren randomisiert arbeitet. Startet man beide Läufe mit dem gleichen Startwert für den Zufallszahlenalgorithmus, werden sie allerdings das exakt gleiche Ergebnis liefern.

Im Rahmen dieser Arbeit hat der Verfasser das Programm *GPU MrBayes* benutzt[20], das Markov-Ketten parallel auf einer *CUDA*-fähigen *NVidia*-Grafikkarte simuliert. Aus Gründen der Effizienz bietet *GPU MrBayes* lediglich das Nukleotidsubstitutionsmodell GTR+I+R (gamma-verteilte Raten und invariante Positionen) an. *MrBayes* kann entweder interaktiv oder über eine Textdatei[21] im Nexus-Format gesteuert werden. Für diese Arbeit wurde das Programm im interaktiven Modus benutzt.

Arbeitsweise von MrBayes

Das Programm benutzt $(MC)^3$, „Metropolis-coupled Markov Chain Monte Carlo“/ „MCMCMC“, eine Variante von MCMC. $(MC)^3$ setzt eine wählbare Anzahl von aufsteigend nummerierten Markov-Ketten ein, von denen alle bis auf eine, die „kalte“ Kette, „erhitzt“ werden. Je höher die Temperatur einer Kette, desto höher ist die Wahrscheinlichkeit dafür, daß sie sich von einem isolierten Gipfel (Extremum) zu einem anderen innerhalb der posterioren Verteilung bewegen wird. Die Hitze von Kette i ist $B = 1 / (1 + i \cdot \lambda)$, wobei λ der von B gesteuerte Anheizkoeffizient *ist.* Ihre posteriore Wahrscheinlichkeit wird in die B-te Potenz erhoben. Für $B = 0$ werden alle Bäume mit gleicher Wahrscheinlichkeit besucht, die Kette besucht Bäume dabei weitgehend beliebig. $B = 1$ ist die „kalte“ Kette bzw. die Verteilung an sich. Nachdem jede Kette einen Schritt in der Suchlandschaft vollzogen hat, werden zwei Ketten zufällig gezogen, und der Versuch wird unternommen, ihren Zustand zu tauschen. Die Akzeptanzwahrscheinlichkeit dafür bestimmt die Gleichung von Nicholas Metropolis (Metropolis et al., 1953). Exzessives Erhitzen kann aber dazu führen, daß die Akzeptanzrate zwischen verschiedenen Ketten allzu gering wird. Während der Burn-in-Phase (s. u.) schwanken die Akzeptanzraten mitunter beträchtlich. Die Autoren von *MrBayes* heben im Handbuch auf S. 21 hervor, daß ihrer Erfahrung nach für die Analyse bestimmter Datensätze Erhitzen (engl. „Heating“) wichtig sei, für andere nicht, ohne jedoch diesbezügliche Regeln angeben zu können.

20 Tatsächlich hat der Verfasser daneben auch das Programm *MrBayes* v3.2.1. X64 benutzt, allerdings nur für Analysen, die nicht in dieser Arbeit gezeigt werden. Für diese Programmversion ist die hier dargestellte Arbeitsweise im übrigen ebenso zutreffend.
21 Die Datei muß einen sog. „MrBayes-Block“ mit dem Alignment, Steuerbefehlen und Parameterbelegungen enthalten. Auskunft zum Aufbau dieses Blocks gibt das Benutzerhandbuch zu MrBayes (http://mrbayes.sourceforge.net/manual.php).

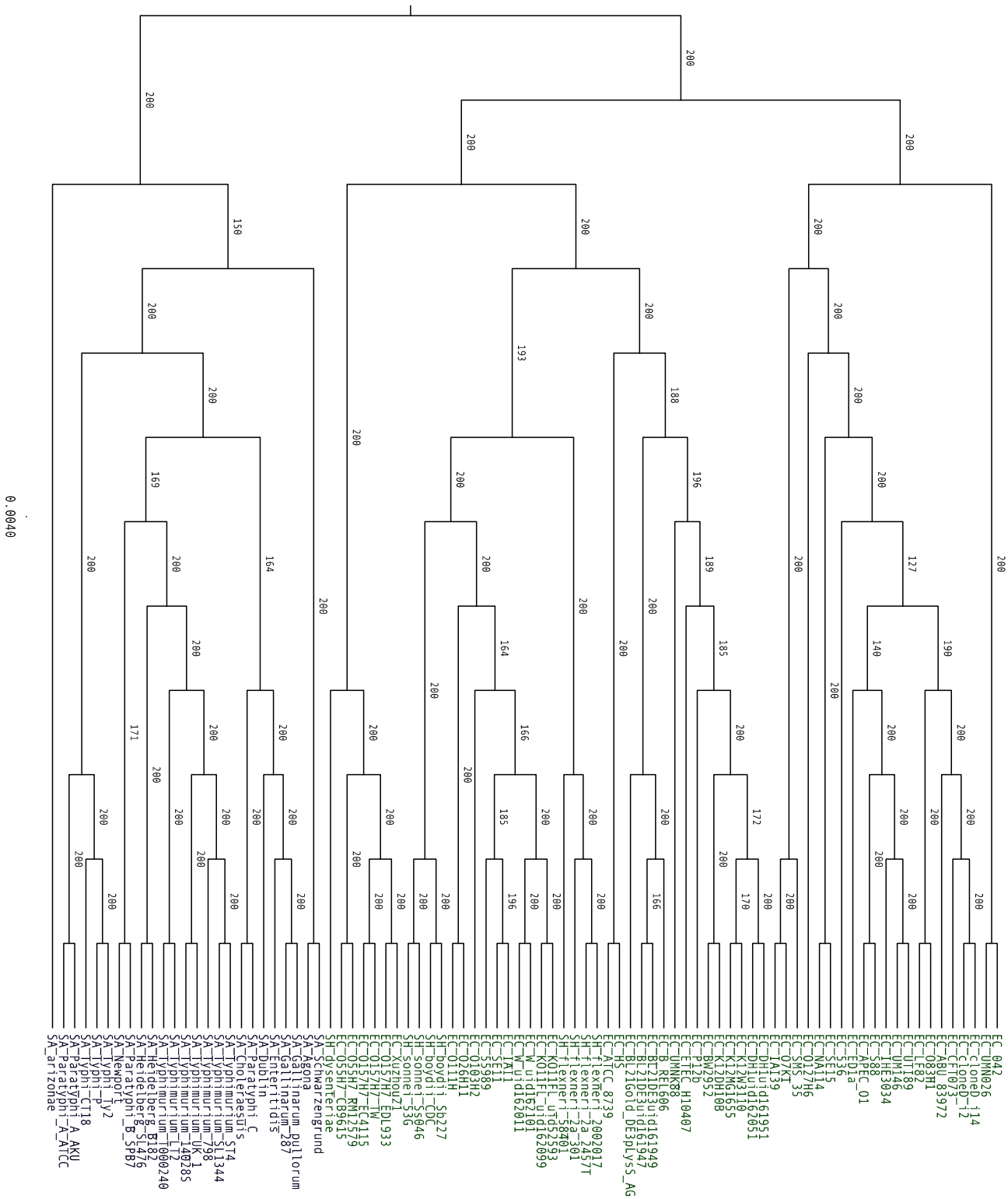


Abbildung 8.3*: Gewurzeltes Kladogramm für 86 E. coli und Salmonella (ECSA-Referenz).* Die Phylogenie beruht auf den ungewurzelten E. coli- und Salmonellen-Teilbäumen. Die Wurzelposition für den E. coli-Teilbaum wurde über RAxML aus einem Baum für alle Taxa bestimmt, in dem der E. coli-Teil fest belassen („Constraint Tree“) und dem die Salmonellen inkrementell angefügt wurden. Die Wurzelposition des E. coli-Teilbaumes ist die Position, an der der Salmonellen-Teilbaum mit ihm verbunden ist. Die Wurzelposition für den Salmonellen-Teilbaum wurde analog bestimmt. Die Äste tragen die für die Teilbäume separat ermittelten Bootstrap-Zahlen, die auf jeweils 200 Wiederholungen basieren.

Für beide Analysen zusammen kann *MrBayes* jeweils nach Verstreichen einer wählbaren Zahl von Generationen eine Diagnostik berechnen, anhand der das Programm entscheiden kann, ob die unterschiedlichen MCMC-Läufe konvergieren, sie also die posteriore Verteilung hinreichend gut abgebildet haben. Am Ende sollten die abgetasteten Bäume weitgehend gleich sein. Das wichtigste Diagnostik-Maß, die Ähnlichkeit der abgetasteten Bäume aus den verschiedenen Läufen, wird am Bildschirm angezeigt.

Immer, wenn die Diagnostik berechnet wird, wird – falls der Benutzer es wünscht - entweder eine feste Zahl oder ein Prozentsatz von Samples vom Beginn einer Kette gelöscht. Die Phase eines MCMC-Laufes, in der Samples gelöscht werden, wird als „Burn-in-Phase“ bezeichnet. Der Burn-in wird allgemein als notwendig erachtet, da eine Markov-Kette gewöhnlich erst nach Verstreichen einer bestimmten Zahl von Generationen einen stationären Zustand mit Likelihood-Werten, die innerhalb eines mehr oder weniger stabilen Bereiches fluktuieren, erreicht[22]. Das bedeutet aber keineswegs, daß die Markov-Kette dieses „Plateau“ nicht doch nach Verstreichen einer hohen Zahl von Generationen verläßt und dann Bäume findet, die eine noch höhere Likelihood haben.

Gegenüber MCMC, das nur eine Kette benutzt, bietet $(MC)^3$ den Vorteil, daß mehrere Ketten lokale Extrema in der Suchlandschaft unabhängig voneinander besuchen können. Ist eine Kette in einem lokalen Optimum „gefangen“, läuft die Analyse trotzdem weiter, da die anderen Ketten noch aktiv sind.

Auswertung der MCMC-Läufe mit MrBayes

Die Stationärität bzw. die sog. *Mixing Time*[23] der in dieser Arbeit durchgeführten Läufe wurde mit den Kommandos `sump` und `sumt` von *MrBayes* und mit dem Programm *Tracer* aus der *BEAST-Suite* (Drummond und Rambaut, 2007) beurteilt. *MrBayes* bietet die Kommandos `sump` und `sumt`, mit denen sich die abgetasteten Parameter, Topologien und Astlängen übersichtlich zusammenfassen lassen. Während eines MCMC-Laufes speichert das Programm die Parameter *Generation, Log Likelihood* der kalten Kette (*LnL*), Summe aller Astlängen (*TL*, engl. „Total tree length“), die sechs GTR-Parameter für die Raten *r(A<->C), r(A<->G)* etc., die vier stationären Nukleotidfrequenzen *pi(A), pi(C)* etc., den Shape-Parameter *alpha* für die Variation der Raten sowie den Anteil invariabler Position *pinvar* in Dateien mit der Endung `.p`. In dieser Arbeit wurde ausschließlich das Modell GTR+I+R verwendet; für ein anderes Modell weicht der Inhalt der Datei selbstverständlich ab.

Das Kommando `sump` verwendet standardmäßig die bei der Berechnung der Markov-Kette benutzten Burn-in-Einstellungen. Es generiert einen Graphen, in dem die Logarithmen der Wahrscheinlichkeit, die Daten zu beobachten (lnL-Werte), je Generation aufgetragen sind. Hat ein Lauf hinsichtlich der lnL-Werte innerhalb der dargestellten Generationen einen stationären Zustand erreicht, zeigen die Werte keinerlei Tendenz zu steigen oder zu fallen[24]. Sind solche Trends trotzdem zu sehen, sollte die Analyse nochmals mit höherer Kettenlänge ausgeführt werden. Zusätzlich liefert `sump` eine Tabelle der Mittelwerte und Varianzen der oben genannten Parameter, die oberen und unteren Grenzen des 95 % Konfidenzintervalls und den Median der abgetasteten Werte. Die letzte Spalte der Tabelle enthält den PSRF (engl. *„Potential Scale Reduction Factor“*), der als Diagnosemaß für die Konvergenz benutzt werden kann. Wurde die posteriore Wahrscheinlichkeitsverteilung hinreichend gut abgetastet, bewegt sich der PSRF zwischen 1.00 und 1.02[25].

22 Neben Befürwortern des Burn-Ins gibt es auch Wissenschaftler, die seinen Nutzen anzweifeln. Für eine Diskussion zu „Fallstricken“ bei der Arbeit mit *MrBayes* sowie zur Bedeutung der Burn-In-Phase siehe http://treethinkers.blogspot.de/2009/05/for-this-unauthorized-installment-of.html.

23 Im Zusammenhang mit MCMC bezeichnet die *Mixing Time* die verstrichene Zahl von Generationen, die eine Markov-Kette benötigt hat, um, ausgehend von einer beliebigen Startposition, einen annähernd stationären Zustand zu erreichen. Eine „gute“ Kette weist „schnelles Mixing“ auf.

24 Die Autoren von *MrBayes* vergleichen das Aussehen des Plots im Fall der Stationarität anschaulich mit „weißem Rauschen“.

25 Es ist nicht unumstritten, den PSRF hier als Diagnosemaß einzusetzen. Leider kann hierzu keine Quelle benannt werden.

Das Kommando `sumt` liefert neben einer mit Indices versehenen Liste aller Bipartitionen insbesondere eine Tabelle der informativen Bipartitionen (solche Partitionen, die mehr als ein Taxon beinhalten). Für jede Partition sind folgende Informationen enthalten: ihr Index (ID), wie oft sie abgetastet wurde (*#obs*), ihre Wahrscheinlichkeit (*Probab.*), Standardabweichung (*Sd(s)+*), Minimum (*Min(s)*) und Maximum (*Max(s)*) ihrer Häufigkeit über die Läufe sowie die Zahl der unabhängigen Analysen (*Nruns*), in welchen die Partition angetroffen worden ist. Man erhält eine weitere Tabelle mit den Parameter der Äste, z. B. Astlängen und Knoten, und neben anderen Größen wieder den PSRF, für den das bereits Gesagte gilt. Ferner gibt `sumt` ein Phylogramm aus, das auf den durchschnittlichen Astlängen beruht, sowie ein einen Baum mit *Bayesian Posterior Probabilities* (BPP) an den Ästen. Dieser Baum kann nach den gleichen Gesichtspunkten ausgewertet werden, wie ein Baum mit Bootstrap-Zahlen. Im Hintergrund produziert `sumt` mehrere Dateien: Taxon-Bipartitionen und ihre Indices/IDs (**.parts*), zusammenfassende Statistik der Partitionen (**.tstat*) bzw. der Astlängen (**.vstat*), Consensus-Baum mit allen Konfidenzwerten und Astlängen (**.con*).

Auswertung der MCMC-Läufe mit Tracer

Das Java-Programm *Tracer* aus dem *BEAST*-Paket bietet unter einer grafischen Benutzeroberfläche eine ähnliche Auswertungsfunktionalität wie *MrBayes* mit `sumt`/`sump`. Es erlaubt dem Benutzer, die Ausgaben von einem der Bayesianischen MCMC-Programme *BEAST, MrBayes* und *LAMARC* grafisch darzustellen und statistisch zu analysieren. Entsprechende Details können der Datei *README.txt* im Installationsordner von BEAST entnommen werden. Für diese Arbeit wurden die sog. „Likelihood-Traces“ (siehe **Abb. 8.4**) ausgewertet, die inhaltlich dem Graph entsprechen, den `sump` produziert. *Tracer* informiert den Benutzer mittels einer Farbcodierung darüber, ob die für den MCMC-Lauf statistisch relevanten Größen genügend abgetastet worden sind. Die in der Spalte „ESS“ („Estimated Sample Size“) rot hervorgehobenen Größen wurden in dem dargestellten MCMC-Lauf unterabgetastet, die blaßroten noch genügend, die schwarzen ausreichend. Der Lauf mit 30 Mio. Generationen und einer Burn-in-Phase von 50 % sollte also mit einer höheren Generationenzahl wiederholt werden. Die Log-Likelihood (LnL) der gezogenen Bäume, für die genug Stichproben vorhanden sind, bewegt sich nach der Burn-in-Phase um den Mittelwert -4554740 herum. Ein Abwärts- oder Aufwärtstrend der Log-Likelihoods ist nicht zu erkennen (obwohl der Lauf insgesamt zu kurz ist).

Bewegt sich die LnL bei einem anderen MCMC-Lauf auf den gleichen Daten und gleichen Parametern um den gleichen Mittelwert, können die Läufe zusammengefaßt werden, wodurch im Endeffekt eine höhere, für die hier betrachteten Daten ausreichende Abtastung der Suchlandschaft erzielt wird. Der Verfasser hat sich dieser Möglichkeit bedient und verschiedene *MrBayes*-Läufe für *E. coli* und *Salmonellen* kumuliert betrachtet und ausgewertet.

MrBayes-Läufe für den E. coli-Datensatz

Für den *E. coli*-Datensatz wurden drei Programmläufe mit jeweils 15 Mio. Samples, einer Burn-In-Phase von 3,75 Mio. Samples und jeweils unterschiedlichem Startwert für den Zufallszahlengenerator durchgeführt. Es wurde die Parameterbelegung verwendet, die aus der untenstehenden Auflistung ersichtlich ist. Das zugrunde liegende Verzeichnis ist *EC/ohne_009801_002695*. Da *MrBayes* lediglich Alignments im „neuen“ Nexus-Format (**.nnex*) akzeptiert, wurde das Original-Alignment zunächst mittels der Exportfunktionalität von *Mesquite* in dieses Format konvertiert (Datei *ec_modified.nnex*). Die Dateien für die drei Läufe befinden sich in den Ordnern *lauf1, lauf2* und *lauf3*. Screenshots der mit Tracer erzeugten Likelihood-Traces sind im Ordner *LikelihoodTraces* zu finden.

Während der drei Läufe wurden jeweils neun unterschiedliche Bäume gefunden (vgl. Datei *ec_modified.nnex.trprobs* und **Tab. 8.4**). In Lauf 1 und Lauf 3 besitzt der Baum mit der höchsten Wahrscheinlichkeit bzw. Posterior Probability (PP), in Lauf 2 der Baum mit der zweithöchsten Wahrscheinlichkeit die gleiche Topologie wie der Baum *EC-Referenz*[26]. Der maximale *RF-Abstand*[27] zwischen einem Baum und dem Baum *EC-Referenz* ist, unabhängig von der Posterior Probability des Baumes, entweder 2 oder 4.

Die Unterschiede in den Bäumen ergeben sich größtenteils aus dem Platztausch verschiedener Stämme im Teilbaum der *E. coli K12*- und *B*-Stämme. Die evolutionären Abstände der Bakterien voneinander sind gering, das phylogenetische Signal daher kaum ausgeprägt. Die *RF-Distanze*n zwischen den Bäumen wurden mit dem Programm *TOPD/FMTS* (Puigbo et al., 2007) berechnet. Neben der *RF-Distanz* kann *TOPD/FMTS* die Taxa ausgeben, deren Positionen in den Bäumen voneinander abweichen („Disagree“-Maß).

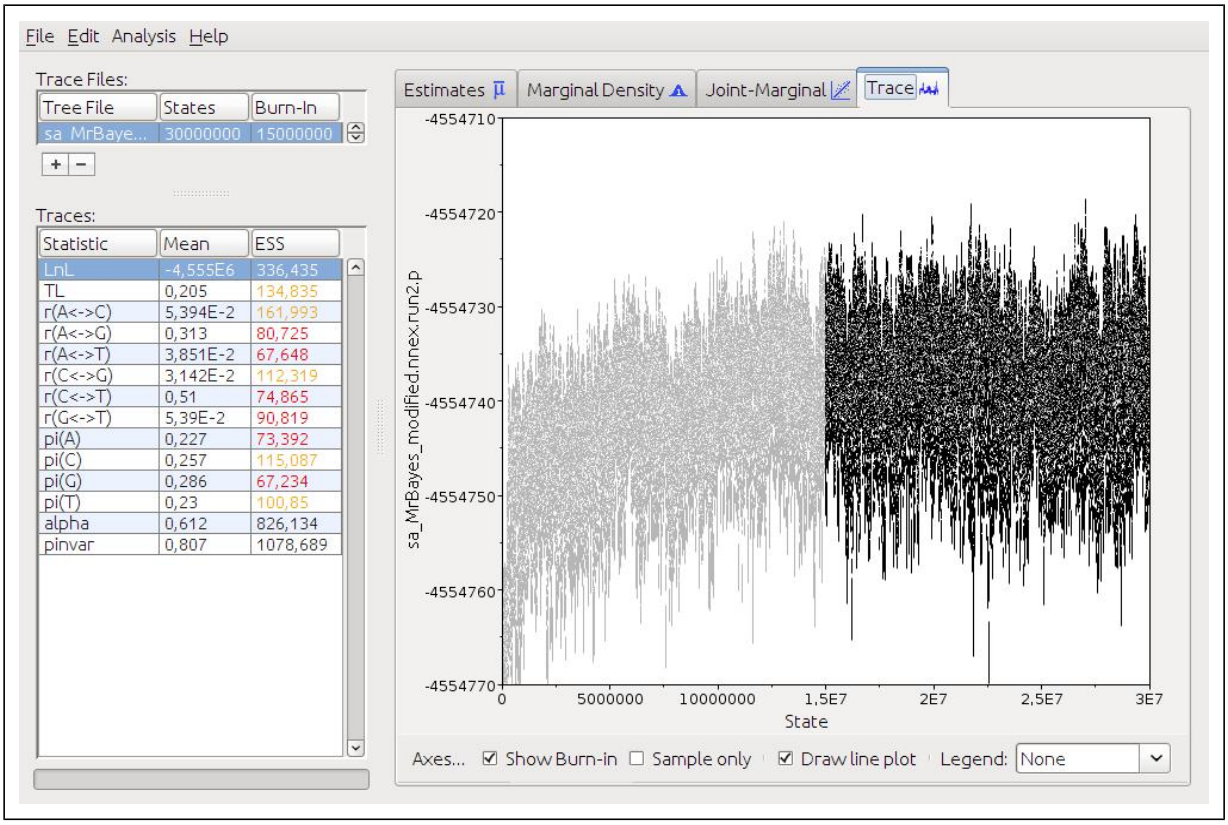


Abbildung 8.4*: Likelihood-Trace für einen MrBayes-Lauf*

Berechnung der symmetrischen Differenz (Split-Abstand) und den Unterschieden nach der Disagree-Methode von *topd*

```
perl topd_v3.3.pl -f tree8_vs_referenz.txt -r no -m disagree -l 4
```

Sind die IDs für zwei Bäume in der Tabelle identisch, handelt es sich um den gleichen Baum, der in mindestens zwei verschiedenen Läufen aufgefunden worden ist. Das Programm *PHYLIP treedist* wurde zur Erstellung einer Distanzmatrix aller Bäume benutzt, aus der die Gleichheit von Bäumen

26 Bei EC-Referenz handelt es sich um den gewurzelten *E. coli*-Baum, hier wurde aber der ungewurzelte *E. coli*-Baum verwendet. Der Einfachheit halber werden die Begriffe EC-Referenz und SA-Referenz in diesem Kapitel weiterhin in dieser Form verwendet.
27 Der Split-Abstand ist identisch mit dem Robinson-Foulds-Abstand (Robinson und Foulds, 1981).

direkt abgelesen werden konnte.

Die Ergebnisse der drei Läufe für den *E. coli*-Datensatz wurden kumuliert betrachtet (Ordner *EC/kumulierteLaeufe*). Zu diesem Zweck wurde die Datei *cumulated_renum.nnex.t* manuell aus den *.t*-Dateien der einzelnen Läufe zusammengesetzt, wobei wie folgt vorgegangen wurde:

- Header sowie „End;“ aus den *.t*-Dateien der Läufe 1 und 2 entfernen
- Die 3749999 Bäume aus der Burn-in-Phase entfernen
- Modifizierte *.t*-Dateien konkatenieren
- Numerierung der Bäume und Parametersätze mit den Skripten `replaceFirstColumn_t.pl` und `replaceFirstColumn_p.pl` erneuern.
- Resultierende Datei mit *MrBayes* analysieren: `execute cumulated_renum.nnex`, dann `sump Relburnin=No Nruns=1.`

Der von `sump` ausgegebene Graph der LnL pro Generation zeigt „weißes Rauschen“ mit einem minimalen Aufwärtstrend. Es empfiehlt sich daher, die Burn-in-Phase etwas zu verlängern. Die ESS („Estimated sample size“) für alle abgetasteten Parameter sind alle deutlich größer als 100 und weisen damit auf eine ausreichende Abtastung hin. Der Baum mit der höchsten *PP* ist identisch mit dem Baum aus Lauf 1 mit der höchsten *PP* und gleicht damit dem Baum EC-Referenz. Der MrBayes-Consensus-Baum für die neun Bäume hat *RF-Abstand* 1 von EC-Referenz: die sehr nah verwandten *E. coli-B*-Stämme *BL21 DE3 uid161947*, *BL21 DE3 uid161949* und *B REL606* bilden einen multifurzierenden Teilbaum (siehe Datei *cumulated_renum.nnex.con.tre*).

MrBayes-Läufe für den Salmonellen-Datensatz

Mit dem *Salmonellen*-Alignment mit 25 Taxa und 2.230,097 Zeichen wurden sechs Läufe mit jeweils zwei unabhängigen Analysen („Runs“), fünf Markov-Ketten, Temperaturkoeffizient=3, Swap-Frequenz=2 (`mcmcp Nchains=4 Temp=0.3 Swapfreq=2`) mit unterschiedlicher Generationenzahl und variierender Abtastrate durchgeführt. Alle notwendigen Quell- und Ergebnisdateien befinden sich im Unterordner *SA*.

In Lauf 1 wurde ein Baum mit PP = 1 gefunden, der identisch mit dem Baum *SA-Referenz* ist. In den Läufen 2 – 6 wurden jeweils drei Bäume gefunden, jeweils einer mit PP = 0,5. Einer unter den drei Bäumen war jeweils identisch mit dem Baum *SA-Referenz,* allerdings nicht immer der mit der höchsten PP, die anderen Bäume wiesen jeweils einen RF-Abstand von 6 zu *SA-Referenz* auf. In diesen Bäumen tritt das Serovar *Paratyphi B* als Outgroup zu *Paratyphi C* und *Choleraesuis SC B67* auf, während es im Referenzbaum einen Teilbaum mit *Newport SL254* bildet (siehe **Tab. 8.6**). In den beiden unabhängigen, parallel ausgeführten Analysen der Läufe 2, 5 und 6 haben die Markov-Ketten nicht konvergiert, daher sind die Ergebnisse als Einzelresultate nur eingeschränkt brauchbar. *MrBayes* empfiehlt entweder eine höhere Zahl von Generationen oder eine Modifikation der Heating-Parameter (siehe Textdateien *MrBayes_Warnung.txt* im Unterordner *Laeufe)*.

Wie zuvor die Ergebnisse der drei MCMC-Läufe für den *E. coli*-Datensatz wurden die Läufe für *Salmonella* kumuliert betrachtet. Alle Dateien befinden sich im Ordner *MrBayes_Laeufe/SA/ kumulierteExperimente/Laeufe2bis5*. Es wurden allerdings nicht jeweils beide Runs aller sechs Läufe benutzt: Lauf 1 wurde komplett ausgespart, da eine Begutachtung des „Likelihood Trace“-Diagramms mit *Tracer* auf eine Unterabtastung diverser Parameter hinwies. Von den Läufen 2 und 5 wurden die Bäume aus Run2, von den Läufen 3 und 4 die Bäume aus Run1 gewählt, da die jeweils letzten 50 gesampleten Bäume hinsichtlich ihrer negativen Log-Likelihood eine Gruppe der

am besten bewerteten Bäume bilden (vgl. **Tab. 8.5**). Die restlichen Runs, ausgenommen Run2 von Lauf 6, bilden eine Gruppe mit weniger hoch bewerteten Bäumen. Run2 von Lauf 6 liegt bezüglich seiner Bewertung zwischen den Gruppen. In den Läufen 2, 3, 5 und 6 besitzt der am höchsten bewertete Baum jeweils eine PP und eine kumulierte PP von 0,500. Dieses auffällige Ergebnis ist dadurch zustande gekommen, daß in einem Run immer ein einziger Baum aufgefunden worden ist, der keinem der zwei im anderen Run aufgefundenen Bäume gleicht. Das weist darauf hin, daß in den Runs der Suchraum wahrscheinlich nicht ausreichend abgetastet worden ist. Es sind weitere Experimente notwendig, die Hinweise zu möglichen Ursachen der Unterabtastung geben können.

MrBayes GPU: Parameter für den E. coli-Datensatz

Nicht genannte Parameter verbleiben in ihrer Standardeinstellung.

Filename	
exe ec_modified.nnex	Anmerkung: Alignment mit 61 Taxa, 1.296.988 Zeichen
Evolutionäres Modell	
lset nst=6 rates=invgamma	6 Substitutionstypen (GTR), Raten sind gammaverteilt, invariante Positionen
nucmodel=4by4	Modell für Nukleotidsubstitution wird über eine 4x4-Matrix ausgedrückt
Priors	
Standardeinstellung	
Parameter für den MCMC-Lauf	
Filename	*ec_modified.nnex*
NGen=15000000	Anzahl Generationen für den MCMC-Algorithmus
NRuns=1	Zahl der unabhängigen, gleichzeitig zu startenden Analysen
NChains=2	Zahl der Ketten pro Analyse (1 kalte Kette, Rest warme Ketten)
Relburnin=Yes Burninfrac=0.25	Die ersten 100·*Burninfrac* % der Samples werden als Burn-in gelöscht
Stoprule=No	Es wird kein Konvergenzkriterium benutzt. Stattdessen läuft eine Kette *NGen* Generationen und wird danach gestoppt.
DiagnFreq=1000	Zahl von Generationen zwischen der Berechnung von MCMC-Diagnostiken. Ausgabedatei ist *Filename.mcmc.*
Metropolis-Coupling	
Temp=0.1	Temperatur-Parameter für das Erhitzen der Ketten
SwapFreq=1 NSwaps=1	Nach jeweils *SwapFreq* Generationen wird versucht, *NSwaps* Zustände zwischen Paaren von Ketten auszutauschen
SampleFreq=500	Bäume werden alle *SampleFreq* Generationen gesammelt. Ausgabedateien sind *Filename.p* für die Parameter des Substitutionsmodells und *Filename.t* für die Topologien und Astlängen.

Tabelle 8.3*: MrBayes-Parameter für den E. coli-Datensatz*

Die ausgewählten Läufe wurden auf die gleiche Weise zusammengefaßt wie zuvor die *E. coli*-Läufe, indem die Einzeldateien wie dort beschrieben modifiziert worden sind. Die entsprechende Datei *cumulatedRun_renum.nnex.t* enthält 771.061 Samples (d. h. untersuchte Baumtopologien).

E. coli-Datensatz: 15 Mio. Generationen, 300.000 Bäume

Lauf 1

Baum	PP	Kumulierte PP	Split-Distanz zu EC-Referenz	ID
1	0,344	0,344	0	A
2	0,307	0,651	2	B
3	0,299	0,651	2	C
4	0,010	0,960	2	D
5	0,010	0,969	2	E
6	0,008	0,977	4	F
7	0,008	0,985	4	G
8	0,008	0,993	4	H
9	0,007	1,000	4	I

Lauf 2

Baum	PP	Kumulierte PP	Split-Distanz zu EC-Referenz	ID
1	0,338	0,338	2	C
2	0,319	0,656	0	A
3	0,287	0,943	2	B
4	0,011	0,954	4	I
5	0,011	0,965	4	G
6	0,009	0,974	2	E
7	0,009	0,983	4	H
8	0,009	0,992	2	D
9	0,008	1,000	4	F

Lauf 3

Baum	PP	Kumulierte PP	Split-Distanz zu EC-Referenz	ID
1	0,329	0,329	0	A
2	0,310	0,640	2	C
3	0,308	0,947	2	B
4	0,009	0,956	4	I
5	0,009	0,965	2	D
6	0,009	0,974	4	H
7	0,009	0,983	4	G
8	0,008	0,992	2	E
9	0,008	1,000	4	F

Läufe 1 – 3 kumuliert

Baum	PP	Kumulierte PP	Split-Distanz zu EC-Referenz	ID
1	0,331	0,331	0	A
2	0,315	0,646	2	C
3	0,300	0,947	2	B
4	0,009	0,956	2	D
5	0,009	0,965	4	H
6	0,009	0,974	2	E
7	0,009	0,983	4	G
8	0,008	0,992	4	H
9	0,008	1,000	4	

Tabelle 8.4: *E. coli-Datensatz: In den Läufen 1 – 3 aufgefundene Bäume*

MrBayes-Lauf	Generationen-zahl	Anzahl Bäume pro Run	Durchschnittl. Log-Likelihood der letzten 50 gefundenen Bäume	
			Run1	Run2
1	1,0 Mio.	20000,0	-4658978,99280	-4658963,93176
2	30,0 Mio.	300000,0	-4554918,65364	-4554736,94974
3	12,8 Mio.	85378,0	-4554731,38738	-4554923,89032
4	40,0 Mio.	400000,0	-4554735,09700	-4554924,62742
5	20,0 Mio.	400000,0	-4554924,17728	-4554732,21962
6	15,0 Mio.	300000,0	-4554925,87672	-4554741,23126

Tabelle 8.5: *MrBayes-Läufe und Log-Likelihoods der letzten 50 Bäume*

Der Likelihood-Trace-Graph, angezeigt mittels `sump`, wies drei „Plateaus“ verschiedener Log-Likelihoods auf und unterstützte damit die Vermutung, daß das Mixing wahrscheinlich nicht ausreichend gewesen ist (siehe Dateien *cumulatedRun_renum.nnex.stat* und **.parts*, **.con*, **.trprobs*). Die einzelnen Läufe sollten mit einer höheren Generationenzahl wiederholt werden.

MrBayes liefert für die kumulierten Daten drei Bäume mit PPs = 0,828, 0,163 und 0,010. Der erste Baum (mit der höchsten PP) ist hinsichtlich des Verzweigungsmusters identisch mit dem Baum *SA-Referenz*, der zweite und dritte Baum haben zu diesem *Split-Distanz* 6 (siehe Ordner *Laeufe2bis5/Vergleiche*).

Salmonellen-Datensatz

Lauf 1

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	1,000	1,000	0	A

Lauf 2

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,500	0,500	0	A
2	0,371	0,871	6	B
3	0,129	1,000	6	C

Lauf 3

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,500	0,500	0	A
2	0,441	0,941	6	B
3	0,059	1,000	6	C

Läufe 2-5 kumuliert

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,828	0,828	0	A
2	0,163	0,990	6	B
3	0,010	1,000	6	C

Lauf 4

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,868	0,868	6	B
2	0,089	0,957	6	C
3	0,043	1,000	0	A

Lauf 5

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,500	0,500	0	A
2	0,439	0,939	6	B
3	0,061	1,000	6	C

Lauf 6

Baum	PP	Kumulierte PP	Split-Distanz zu SA-Referenz	ID
1	0,500	0,500	0	A
2	0,472	0,472	6	B
3	0,028	1,000	6	C

Tabelle 8.6*: Salmonellen-Datensatz: In den Läufen 1 – 6 aufgefundene Bäume*

8.2.2.3 Phylogenien mit Neighbor-Joining

Bäume nach der Neighbor-Joining-Methode (Saitou und Nei, 1987) wurden auf der Grundlage der Alignments, auf denen die Referenzphylogenien beruhen, mit der PHYLIP-Suite erstellt. Aus dem Alignment wurde dazu jeweils mit *seqboot* ein Bootstrap-Datensatz aus 200 Wiederholungen generiert, für die dann *dnadist* jeweils die Distanzen zwischen den Spezies schätzte (siehe **Tab. 8.7**). Aus der resultierenden Datei mit den Distanzmatrizen erstellte *neighbor* Neighbor-Joining-Bäume, für die *consensus* zuletzt einen Consensus-Baum konstruierte (Datei *outtree* im Ordner *Kapitel_Phylogenien/Referenzbaeume/MMN_ALL_EC_ohne_009801_002695/MMN/data/alignments/gblocks_reverse/concatenated/nj_Baum/seqboot/dnadist/neighbor/consense*). Der Neighbor-Joining-Baum für den *Salmonellen*-Datensatz liegt im Ordner *Kapitel_Phylogenien/Referenzbaeume/MMN_ALL_SA_OKT12/MMN/data/alignments/gblocks_reverse/concatenated/nj_Baum/seqboot/dnadist/neighbor/consense* in der Datei *outtree*.

Erstellung eines Neighbor-Joining-Baumes mit den Programmen der PHYLIP-Suite
Alignment → *seqboot* → *dnadist* → *neighbor* → *consense* → Neighbor-Joining-Consensus

Verwendete Parameter

seqboot
Random number seed: 3, sequence: Molecular sequences, Bootstrap, regular or altered sampling fraction: regular, block size for block-bootstrapping: 1, how many replicates: 200 , read weights of characters: no, read categories of sites: no

dnadist
Distance: F84, gamma distributed rates across sites: no, transition/transversion ratio: 2.0
one category of substitution rates: yes, use weights for sites: no, use empirical base frequencies: yes, form of distance matrix: square, analyze multiple data sets: yes (200)*neighbor*
Neighbor-joining or UPGMA tree: neighbor-joining, outgroup root: no, use as outgroup species 1, lower-triangular data matrix: no, upper-triangular data matrix: no, subreplicates: no, randomize input order of species: yes (random number seed = 3), analyze multiple data sets: yes, 200 sets

consense
Consensus type: majority rule (extended) , outgroup root: no, use as outgroup species 1, trees to be treated as rooted: no

Tabelle 8.7*: Zur Berechnung der Neighbor-Joining-Bäume verwendete Parameter*

8.2.2.4 Phylogenien auf Grundlage hochkonservierter Genfamilien

Der primäre phylogenetische Baum für 191 Spezies, „Tree of Life v1.0“, auf dem das Projekt „Interactive Tree of Life“ (kurz „iTol“) beruht (www.iTol.embl.de), wurde mit der in Ciccarelli et al. im Jahre 2006 beschriebenen Methodik generiert.

Der Baum basiert auf einer Konkatenation von 31 stark konservierten Orthologen, „COGs“ (siehe **Tab. 8.8**): Von den ursprünglich entdeckten 36 Familien entfernten die Autoren im Vorfeld fünf Familien mit multiplen horizontalen Gentransfers oder solche, die schwierig zu alignieren waren. Ciccarelli et al. schätzten eine ML-Phylogenie mit *PhyML*. Für den „iTol-Baum“ benutzten Ciccarelli und Kollegen die in **Tabelle 8.7** aufgeführten 31 COGs.

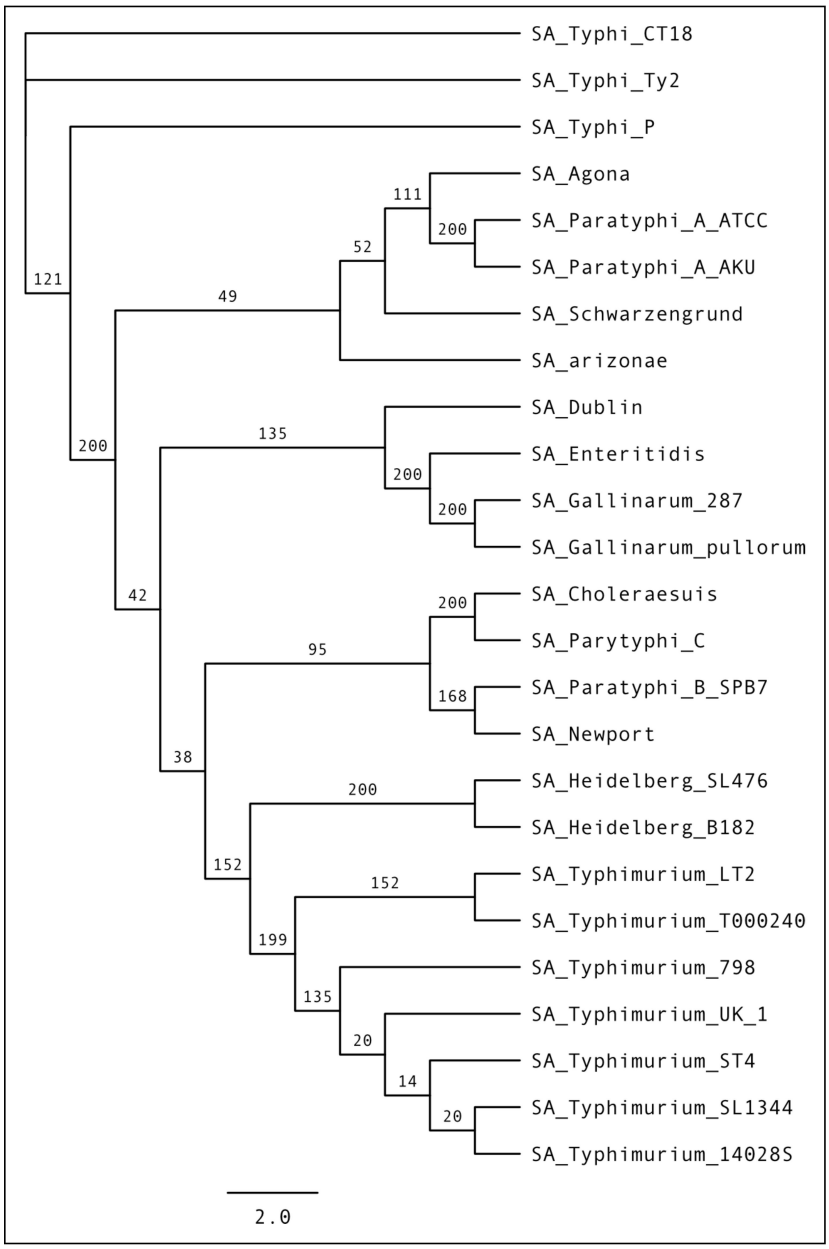
Identifikation der zu den COGs äquivalenten OGs in den eigenen Datensätzen

Die Schnittmenge der für den „iTol-Baum“ verwendeten vollständig sequenzierten 191 Genome und der in dieser Arbeit eingesetzten 86 Genome ist klein: lediglich die Taxa *E. coli K-12 MG1655*, *E. coli O157:H7 EDL933* und *Salmonella Typhimurium LT2* wurden in beiden Arbeiten verwendet. Um die den COGs entsprechenden universellen Genfamilien im EC-, SA- bzw. ECSA-Datensatz des Verfassers zu identifizieren, wurde auf Kopien der *MMN*-Projekte *MMN_EC_ohne009801_002695*, *MMN_SA_OKT12* und *MMN_ECSA_ohne009801_002695* gearbeitet, nämlich *EC_iTol*, *ECSA_iTol* und *SA_iTol* im Ordner *iTol* im Projektordner *Kapitel_Phylogenien*.

Jeweils im Unterordner *data/alignments/gblocks_reverse* sind alle Nukleotidalignments der universellen orthologen Gruppen im FASTA-Format gespeichert. Diese werden innerhalb der *MMN-Pipeline* vor der Berechnung eines Baumes mittels *PhyML* konkateniert. Für die Konstruktion eines konkatenierten MSAs der zu den iTol-COGs äquivalenten universellen

Genfamilien wurden alle Orthogruppen gelöscht, die keines der Gene in den iTol-COGs enthalten[28].

Abbildung 8.5*: Kladogramm für die Salmonellen aus* ***Tab. 8.1*** *auf Grundlage von 26 iTol-COGs.* Die Zahlen an den Ästen sind Bootstrap-Werte aus 200 Wiederholungen.



E. coli-, Salmonellen- und E. coli/Salmonellen-Phylogenien

Mit *MMN* wurde ein konkateniertes MSA der verbliebenen 26 orthologen Gruppen hergestellt und danach auf ihnen ein Bootstrap-Baum geschätzt (Verzeichnis *concatenated*, Datei *fgRevNuc Genome_ALL.faa.phy_phyml_tree.txt* im jeweiligen Ordner*)*, der hier nicht dargestellt ist. Fünf der von Ciccharelli et al. benutzten COGs waren im Datensatz des Verfassers nicht universell.

Bei beiden Teilbäumen (siehe **Abb. 8.5** und **8.6**) fällt auf, daß die statistische Unterstützung vieler Splits durch den Bootstrap sehr unausgewogen ist.

28 Die Gene der drei Spezies in der Schnittmenge, die jeweils in einem iTol-COG sind, sind in den *E. coli*- bzw. *Salmonellen*-Datensätzen ebenfalls in einer OG, was angesichts der hohen Konserviertheit der COGs auch der Erwartung entspricht.

Aus den Teilbäumen zusammengesetzte E. coli/Salmonellen-Phylogenie

Mit der in *Abschnitt 8.2.2.1* beschriebenen Technik wurde aus den ungewurzelten, auf den wie oben beschrieben geschätzten *E. coli*- und *Salmonellen*-Teilbäumen ein zusammengesetzter, gewurzelter Baum mit Astlängen (Datei *ecsa_iTol_zusammengesetzt.tre*) konstruiert. Die Bootstrap-Werte an seinen Ästen sind hier aus Platzgründen nicht dargestellt, sie stammen von den Teilbäumen. Die entsprechenden Dateien befinden sich innerhalb des Projektordners im Unterordner *iTol/ECSA_kombiniert_iTol/Konstruktion_kombinierterECSA_iTol_Baum.* Das Alignment, aus dem die Distanzen für den Baum geschätzt wurden, stammt aus dem *MMN*-Projekt im Ordner *Referenzbaeume/MMN_ALL_ECSA_ohne009801_002695*. Die *MMN*-Projekte für die Berechnung der Constraint-Bäume sind in den Unterordnern *EC_itol_mitSAConstraintTree* und *SA_itol_mitECConstraintTree* innerhalb des Ordners *ECSA_kombiniert_iTol* zu finden. Der aus den Teilbäumen zusammengesetzte *iTol-E. coli/Salmonellen*-Baum (**Abb. 8.7**) hat zu dem direkt berechneten *iTol-E. coli/Salmonellen*-Baum, der hier nicht abgebildet ist, einen Robinson-Foulds-Abstand 66 und einen *ssRF-Abstand* 0 bei Bootstrap-Schwelle 129 (entspricht 65 %, vgl. *Verzeichnis iTol/ECSA_kombiniert_iTol/Konstruktion_kombinierterECSA_iTol_Baum/Vergleich_ecsa_vs_ecsaZusammengesetz*t).

Verwendete Clusters of Orthologous Groups (COGs)

COG	Bezeichnung	Ciccarelli et al. („iTol-Baum“)	EC	SA	ECSA
0012	Predicted GTPase, probable translation factor				
0016	Phenylalanyl-tRNA synthetase alpha subunit				
0018	Arginyl-tRNA-synthetase	**x**	**x**	**x**	**x** (n. u.)
0048	Ribosomal protein S12				**x** (n. v.)
0049	Ribosomal protein S7				
0052	Ribosomal protein S2				
0060	Isoleucyl-tRNA-Synthetase	**x**	**x**	**x**	**x**
0080	Ribosomal protein L11				
0081	Ribosomal protein L1				
0085	DNA-directed RNA polymerase, beta subunit/140 kd subunit	**x**	**x**	**x**	**x**
0087	Ribosomal protein L3				
0091	Ribosomal protein L22				
0092	Ribosomal protein S3				
0093	Ribosomal protein L14			**x** (n. u.)	
0094	Ribosomal protein L5				
0096	Ribosomal protein S8				
0097	Ribosomal protein L6P/L9E				
0098	Ribosomal protein S5			**x** (n. u.)	
0099	Ribosomal protein S13				**x** (n. v.)
0100	Ribosomal protein S11				**x** (n. v.)
0102	Ribosomal protein L13				
0103	Ribosomal protein S9				
0124	Histidyl-tRNA-synthetase	**x**	**x**	**x**	
0143	Methionyl-tRNA-synthetase	**x**	**x** (n. v.)	**x**	**x** (n. v.)
0172	Seryl-tRNA-synthetase		**x** (n. u.)		**x** (n. v.)
0184	Ribosomal protein S15P/S13E				
0186	Ribosomal protein S17				
0197	Ribosomal protein L16/L10E				
0200	Ribosomal protein L15			**x** (n. u.)	**x** (n. u.)
0201	Preprotein translocase subunit SecY				
0202	DNA-directed RNA polymerase, alpha subunit/40 kd subunit				
0256	Ribosomal protein L18				
0495	Leucyl-tRNA-synthetase				**x** (n. u.)
0522	Ribosomal protein S4 and related proteins				
0525	Valyl-tRNA-synthetase			**x** (n. v.)	**x** (n. v.)
0533	Metal-dependent proteases with possible chaperone activity				

x nicht verwendet
n. u.: Genfamilie ist im Datensatz des Verfassers nicht universell
n. v.: Mindestens ein Gen der Genfamilie ist im Datensatz des Verfassers nicht vorhanden

Tabelle 8.8*: Einsatz der COGs bei Ciccarelli et al. und in dem hier dargestellten Projekt*

8.2.2.5 Phylogenien nach einer Methode von Sergei Maslov

Dieser Abschnitt stellt eine neuartige Methode (Dixit et al., 2013) vor, die auf Ideen von Sergei Maslov, Wissenschaftler am Brookhaven National Laboratory, Long Island, New York, USA, zurückgeht. Sie benutzt Neighbor-Joining (Saitou und Nei, 1987).

Idee des Verfahrens

Sergei Maslovs Methode fußt auf der Bobachtung, daß viele Regionen auf dem Genom in *E. coli* von Rekombination geprägt sind, sodaß sie sich nicht für Vorhersagen zur vertikalen Evolution der Spezies eignen. Chromosomenfragmente, die auf verschiedene Art in eine Bakterienzelle gelangen, können über homologe Rekombination mit anderen Chromosomen Anteile austauschen. Maslov definiert Gene als klonal, wenn sie weniger als drei Substitutionen per 1000 bp aufweisen. Alle anderen Gene wertet er als rekombinant und spricht ihnen dabei jedes vertikale Signal ab. Darüberhinaus finde Rekombination nur zwischen *E. coli*-Stämmen statt, deren genetisches Material sich um höchstens zwei Prozent unterscheidet. Darüber sei lediglich lateraler Gentransfer, aber keine Rekombination mehr möglich. Dieser Schwellenwert könne auch einer Abgrenzung zwischen den Arten dienen. Eben Gesagtes läßt den Schluß zu, daß die Distanz zwischen zwei molekularen Sequenzen möglicherweise proportional zu der Zahl der Regionen auf einem Chromosom ist, die durch Rekombination ausgetauscht worden sind. Dementsprechend könnte die Distanz auch proportional zur evolutionären Distanz sein.

Realisation

Die vom Verfasser dieser Arbeit ausprogrammierte Methode wurde in Form des Perl-Programmes `computePairwiseDistances.pl` realisiert (vgl. **Algorithmus 8.1**). Das Programm benutzt die mit *Gblocks* vorbehandelten MSAs der einzelnen universellen Genfamilien. Mit dem unveröffentlichten C-Programm *PowerNeedle*, einer schnellereren Version des Programms *EMBOSS* für globale Alignments (Rice et al., 2000), wird die paarweise globale Sequenzähnlichkeit zwischen allen Genen innerhalb einer Genfamilie berechnet. Der dazu benutzte Algorithmus für die Berechnung eines globalen Alignments zweier Sequenzen wurde seinerzeit von Saul B. Needleman und Christian D. Wunsch eingeführt (Needleman und Wunsch, 1970). Als Nächstes wird eine Distanzmatrix für den zu berechnenden Neighbor-Joining-Baum konstruiert.

Die Distanz zwischen zwei *E. coli*-Stämmen ist gleich dem Prozentsatz rekombinanter Regionen. Genpaare mit einer paarweisen Sequenzidentität über 99,6% werden als klonal, Genpaare unter dieser Schranke als rekombinant gewertet. Für jedes Paar von Stämmen werden die rekombinanten Regionen, also die rekombinanten Genpaare zwischen den Stämmen, gezählt. Paare von Genen aus unterschiedlichen Genfamilien werden nicht als rekombinant behandelt. Der Prozentsatz rekombinanter Regionen für ein konkretes Paar von Stämmen wird an der entsprechenden Stelle in die Distanzmatrix eingetragen. Für die Berechnung von 200 Bootstrap-Wiederholungen wird für jeden Stamm jeweils ein neuer Genpool geschaffen, indem aus dem tatsächlichen Genpool mit Zurücklegen gezogen wird. Für jeden Bootstrap-Lauf wird eine neue Distanzmatrix berechnet. Die resultierenden 200 Distanzmatrizen werden hintereinander in einer Datei gespeichert. Aus ihnen generiert P*HYLIP neighbor* einen Neighbor-Joining-Baum mit Bootstrap-Werten ohne Outgroup. Abschließend wird mittels *PHYLIP consense* ein Consensus-Baum berechnet.

Inferenz von Phylogenien nach der Methode von Sergei Maslov

Der folgende Absatz bezieht sich auf das Verzeichnis *Maslovs_Methode* innerhalb des Projektordners *Kapitel_Phylogenien*. Mithilfe des oben beschriebenen Verfahrens wurden jeweils Phylogenien für die *E. coli*- bzw. *Salmonellen*-Taxa aus **Tabelle 8.1** auf Basis der universellen Genfamilien berechnet (Ordner *CoreGenes*).

Die Daten für die Berechnung des *E. coli*-Teilbaumes befinden sich im Unterordner *EC,* zusammen mit dem Programm `ec_computePairwise Distances.pl`. In *EC* befindet sich ein weiterer Unterordner, *pw,* mit den mit einer Variante des Programmes *EMBOSS Needle* berechneten paarweisen Sequenzidentitäten zwischen den *E. coli*-Genen. Für den Ordner *SA* gilt Äquivalentes.

Die mit dem Verfahren von Sergei Maslov berechneten Bäume für den *E. coli*- und den *Salmonellen*-Datensatz sind in den Dateien *ec_Maslov_PHYLIPConsensus.tre* im Ordner *ec/pw/neighbor/consense* und in *sa_Maslov_PHYLIPConsensus.tre* im Ordner *sa/pw/neighbor/con sense* zu finden.

Algorithmus 8.1: Paarweise Distanzen nach einer Idee von Sergei Maslov

Eingabe Paarweise Sequenzidentitäten je Genfamilie, #Bootstrap-Wiederholungen, Ausgabedatei
Ausgabe Neighbor-Joining-Baum

For (1 .. #Bootstrap-Wiederholungen)
Foreach Bakterium
Erstelle einen neuen Genpool: ziehe so oft Gene mit Zurücklegen aus dem Ursprungs-Genpool, wie das Bakterium Gene besitzt
EndFor
Foreach Paar { Bakterium1, Bakterium2 }
ZahlRekombinanterPaare = ZahlKlonalerPaare = ZahlPaareGesamt := 0
If Bakterium1 = Bakterium2
Then Distanz(Bakterium1, Bakterium2) = Distanz(Bakterium2, Bakterium1) := 0
Else Foreach Paar { GI-Nummer1 aus Bakterium1, GI-Nummer2 aus Bakterium2 }
If paarweise Sequenzidentität(GI-Nummer1, GI-Nummer2) > 99,6 %
Then ZahlKlonalerPaare++
Else ZahlRekombinanterPaare++
EndFor
ZahlPaareGesamt := ZahlKlonalerPaare + ZahlRekombinanterPaare
ProzentsatzRekombinanterPaare := 100 · (ZahlRekombinanterPaare / ZahlPaareGesamt)
Distanz(Bakterium1, Bakterium2) := Distanz(Bakterium2, Bakterium1) = ProzentsatzRekombinanterPaare
EndFor
EndFor
Konstruiere eine Distanzmatrix für das Paar (Bakterium1, Bakterium2)
Füge die Distanzmatrix der Ausgabedatei für Distanzmatrizen hinzu
EndFor
Erstelle für die Distanzmatrizen in der Ausgabedatei einen Consensus-Baum

Die Splits im *E. coli*-Baum (**Abb. 8.8**) werden insgesamt nur schlecht unterstützt. Die Differenzierung der einzelnen phylogenetischen Gruppen von *E. coli* ist nur an zwei Stellen statistisch ausreichend, nämlich bei dem Teilbaum aus *Shigella sonnei* und *Shigella boydii* sowie dem Teilbaum aus den *E. coli*-Stämmen *IAI39, O7:K1, SMS 3-5,* 042 und *UMN26,* die gewöhnlich den phylogenetischen Gruppen *D1* und *D2* zugeschlagen werden (siehe z. B. Chaudhuri und

Henderson, 2012). Die statistische Unterstützung der Splits im *Salmonellen*-Baum (**Abb. 8.9**) ist insgesamt gut, sieht man von einem Teilbaum mit *Typhimurium*-Serovaren ab, bei denen es der Methode nicht gelungen ist, diese genügend zu differenzieren.

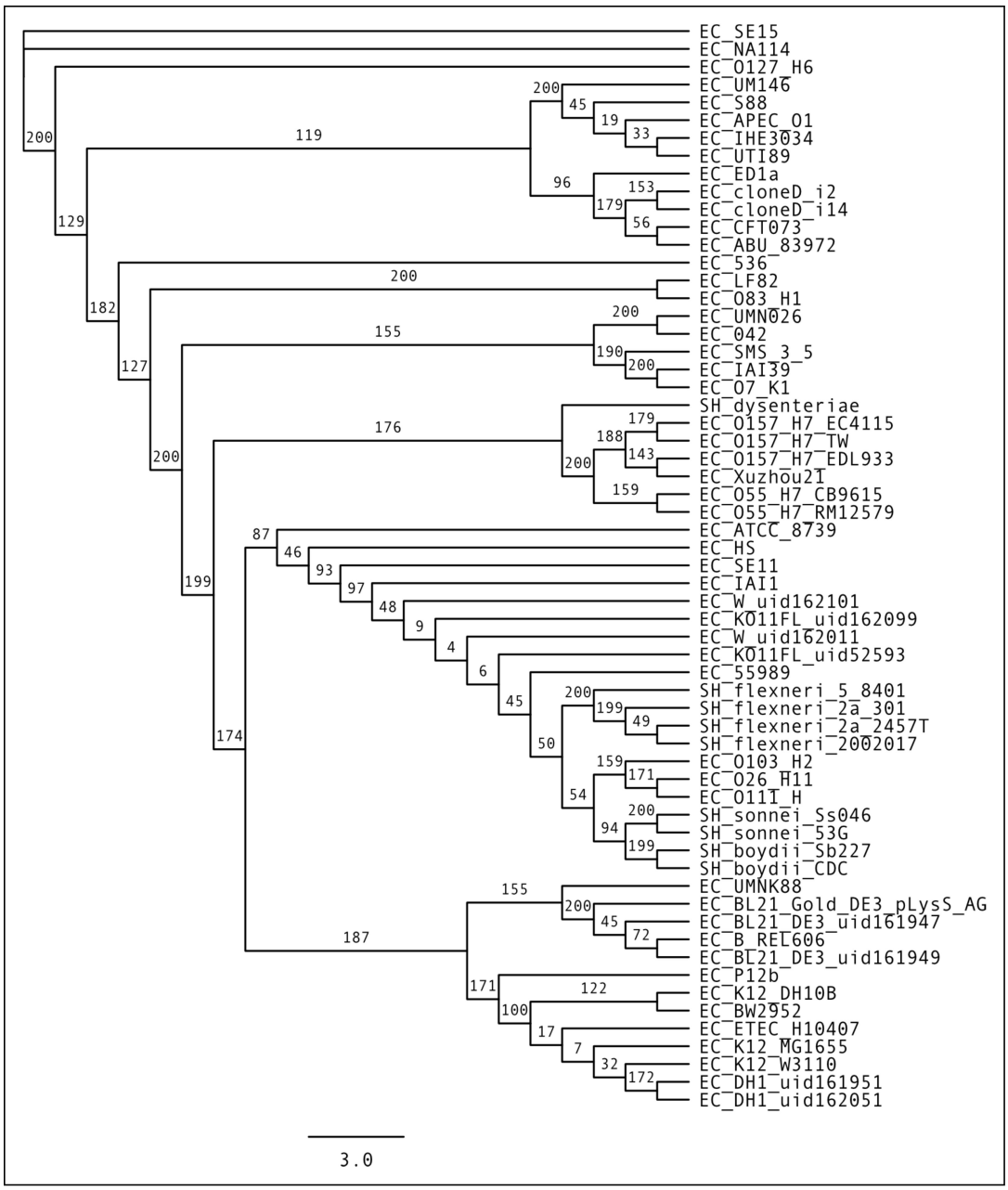


Abbildung 8.6*: Kladogramm für die E. coli-Stämme aus Tab. 8.1 auf Grundlage von 26 iTol-COGs.* Die Zahlen an den Ästen sind Bootstrap-Werte aus 200 Wiederholungen.

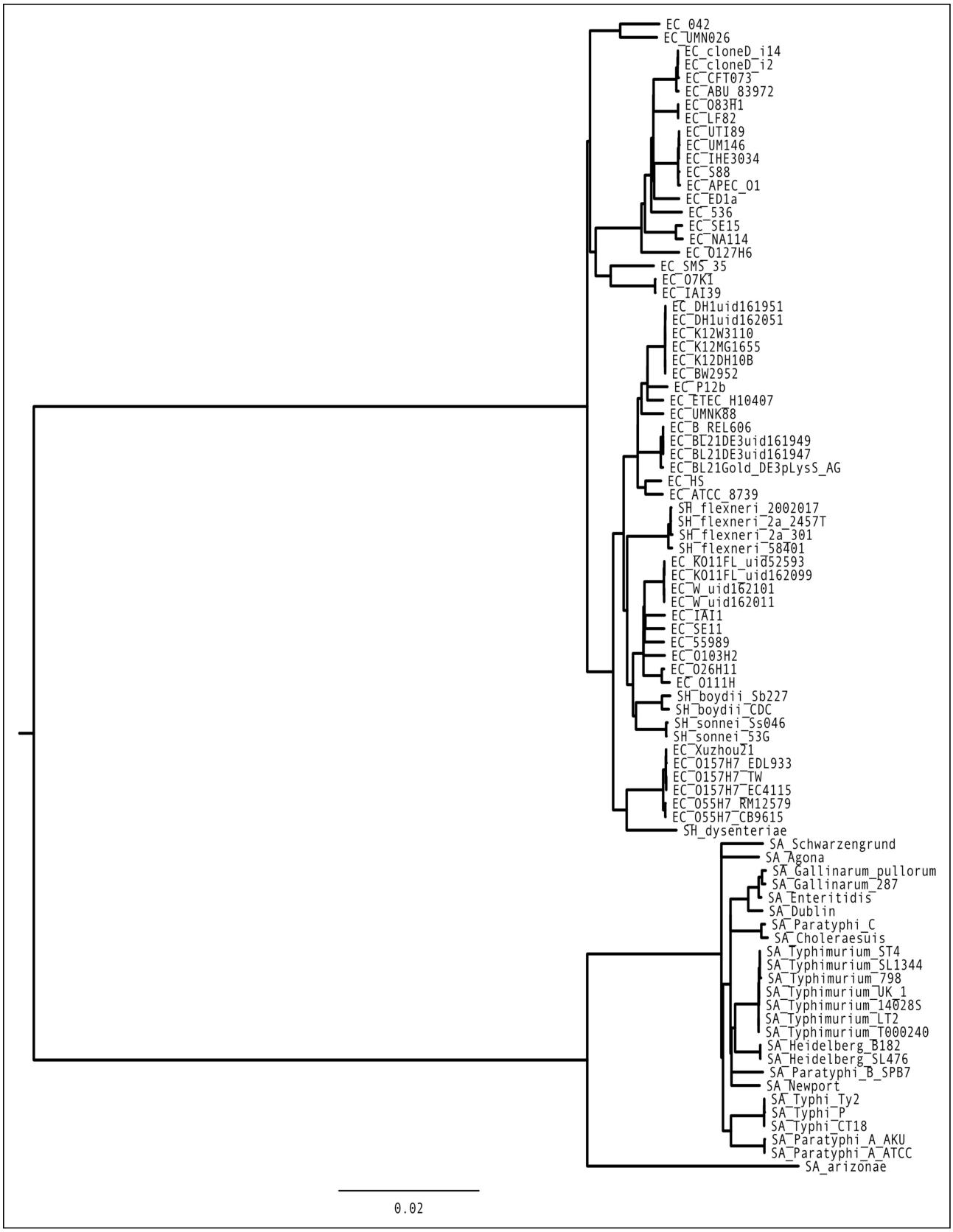


Abbildung 8.7*: Aus den Teilbäumen in den* ***Abb. 8.5*** *und* ***8.6*** *zusammengesetzter iTol-E. coli/Salmonellen-Baum.* Die beiden Äste an der Wurzel sind verkürzt dargestellt. Die Astlängen wurden aus einem Alignment der konkatenierten universellen Genfamilien der E. coli- und Salmonellen-Stämme geschätzt.

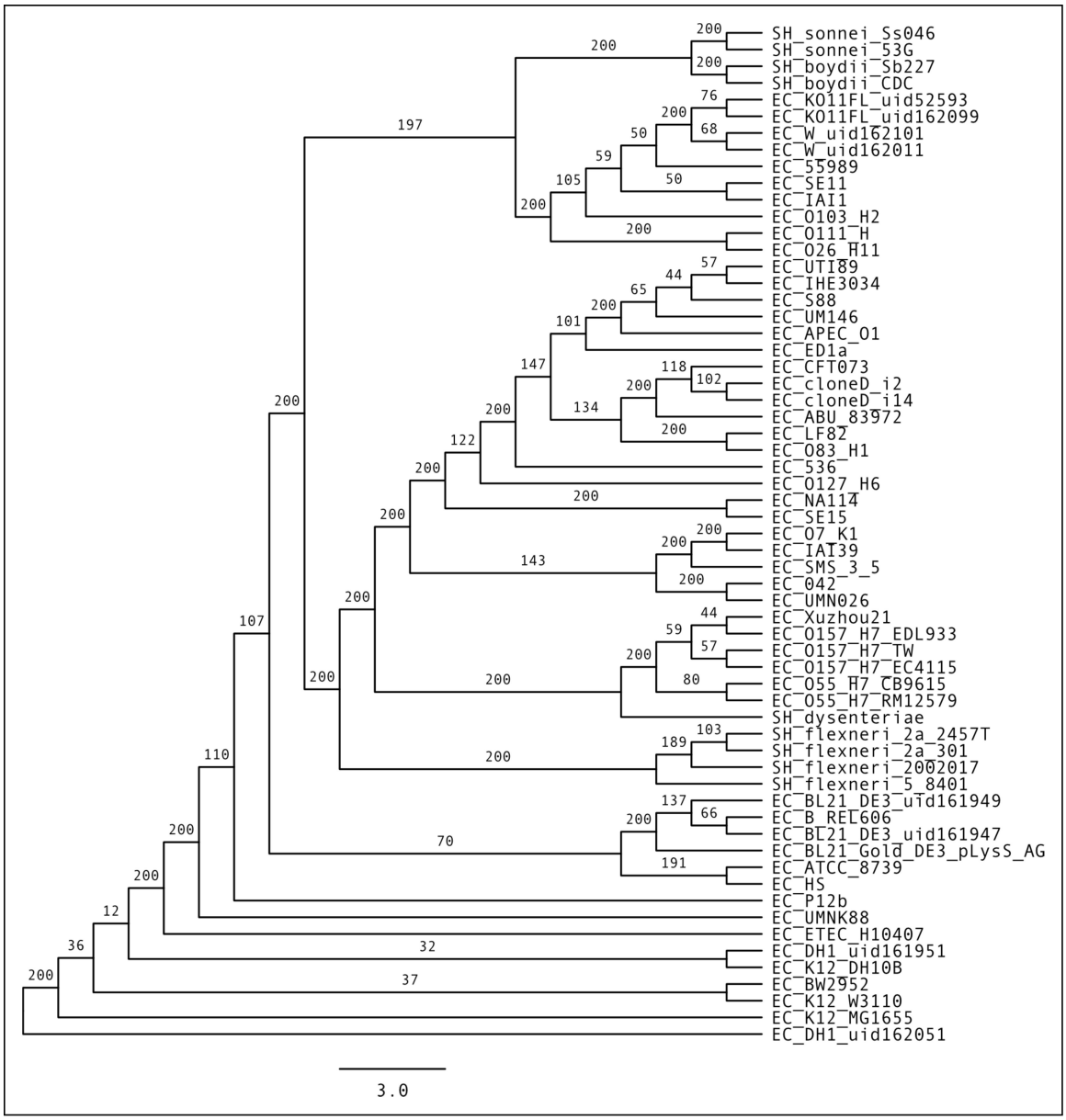


Abbildung 8.8*: Kladogramm der mit der Methodik nach Maslov konstruierten Phylogenie für die E. coli-Stämme aus* ***Tab. 8.1***. Die Zahlen an den Ästen sind Bootstrap-Werte aus 200 Wiederholungen.

8.3 Ergebnisse

In den vorhergehenden Abschnitten dieses Kapitels wurde dargestellt, wie unter einem ML-Framework (**Abschnitt 8.2.2.1**) und einem bayesianischen Framework (**Abschnitt 8.2.2.2**) auf einem MSA universeller Genfamilien Speziesbäume für die in dieser Arbeit verwendeten (siehe ***Tabelle 8.1***) *E. coli*-Stämme, für die Salmonellen und für *E. coli* und *Salmonellen* gemeinsam geschätzt wurden. Desweiteren wurden Speziesbäume aus den *E. coli*- und *Salmonellen*-Teilbäumen zusammengesetzt. Daneben wurden *E. coli*- und *Salmonellen*-Speziesbäume noch auf der Grundlage ausgewählter hochkonservierter Genfamilien ebenfalls unter ML (**Abschnitt 8.2.2.4**) und über Neighbor-Joining (**Abschnitt 8.2.2.3**) geschätzt. Abschließend wurde gezeigt, wie Speziesbäume für *E. coli* und *Salmonellen* über eine bisher unveröffentlichte, auch auf Neighbor-Joining basierende Methode inferiert wurden, die als Distanzmaß den Prozentsatz rekombinierter Regionen zwischen einem Paar von Spezies benutzt (**Abschnitt 8.2.2.5**). Dieser Abschnitt stellt die mit den verschiedenen Methoden berechneten Phylogenien in einen quantitativen Zusammenhang. Zunächst wird gezeigt, wie gut die statistische Absicherung der Splits der *E. coli*- und *Salmonellen*-Referenzbäume ist. Dazu wurden Bootstrap-Werte, BPPs und die Split-Frequenzen in den Genbäumen verglichen. Ein Vergleich der Distanzen zwischen den mit verschiedenen Methoden berechneten Phylogenien schließt sich an, und zuletzt werden die *E. coli*- und *Salmonellen*-Referenzbäume mit von anderen Gruppen veröffentlichten Phylogenien verglichen.

8.3.1 Statistischer Support der Splits von Escherichia coli- und Salmonellen-Phylogenie

Aus den Alignments der konkatenierten universellen Genfamilien für die 25 *Salmonellen* bzw. 61 *Escherichia coli* (vgl. **Tabelle 8.1**) wurden mit verschiedenen Methoden drei Phylogenien geschätzt:

a) ein Speziesbaum mit *PhyML/RAxML* (**Abb. 8.1 und 8.2**),

b) ein Consensusbaum der Genbäume (siehe **Abschnitt 8.2.2.1**) und

c) ein Consensusbaum mehrerer mit *MrBayes* aufgefundener Bäume (vgl. **Abschnitt 8.2.2.2**).

Die **Abb. 8.10** und **8.11** zeigen die Bäume SA-Referenz und EC-Referenz. Jeder Ast trägt drei Zahlen, die darüber Auskunft geben, wie hoch seine statistische Unterstützung in Prozent unter den drei Methoden ausgefallen ist. V. l. n. r. sind dies Bootstrap-Support, Bayesian Posterior Probability, Häufigkeit des Splits in den Genbäumen.

E. coli-Datensatz: **Abb. 8.10** zeigt die E. coli-Phylogenie aus **Abb. 8.2**. Die Bäume a) und c) haben, bis auf eine Abweichung innerhalb der phylogenetischen Gruppe *B*, das gleiche Verzweigungsmuster, Baum b) weicht signifikant ab. Praktisch alle Äste besitzen einen Bootstrap-Support oberhalb 70 %, lediglich ein Ast weist einen Bootstrap-Wert von 64 % auf. Alle Splits bis auf einer werden durch Posterior Probabilities von 100 % unterstützt. Der Split, der zwischen den zwei Laborstämmen *K-12 W3110* und *K-12 MG1655* und den beiden von *K-12 MG1655* abstammenden *DH1*-Stämmen unterscheidet, wird durch eine immer noch sehr gute Posterior Probability von 95 % unterstützt. Unzureichend durch die Genbäume unterstützte Splits finden sich sowohl tief im Baum als auch in der Nähe der Blätter. Die sehr hohe Verwandtschaft der *E. coli*- erschwert wahrscheinlich deren phylogenetische Unterscheidung.

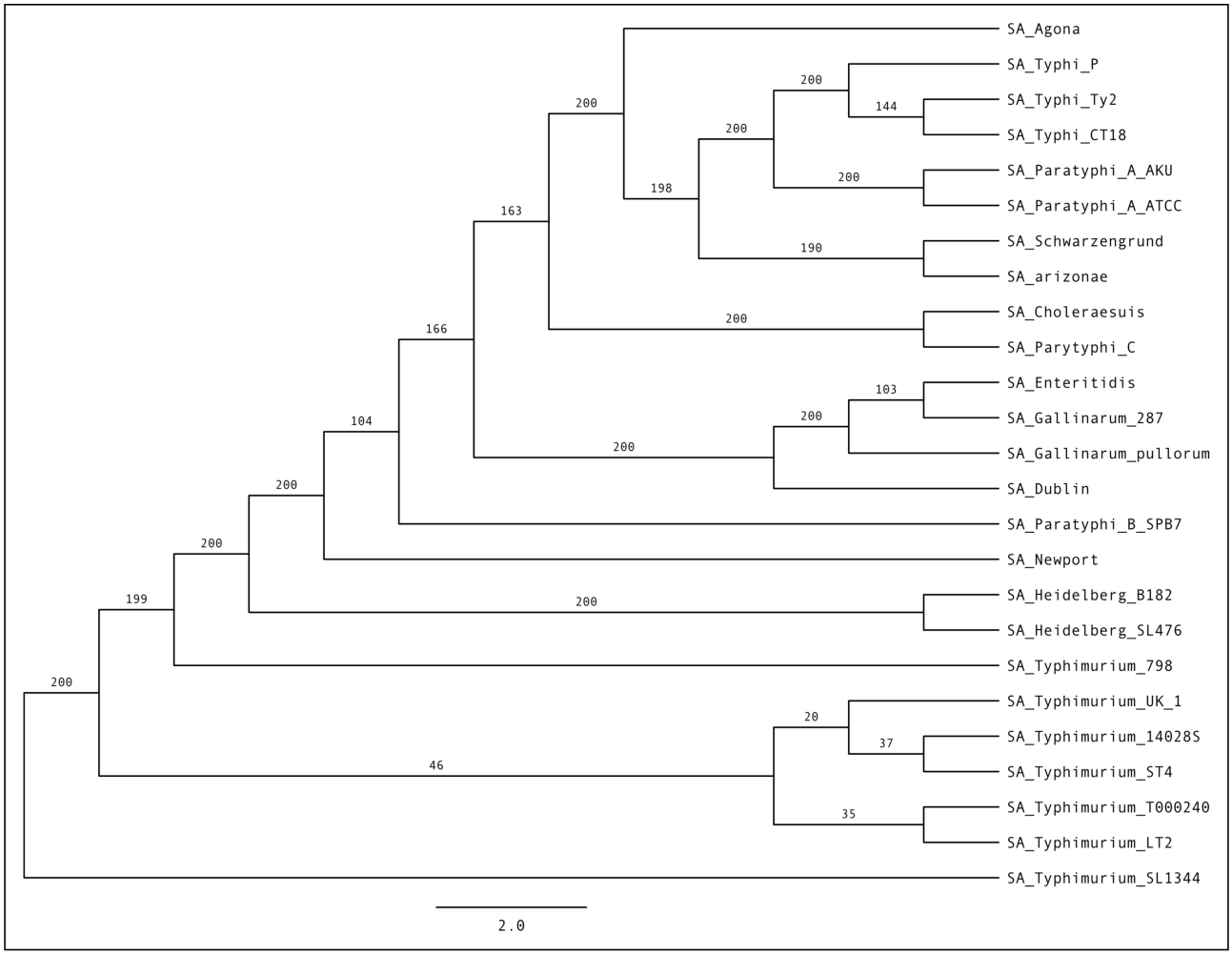


Abbildung 8.9*: Kladogramm der mit der Methodik von Maslov konstruierten Phylogenie für die Salmonellen-Stämme aus* ***Tab. 8.1****.* Die Zahlen an den Ästen sind Bootstrap-Werte aus 200 Wiederholungen.

Salmonellen-Datensatz: Abb. 8.11 zeigt die *Salmonellen*-Phylogenie aus **Abb. 8.1**. Die Bäume a) und c) besitzen die gleiche Topologie. Ihre Äste sind statistisch durchgängig gut durch Bootstrap-Werte über 70 % und BPPs von 100 % abgesichert. Baum b) weicht allerdings signifikant von ihnen ab, was sich unmittelbar in einem größtenteils niedrigen Split-Support manifestiert. Für die Splits, die zwischen sehr nah verwandten *Salmonellen* unterscheiden, ist der Split-Support besonders schlecht, so gesehen beispielsweise bei den *Typhimurium*-Serotypen. Da sie erst in jüngerer Zeit divergiert sind und damit die diesbezüglichen Astlängen sehr kurz sind, reichen die Unterschiede zwischen ihnen nicht zu einer phylogenetischen Differenzierung aus. Viele Splits tiefer im Baum werden durch die Genbäume allerdings vergleichbar schlecht unterstützt. Eine nahe liegende Erklärung könnte über Plasmide vermittelter häufiger lateraler Gentransfer auf Spezies-Niveau sein (Kloesges et al., 2010).

8.3.2 Distanzen zwischen den mit verschiedenen Methoden berechneten Phylogenien

Die Abstände zwischen den Phylogenien hat der Verfasser mit dem Perl-Programm *computeSSRF* berechnet. *RF-Distanzen* zwischen dem mit *MrBayes* berechneten Consensus-Baum („Bayes-Baum“) und den anderen Bäumen wurden mit dem Programm *treedist* aus der *PHYLIP*-Suite

berechnet (siehe Anmerkung unterhalb von **Tab. 8.9**). Die Ergebnisse für die EC, SA-, ECSA- und die aus EC und SA zusammengesetzten Phylogenien sind in entsprechend benannten Unterordnern im Ordner *Baumvergleich* gespeichert. Innerhalb der Unterordner sind die eigentlichen Distanzberechnungen in Ordnern abgelegt, deren Name die verglichenen Methoden[29] widerspiegelt. Das Ergebnis einer Distanzberechnung mit *computeSSRF* findet sich immer in einer Datei namens *ssrfDistances.txt*.

Tab. 8.9 zeigt die Robinson-Foulds-Abstände (schwarz auf weiß) und die *ssRF-Abstände* (weiß auf schwarz) bei Signifikanzschwelle 70 % (entspricht Bootstrap-Wert 140 auf einer Skala bis 200) für die in diesem Kapitel gezeigten Phylogenien. Für den Fall, daß bei einem Vergleich kein Bootstrap-Wert 140 präsent gewesen ist, wurde die *ssRF-Distanz* für die nächsthöhere Bootstrap-Schwelle genommen und im Text jeweils hinter den ssRF-Abstand in Klammern gesetzt.

E. coli: Die mit der Bayesianischen Methode und der Maximum-Likelihood-Methode inferierten *E. coli*-Phylogenien weisen nur einen Unterschied in einem Teilbaum mit drei sehr nah verwandten *E. coli B*-Taxa auf (vgl. dazu den Kommentar unter **Tab. 8.9**). Die ML-Phylogenie hat *ssRF-Abstand* 10 von der iTol- und *ssRF-Abstand* 6 von der mit der Maslov-Methode erzeugten Phylogenie, während diese untereinander einen *ssRF-Abstand* 3 haben. Die iTol- und Maslov-Methoden erzeugen demzufolge einen ähnlicheren Baum.

Salmonellen: Die mit der Bayesianischen Methode und der Maximum-Likelihood-Methode inferierten *Salmonellen*-Phylogenien sind in ihrem Verzweigungsmuster identisch. Die mit der iTol-Methode generierte Phylogenie hat von der ML-Phylogenie *ssRF-Abstand 0* (Signifikanzschwelle 150), von der mit Maslovs Methode erzeugten Phylogenie *ssRF-Abstand* 11 (Schwelle 144). Die ML-Phylogenie hat ebenfalls ssRF-Abstand 11 von der Maslov-Phylogenie (Schwelle 144). Die in den folgenden beiden Absätzen vorgestellten Phylogenien wurden mit den anderen Phylogenien dieses Kapitels verglichen, erscheinen aber aus Platzgründen nicht in **Tabelle 8.9**.

iTol-Datensatz für 29 COGs

Im Vorfeld dieser Arbeit wurden drei Phylogenien mit den oben verwendeten COGs sowie den COGs 060, 085 und 124, sofern sie jeweils universell waren, mit der gleichen Technik berechnet. Dies geschah in der Erwartung, die statistische Unterstützung der Splits insgesamt verbessern zu können, was allerdings nicht der Fall war. Francesca Ciccarelli und Kollegen, Urheber des „iTol-Baumes“ (Ciccarelli et al., 2006), haben bei ihren Recherchen die o. g. COGs ausgeschlossen, da sie die COGs 060 und 124 in Verdacht haben, zweimal bzw. dreimal horizontal in mehrere Spezies transferiert worden zu sein. Für COG 085 konnten sie kein eindeutiges MSA herstellen.

Die *Salmonellen*-Teilbäume, die auf der Grundlage von 26 bzw. 29 COGs erzeugt worden sind, haben *RF-Distanz* 24 und *ssRF-Distanz* 2 (Signifikanzschwelle 142) und die entsprechenden *E. coli*-Teilbäume *RF-Distanz* 44 und *ssRF-Distanz* 6 (Schwelle 142). Die erzeugten Phylogenien sind nicht identisch, worauf man hätte hoffen können, da mit hochkonservierten COGs gearbeitet worden ist. Womöglich stören die hinzugenommenen COGs 060 und 124 das phylogenetische Signal tatsächlich. Da sich der Bootstrap-Support vieler Äste auch in den auf mehr COGs beruhenden Phylogenien nicht verbessert hat, ist es zusätzlich denkbar, daß das in den COGs vorhandene phylogenetische Signal nicht zur Differenzierung der nah verwandten Taxa ausreicht.

Die für die Berechnung der Phylogenien relevanten Dateien sind im Ordner *iTol_29COGs* zu finden, die vorgenommenen Distanzberechnungen befinden sich im Ordner *Ordnervergleich*.

29 Da mit der Methode von Maslov (vgl. **Abschnitt 8.2.2.5**) mangels Verfügbarkeit der paarweisen Sequenzidentitäten kein Baum für *E. coli* und *Salmonellen* berechnet werden konnte, fehlen die entsprechenden Vergleiche.

Neighbor-Joining-Bäume

Die mit der mit dem Neighbor-Joining-Verfahren erzeugte *E. coli*-Phylogenie aus **Abschnitt 8.2.2.3** hat von der mit der Methode von Maslov generierten Phylogenie einen *RF-Abstand* 30 (Signifikanzschwelle 140) und einen *ssRF-Abstand* 4 (Schwelle 143). Sie hat *RF-Abstand* 4 und *ssRF-Abstand* 5 von EC-Referenz (Schwelle 144). Der Neighbor-Joining-Baum für die Salmonellen, dessen Konstruktion ebenfalls in **Abschnitt 8.2.2.3** erläutert wurde, hat von der mit Maslovs Methode erzeugten Phylogenie *RF-Abstand* 14 und einen *ssRF-Abstand* 5 (Schwelle 144). Von SA-Referenz hat er *RF-Abstand* 6 und einen *ssRF-Abstand* 6 (Schwelle 150).

8.3.3 Vergleich der ML-Bäume mit Literaturbäumen

In diesen Abschnitt wird die *E. coli*-Phylogenie EC-Referenz mit zwei Bäumen aus im Jahr 2011 veröffentlichten Arbeiten von Chaudhuri et al. und Reeves et al. verglichen. Es bot sich darüberhinaus an, die Bäume im Hinblick auf deren Übereinstimmung mit den von Selander (Selander et al., 1987) definierten großen phylogenetischen Gruppen der Spezies *Escherichia coli* zu untersuchen. Die *Salmonellen*-Phylogenie SA-Referenz wird mit einer von Timme (Timme et al., 2013) veröffentlichten Phylogenie verglichen.

Phylogenie für 24 E. coli von Chaudhuri und Henderson

Chaudhuri und Henderson (Chaudhuri und Henderson, 2011) haben eine Phylogenie für 24 *E. coli* mit *RAxML* auf Grundlage der konkatenierten universellen Genfamilien berechnet. Ihre Phylogenie ist entsprechend der in der Literatur verwendeten Klassifikation der *E. coli*-Stämme entlang der großen phylogenetischen Gruppen (Selander et al., 1987) eingefärbt. Alle Splits im Baum werden von allen 100 Bootstrap-Replikaten unterstützt. Die für diese Arbeit berechnete EC-Referenzphylogenie (siehe **Abschnitt 8.2.2.1**) basiert - ebenfalls wie die von Chaudhuri und Henderson geschätzte Phylogenie für 24 Stämme (siehe **Abb. 8.12**) - auf einem multiplen Sequenzalignment der universellen konkatenierten Genfamilien.

Beide Bauminferenzen erfolgten unter einem ML-Framework; erstere wurde mittels *PhyML* mit dem Nukleotidsubstitutionsmodell TN93 berechnet, letztere mittels *RAxML* mit dem Modell GTR. **Abbildung 8.13** zeigt die gewurzelte *E. coli*-Referenzphylogenie *EC-Referenz* für 61 Stämme mit geschätzten Astlängen und analog zum Baum von Chaudhuri kolorierten phylogenetischen Gruppen. Die Stämme *TW 10509, B171, TY 2482, O104:H4, E1167* und *B7A* im Chaudhuri-Baum sind im Baum *EC-Referenz* nicht vorhanden. Der Bootstrap-Support der Splits in *EC-Referenz* ist mit einer Ausnahme in der Gruppe *B2* mindestens 70 %. Der Baum spiegelt ebenfalls die für *E. coli* definierten großen phylogenetischen Gruppen wider. Die Taxa, die gemäß ihrer elektrophoretischen Eigenschaften den verschiedenen phylogenetischen Gruppen zugerechnet werden, clustern in *EC-Referenz* auch in eben diesen Gruppen. Der zur Gruppe *B1* gehörende Stamm *O103:H2* befindet sich in *EC-Referenz* im Vergleich zum Chaudhuri-Baum allerdings an einer anderen Stelle innerhalb von *B1*.

Chaudhuri und Mitarbeiter haben bei der Auswahl der *E. coli*-Stämme für ihre Phylogenie solche ausgewählt, die sich gut differenzieren lassen bzw. die nicht zu nah verwandt sind. Im Ergebnis konstruierten sie damit eine robuste, wenn auch nicht hoch aufgelöste Phylogenie für *E. coli*.

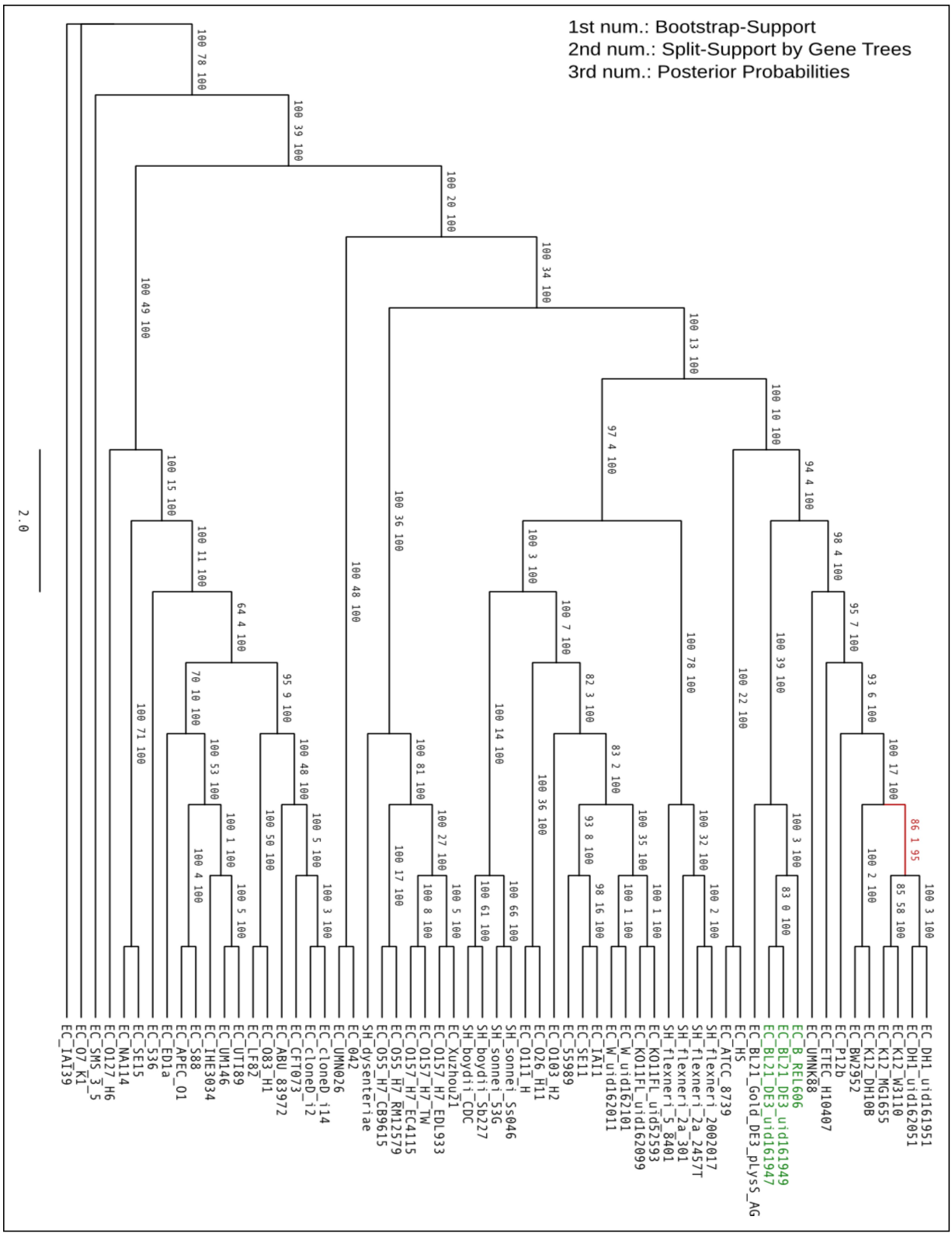


Abbildung 8.10*: E. coli-Phylogenie mit Support durch Bootstrap, Häufigkeit der Splits in den Genbäumen und Posterior Probabilities.* Die Zahlen sind Prozentwerte.

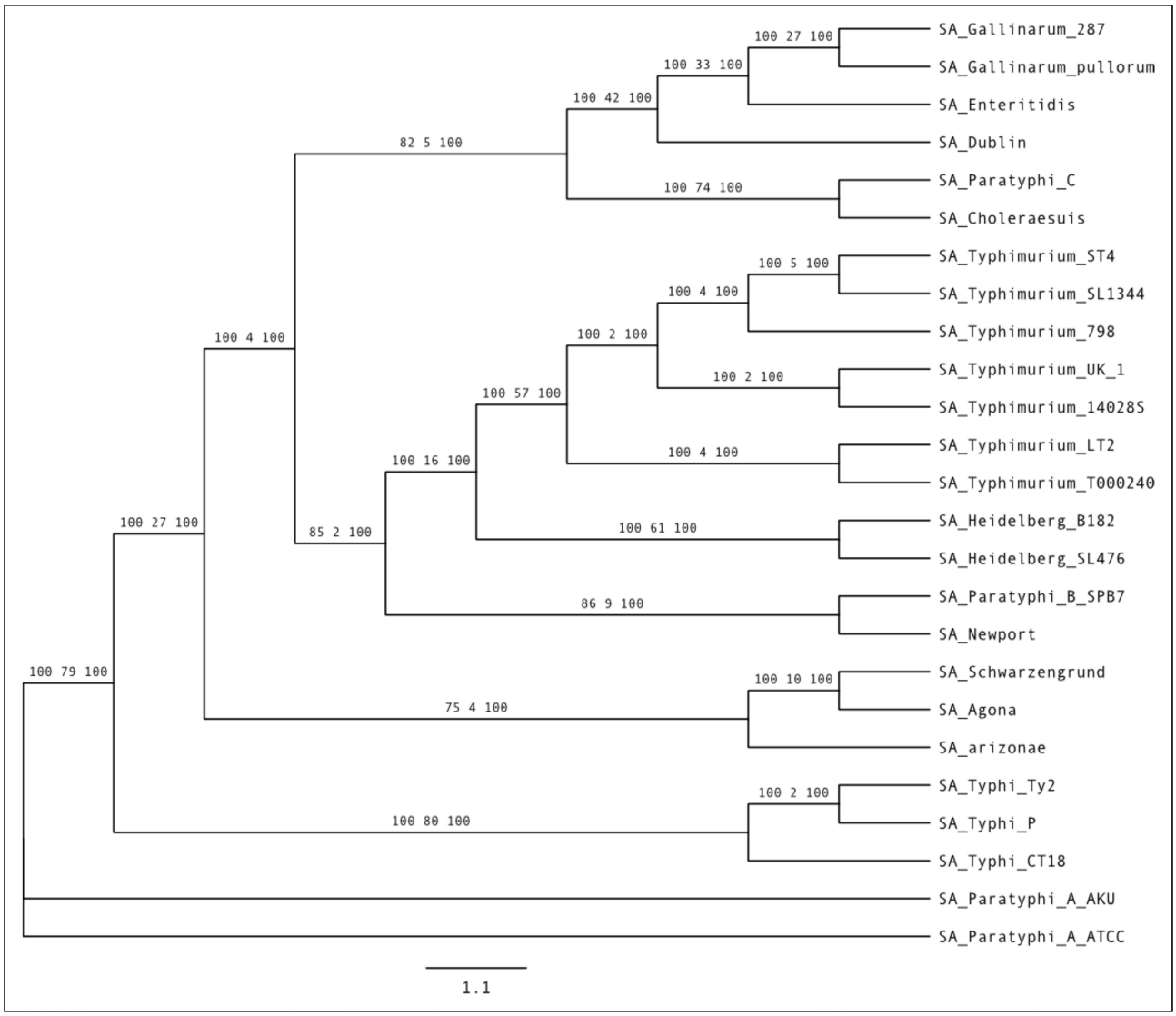


Abbildung 8.11: *Salmonellen-Phylogenie mit Support durch Bootstrap, Häufigkeit der Splits in den Genbäumen und Posterior Probabilities*. Die Zahlen sind Prozentwerte.

E. coli/Shigellen-Phylogenie von Reeves et al.

Von der Arbeitsgruppe um Peter Reeves stammt eine Phylogenie für 56 vollständig sequenzierte *Escherichia coli*- und *Shigellen*-Genome (Reeves et al., 2011). Aufgrund der nahen Verwandtschaft der *K-12*-, *O157:H7*- und *B*-Stämme untereinander haben sich die Autoren entschlossen, diese jeweils als einzelnes Taxon im Baum darzustellen. Das Genom von *E. fergusonii* wurde als Outgroup benutzt. Die Phylogenie wurde ebenfalls auf Grundlage der Alignments der Gene des Kerngenoms von *E. coli* und *Shigellen* konstruiert. Viele Bootstrap-Zahlen an den Ästen ihres Baumes (vgl. dazu Abb. 1 aus Reeves et al., 2011) sind aber sehr niedrig, dessen Verzweigungsmuster ist daher nicht sehr sicher. Reeves Phylogenie und *EC-Referenz* haben 41 Taxa gemeinsam.

Da der Reeves-Baum nicht in einem maschinenlesbaren Format vorlag, wurde die Abbildung aus Reeves et al. mit dem Java-Programm *TreeSnatcher Plus* in einen Newick-String übersetzt (für eine Erklärung des Programms siehe **Kapitel 5**). Nachdem aus den Bäumen mit *Mesquite* jeweils die Taxa entfernt worden waren, die nicht in beiden Bäumen vorkommen, wurden für Reeves Baum die Bootstrap-Zahlen manuell an die Skalierung der Bootstrap-Zahlen von *EC-Referenz* angepaßt. Dann wurden *RF*- und *ssRF-Distanzen* mittels *computeSSRF* ermittelt. Das Entfernen von Taxa aus den

Bäumen führte naturgemäß zu einer veränderten Konfidenz der Splits, daher sind die *ssRF-Distanzen* nur eingeschränkt gültig[30]. Die *RF-Distanz* von *EC-Referenz* zum Baum von Reeves ist 8, die *ssRF-Distanz* bei Signifikanzschwelle 140 ist 4 und 0 bei Schwelle 172. Unter besagtem Vorbehalt wird insgesamt aber deutlich, daß die Verzweigungsstruktur der Bäume, die beide fast alle heute sequenzierten *E. coli*-Genome berücksichtigen und beide aus Alignments konkatenierter Genfamilien konstruiert worden sind, sehr ähnlich ist. Die Bäume sind in **Abb. 8.14** dargestellt.

Die editierten Newick-Ausdrücke für die Phylogenien sind in den Dateien *ec_PHYLIPConsensus_final.tre* und *ec_reeves_final.tre* im Ordner *Reeves* zu finden.

RF- und ssRF-Distanzen zwischen den mit verschiedenen Methoden rekonstruierten Phylogenien

Methode		ML				Bayes		iTol: 26 COGs				Maslovs Methode	
	Spezies	EC	SA	ECSA	ECSA komb.	EC	SA	EC	SA	ECSA	ECSA komb.	EC	SA
ML	EC					0*		10				6	
	SA						0		0				11
	ECSA				4					14	15		
	ECSA komb.			10						9	13		
Bayes	EC	1						10*				6*	
	SA		0						0				10
iTol: 26 COGs	EC	52				51*						3	
	SA		18				18						11
	ECSA			78	74						0		
	ECSA komb.			72	72					66			
Maslovs Methode	EC	32				31*		60					
	SA		20				20		20				

SSRF-Distanz bei Signifikanzschwelle 70 %
RF-Distanz

* Der bayesianische E. coli-Baum unterscheidet sich nur durch einen Teilbaum mit drei B-Stämmen vom ML-Baum. In letzterem fungiert einer der Stämme als Outgroup zu den beiden anderen, im Bayes-Baum sind alle drei Stämme gleichberechtigt (der Teilbaum ist multifurzierend). Da sich weder mit computeSSRF noch mit TOPD der Abstand zwischen multifurzierenden Bäumen berechnen läßt, wurde das Programm treedist aus dem PHYLIP-Paket benutzt, um den RF-Abstand zu berechnen. Da ssRF-Abstand zwischen dem Bayes- und dem ML-Baum lediglich 1 ist, wurde der Abstand zum ML-Baum anstelle des Abstandes zum Bayes-Baum eingesetzt. Der wahre Abstand dürfte nur geringfügig abweichen.

Tabelle 8.9*: Distanzen zwischen den rekonstruierten Phylogenien bei Bootstrap-Schwelle 140 (von 200) bzw. dem nächsthöheren Wert*

Phylogenie für 13 Salmonellen von Timme et al.

Ruth Timme und ihre Mitautoren haben die erste Phylogenie (Abb. 1 in Timme et al., 2013) für *Salmonella enterica subsp. enterica* vorgelegt (Timme et al., 2013), die auf einem neuartigen, referenzlosen *k-mer-Ansatz* beruht, der genomweit gesammelte SNPs konkateniert. Ihre Phylogenie umfaßt 156 annotierte *Salmonellen*-Genome mit 78 Serovaren, von denen 102 erst kurze Zeit vor Veröffentlichung ihrer Arbeit sequenziert worden sind. 59 vollständig oder teilweise sequenzierte Genome stammen vom NCBI. Die Inferenz eines Maximum-Likelihood-Baumes wurde von ihnen auf den konkatenierten SNPs mittels *RAxML* 7.3.2 ausgeführt, wobei Rapid Bootstrapping zum Einsatz kam. Zum Wurzeln des Baumes wurde eine Outgroup, bestehend aus vier Vertretern der Subspezies von *Salmonella*, *S. e. diarizonae*, *S. e. houtenae*, *S. e. indica* und *S. e. salamae*, benutzt. In der ursprünglichen Phylogenie wurden die Bootstrap-Werte an den Ästen indirekt über deren

30 Zum Zwecke einer genaueren Aussage könnte man die Konfidenz der Splits ausrechnen, indem man die jeweiligen Taxa aus den Bäumen entfernt.

Dicke und Färbung visualisiert. Aus der Abbildung ist ersichtlich, daß zahlreiche endständige Äste einen sehr guten, viele Äste tiefer im Baum aber auch einen unzureichenden Bootstrap-Support besitzen. Beide Bäume wurden auf ihr gemeinsames Taxon-Set mit 13 Spezies reduziert (siehe Dateien *timme_13Taxa.phy* und *sa_PHYLIPConsensus_13Taxa.phy* im Verzeichnis *Timme*).

Da aus beiden Phylogenien, insbesondere aus der von Timme, viele Taxa entfernt wurden, wurde auf die Berechnung des *ssRF-Abstands* verzichtet. Der *RF-Abstand* zwischen beiden Bäumen ist 6 (vgl. Abb. 16), er wurde mit *treedist* aus der *PHYLIP*-Suite berechnet (vgl. **Abb. 8.15**). Im Hinblick auf die Größe der Phylogenie sind die Unterschiede beträchtlich, trotz gemeinsamer Struktur sowohl im Bereich der Blätter als auch tief im Baum. Die Dissonanz zwischen den Bäumen könnte zwei Hauptursachen haben: zum einen ist der Baum von Timme deutlich feiner aufgelöst als *SA-Referenz*, was es schwieriger macht, für ihn die Signale vertikaler Vererbung herauszuarbeiten.

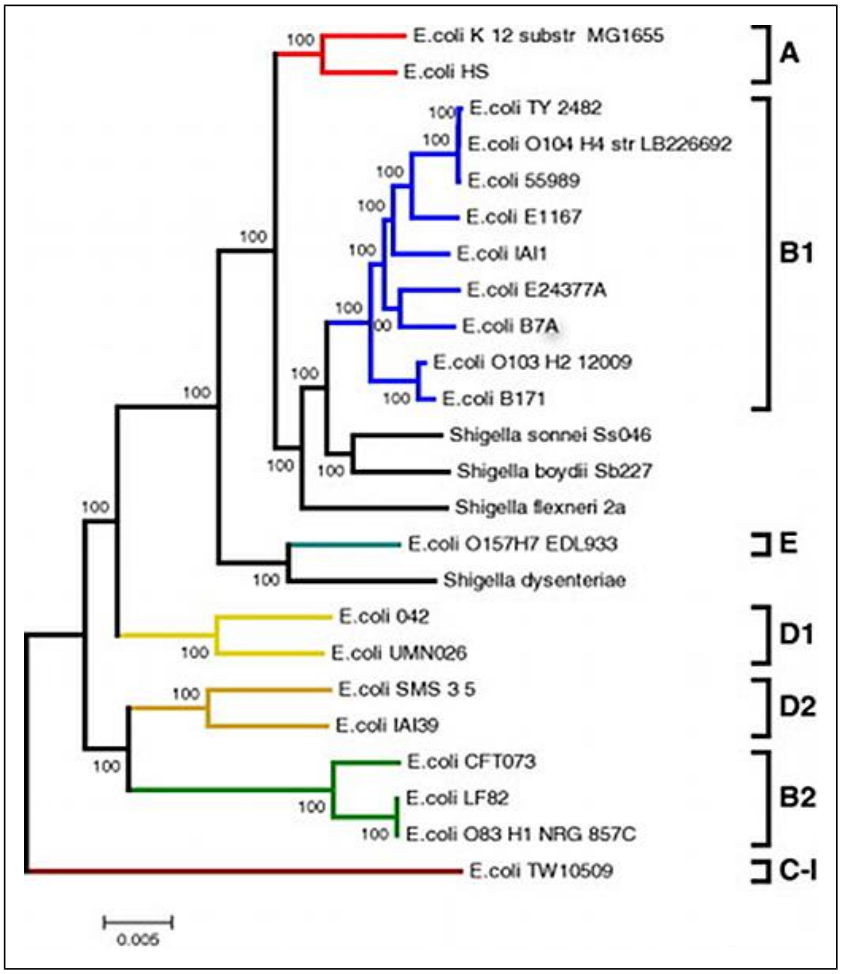


Abbildung 8.12: *E. coli-Phylogenie aus Chaudhuri et al., Abb. 4, auf Grundlage konkatenierter universeller Genfamilien. Die Phylogenie wurde mit RAxML v7.04 unter Benutzung des GTR-Modells für Nukleotidsubstitution und der CAT-Approximation für Ratenheterogenität geschätzt. Die Zahlen an den Ästen repräsentieren den Bootstrap-Support in Prozent.*

Zum anderen weisen Timme et al. selbst auf mögliche Probleme im Zusammenhang mit ihrer Methode hin (Timme et al., 2013, S. 2): ihre neuartige Methode baut eine Matrix mit Polymorphismen der einzelnen Nukleotide (SNPs), wendet darauf aber traditionelle Modelle unter einem Maximum-Likelihood-Framework an. Außerdem läßt ihre Methode eine separate

Betrachtung von Genbäumen nicht zu. Das macht es unmöglich, inkongruente, durch HGT hervorgerufene phylogenetische Signale zu identifizieren und zu analysieren. Da der Verfasser zur Konstruktion von *SA-Referenz* keine Outgroup benutzt hat, stand wahrscheinlich mehr innerartliches phylogenetisches Signal zur Differenzierung der Serovare zur Verfügung.

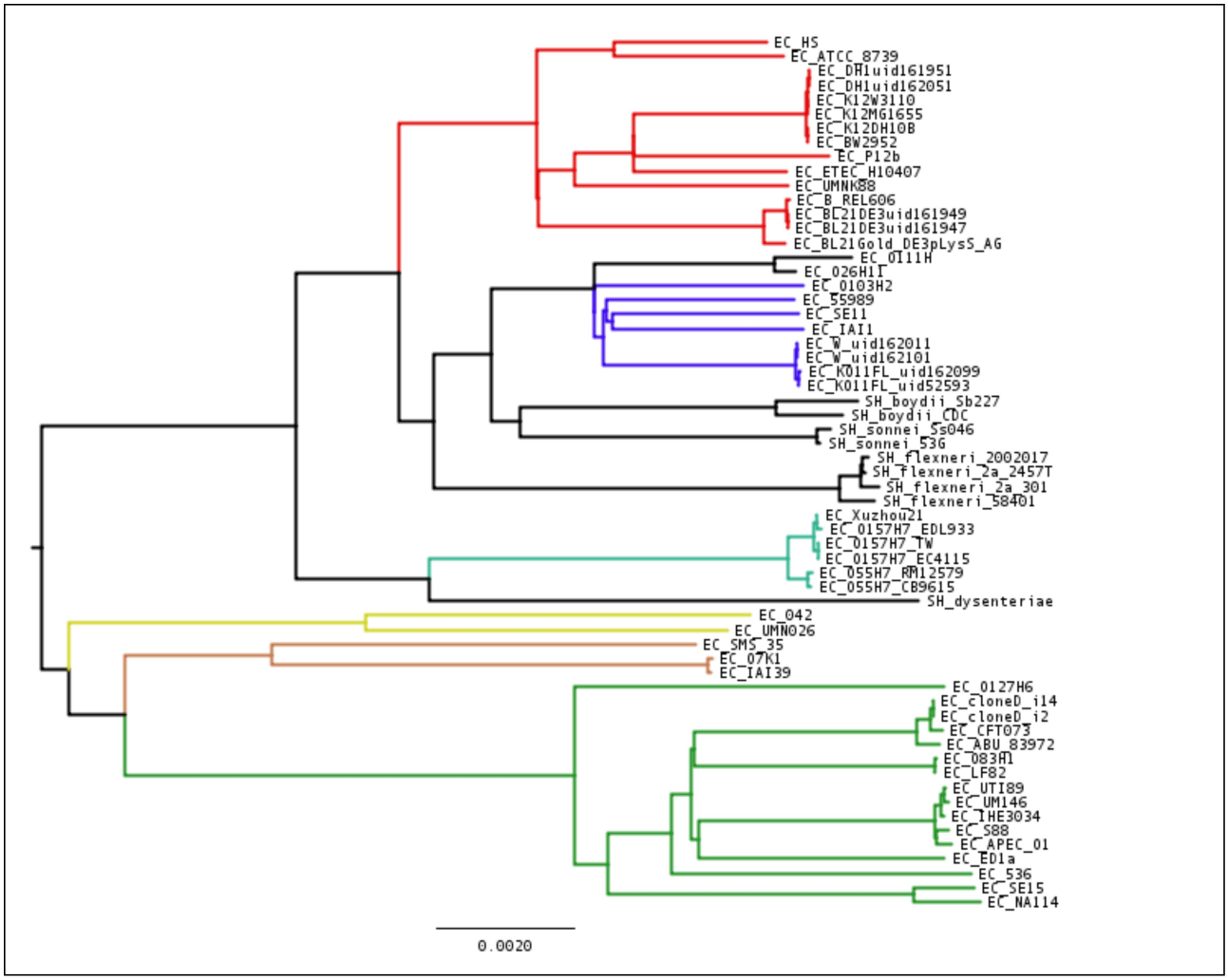


Abbildung 8.13*: Phylogenetische Beziehungen zwischen den 61 E. coli im EC-Referenzbaum.* Bei diesem Baum handelt es sich um den *E. coli*-Teilbaum aus **Abb. 8.2** mit über *RAxML* geschätzten Astlängen. Die Einfärbung nach phylogenetischen Gruppen wurde analog zum Baum in **Abb. 8.12** vorgenommen. Die Taxa, die nach der Klassifikation von Selander et al., 1987 verschiedenen phylogenetischen Gruppen zugerechnet werden, clustern auch in diesen Gruppen.

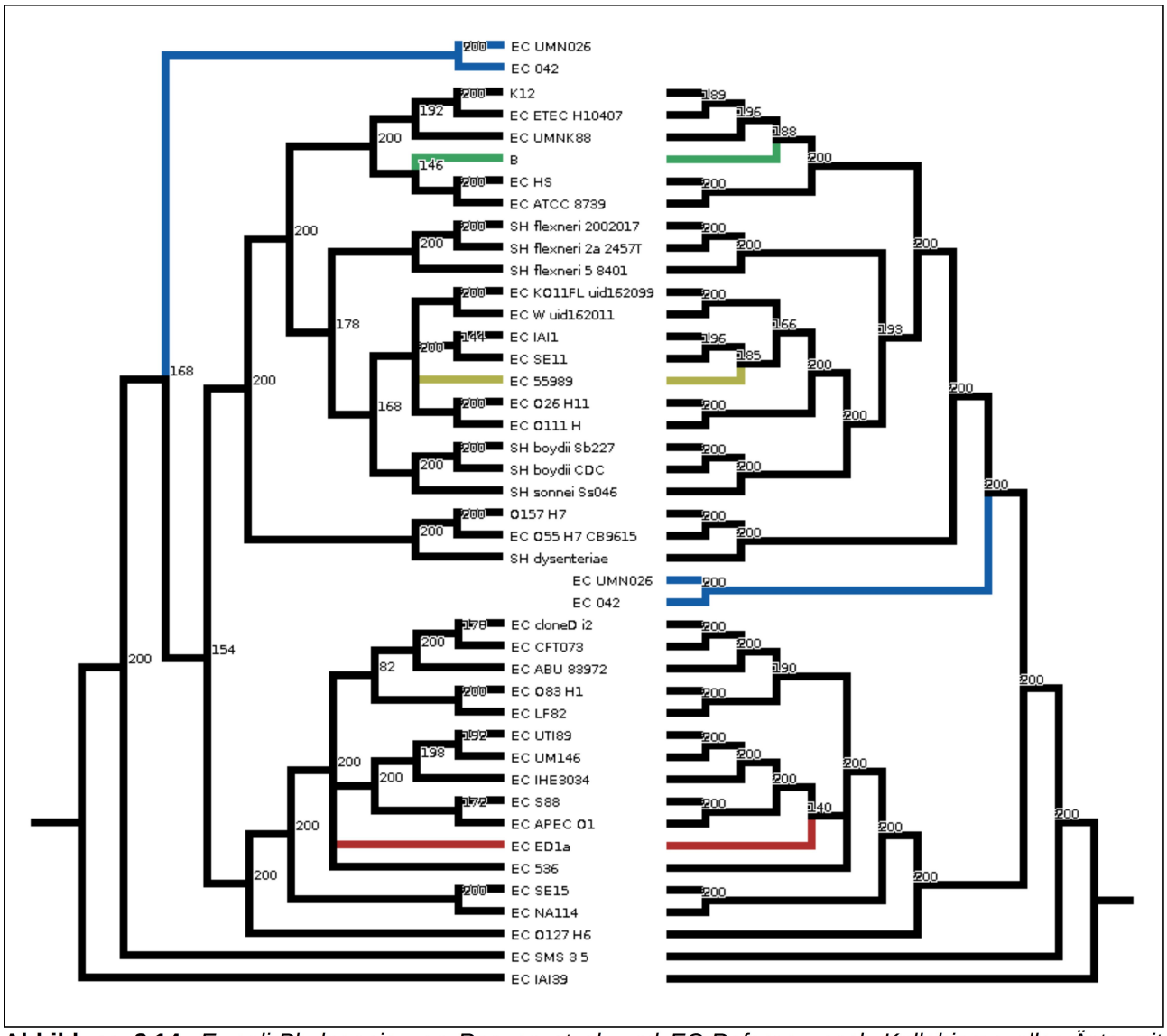


Abbildung 8.14*: E. coli-Phylogenie von Reeves et al. und EC-Referenz nach Kollabieren aller Äste mit Bootstrap-Wert < 140.* Der Baum links entspricht dem Baum aus Abb. 1 von Reeves et al., 2011, aus dem die Taxa entfernt wurden, die nicht in *EC-Referenz* enthalten sind. Aus *EC-Referenz*, rechts im Bild, wurden die Taxa entfernt, die nicht im Baum von Reeves et al. enthalten sind. Beide Bäume besitzen 56 Taxa. Die Zahlen an den Ästen stammen jeweils aus einem Bootstrap mit 200 Wiederholungen. Die farbig hervorgehobenen Äste repräsentieren die Unterschiede zwischen den Bäumen, deren *ssRF-Distanz* 4 ist.

8.4 Diskussion

In diesem Kapitel wurden Methoden beschrieben, die es zum Ziel hatten, in statistischer Hinsicht robuste Phylogenien für die im Rahmen dieser Arbeit verwendeten 86 *E. coli/Shigellen*- und *Salmonellen*-Stämme (siehe **Tab. 8.1**) zu schätzen. Rekonstruktionen, die möglichst den „wahren“ Phylogenien entsprechen, sind unabdingbar für Studien, die den anzestralen Lebensstil, den Metabolismus oder die Simulation evolutionärer Prozesse bei den betrachteten Bakterien zum Inhalt haben. Wurde in den Arbeiten, die dieses Kapitel beschreibt, die "wahre" Phylogenie von *E. coli* bzw. *Salmonellen* gefunden?

Zunächst empfiehlt es sich, die rekonstruierten Phylogenien nochmals einer Betrachtung zu unterziehen. Die ML-Methode und die Bayesianische Methode haben einen beinahe identischen *E. coli*-Baum geliefert, die *Salmonellen*-Bäume sind sogar völlig identisch. Die entsprechenden Neighbor-Joining-Bäume weisen lediglich geringe Unterschiede zu ihnen auf. Den Bäumen ist gemeinsam, daß ihre Splits bis auf wenige Ausnahmen durchgängig gut statistisch abgesichert sind. Die *E. coli*-Bäume zeigen die für diese Spezies definierten großen phylogenetischen Gruppen (Selander et al., 1987) und entsprechen damit im Wesentlichen bereits publizierten Bäumen (Chaudhuri et al., 2011; Reeves et al., 2011), sie sind allerdings höher aufgelöst. Der *Salmonellen*-Baum teilt seine Grundstruktur mit einem von Timme et al. im Jahr 2013 veröffentlichten Baum, weist aber einige Unterschiede zu ihm auf.

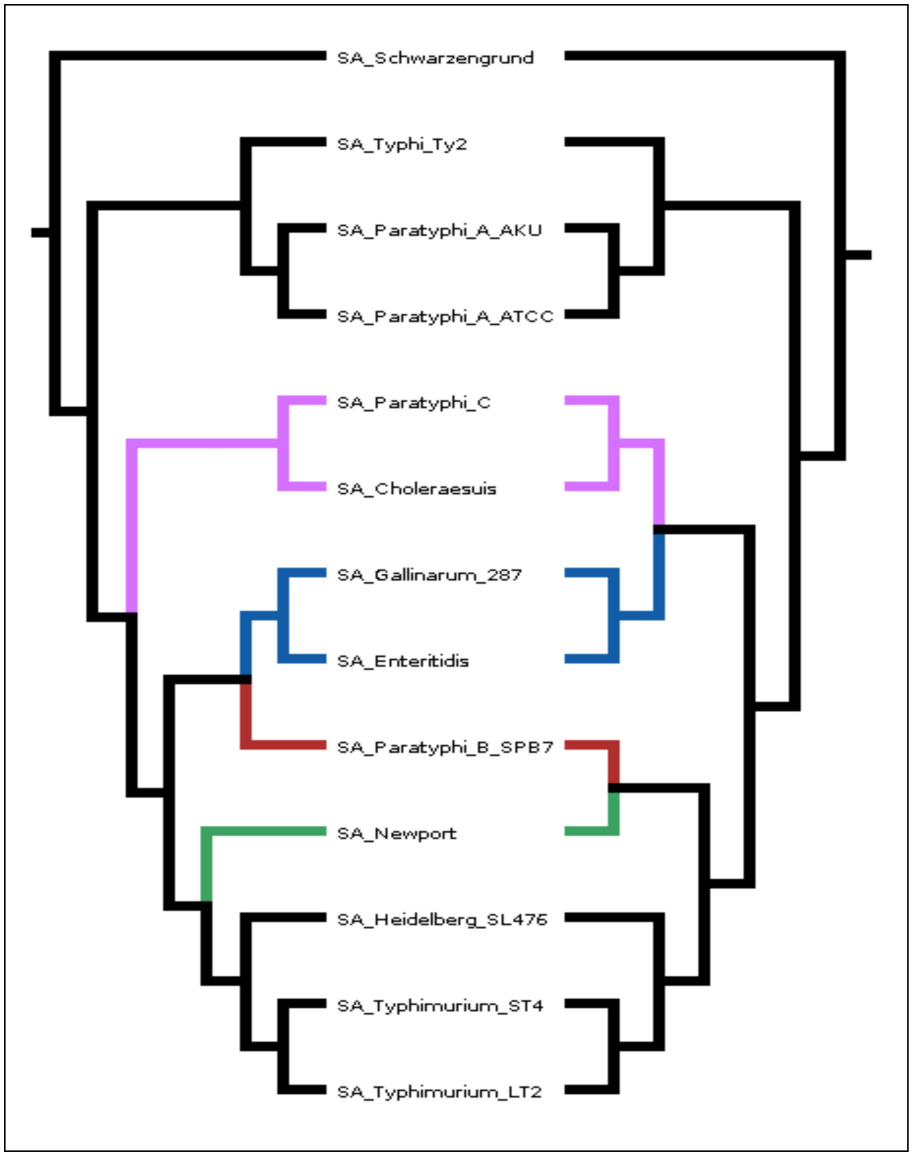


Abbildung 8.15*: Salmonellen-Phylogenie von Timme et al. und SA-Referenz.* Der Baum links ist der Baum aus Abb. 1 von Timme et al., 2013, aus dem die Taxa entfernt wurden, die nicht in SA-Referenz enthalten sind. Aus *SA-Referenz*, rechts im Bild, wurden die Taxa entfernt, die nicht im Baum von Timme et al. enthalten sind. Die farbig hervorgehobenen Äste repräsentieren die Unterschiede zwischen den Bäumen, deren *RF-Distanz* 6 ist.

Neben dieser ersten Gruppe von Bäumen existiert eine zweite Gruppe von Bäumen, die zum einen auf Grundlage hochkonservierter Gene („iTol-Methode“) erzeugt wurden, zum anderen mit einer bis dato unveröffentlichten, auf Neighbor-Joining basierenden Methode („Maslov-Methode“), die den Prozentsatz rekombinanter Regionen zwischen zwei Genomen als Distanz benutzt. Die von den beiden Methoden rekonstruierten *E. coli*-Bäume sind sich einander ähnlicher, als sie den *E. coli*-Bäumen der ersten Gruppe sind. Die *Salmonellen*-Bäume der zweiten Gruppe andererseits sind sich einander ebenso unähnlich wie den Bäumen der ersten Gruppe. Die statistische Unterstützung der Splits in den Bäumen der zweiten Gruppe ist grundsätzlich schlechter als in der ersten Gruppe, am schlechtesten ist sie bei der „iTol-Methode“.

Aufgrund dieser Überlegungen spräche einiges dafür, eher den Bäumen der ersten Gruppe zuzubilligen, womöglich eine Annäherung an die „wahren“ Phylogenien darzustellen. Allerdings können die verwendeten Methoden eine hohe statistische Unterstützung der Splits in den Bäumen selbst in Anwesenheit von systematischen und anders gearteten Fehlern im Vorfeld produzieren. Es kommen gleich mehrere Fehlerquellen in Frage:

Die Bestimmung von Orthologen ist eine Hypothese gemeinsamer Verwandtschaft (Timme et al., 2013). Falls sie inkorrekt ist, führt sie unweigerlich einen systematischen Fehler in die Bauminferenz ein. Die in dieser Arbeit eingesetzte Methode zur Orthologenbestimmung über Sequenzähnlichkeit und Syntenie (Eßer, 2010) wurde bisher noch nicht peer-reviewed. Es ist u. a. diskutabel, ob sie zur Detektion lateralen Gentransfers zwischen weniger nah verwandten Spezies eingesetzt werden kann. So erwartet sie neben einem hohen Grad an Sequenzidentität noch eine ausgeprägte Syntenie, die bei weniger nah verwandten Spezies aber gar nicht vorhanden ist. Für die Arbeiten zu den Kapiteln 6 und 7 wurden Orthogruppen mit *Proteinortho* bestimmt, da der Verfasser vermutet, daß die Methode von Christian Eßer in Anwesenheit lateralen Gentransfers ausgerechnet die syntenischen Gene nicht liefert.

Die in vielen Studien und auch vom Verfasser eingesetzte Konkatenationsmethode konstruiert ein multiples Sequenzalignment aus den universellen Genfamilien, auf dem die Phylogenie inferiert wird. Es werden also die Sequenzen aller Loci in einem gemeinsamen Datenquelle gebündelt, d. h., die gesamte zur Verfügung stehende Information wird zur Inferenz eines Baumes verwendet. Aus praktischen Gründen wird aber lediglich ein Modell für die Substitution der Nukleotide in den Sequenzen eingesetzt, obwohl davon auszugehen ist, daß die einzelnen Gene unterschiedlich evolvieren. Nur falls die evolutionären Raten aller Loci niedrig sind, kombinieren Maximum Parsimony- und Maximum Likelihood-Verfahren zuverlässig alle vorhandenen phylogenetischen Signale (Felsenstein, 2004, Kapitel 9).

Trotz dieser Einwände halten Touchon und Kollegen (Touchon et al., 2009) Whole-Genome-Analysen, d. h. Analysen, in deren Verlauf die gesamte in einem Genom enthaltene Information benutzt wird, nach wie vor für den „Gold-Standard“, wobei sie sich aber nur auf *Escherichia coli* beziehen. Salichos und Rokas (2013) bezeichnen Konkatenation und das Entfernen von Taxa aus einer Phylogenie demgegenüber als „Brute-Force“-Methode. Ihre Untersuchung von 1.070 Orthologen aus 23 Hefe-Genomen identifizierte 1.070 unterschiedliche Genbäume, von denen kein einziger mit dem Speziesbaum kongruent war, den sie über eine Konkatenationsmethode bestimmt hatten. Die Inkongruenz war für kürzere Äste tiefer in der Phylogenie größer. Die Wahl von Genen oder Knoten mit durchschnittlich hoher statistischer Unterstützung bei der Verwendung derartiger Methoden verbesserte insgesamt die Robustheit ihrer Bauminferenz. Sie erhielten ähnliche Ergebnisse bei Analysen von Vertebraten- und Metazoen-Phylogenien. Im Lichte ihrer Ergebnisse raten sie von der Verwendung von Konkatenation ab und befürworten stattdessen die Auswahl von Genen mit starkem phylogenetischem Signal.

Consensus-Bäume umgehen diese Probleme, berücksichtigen aber nicht alle in den konsolidierten Bäumen vorhandenen Signale und insbesondere nicht die Stärke der Signale. Eine

Charakterisierung der unter dem Begriff „Total Evidence Debate“ geführten Debatte für und wider Consensus bzw. Konkatenation ist nachzulesen bei Felsenstein, 2004. Steel und Böcker (2000) lieferten mathematisch begründete Einwände gegen die Consensus-Methode.

Die *E. coli*- und *Salmonellen*-Teilbäume wurden für die Verwendung mit dem Programm GLOOME, das anzestrale Muster auf der Phylogenie schätzt, gewurzelt. Die Bestimmung geeigneter Wurzelpositionen ging wie in **Abschnitt 8.2.2.1** erläutert vonstatten. Dabei wurde die Option „Constraint Tree/Guide Tree“ des Programms *RAxML* verwendet. Mit Angabe eines „Constraint-Trees“ wird allerdings eine – wenn auch biologisch motivierte - Grundannahme über die Gestalt der Phylogenie in die Bauminferenz eingebracht, die unter Umständen keine Entsprechung in den Daten besitzt[31].

Die Bäume ECSA-Referenz aus **Abschnitt 8.2.2.1** und der kombinierte „iTol-Baum“ aus **Abschnitt 8.2.2.4** wurden aus den gewurzelten Teilbäumen zusammengesetzt, und die Bootstrap-Zahlen aus den Teilbäumen wurden übernommen. Die Bäume wurden allerdings in unterschiedlichen Sitzungen auf verschiedenen Alignments berechnet, wenn auch mit den gleichen Parametern und der identischen Zahl von Bootstrap-Wiederholungen. Auf diesen Sachverhalt wird durch die unterschiedliche Einfärbung der Speziesnamen in **Abb. 8.3** hingewiesen. Insgesamt erscheint dem Verfasser dieses Vorgehen vertretbar, da es sich bei *E. coli* und *Salmonellen* nach heutigem Kenntnisstand um Kladen (Porwollik et al., 2002; Elena et al., 2005) handelt, die einen gemeinsamen Vorfahren haben[32]. Einen etwas anderen Ansatz, Teilbäume zu wurzeln und zu einem Baum zusammenzufassen, haben Tal Dagan, Yael Artzy-Randrup und William Martin (Dagan et al., 2008) benutzt: Von den Alignments der Teilbäume werden jeweils Consensus-Sequenzen derartig konstruiert, daß die am häufigsten in einer Spalte vorkommenden Nukleotide in eine Sequenz übernommen werden. Diese Sequenzen werden dazu benutzt, den Baum der Teilbäume zu schätzen und die Teilbäume zu wurzeln.

Der Ansatz der Gruppe hinter dem „Interactive Tree of Life“-Baum (Ciccarelli et al., 2006) erscheint geeignet, einen „Stammbaum des Lebens“ (engl. *„Tree of Life“)* hervorzubringen, da er auf Genfamilien beruht, die über das ganze zelluläre Leben verbreitet und hochkonserviert sind. Nach Auswahl universeller Proteine und Ausschluß evtl. lateral transferierter Proteine lieferte der Ansatz 31 orthologe Gruppen, die in 191 Genomen vorhanden sind. Auf ihnen errechnete er einen Baum mit durchweg guten Bootstrap-Werten. Dagan und Martin (Dagan und Martin, 2006) kritisierten, daß Ciccarelli et al. mit ihrer Technik nur lediglich ein Prozent der Positionen in einem durchschnittlichen prokaryotischen Genom konkatenierten. Demnach erhielten Ciccarelli und ihre Mitautoren zwar einen Baum, jedoch einen, den Dagan und Martin als „Tree of one percent of life“ bezeichneten. Dessen ungeachtet – da für diese Arbeit nicht relevant - hat der Verfasser die zu den von Ciccarelli et al. benutzten COGs äquivalenten Orthogruppen in seinen *E. coli*- und *Salmonellen*-Datensätzen bestimmt und auf ihnen Phylogenien geschätzt. Warum weisen die Äste dieser Phylogenien beinahe durchgehend einen schlechten Bootstrap-Support auf? Der Verwandtschaftsgrad der Organismen im „Tree of Life“-Baum ist vermutlich gering genug, daß diese sich über die COGs genügend differenzieren lassen. Der *E. coli*-Teilbaum im „Tree of Life“ umfaßt sechs Stämme, die sich auch in den höher aufgelösten Bäumen des Verfassers gut differenzieren ließen. Für diese Arbeit wurde der von Ciccarelli et al. erdachte Algorithmus nicht nachgebaut, sondern es wurden bestimmte Orthogruppen unter den bereits existierenden, mit dem Verfahren von Christian Eßer (Eßer, 2010) erzeugten Gruppen ausgewählt. Der Verfasser geht

31 Der Autor des Programms *RAxML* weist in einem Tutorium (http://sco.h-its.org/exelixis/web/software/raxml/hands_on.html) auf ein Szenario hin, in dem der Einsatz eines Guide Trees seiner Ansicht nach vertretbar ist: „The only purpose for which they may be useful is to assess various hypotheses of monophyly by imposing constraint trees and then conducting likelihood-based significance tests to compare the trees that were generated by the various constraints“.

32 Die phylogenetische Einordnung der *Shigellen* hingegen wird mittlerweile wieder diskutiert. Zu Beginn des 21. Jahrhunderts schien sicher, daß sie ein Stamm von *E. coli* ist und kein eigenes Genus (siehe z. B. Lan und Reeves, 2002). Eine jüngere Studie von Zuo und Kollegen (Zuo et al., 2013) kommt zu dem Schluß, daß die *Shigellen* eine Schwesterspezies von *E. coli* innerhalb des Genus *Escherichia* sind.

jedoch davon aus, daß dieser Unterschied sich nicht oder nur kaum auswirkt. Er hat allerdings die Vermutung, daß sich die Methode so, wie sie hier beschrieben ist, nicht uneingeschränkt für nah verwandte Organismen verwenden läßt. Es empfiehlt sich, eine größere Zahl von COGs zu verwenden, um den zufälligen Fehler bei der Schätzung der Phylogenie zu verkleinern.

Die von Sergei Maslov vorgeschlagene Methode produziert Phylogenien, die statistisch besser unterstützt, aber nicht völlig robust sind. Das verwendete Distanzmaß ist biologisch sinnvoll. Maslov hat seine Beobachtungen allerdings vorerst nur für *E. coli* aufgestellt. Zur Berechnung der Distanz zwischen zwei Spezies werden die rekombinanten Regionen jeweils innerhalb der Genfamilien bestimmt und aufaddiert. Der Verfasser hat in seinem Algorithmus für die Distanzberechnung zwischen einem Paar von Spezies rekombinante Regionen auch zwischen Genen aus verschiedenen Genfamilien zur Distanz addiert. Dieses Vorgehen führt möglicherweise zu den beobachteten, sehr niedrigen Bootstrap-Zahlen an den Ästen. In Wirklichkeit sind die Distanzen vielleicht kürzer. Ciccarelli und Kollegen und Sergei Maslov haben ihre Ansätze zur Ableitung einer Phylogenie so ausgelegt, daß diese eine robuste Phylogenie liefern, ohne dabei von Ereignissen lateralen Gentransfers beeinflußt zu werden. Aus diesem Grund erwartete der Verfasser, daß die von den Methoden erzeugten Bäume ähnlicher sind, als es tatsächlich der Fall ist. Da die Methode von Maslov für eine unabhängige Bewertung bisher noch nicht zur Verfügung stand, wurden etwaige Stärken und Schwächen noch nicht untersucht.

Der Verfasser schlägt insgesamt vor, für phylogenetische Studien zur Diversität von *E. coli* bzw. *Salmonellen* die in **Abschnitt 8.2.2.1** beschriebene Methodik zu benutzen: Konkatenation der universellen Genfamilien, Schätzen eines Baumes, besser mehrerer Bäume unter einem Maximum-Likelihood-Framework, Schätzen eines Consensus-Baumes für die „besten“ Bäume aus mehreren unabhängigen Experimenten unter einem Bayesianischen Framework. Die solchermaßen gewonnene *E. coli*-Phylogenie gibt, wie die bisher veröffentlichten Phylogenien, die für *Escherichia coli* definierten großen phylogenetischen Gruppen wieder, was ihr (subjektiv) eine gewisse Glaubwürdigkeit verleiht. Selbstverständlich brauchen aber auch diese Phylogenien nicht der „wahren“ Phylogenie von *E. coli* zu entsprechen. Es wird kritisiert, daß die im Jahre 1987 vorgenommene Einteilung in phylogenetische Gruppen maßgeblich auf *E. coli*-Isolaten von Menschen, Zootieren und domestizierten Tieren stammt, aber nicht von Tieren aus natürlichen Habitaten (Pupo et al., 2000).

Wahrscheinlich funktioniert die in **Abschnitt 8.2.2.1** skizzierte Vorgehensweise auch bei weniger nah verwandten Bakterienspezies. Es besteht dann aber die Gefahr, daß mit abnehmender Verwandtschaft der Organismen die vertikalen phylogenetischen Signale durch die Zunahme lateralen Gentransfers immer stärker verwaschen werden und dadurch schwerer zu entdecken sind (Kloesges et al., 2011).

Wie visualisiert man die Diversität von Bakterien angesichts lateralen Gentransfers am besten – durch einen Baum oder durch ein Netz? Bakterien vermehren sich durch Verdopplung, wobei ihr Erbmaterial mit den bekannten Einschränkungen nahezu unverändert dupliziert wird. Dieser Prozeß hat binären Charakter und wird daher folgerichtig als bifuzierender Baum dargestellt. LGT-Ereignisse während der Bakterienevolution verwässern dieses Signal allerdings, ihr Effekt darf daher nicht unberücksichtigt bleiben (Dagan und Martin, 2006; Dagan und Martin, 2008). Es ist für die Fragestellung nicht entscheidend, in welchem Maße LGT die möglicherweise dominanten vertikalen Signale stört. Vielmehr müssen, wenn möglich, beide Arten von Signalen integriert werden. Eine geeignete Darstellung für die Prokaryotenevolution wäre ein Netz, in welchem die vertikalen Kanten sehr viel höhere Gewichte besitzen als die horizontalen bzw. gerade die Gewichte, die den individuellen Stärken der Vererbungspfade entsprechen. Phylogenetische Bäume würden ihren Sinn nicht einbüßen, wenn man sie für die Darstellung evolutionärer Prozesse verwendete, in denen lateraler Gentransfer nur eine sehr eingeschränkte Rolle spielte.

Gewöhnlich betrachten Phylogenetiker LGT immer als „Störenfried“, der die Rekonstruktion von Stammesgeschichten unnötig erschwert, und versuchen daher, den Effekt von LGT auf Phylogenien zu minimieren. Damit ignorieren sie allerdings einen Großteil der in den einzelnen Genen steckenden Information. Sollte LGT im Verlauf der Evolution tatsächlich jedes Gen tangieren, sind dann molekulare Marker überhaupt ein probates Werkzeug zur Rekonstruktion der Beziehungen zwischen Spezies (Abby et al., 2011)? Sophie S. Abby und Kollegen schlagen ein phylogenetisches Modell für Gentransfers vor, das die in den Gene enthaltene Information mit dem Baum für die untersuchten Arten abgleicht. Für ihren Datensatz aus 16 Bakterien- und Archaeen-Phyla mit 12,000 Genfamilien in 336 Genomen stellten sie fest, daß LGT vermochte, bei den meisten Phyla die Diversifizierungsmuster der Spezies hervorzuheben, und daß die Ergebnisse robust gegenüber der Wahl der verwendeten Genfamilien war. Ebenso lieferte ihnen LGT ein ergiebiges Signal für die Wurzelung der Speziesbäume. Die Idee von Abby und Kollegen verdient breite Beachtung und sollte in der Forschungsgemeinde weiter verfolgt werden.

Die eingangs gestellte Frage, ob in dieser Arbeit die wahre Phylogenie für *E. coli* oder *Salmonellen* berechnet worden sind, ist rein abstrakter Natur und muß selbstverständlich unbeantwortet bleiben. Anhand von Experimenten, die auf Grundlage einer geschätzten Phylogenie weitere Daten produzieren, aus denen überprüfbare Ergebnisse gewonnen werden, läßt sich ggf. ermessen, wie präzise eine Phylogenie, die als Hypothese in die Arbeit eingeht, die Wirklichkeit annähert.

9 Identifikation biologisch assoziierter Gen/Nährstoff-Paare in der Evolution von Escherichia coli und Salmonellen

9.1 Einleitung

In der zweiten Hälfte des 20. Jahrhundert hat sich die Wissenschaftsgemeinde intensiv darum bemüht, das Erbmaterial vieler Lebewesen möglichst vollständig zu sequenzieren, zu kartografieren und jedes einzelne Gen zu annotieren, um dessen Funktion zu bestimmen.

Die Genomannotation allein vermag allerdings nicht die Frage zu klären, welche Rolle die Genprodukte innerhalb des Gesamtsystems Zelle und darüber hinaus haben. Dazu ist ein ganzheitlicher Ansatz erforderlich, in dessen Rahmen die Interaktion der verschiedenen Teilsysteme des Organismus untersucht wird.

Die Systembiologie ist ein solcher ganzheitlicher Ansatz. Sie entstand in den 1990er Jahren, als die Rechenleistung der Computer bereits genügend hoch war, um in-silico-Simulationen metabolischer Netzwerke von Prokaryoten in vertretbarer Zeit zuzulassen (Varma et Palsson, 1995; Edwards et al., 2001). Vor allem die Spezies *Escherichia coli* war oft Ziel metabolischer Rekonstruktionen, da sie zu den am besten untersuchten Prokaryoten zählt. Das metabolische Netzwerk von Bakterien ist nicht statisch, sondern evolvierte im Laufe der Jahrmillionen, indem an seiner Peripherie neue Gene integriert wurden, während die zentralen Bereiche konserviert blieben (Pal et al., 2005). Für den Großteil der Änderungen am metabolischen Netzwerk von *E. coli* wird lateraler Gentransfer verantwortlich gemacht.

Da der gesamte Genpool von *E. coli* sehr viel größer als sein Kerngenom ist, fällt es zunehmend schwer, die einzelnen *E. coli*-Stämme voneinander abzugrenzen und zu definieren, welches überhaupt die definierenden Eigenschaften der Spezies *E. coli* sind. Monk und Kollegen (Monk et al., 2013) haben den Metabolismus für 55 lebende *E. coli*- und *Shigella*-Stämme in-silico rekonstruiert und die Ergebnisse durch Laborexperimente unterstützt. Sie haben herausgefunden, daß sich die einzelnen Stämme offenbar durch individuelle, für die besetzte ökologische Nische charakteristische Wachstumsmöglichkeiten unterscheiden. Die Spezies *E. coli* charakterisieren sie über gemeinsame metabolische Fähigkeiten.

Möchte man die Dynamik der metabolischen Fähigkeiten von *E. coli* und anderer Organismen in die Zukunft extrapolieren oder vorhersagen können, welches Nährstoffangebot in der Umwelt zum Überleben einer Spezies vorhanden sein muß, ist es notwendig, den Einfluß sich verändernder Nährstoffangebote auf den Umbau des metabolischen Netzwerkes im Laufe der Evolution zu betrachten. Ein solcher Umbau kann aber nur geschehen, wenn zum passenden Zeitpunkt die notwendigen Genprodukte bereitstehen: der Transport eines Nährstoffs durch die Zellmembran erfordert beispielsweise mindestens einen entsprechenden Transporter. Evtl. ist auch die Expression mehrerer gekoppelter Gene notwendig, um einen bestimmten metabolischen Prozeß zu ermöglichen.

Für dieses Projekt wurden derartig assoziierte Paare von Genen und Nährstoffen gesucht, die auf der Phylogenie entweder gleichzeitig gewonnen oder verloren wurden. Dazu wurden der anzestrale Gengehalt auf Basis orthologer Gruppen sowie die essentiellen Nährstoffe in den anzestralen Umgebungen, die *E. coli* zum Überleben metabolisieren können mußte, mithilfe des Verfahrens von Borenstein (Borenstein et al. 2008) bestimmt. Eine etwaige mathematische Assoziation zwischen

den Verteilungen der Gewinn- und Verlustereignisse wurde über *Fishers Exakten Test* (Fisher, 1923) berechnet und, wo möglich, auch biologisch untermauert. Eine interessante Fragestellung im Zusammenhang mit diesem Projekt ist es, nach Genen zu fahnden, die meistens zusammen gewonnen oder verloren worden sind und herauszufinden, welcher Nährstoff zeitgleich aufgenommen oder verloren worden ist.

Für dieses Kapitel werden folgende Konventionen getroffen: Jede Aussage zu *Salmonella* bezieht sich auf den von der NCBI-Seite heruntergeladenen Datensatz von 25 *Salmonellen*-Genomen (SA) (Stand Oktober 2013), der der vorliegenden Arbeit zugrunde liegt. Jede Aussage zu *Escherichia coli* (EC) bezieht sich auf den von dort zum gleichen Zeitpunkt heruntergeladenen Datensatz von 61 *E. coli*- und *Shigellen*-Genomen (siehe **Tab. 8.1**). Der Datensatz aus den 61 *E. coli/Shigellen*- und 25 *Salmonellen*-Genomen wird im Folgenden als ECSA bezeichnet. Obwohl inhaltlich nicht völlig korrekt, werden die Begriffe Nährstoff und Compound der Einfachheit halber äquivalent benutzt. Äquivalent verwendet werden ebenfalls die Begriffspaare *gain* und *Gewinn*, *loss* und *Verlust* sowie *retained* und *keine Zustandsänderung*. Das Gleiche gilt für den Gebrauch von *Genfamilie* und *Proteinfamilie*.

9.2 Material und Methoden

Die für das in diesem Kapitel vorgestellte Projekt benötigten Dateien befinden sich innerhalb der Ordnerstruktur im Projektordner *Nährstoffe*. Die Daten im Ordner *Proteinfamilien_Ecoli_ ProteinOrtho* sind dort der Übersicht halber redundant abgelegt. Sie befinden sich auch im Ordner *Proteinfamilien_ProteinOrtho*.

9.2.1 Berechnung von Proteinfamilien

Für dieses Projekt wurden die *E. coli*-Proteinfamilien verwendet, die auch im Rahmen des Lebensstil-Projekts (**Kap. 10**) verwendet wurden. Nach dem gleichen Verfahren und mit identischer Parameterbelegung wie dort beschrieben wurden mittels *proteinOrtho* Proteinfamilien für die 25 *Salmonella*-Stämme konstruiert.

Die von *proteinOrtho* produzierte Ausgabedatei *proteinOrtho_output_SA_75.txt* enthält 11.177 Proteinfamilien und befindet sich im Ordner *Proteinfamilien_Salmonella_proteinOrtho*, zusammen mit der Datei *gi_matrix_sa_75.txt*. Ebenfalls auf die gleiche Weise wurden die kombinierten Familien für die ECSA-Proteine berechnet. Die Ausgabedateien *proteinOrtho_output_ECSA_75.txt* mit 25.955 Proteinfamilien und *gi_matrix_ecsa_75.txt* sind im Ordner *Proteinfamilien_ EcoliSalmonella_proteinOrtho* abgelegt.

9.2.2 Anzestraler Gengehalt

Für dieses Projekt wurde der anzestrale Gengehalt von *E. coli* benutzt, dessen Konstruktion an anderer Stelle beschrieben worden ist (siehe **Kap. 10.2.4**). Die dazugehörigen Dateien sind redundant im Ordner *AnzestralerGengehalt/EC* abgelegt. Mit der gleichen Prozedur wurden der anzestrale Gengehalt von SA (Ordner SA) auf der *Salmonellen*-Phylogenie (*sa_tree.tre*) und der von ECSA (Ordner ECSA) auf der ECSA-Phylogenie (Datei *ecsa_tree.tre*) geschätzt.

9.2.3 Anzestrale Umgebungen mit essentiellen Nährstoffen

Aufbauend auf dem Gengehalt der anzestralen *E. coli* (s. o.) wurden mit *KEGG* (Kanehisha, 2004), der *Kyoto Encyclopedia of Genes and Genomes*, deren metabolische Netzwerke vorhergesagt. Für diese Netzwerke wurden die essentiellen Nährstoffe, die in der Vergangenheit in der Umwelt von *E. coli* wahrscheinlich vorhanden waren und die *E. coli* nicht selbst synthetisieren konnte, von Jonathan Fritzemeier, einem Kollegen am Lehrstuhl für Bioinformatik, mit dem von Borenstein und Kollegen (Borenstein et al., 2008) vorgestellten Verfahren bestimmt. Dazu hat er Programme modifiziert, die Benjamin Braasch im Rahmen seiner Masterarbeit am gleichen Lehrstuhl angefertigt hat (Braasch, 2012). Folgende Schritte sind notwendig:

Die Dateien *EC_orthogroupsPresentPerNode.txt* und *gi_matrix_ec_75.txt* im Ordner *Anzestraler Nährstoffgehalt/Fritzemeier/* verknüpfen die Informationen, welche Genfamilien jeweils an einem Knoten in der Phylogenie präsent sind und welche Gene, identifiziert über ihre GI-Nummer, in jeder Genfamilie vertreten sind. Für jeden anzestralen *E. coli*-Repräsentanten werden die GI-Nummern über die von *KEGG* angebotene Enzym-Datenbank mit einer Reaktion verknüpft. In der Regel werden dabei immer mehrere Gene mit einer Reaktion verknüpft. Aus der Menge der gefundenen Reaktionen wird ein metabolisches Netzwerk erstellt. Für jedes Netzwerk wird mit dem Borenstein-Algorithmus auf Basis der Netzwerktopologie die Menge der exogenen Nährstoffe errechnet. Diese Menge wird von Borenstein als „Seed Set“ bezeichnet.

Die Implementierung des Verfahrens von Borenstein in der Programmiersprache R war u. a. Thema der unveröffentlichten Masterarbeiten von Jonathan Fritzemeier (Fritzemeier, 2012) und Rafael Dellen (Dellen, 2012) am Lehrstuhl für Bioinformatik, Heinrich-Heine-Universität Düsseldorf. Es wird daher an dieser Stelle nicht erläutert. Die grundlegenden Schritte des Algorithmus sehen allerdings wie folgt aus:

1. Dem rekonstruierten metabolischen Netzwerk liegt eine sog. stöchiometrische Matrix zugrunde, die die Stöchiometrie aller chemischen Vorgänge im Netzwerk, also Reaktanten, Produkte, Reaktionsrichtungen und Stoffmengen, bündelt. Aus der Matrix und den Reaktionsrichtungen wird ein bipartiter Graph aufgebaut, dessen eine Sorte Knoten Metabolite und dessen andere Sorte Knoten Reaktionen sind, die jeweils zwei Metabolite verbinden.

2. Der Graph wird in starke Zusammenhangkomponenten zerlegt. In einer starken Zusammenhangskomponente ist jeder Knoten über einen Pfad von jedem anderen Knoten erreichbar.

3. Starke Zusammenhangkomponenten enthalten dann Metabolite, die alle aus den anderen synthetisiert werden können. Wenn eine starke Zusammenhangskomponente keine eingehende Kante hat, dann muss mindestens ein Metabolit, der in dieser enthalten ist, vom Organismus aufgenommen werden, also exogen sein.

Die hier verwendete Implementierung stellt eine Modifikation der o. g. Algorithmen dar. Sie betrachtet „Inseln“ in den rekonstruierten Netzwerken mit maximal 10 Metaboliten nicht weiter, falls sie nicht mit der größten starken Zusammenhangskomponente verbunden sind[33].

Für die SA- und ECSA-Datensätze wurde analog vorgegangen.

Alle im Zusammenhang mit der Rekonstruktion des anzestralen Nährstoffgehaltes verwendeten Dateien befinden sich im Verzeichnis *AnzestralerNährstoffgehalt/Fritzemeier.*

33 J. Fritzemeier schreibt dazu sinngemäß in seiner Masterarbeit: Während der Borenstein-Rekonstruktion kann es, je nach Vollständigkeit der Netzwerke, vorkommen, daß Graphen nicht zusammenhängend sind. Ein partitionierter Graph ist biologisch sinnlos, da es dann eine Gruppe von Metaboliten gäbe, die niemals aus anderen Metaboliten, z. B. Grundbausteinen, hergestellt werden könnten. Borenstein et al. schlagen vor, Knoten und Kanten aus dem Graphen zu entfernen, die nicht mit der größten starken Zusammenhangskomponente verbunden waren und nicht mehr als 10 Knoten enthalten. Falls der Schwellenwert von 10 Knoten überschritten wird, sollte das Netzwerk manuell kontrolliert werden.

9.2.4 Test auf assoziierte Gen/Nährstoff-Paarungen

Im folgenden wird ein Verfahren erläutert, das sich auf die *E. coli/Shigella*-Phylogenie und die mittels GLOOME rekonstruierten anzestralen Gengehalte und Nährstoffumgebungen von *E. coli* bezieht. Die dazu nötigen Quelldateien befinden sich in den Ordnern A*nzestralerGengehalt/EC, Proteinfamilien_Ecoli_proteinOrtho* und dem Projektordner *Kapitel_Nährstoffe*. Das Verfahren speichert seine Ausgaben im Verzeichnis *FisherTests/EC*. Zur Durchführung des Verfahrens wurde das Perl-Programm `getSignificantGeneNutrientPairs_EC.pl`, das sich im Ordner *FisherTests/EC* befindet, erstellt. Außer der Version für EC existieren noch Varianten des Verfahrens für SA und ECSA. Die nötigen Daten sind in ähnlich lautenden Dateien und Verzeichnissen gespeichert. Da die entsprechenden Perl-Programme inhaltlich identisch zu `getSignificantGeneNutrientPairs_EC.pl` sind, werden sie nicht zusätzlich zu diesem besprochen.

Neben den vom Programm verwendeten Pfaden setzt der Benutzer zwei Konstanten: SIGN_THRESHOLD ist die Signifikanzschwelle, die für die Bewertung der Ergebnisse der *Fisher-Tests* (*Exakter Test nach Fisher*) herangezogen wird. SUM_GAIN_LOSS_DEMANDED wird verwendet, um die Zahl der Gen/Nährstoff-Paare (ab jetzt als „GN-Paare“ bezeichnet) zu beschränken, für die ein Fisher-Test durchgeführt wird. Wird die Konstante beispielsweise auf 3 gesetzt, werden nur Paare aus Gen und Nährstoff weiter betrachtet, für die gilt: Gewinn+Verlust des Gens >= 3 und Gewinn+Verlust des Nährstoffs >=3.

Vorbereitende Arbeitsschritte

Die im Perl-Programm realisierten vorbereitenden Arbeitsschritte 1 - 8 sind:

1. Die Datenstruktur, in der die *E. coli*-Phylogenie vorgehalten wird, wird aufgebaut. Ihre Konstruktion wird in **Abschnitt 10.2.5** beschrieben.

2. Die Zuordnung zwischen GI-Nummern und Proteinfamilien wird aus der Datei *AnzestralerGengehalt/EC/gi_matrix_ec_75.txt* gelesen.

3. Die Zuordnung zwischen GI-Nummern und den Klarnamen der entsprechenden Gene sowie der Spezies-Bezeichner wird aus der Datei *FisherTests/EC/geneTable.txt* gelesen.

4. Der mittels GLOOME rekonstruierte anzestrale Gengehalt wird aus der Datei *AnzestralerGengehalt/EC/Parsimony_count_of_events_per_site_per_branch.txt* eingelesen. GLOOME unterscheidet zwischen den Zuständen *gain* - ein Gen wurde auf einem Ast hinzugewonnen - und *loss* - ein Gen wurde auf einem Ast verloren. Das Programm fügt einen dritten Zustand, *retained* (etwa „beibehalten“), auf den Ästen hinzu, auf denen das betreffende Merkmal weder gewonnen noch verloren wurde. Zweck ist es, diese in den Daten bereits vorhandene Mehrinformation zu nutzen.

5. Aus der Datei *AnzestralerNährstoffgehalt/EC/KEGG_CompoundClearnames.txt* wird die Zuordnung der von *KEGG* verwendeten Compound-Abkürzungen zu ihren Klarnamen eingelesen[34].

6. Den Compound-Abkürzungen wird eine laufende Nummer zugeordnet. Eine Liste der Compound-Abkürzungen, der entsprechenden laufenden Nummer und der Klarnamen wird in der Datei *FisherTests/EC/ec_FisherTests_X_compoundMappings.txt* abgelegt, wobei X ein Parameter ist, der als Kommandozeilen-Parameter spezifiziert ist.

7. Mit dem Borenstein-Algorithmus wurde bereits der Gehalt an essentiellen Nährstoffen an den Knoten der *E. coli*-Phylogenie berechnet (siehe **Absatz 9.2.3**). Die Zustände wurden als 1 - *Nährstoff anwesend* und 0 - *Nährstoff abwesend* codiert. Diese Informationen werden eingelesen.

34 Für die Rekonstruktion des anzestralen Nährstoffgehalts für *E. coli* hat der dafür verwendete Borenstein-Algorithmus nicht alle KEGG für *E. coli* bekannten Compounds verwendet, sondern eine Teilmenge, ebenso wie für die beiden anderen Rekonstruktionen.

Für jeden Ast zwischen zwei Knoten werden die Zustände *gain*, *loss* und *retained* folgendermaßen abgeleitet: Ist der Zustand am Vorgänger-Knoten *abwesend* und der Zustand am Nachfolger-Knoten *anwesend*, wähle den Zustand *gain* für den Ast. Ist der Zustand am Vorgänger-Knoten *anwesend* und am Nachfolger-Knoten *abwesen*d, wähle den Zustand *loss* für den Ast. Anderenfalls wähle den Zustand *retained*.

An die vorbereitenden Schritte schließt sich der Algorithmus an, der mathematisch assoziierte Paare, bestehend aus einem Gen[35] und einem Nährstoff, identifiziert.

Algorithmus zur Identifikation assoziierter Gen-/Nährstoff-Paare

Die anzestralen Gengehalte und die Umgebungen essentieller Nährstoffe auf der *E. coli*-Phylogenie wurden wie oben beschrieben rekonstruiert, und ihr Zustandekommen kann über *gain*- und *loss*-Ereignisse sowie den später ergänzten Zustand *retained* auf den Ästen nachgezeichnet werden.

Für ein bestimmtes *E. coli*-Gen und einen von *E. coli* verwerteten Nährstoff soll festgestellt werden, ob ein statistisch signifikanter Zusammenhang zwischen den auf der Phylogenie rekonstruierten gain- und loss-Mustern existiert. Idealerweise liegt diesem Zusammenhang aber auch eine plausible biologische Erklärung zugrunde. So könnte beispielsweise ein statistischer Zusammenhang zwischen den Mustern für ein Gen und einem Nährstoff darauf hindeuten, daß dieses Gen für einen Transporter codiert, der den betreffenden Stoff durch die Zellmembran schleust. Der *Exakte Test nach Fisher* (Fisher, 1923) wird eingesetzt, um für ein Gen/Nährstoff-Paar einen etwaigen nichtzufälligen, statistisch signifikanten Zusammenhang festzustellen.

Anmerkung: Generell wird ein bestimmtes Gen auf einem Ast nur einmal entweder gewonnen oder verloren, das Gleiche gilt für einen bestimmten Nährstoff.

Es werden drei Fälle unterschieden, in denen die Verteilungen von Gen- und Nährstoff-Ereignissen für ein gegebenes Gen/Nährstoff-Paar miteinander verglichen werden. Die Fälle sind:

Direkt: Die drei Ereignisse *gain*, *loss* und *retained* für Gen und Nährstoff werden verglichen. Es sollen Paare von Genen und Nährstoffen identifiziert werden, die sowohl gleichzeitig gewonnen als auch verloren worden sind.

Co-Gain: Die zwei Ereignisklassen *gain* und *non-gain* (eines der Ereignisse *loss* oder *retained*) werden gegenübergestellt. Hier steht die Frage im Vordergrund, ob Gen und Nährstoff zusammen auf der Phylogenie gewonnen wurden. Ob der Verlust auch synchron erfolgte, ist zweitrangig.

Co-Loss: Die zwei Ereignisklassen *loss* und *non-loss* (eines der Ereignisse *gain* oder *retained*) werden gegenübergestellt. Hier steht die Frage im Vordergrund, ob Gen und Nährstoff zusammen auf der Phylogenie verloren wurden. Ob der Gewinn auch synchron erfolgte, ist zweitrangig.

Für den Fall **Direkt** wird eine 3x3-Kontingenztafel als Eingabe für den *Fisher-Test* mit den numerischen Variablen *a* - *i* konstruiert (siehe **Abb. 9.1**). Beispiel zur Interpretation: die Variable *a* zählt, wie oft auf allen Ästen die Paarung „Gengewinn bei gleichzeitigem Nährstoffgewinn“ rekonstruiert wurde. Die Variable *i* zählt die Paarungen „keine Veränderung - *retained* - beim Gen und keine Veränderung beim Nährstoff“.

Für die Fälle **Co-Gain** (siehe **Abb. 9.2**) und **Co-Loss** (siehe **Abb. 9.3**) werden jeweils 2x2-Kontingenztafeln mit den numerischen Variablen *j* - *m* respektive *n* - *q* konstruiert. Beispiel zur Interpretation: die Variable *k* zählt, wie oft auf allen Ästen die Paarung „Gengewinn bei gleichzeitigem Nährstoffverlust oder keiner gleichzeitigen Veränderung beim Nährstoff“ rekonstruiert wurde.

35 Anstelle des Begriffs „Gen“ müßte richtigerweise eher „Genfamilie“ verwendet werden, da das Gen in diesem Zusammenhang als Repräsentant „seiner“ Genfamilie anzusehen ist, welche sich aus den orthologen Genen der verschiedenen *E. coli* zusammensetzt.

Neben den Summen, die in den Variablen *a* - *q* gespeichert werden, werden für jedes Gen/Nutrient-Paar weitere Größen erfaßt, die sich ebenfalls jeweils auf die Gesamtheit aller Äste beziehen: *Zahl der Gengewinne oder -verluste*, *Zahl der Nährstoffgewinne oder -verluste*.

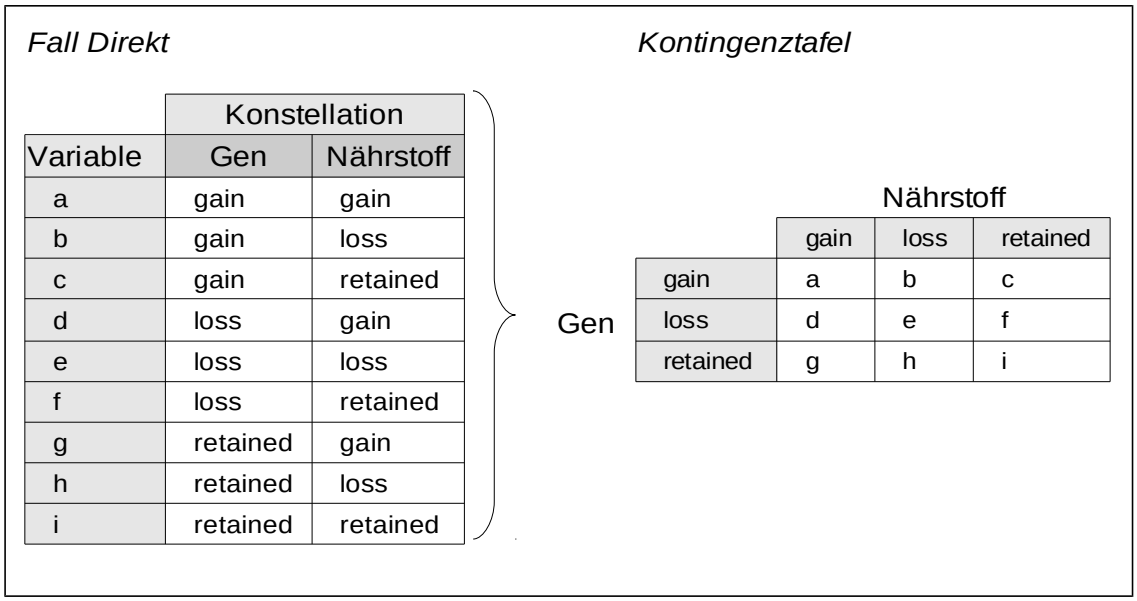


Abbildung 9.1: *Variablenbelegung der Kontingenztafel für den Fall Direkt*

Um die FDR-Korrektur (engl. „False Discovery Rate“) in Grenzen zu halten und eine genügende Genauigkeit des *Fisher-Tests* zu gewährleisten, werden nur für solche Paarungen aus Gen und Nährstoff Fisher-Tests durchgeführt, für die gilt (*Zahl der Gengewinne oder -verluste* >= SUM_GAIN_LOSS_DEMANDED) und (*Zahl der Nährstoffgewinne oder -verluste >= SUM_GAIN_LOSS_DEMANDED).* Die theoretisch durchzuführende Zahl von Tests für EC ist (Anzahl Gene/Genfamilien) · (Anzahl Nährstoffe) = 19.400 · 127 = 2.463.800.

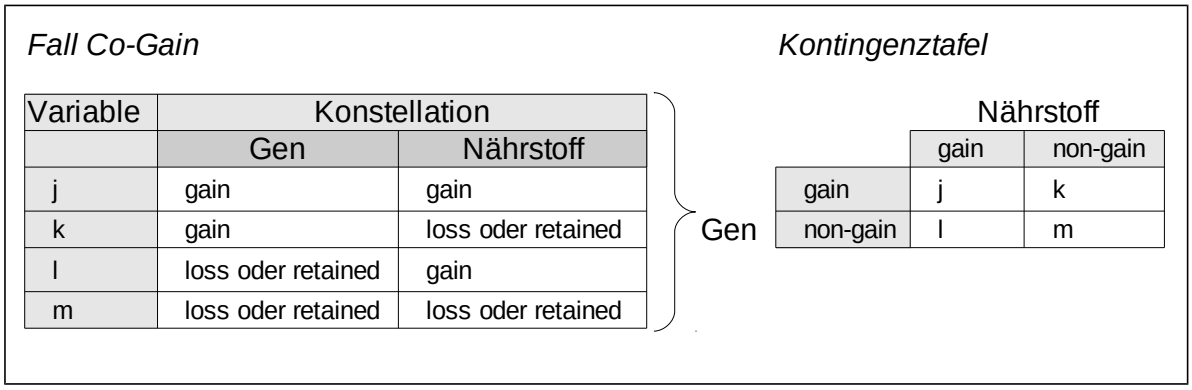


Abbildung 9.2*: Variablenbelegung der Kontingenztafel für den Fall Co-Gain*

Im Anschluß werden für das Gen/Nährstoff-Paar drei *Fisher-Tests* in der Statistiksprache R für jeden der drei Fälle durchgeführt. Das Programm benutzt die Methode *fisher.test*, die bereits im R-Grundpaket enthalten ist. Die Kontingenztafel wird der Methode jeweils als R-Matrix übergeben. Für jeden *fisher.test* werden der *p-Wert* und das *odds-ratio* erfaßt und gespeichert. Der Test wird in seiner einseitigen Variante und einem Konfidenzniveau von 0,95 ausgeführt. Für 2x2-Matrizen, das betrifft die Fälle **Co-Gain** und **Co-Loss**, verwendet *fisher.test* den Parameter „alternative“, der die Richtung der alternativen Hypothese für den Fisher-Test angibt.

Empirische Experimente mit verschiedenen Kontingenztafeln haben im konkreten Fall die Richtung „greater“ als geeignet ausgewiesen. *Fisher.test* kann lediglich für 2x2-Matrizen die *Grenzen des Kontingenzintervalls* ausgeben. Für größere Matrizen verwendet *fisher.test* ein anderes Verfahren. Für den Fall *Direkt*, das eine 3x3-Matrix erfordert, ist es erforderlich, den Parameter *simulate.p.value* auf TRUE zu setzen. Hier werden die *p*-Werte über eine Monte Carlo-Simulation ermittelt. In allen anderen Fällen ist er auf FALSE voreingestellt; dort werden die *p*-Werte aus einer hypergeometrischen Verteilung gewonnen.

Fall Co-Loss

Variable	Konstellation	
	Gen	Nährstoff
n	loss	loss
o	loss	gain oder retained
p	gain oder retained	loss
q	gain oder retained	gain oder retained

Kontingenztafel

Gen	Nährstoff: loss	Nährstoff: non-loss
loss	n	o
non-loss	p	q

Abbildung 9.3*: Variablenbelegung der Kontingenztafel für den Fall Co-Loss*

Die Ergebnisse der Fisher-Tests werden in die drei Dateien

ec_FisherTest_SUM_GAINLOSS_DEMANDED_3V_notCorrected.txt ***(Direkt)***

ec_FisherTest_SUM_GAINLOSS_DEMANDED_GN_notCorrected.txt ***(Co-Gain)***

ec_FisherTest_SUM_GAINLOSS_DEMANDED_LN_notCorrected.txt ***(Co-Loss)***

im Ordner *FisherTests/EC* geschrieben. Dabei wird die Konstante SUM_GAINLOSS _DEMANDED durch ihre im Programmlauf verwendete Belegung ersetzt. Die anderen Ergebnisdateien tragen ebenfalls diese Namen, jedoch mit anderem Suffix.

Die Belegungen der Variablen *a* - *q*, der unkorrigierte *p*-Wert sowie die Grenzen des Konfidenzintervalls für die *Fisher-Tests* pro Gen/Nährstoff-Paar werden in der Datei *ec_FisherTest_SUM_GAINLOSS_DEMANDED_3V_rawData.txt* im gleichen Ordner gespeichert. Da im Fall *Direkt* die Grenzen des Konfidenzintervalls nicht berechnet bzw. ausgegeben werden, werden diese auf 0 gesetzt. Für Debugging-Zwecke werden die Belegungen der Variablen *a* - *q* redundant ohne andere Daten in der Datei *ec_FisherTest_SUM_GAINLOSS_DEMANDED_3V_ params.txt* ebenfalls im gleichen Verzeichnis abgelegt.

Im Rahmen dieses Projektes werden z. B. für EC maximal 2,47 Millionen Tests für die verschiedenen Gen/Nährstoff-Paarungen durchgeführt. Da es sehr wahrscheinlich ist, daß sich unter diesen Tests diverse Falsch-Positiv-Resultate befinden, werden die *p*-Werte der Tests noch mit der Methode von Benjamini-Hochberg (auch als „FDR-Korrektur“ bezeichnet) für multiples Testen korrigiert (Benjamini und Hochberg, 1995). Die Tests mit den korrigierten *p*-Werten („*q*“ im FDR-Jargon) werden in Dateien mit dem Suffix „*_corrected.txt*“ im Ordner FisherTests/EC abgelegt.

Alle Tests, die im Sinne der geforderten Schwelle für den *p*-Wert, SIGN_THRESHOLD, signifikant sind, werden ebenfalls dort in Dateien mit dem Suffix „*_significant.txt*“ abgelegt, alle anderen in gleichlautenden Dateien mit dem Suffix „*_nonSignificant.txt*“.

Abschließend werden die IDs für die KEGG-Compounds in die entsprechenden Klarnamen der Nährstoffe übersetzt. Ebenso wird jede Gen-ID in eine Liste der Gene in der entsprechenden

Genfamilie übersetzt, wobei die Liste zusätzlich die Annotation der Gene beherbergt. Die solchermaßen übersetzten Dateien tragen das Suffix „*_clearnames.txt*“.

9.3 Ergebnisse

Für dieses Projekt wurden mit dem im vorhergehenden Abschnitt dargestellten Verfahren assoziierte *GN-Paare* für den EC-, den SA- bzw. den ECSA-Datensatz gesucht. Die Informationen zur physiologischen Rolle der untersuchten Gene und Nährstoffe stammen aus der *EcoCyc*-Datenbank (Keseler et al., 2012) und aus der *KEGG*-Datenbank (Kanehisa, 2004), sofern nicht Gegenteiliges gesagt wird.

Es wurden zunächst Fisher-Tests nur für solche *GN-Paare* durchgeführt, die das Kriterium SUM_GAINLOSS_DEMANDED ≥ 3 erfüllen (siehe vorheriger Abschnitt). Generell wurden Tests mit $p \leq$ SIGN_THRESHOLD = 0.05 als signifikant eingestuft. Im Anschluß wurden die Tests für unterschiedliche Obergrenzen für q (Tests mit p-Wert < q werden als signifikant gewertet) im Rahmen der FDR-Korrektur nochmals durchgeführt, um ggf. weitere untersuchenswerte Paare zu erhalten. Da diese Testergebnisse aus statistischen Gründen nicht mit den übrigen zusammengefaßt werden dürfen, werden sie gesondert aufgelistet. Die gezeigten p-Werte wurden für multiples Testen korrigiert.

9.3.1 Salmonellen-Datensatz

Der SA-Datensatz enthält 11.178 Proteinfamilien. Die mit dem Borenstein-Algorithmus (Borenstein et al., 2010) konstruierten anzestralen Nährstoffumgebungen basieren auf einem Reservoir von 139 KEGG-Compounds. Es wurden insgesamt 335.823 Fisher-Tests durchgeführt. Nach der im Anschluß erfolgten Benjamini-Hochberg-Korrektur für multiples Testen der p-Werte war kein Testergebnis in einem der drei Fälle **Direkt**, **Co-Gain** oder **Co-Loss** signifikant.

9.3.2 Escherichia coli-Datensatz

Der EC-Datensatz enthält 19.400 Proteinfamilien. Die anzestralen Nährstoffumgebungen basieren auf 127 KEGG-Compounds.

`Fall Co-Gain`

Unter 876.204 Fisher-Tests waren folgende drei Tests nach erfolgter Benjamini-Hochberg-Korrektur signifikant (vgl. Datei *ec_FisherTest_3_GN_significant_clearnames.txt*):

1 EC: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	phytic acid (myo-inositol hexakisphosphate)	(KEGG-Compound C01204)
Genfamilie	hypothetical protein b4592	(GI 145698240)
	hypothetical protein SSON_0986	(GI 74311536)
	yccB gene product (appX)	(GI 82776269)
	membrane protein	(GI 82544697)
	…	
	unnamed protein product	(GI 386708791)

p-Wert 0.0147

*Phytinsäure (*auch: *Myoinositolhexakisphospat)* speichert Phosphat und kommt in der Natur als Anion, dann als Phytat bezeichnet, vor. Lebewesen nehmen Phytinsäure nicht über die Nahrung auf, sondern müssen es selbst aus Phosphat und Inositol bzw. dessen Vorläufer Glukose synthetisieren (Hanakahi et al., 2000; Ungewickell et al., 1995).

*AppX (*auch: *yccB)* ist ein überwiegend in der stationären Phase (Hemm et al., 2008) exprimiertes Gen, das für ein vorhergesagtes Segment der äußeren Membran bei *E. coli* codiert.

Putative biologische Assoziation:
Die Gene appX und *appA* sind beide auf der Transkriptionseinheit *appCBA-yccB* lokalisiert. Die Expression des Phytase-Gens *appA* (Greiner et al., 1993) ist u. a. von der Phosphatkonzentration abhängig. Es wird unter anaeroben Bedingungen und spät in der stationären Phase exprimiert. Aufbereitete *E. coli*-Phytase dephosphoryliert Myoinositolhexakisphosphat (Greiner et al., 1993; Wyss et al. 1999; Greiner et al., 2000).

Abb. 9.4 zeigt die Gengewinne, Genverluste, Nährstoffgewinne und Nährstoffverluste auf der *E. coli*-Phylogenie.

2 EC: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	2-Trimethylaminoethylphosphonate (auch N-Trimethyl-2-aminoethylphosphonate)	 (KEGG-Compound C06459)
Genfamilie	flgC gene product	(GI 170021362)
	flagellar basal body rod protein FlgC	(GI 170683008)
	flagellar basal body rod protein FlgC	(GI 218703523)
	putative lateral flagellar component of cell-proximal portion of basal-body rod	(GI 260866411)
	unnamed protein product	(GI 378714340)
	…	
	flgC2 gene product	(GI 386708052)

p-Wert 0.0147

N-Trimethyl-2-aminoethylphosphonat ist am Phosphonat- und Phosphinatmetabolismus beteiligt.
FlgC (auch: *flaFIII* und *flaW*) ist ein Bestandteil des Flagellenmotor-Komplexes bei *E. coli*.

Putative biologische Assoziation: keine gefunden

3 EC: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	N-Trimethyl-2-aminoethylphosphonate (2-Trimethylaminoethylphosphonate)	(KEGG-Compound C06459)
Genfamilie	unnamed protein product	(GI 170021368)
	glycosyl transferase, group 2 family protein	(GI 170680272)
	putative glycosyltransferase	(GI 218703517)
	…	
	unnamed protein product	(GI 386708046)

p-Wert 0.0147

N-Trimethyl-2-aminoethylphosphonat ist am Phosphonat- und Phosphinatmetabolismus beteiligt. Das Enzym *Glycosyltransferase* katalysiert die Übertragung von Glycosylresten auf auf ein Akzeptor-Molekül (Breton et al., 2005).

Putative biologische Assoziation: keine gefunden

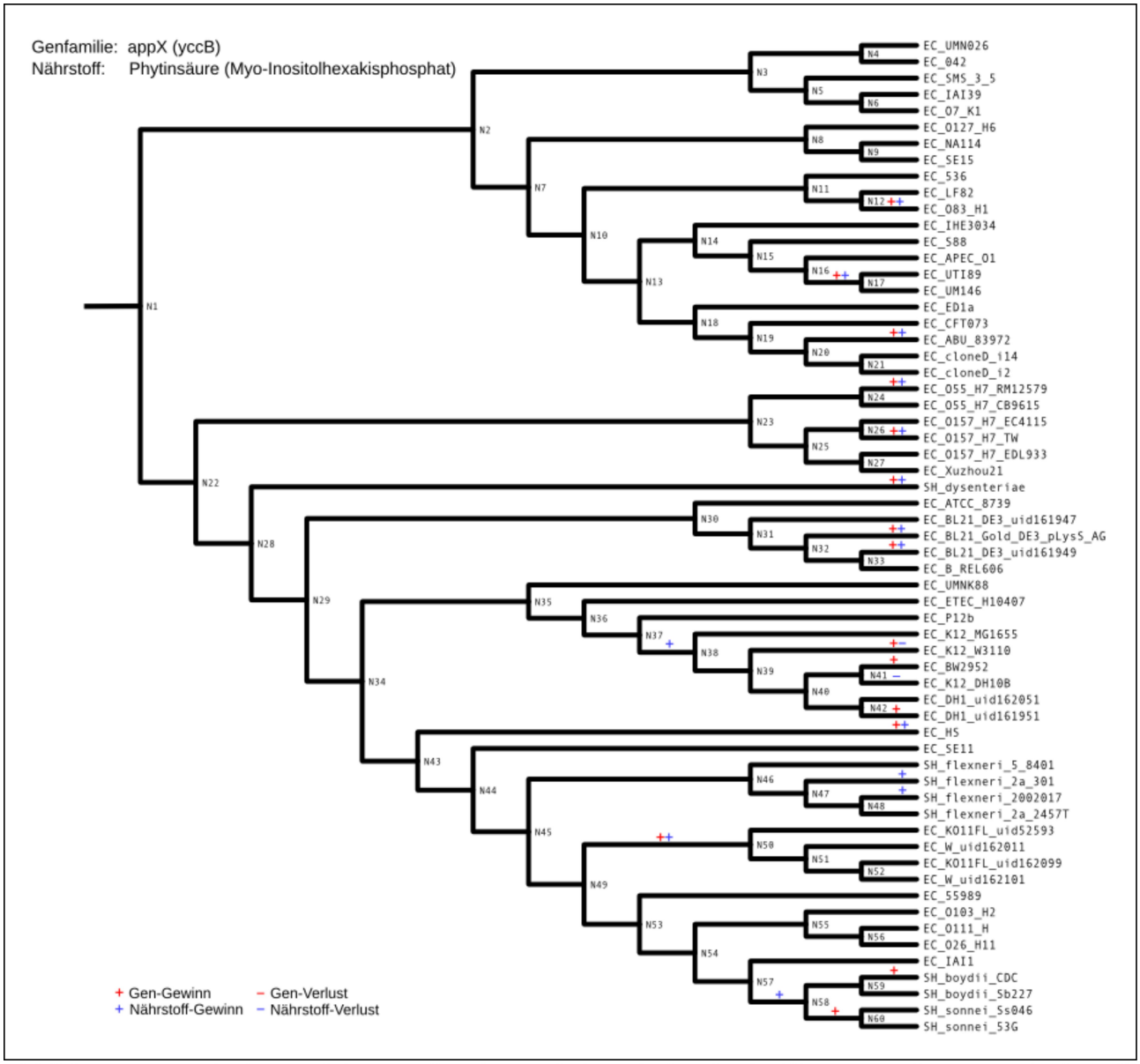


Abbildung 9.4*: Ereignisse auf der E. coli-Referenzphylogenie für ein signifikant assoziiertes Gen/Nährstoff-Paar*

9.3.3 Escherichia coli/Salmonellen-Datensatz

Der ECSA-Datensatz enthält 25.955 Proteinfamilien. Die mit dem Borenstein-Algorithmus (Borenstein et al., 2008) konstruierten anzestralen Nährstoffumgebungen basieren auf 139 KEGG-Compounds.

Fall Direkt

In diesem Fall gab es kein signifikantes Gen/Nährstoff-Paar.

Fall Co-Loss

In diesem Fall gab es sechs signifikant assoziierte GN-Paare. Bei drei von ihnen ist der Metabolit eine tRNA. Diese Paare werden hier nicht gezeigt, da tRNA kein in einem metabolischen Netzwerk hergestellter Compound ist[36].

Fall Co-Gain

In diesem Fall gab es 42 signifikant assoziierte GN-Paare. Bei 40 Paaren handelte es sich beim Nährstoff um N-Trimethyl-2-aminoethylphosphonate.

1 ECSA: Co-Loss, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	6-Deoxy-L-galactose (L-Fucose)	(KEGG-Compound C01019)
Genfamilie	L-fucose transporter	(GI 16130708)
	fucP gene product	(GI 15803323)
	...	
	L-fucose transporter	(GI 387883981)

p-Wert 0.0023

6-Deoxy-L-galactose (auch: *Fucose*) ist eine der essentiellen Zuckerarten, die für die Zell-Zell-Kommunikation benötigt wird. *Fucose* ist Bestandteil der bakteriellen Zellwand. *FucP* ist ein L-Fukose/Proton-Symporter

Putative biologische Assoziation: Stoff und Stoff-Transporter

2 ECSA: Co-Loss, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	Bilirubin beta-diglucuronide (Bilirubin-bisglucuronoside)	(KEGG-Compound C05787)
Genfamilie	glucuronide transporter	(GI 16129574)
	uidB gene product	(GI 15802031)
	...	
	glucuronide transporter	(GI 387882721)

p-Wert 0.0063

Bilirubinbetadiglucuronid ist an der Herstellung von Porphyrin beteiligt. Das von Bakterien hergestellte Vitamin B_{12} enthält Corrin, das ähnlich wie Porphyrin aufgebaut ist. Das Protein U*idB* (auch: *GusB*) ist ein Transporter für Alpha- und Beta-Glucoronide (Poolman et al., 1996).

Putative biologische Assoziation: Stoff und Stoff-Transporter

3 ECSA: Co-Loss, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	2-(alpha-D-Mannosyl)-3-phosphoglycerate	(KEGG-Compound C11516)
Genfamilie	predicted protein	(GI 90111359)
	hypothetical protein	(GI 15802388)
	…	

36 Das Borenstein-Verfahren behandelt tRNAs fälschlich als essentielle Nährstoffe, da unbeladene tRNAs von keiner metabolischen Reaktion erzeugt werden, sie jedoch an vielen Reaktionen partizipieren. In diesen Reaktionen wird die unbeladene tRNA in die geladene umgewandelt bzw. andersherum.

	hypothetical protein STM1984	(GI 16765321)
	...	
	yodD gene product	(GI 62180559)
	…	
	hypothetical protein CDCO157_2486	(GI 387883054)

p-Wert 0.0063

Das Enzym 2-(alpha-D-Mannosyl)-3-Phosphoglycerat ist eine Glykotransferase, die am Fruktose- und Mannosemetabolismus beteiligt ist. Das Gen *yodD* wird als Streßantwort auf Azidität und Wasserstoffperoxid exprimiert (Lee et al., 2010). Das Genprodukt *YodD* ist an der Bildung von Biofilmen beteiligt.

Putative biologische Assoziation: keine gefunden

1 ECSA: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	*N-Trimethyl-2-aminoethylphosphonate*	(KEGG-Compound C06459)
Genfamilie	unnamed protein product	(GI 170021369)
	cytidyltransferase-like protein	(GI 170681619)
	glycerol-3-phosphate cytidylyltransferase	(GI 218703516)
	...	
	unnamed protein product	(GI 386707990)

p-Wert $5.4212 \bullet 10^{-5}$

N-Trimethyl-2-aminoethylphosphonat ist am Phosphonat- und Phosphinatmetabolis-mus beteiligt. Das Enzym *Cytidyltransferase* überträgt innerhalb des Glycerophospho-lipid-Metabolismus phosphorhaltige Gruppen auf Akzeptor-Moleküle.

Putative biologische Assoziation: keine gefunden

2 ECSA: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	N-Trimethyl-2-aminoethylphosphonate	(KEGG-Compound C06459)
Genfamilie	glycosyl transferase, group 2 family protein	(GI 170680272)
	putative glysosyltransferase	(GI 218703517)
	putative glycosyltransferase	(GI 260866405)
	unnamed protein product	(GI 386607617)
	…	
	unnamed protein product	(GI 386708046)

p-Wert $5.4212 \bullet 10^{-5}$

N-Trimethyl-2-aminoethylphosphonat ist am Phosphonat- und Phosphinatmetabo-
lismus beteiligt. Das Enzym *Glycosyltransferase* katalysiert die Übertragung von Glycosylresten auf auf ein Akzeptor-Molekül (Breton et al., 2005).

Putative biologische Assoziation: keine gefunden

3 ECSA: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	D-Galactonate (Galactonate)	(KEGG-Compound C00880)
Genfamilie	SopD-like protein	(GI 16764332)
	sopD gene product	(GI 62179496)
	…	
	unnamed protein product	(GI 386590837)

p-Wert $5.4212 \cdot 10^{-5}$

D-Galaktonat (auch: *Galaktonat*) kann *E. coli* als alleinige Kohlenstoff- und Energiequelle dienen (Deacon et al., 1977). S*opD* ermöglicht es dem Bakterium, in Epithelzellen einzudringen.

Putative biologische Assoziation: keine gefunden

4 ECSA: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	Phytic acid (myo-inositol hexakisphosphate)	(*KEGG-Compound C01204)*
Genfamilie	hypothetical protein b4592	(GI 145698240)
	membrane protein	(GI 16765135:)
	…	
	yccB gene product	(GI 387506090)
	…	
	unnamed protein product	(GI 386708791)

p-Wert $5.8641 \cdot 10^{-7}$

Phytinsäure (siehe Co-Gain, EC, Fall1)
appX/yccB (siehe Co-Gain, EC, Fall 1)
Biologische Assoziation (siehe Co-Gain, EC, Fall 1)

Die übrigen Ergebnisse 5 – 42 assoziieren N-Trimethyl-2-aminoethylphosphonat mit folgenden Proteinen, die am Betrieb oder am Aufbau des Flagellen-Komplexes beteiligt sind:

5 - 42 ECSA: Co-Gain, SUM_GAINLOSS_DEMANDED ≥ 3

Nährstoff	N-Trimethyl-2-aminoethylphosphonate
Genfamilien	flagellar biosynthesis sigma factor
	lateral flagellar flagellin LafA
	lateral flagellar hook length control proteins LafC, LafD, LafE
	lateral flagellar basal body-associated protein LafF
	lateral flagellar RpoN-interacting regulatory protein LafK
	lateral flagellar associated protein LafV
	lateral flagellar hook associated protein LafW
	lateral flagellar hook associated proteins 1, 2, 3
	lateral flagellar peptidoglycan hydrolase LfgJ
	flagellar basal body P-ring/L-ring proteins
	flagellar basal body rod proteins FlgB, FlgC, FlgD, FlgF, FlgG
	lateral flagellar hook protein LfgE
	lateral flagellar P-ring addition protein LfgA
	lateral flagellar anti-sigma factor 28 protein LfgM
	lateral flagellar chaperone protein LfgN
	lateral flagellar basal body component protein LfiE
	lateral flagellar export/assembly proteins LfiI, LfiJ, LfiM, LfiN, LfiQ, LfiR
	flagellar biosynthesis protein FliP
	flagellar assembly protein H
	flagellar motor switch protein G

flagellar MS-ring protein
flagellar biosynthesis protein FlhB
flagellar motor protein MotA

Putative biologische Assoziation: keine gefunden

Die Ergebnisse und ihre *p*-Werte sind in der Datei *ecsa_FisherTest_3_GN_significant_clear names.txt* im Ordner *FisherTests/ECSA/* nachzulesen.

9.3.4 Was passiert für andere q-Werte?

Durch eine weniger restriktive Wahl des *q*-Wertes bei der FDR-Korrektur läßt sich eine u. U. größere Untermenge von Gen/Nährstoff-Paaren aus der Menge aller Paare auswählen. Tatsächlich erhöht man damit die akzeptierte Obergrenze für die *p*-Werte der Tests, die man als signifikant betrachtet. Man findet auf diese Weise möglicherweise Paare, deren nähere Inspektion sich unter Umständen lohnt.

Für SA wiesen bereits im Ausgangsdatensatz alle Paare *p*-Wert 1 vor FDR-Korrektur aus. Eine Variation von *q* bringt für SA daher keine Änderung. Ebensowenig gab es Paare für einen der drei Datensätze für den Fall *Direkt* bei unterschiedlicher Setzung von *q*. Die Datei *FisherTests/ assoziiertePaare.txt bietet* einen Überblick über alle mathematisch assoziierten *GN-Paare* mit *p*-Werten aus den Fisher-Tests und GI-Nummern. Im Anschluß werden die Veränderungen bei den Paaren für verschiedene Setzungen von *q* dargestellt.

EC

Co-Gain, q = 0,0523
Ein neues Paar gegenüber dem Fall q = 0,0500

Nährstoff Stachyose
Gen cscK (Genprodukt ist eine Fructokinase)
Assoziation Zucker-Metabolismus

Co-Gain, q = 0,0670
Verlust zweier Paare gegenüber dem Fall q = 0,0524

Co-Loss, q = 0,1532
Zwei neue Paare gegenüber dem Fall q = 0,0500

Nährstoff Bilirubin beta-diglucuronide
Gen uidB (Genprodukt ist ein Glucoronid-Carrier)
Assoziation Stoff und Stoff-Transporter

Nährstoff 2-(alpha-D-Mannosyl)-3-phosphoglycerate
Gen yodD, hypothetisches Protein
Assoziation keine gefunden

Co-Loss, q = 0,2379
Verlust aller Paare gegenüber dem Fall q = 0,1532

ECSA

Co-Gain, q = 0,0524
74 neue Paare gegenüber Fall q = 0,0500, darunter 16 ignorierte Paare mit einer tRNA als Nährstoff

Nährstoff Phytic acid
Gen yshB (Genprodukt: Transmembranprotein)
Assoziation keine gefunden

Nährstoff N-Acetylneuraminate
Gen hypothetisches Protein
Assoziation keine gefunden

Gen hypothetisches Protein, sopD-like (Genprodukt ermöglicht es dem Bakterium, in Epithelzellen einzudringen)
Nährstoffe Maltose, dCMP, L-Sorbose, ADP-ribose, Mannitol, L-Cystine, Raffinose, XTP, Pantheteine, Cob(I)alamin, L-Rhamnulose, Aquacob(III)alamin, 6-Deoxy-L-galactose (Fucose), N-Acetyl-D-galactosamine, Inositol 1-phosphate, P1,P4-Bis(5'-adenosyl) tetraphosphate, RX, Butanal, Salicin, Linamarin, L-Fuculose, D-Glucoside, Pseudouridine, D-Galactosamine, D-Phenylalanine, (R)-2-Methylmalate, N-Acetylmuramate, 1-(5-Phospho-D-ribosyl)-ATP, N-((R)-Pantothenoyl)-L-cysteine, N-Succinyl-2-L-amino-6-oxoheptanedioate, 5-Methyltetra-hydropteroyltri-L-glutamate, D-erythro-1-(Imidazol-4-yl)glycerol 3-phosphate, Dhurrin, 5(S)-HPETE, Selenite, Selenate, 15(S)-HPETE, O-Phosphorylhomoserine, Uroporphyrinogen, Cobinamide, L-4-Hydroxyglutamate semialdehyde, Digalactosyl-diacylglycerol, Digalactosylceramide, Arbutin, Amygdalin, Lotaustralin, trans-2-Methyl-5- isopropylhexa-2,5-dienoyl-CoA, cis-2-Methyl-5-isopropylhexa-2,5-dienoyl-CoA, N-Acetyl-L-citrulline, Methylselenic acid, D-glycero-alpha-D-manno-Heptose 1,7-bisphosphate, D-Cysteine, 2-(alpha-D-Mannosyl)-3-phosphoglycerate
Assoziation keine gefunden

Co-Gain, q = 0,0554
Verlust aller Paare gegenüber Fall q = 0,0524

Co-Loss, q = 0,0565
171 neue Paare gegenüber Fall q = 0,05, darunter 9 ignorierte Paare mit einer tRNA als Nährstoff

Gene hcaR, hcaA1 (hcaE), hcaA2 (hcaF), hcaC (bphF), hcaB
Nährstoffe Trans-Cinnamate, Benzene, Toluene, Phenylpropanoate, 4-Chlorobiphenyl, Biphenyl
Assoziation keine gefunden (ausgenommen die nächsten fünf Paare)

Gen hcaA1/hcaE (Genprodukt: 3-phenylpropionate/cinnamic acid dioxygenase subunit alpha)
Nährstoff Trans-Cinnamate
Assoziation Phenylalanin-Metabolismus

Gen hcaA2/hcaF (Genprodukt: 3-phenylpropionate/cinnamic acid dioxygenase subunit beta)
Nährstoff Trans-Cinnamate, Phenylpropanoate
Assoziation Phenylalanin-Metabolismus

Gen hcaC/bphF (Genprodukt: Dioxygenase ferredoxin subunit)
Nährstoff Phenylpropanoate
Assoziation Phenylalanin-Metabolismus

Gen hcaB (Genprodukt: 2,3-dihydroxy-2,3-dihydrophenylpropionate dehydrogenase)
Nährstoffe Trans-Cinnamate, Phenylpropanoate
Assoziation Phenylalanin-Metabolismus

Gen hcaR (Genprodukt: LysR family transcriptional regulator, hca operon transcriptional activator)
Nährstoff Phenylpropanoate
Assoziation HcaR + 3-phenylpropanoate = HcaR transcriptional dual regulator (Quelle: http://ecocyc.org)

Nährstoff D-Allose
Gene ygeX (Genprodukt: diaminopropionate ammonia-lyase)
ygeY (Genprodukt: Peptidase)
yqeB/xdhC (Genprodukt: hypothetisches Protein, xanthine dehydrogenase accessory factor)
yqeC (Genprodukt: 6-phosphogluconate dehydrogenase-like protein, hypothetical protein)
ygfJ (Genprodukt: involved in (molybdenum cofactor) cytidylyltransferase activity, hypothetical protein (Quelle: ECMDB, http://www.ecmdb.ca)
yecD (Genprodukt: isochorismatase family protein)
frV (Genprodukt: putative frV operon regulator (Quelle: www.ecogene.org)
yjcS (hypothetical protein, conserved protein, metallo-beta-lactamase superfamily)
csgA (Genprodukt: major curlin subunit)
yfiN (Genprodukt: transport permease YfiN, ABC-2 type transport system permease protein)
yhiM (Genprodukt: acid resistance protein, inner membrane protein)

	unnamed (Genprodukt: membrane protein, hypothetical protein, ATP binding protein of ABC transporter)
Assoziation	keine gefunden
Nährstoff	D-Allose
Gene	yqeA (Genprodukt: carbamoyl phosphokinase)
Assoziation	Microbial metabolism in diverse environments
Nährstoff	2-(alpha-D-Mannosyl)-3-phosphoglycerat
Gene	ymiA (Genprodukt: Membranprotein, hypothetisches Protein)
	yheV, yciX, ydcA (Genprodukte: hypothetisches Protein)
Assoziation	keine gefunden
Nährstoffe	2',3'-Cyclic AMP, 2',3'-Cyclic GMP, 2',3'-Cyclic CMP, 2',3'-Cyclic UMP
Gene	sipC (Genprodukt: invasin C)
	sipD (Genprodukt: cell invasion protein), yacA (Genprodukt: SecA regulator SecM)
	yaiV (Genprodukt: predicted DNA-binding transcriptional regulator (KEGG), putative transcriptional regulator (Quelle: www.ecogene.org)
	misL (Genprodukt: autotransporter family porin)
	ecnR (Genprodukt: LuxR family transcriptional regulator, putative regulatory protein)
	caiF (Genprodukt: DNA-binding transcriptional activator, transcriptional activator CaiF)
	yajG (Genprodukt: lipoprotein, uncharacterized lipoprotein)
	ahpF (Genprodukt: alkyl hydroperoxide reductase)
	yccU (Genprodukt: hypothetical protein, function unknown (Quelle: www.ecogene.org)
	mltE (Genprodukt: membrane-bound lytic murein transglycosylase E)
	yfcB (Genprodukt: site-specific DNA-methyltransferase (adenine-specific)
	glmY (Genprodukt: tRNA1Val (adenine37-N6)-methyltransferase)
	amiC (Genprodukt: N-acetylmuramoyl-L-alanine amidase)
	rpiA (Genprodukt: ribose 5-phosphate isomerase, constitutive)
	yhaD (Genprodukt: glycerate 3-kinase)
	yrdA (Genprodukt: conserved protein, ferripyochelin-binding protein, transferase)
	yigL (Genprodukt: pyridoxal phosphate phosphatase YigL)
	yijC (Genprodukt: FabR: Putative ABC transport system)
	ytfH (Genprodukt: predicted transcriptional regulator, HxlR-type, DUF24 family)
	unnamed (Genprodukt: transcriptional regulator, hypothetical protein)
	unnamed (Genprodukt: hypothetical protein)
	unnamed (Genprodukt: membrane transporter)
Assoziation	keine gefunden
Nährstoffe	2',3'-Cyclic AMP, 2',3'-Cyclic GMP, 2',3'-Cyclic CMP, 2',3'-Cyclic UMP
Gene	yfcB (Genprodukt: site-specific DNA-methyltransferase (adenine-specific)
	yfiC (Genprodukt: tRNA1Val (adenine37-N6)-methyltransferase
	pnp/PNPT1 (Genprodukt: polyribonucleotide nucleotidyltransferase)
Assoziation	Synthese von Purinen und Pyrimidinen
Nährstoff	Bilirubin beta-diglucuronide
Gen	flhD (Genprodukt: transcriptional activator FlhD)
Assoziation	keine gefunden
Gen	D-ala-D-ala transporter subunit
Nährstoffe	trans-Cinnamate, Benzene, Toluene, D-Allose, Biphenyl
Assoziation	keine gefunden
Gen	orfB (Genprodukt: transposase/IS protein)
Nährstoffe	(Indol-3-yl)acetamide, 4-Guanidinobutanamide
Assoziation	keine gefunden
Gen	fimA (Genprodukt: major type 1 subunit fimbrin (pilin))
Nährstoffe	3-Cyano-L-alanine, 3-Aminopropiononitrile
Assoziation	keine gefunden
Gen	fimC (Genprodukt: fimbrial chaperone protein)
Funktion	biogenesis of type 1 fimbriae. Binds and interact with FimH (Quelle: http://www.uniprot.org)
Nährstoffe	3-Cyano-L-alanine, 3-Aminopropiononitrile
Assoziation	keine gefunden

Co-Loss, q = 0,2379
2 neue Paare gegenüber Fall q = 0,0565, beide ignoriert, da Nährstoff tRNA
Verlust aller anderen Paare

9.3.5 Zusammenfassung

Für viele Gen/Nährstoff-Paare (*GN-Paare*) konnte eine mögliche biologische Assoziation gefunden werden. Die Ergebnisse im einzelnen sind in der Datei *Fisher-Tests/signifikantePaare.txt* nachzulesen; es folgt eine Übersicht:

- Als der Nährstoff Bilirubin beta-diglucuronide in der Umwelt vorhanden war bzw. benötigt wurde, wurde das Gen uidB exprimiert, das für seinen Transporter codiert. Das Gleiche gilt für den Nährstoff 6-Deoxy-L-galactose/L-Fucose und das Gen fucP.
- Das Gen appX wurde exprimiert, als Phytinsäure neu verfügbar war. Die Gene appA und appX liegen auf der gleichen Transkriptionseinheit. Phytase ist das Genprodukt von appA und dephosphoryliert Phytinsäure.
- Zeitlich zusammen lagen die Verfügbarkeit von N-Trimethyl-2-aminoethylphosphonat und die Expression diverser Gene, die in Beziehung zum Flagellen-Komplex von *E. coli* und *Salmonellen* stehen. Möglicherweise ist dieser Nährstoff am Aufbau bzw. Betrieb des Flagellenapparates beteiligt.
- Wenn auf der Phylogenie das Gen sopD, dessen Produkt das Bakterium befähigt, in Epithelzellen einzudringen, exprimiert wurde, war eine Vielzahl von Nährstoffen neu in der Umgebung vorhanden. Zwar konnte keine Begründung gefunden werden, man kann jedoch spekulieren, daß *E. coli* bzw. *Salmonellen* zu diesen Zeitpunkt in eine neue Umgebung eingewandert sind (ggf. in die Epithelzellen), in der sie sich dieser neuen Nährstoffe bedienen konnten.
- Mehrere Gene und Nährstoffe gingen zeitgleich verloren, die im Rahmen der Synthese von Phenylalanin interagieren.
- Als der Zucker D-Allose nicht mehr vorhanden war, gingen diverse Gene verloren, unter denen einige für Transporter codieren. Der genaue Zusammenhang konnte nicht ergründet werden.
- Die Nukleotide 2',3'-Cyclic AMP, 2',3'-Cyclic GMP, 2',3'-Cyclic CMP, 2',3'-Cyclic UMP müssen entweder produziert werden, oder sie sind in der Umwelt vorhanden. Es gibt Hinweise darauf, daß diverse assozierte Genprodukte direkt oder indirekt an der Synthese von Purinen und Pyrimidinen beteiligt sind. Sie wurden möglicherweise dafür benötigt bzw. nicht mehr gebraucht, als die Nukleotide von außerhalb der Zelle bezogen werden konnten.

Mit dem Begriff der Gleichzeitigkeit wurde oben etwas nachlässig umgegangen. Für ein konkretes *GN-Paar* sollte individuell entschieden werden, welche Definition von „Gleichzeitigkeit“ für die Gewinn- und Verlust-Ereignisse für die jeweilige Fragestellung passend ist. **Abb. 9.4** illustriert, daß der Fisher-Test die Verteilungen der Ereignisse für ein konkretes GN-Paar durchaus für signifikant erklärt, selbst wenn die Ereignispaare nicht immer exakt gleichzeitig liegen.

Für den Fall **Direkt,** SUM_GAINLOSS_DEMANDED ≥ 3, gab es keine mathematisch assoziiierte *GN-Paare*. Während für **Co-Gain** der kleinste nichtsignifikante *p*-Wert vor Benjamini-Hochberg-Korrektur $2{,}404 \cdot 10^{-7}$ und nach Korrektur 0,053 war, waren die *p*-Werte in diesem Fall deutlich größer: 0,0005 vor und 0,232 nach der Korrektur. Die *p*-Werte für **Co-Loss** bewegen sich in einer vergleichbaren Größenordnung. Intuitiv sind die niedrigeren *p*-Werte für **Co-Gain** und **Co-Loss** dadurch erklärbar, daß anstatt für drei Zustände die Verteilungen für nur zwei Zustände verglichen werden. Dadurch steigt die Wahrscheinlichkeit, daß die Verteilungen signifikant ähnlich sind. Der Vorteil, niedrigere *p*-Werte und damit mehr signifikante Tests zu erhalten, wird allerdings durch den

Nachteil aufgewogen, daß das Zusammenziehen zweier Variablen einen Genauigkeitsverlust mit sich bringt.

9.4 Diskussion

Eine verfeinerte Suche nach etwaigen physiologischen Beziehungen zwischen solchen assoziierten Genen und Nährstoffen, für die hier keine unmittelbare Erklärung gefunden werden konnte, ist außerhalb der Möglichkeiten dieser Arbeit, zumal eine solche interdisziplinär erfolgen sollte. Das Gleiche gilt für erweiterte Fragestellungen, von denen oben einige skizziert wurden. Eine solche Analyse sollte vielmehr in einem Anschlußprojekt mit einer angemessenen biochemischen Expertise angestellt werden.

In dem hier dargestellten Projekt wurde, teilweise erfolgreich, versucht, biologische Assoziationen zwischen mathematisch assoziierten Paaren von Genen und Nährstoffen manuell über den Vergleich von Annotationen und Beschreibungen in Datenbanken zu bestimmen. Ein erste Verbesserung, die es erlaubt, derartige biologische Assoziationen immerhin halbautomatisch z. B. auf Basis der Informationen in KEGG zu bestimmen, könnte wie folgt aussehen:

1. Modelliere die Vereinigung der metabolischen Netzwerke der betrachteten Spezies.
2. Erstelle einen bipartiten Graphen mit Metaboliten und Reaktionen als Knoten und Stoffwechselflüsse als Kanten.
3. Messe für jedes *GN-Paar* die Länge des Pfades im Netzwerk, der beide verbindet. Diese Zahl repräsentiert den Grad der Assoziation zwischen Gen und Nährstoff.

Mit diesem Ansatz ließen sich indirekte Beziehungen zwischen Genen und Nährstoffen aufdecken. Direkte Beziehungen zwischen Liganden und Genen lassen sich mit der *KEGG* API sehr leicht abfragen.

Keiner der Tests, in denen die Verteilungen der Zustände *gain*, *loss* und *retained* verglichen wurden, war signifikant. Ob mathematische oder methodische Gründe dafür verantwortlich sind, konnte bisher nicht geklärt werden. Darüberhinaus war kein Test im Zusammenhang mit dem SA-Datensatz signifikant. Es ist denkbar, daß dort die Zahl der Ereignisse zu gering gewesen ist. Die SA-Phylogenie besitzt sehr viel weniger Äste als die EC-Phylogenie, und für die ECSA-Phylogenie gab es sehr viel mehr signifikante Ergebnisse als für die EC-Phylogenie allein.

Führt man Ergebnisse mit individuellen Belegungen des Parameters SUM_GAINLOSS_DEMANDED durch, müssen die *p*-Werte aller durchgeführten Tests nachträglich FDR-korrigiert werden, bevor die Testergebnisse bewertet werden. Mit einer steigenden Zahl von Tests werden die *p*-Werte durch die FDR-Korrektur allerdings immer größer, wodurch umgekehrt die Zahl signifikanter Tests sinkt. Möchte man lediglich Paare aus Genen und Nährstoffen identifizieren, deren Untersuchung sich unter Unständen lohnt, ist es vertretbar, verschiedene Programmläufe mit unterschiedlicher Schranke für *q* durchzuführen.

Die putativen biologischen Assoziationen, die für einen beträchtlichen Teil der *GN-Paare* gefunden wurden, erscheinen plausibel. Das spricht dafür, daß die in diesem Kapitel vorgestellte Methode mit Kenntnis der geschätzten Verteilungen der Ereignisse auf der Phylogenie tatsächlich biologisch assoziierte Paare zu liefern vermag. Wenn auch nur implizit, nutzt sie dabei die Ergebnisse von Laborexperimenten, auf denen die von KEGG bezogenen Daten beruhen.

Über den Borenstein-Algorithmus fließt Vorwissen über durch Stoffwechselwege assoziierte Gene und Nährstoffe in das hier dargestellte Verfahren ein. Findet es diese Assoziationen auch wieder? Gibt es die entsprechenden Gen/Nährstoff-Paare als signifikant assoziiert aus? Dieser Frage, die auf die Zirkularität des Ansatzes abzielt, sollte nachgegangen werden. Der Ansatz lieferte signifikant assoziierte Paarungen aus Nährstoff und dem Gen, das für den entsprechenden Transporter codiert, zurück. Der Borenstein-Algorithmus liefert nur Metabolite, jedoch keine Gene und insbesondere keine Transporter. Das läßt vermuten, daß der Ansatz sinnvolle Ergebnisse hervorzubringen vermag, die nicht explizit in den Eingangsdaten vorhanden waren. Unter den signifikant assoziierten Paarungen sind einige hypothetische Proteine. Das vorgestellte Verfahren kann bei der Suche nach den mutmaßlichen Genprodukten hilfreich sein.

Eigene Ergebnisse aus in-silico-Simulationen mit Laborexperimenten abzugleichen, empfiehlt sich, wenn die Möglichkeit besteht, da Simulationen praktisch immer unvollkommen sind. Monk und Kollegen (Monk et al., 2013) haben die unvollständige in-silico-Rekonstruktion metabolischer Netzwerke für 55 lebende *E. coli*- und *Shigella*-Stämme durch Daten aus Laborexperimenten komplettiert. Ihr Interesse galt allerdings nicht der Evolution der metabolischen Fähigkeiten der Spezies *E. coli*, sondern einer Momentaufnahme. Sie war dazu geeignet, die Spezies *E. coli* von anderen prokaryotischen Spezies abzugrenzen und die nischen-spezifischen metabolischen Anpassungen der einzelnen *E. coli*-Stämme herauszuarbeiten.

Für die Borenstein-Rekonstruktionen der anzestralen metabolischen Netzwerke, die für dieses Projekt angefertigt wurden, wurde nicht geprüft, ob sie Biomasse produzieren können, auch wurde nicht versucht, etwaige Lücken in den Stoffwechselwegen zu füllen. Um die Tauglichkeit der Methode weiter zu verbessern, sollte beides in einem Anschlußprojekt nachgeholt werden.

10 Gingen verstärkte Gengewinne oder -verluste mit einer Änderung in der Lebensweise bei E. coli einher?

10.1 Einleitung

Die Vertreter der Klade *Escherichia coli-Shigella* besetzen verschiedene ökologische Nischen und weisen einen sehr unterschiedlichen Lebensstil auf. So können sie unter extremen Bedingungen überleben und auch ein Leben in Frischwasser und in Böden führen (Winfield und Groisman, 2003). Mehrere Stämme leben als Kommensalen im Darm von Warmblütern, andere sind Pathogene, die für zum Teil lebensbedrohliche Krankheiten beim Menschen verantwortlich sind. Aus diesem Grund beschäftigt sich die Forschung seit langem mit den möglichen Ursprüngen ihrer Pathogenität. Pathogene *E. coli* haben sich mehrfach auf unterschiedlichen Teilen der Phylogenie aus apathogenen Stämmen entwickelt (Ogura et al., 2009). Der Wechsel von einer apathogenen zu einer pathogenen Lebensweise, oder andersherum, erfordert es, sich schnell an wechselnde Lebensbedingungen anpassen zu können. Es wurde vielfach die These aufgestellt, daß Bakterien die dazu notwendigen neuen Fähigkeiten nicht durch vertikale Vererbung, sondern durch horizontalen Gentransfer (HGT) erworben haben (Ochman et al., 2000).

Verschiedene Arbeiten haben sich zum einen mit der Frage beschäftigt, zwischen welchen Prokaryoten wie intensiv lateraler Gentransfer betrieben wurde (Skippington und Ragan, 2012), und zum anderen, ob ursächlich eine phylogenetische Beziehung, die gemeinsame ökologische Nische oder gleicher Lebensstil dafür verantwortlich sind (Skippington und Ragan, 2012; Popa et al., 2011). Skippington und Ragan haben 27 *Escherichia coli*-Stämme untersucht und herausgefunden, daß *E. coli* lateralen Gentransfer überwiegend innerhalb der eigenen Klade betrieben hat, während Transfers über Speziesgrenzen hinweg zwar unbehindert möglich waren, aber sehr viel seltener stattgefunden haben. Dabei waren weder Frequenz noch Stärke lateralen Gentransfers uniform. Bemerkenswert ist hierbei, daß sie keinen Hinweis darauf gefunden haben, daß *E. coli*, die den gleichen Lebensstil pflegen oder in vergleichbarer Umgebung leben, intensiver oder häufiger HGT betrieben haben als ihre Verwandten, die beides nicht tun.

Popa et al. (Popa et al., 2011) haben beobachtet, daß bei nah verwandten Bakterien pathogene Spezies häufiger Gene austauschen als nicht pathogene. Der phylogenetischen Beziehung wird generell eine entscheidende Rolle zugedacht, da nah verwandte Spezies ausgetauschtes genetisches Material leicht durch homologe Rekombination in das Genom integrieren können (Lawrence und Retchless, 2009). Im Gegensatz dazu berichten Smillie und Kollegen (Smillie et al., 2011), dass zwischen fern verwandten Genomen der genetische Austausch mehr über die Ökologie als über eine phylogenetische Beziehung strukturiert wird, und dies vorzugsweise zwischen Isolaten, die in einer ökologisch vergleichbaren Umgebung leben.

In der vorliegenden Arbeit wurde die verwandte Frage untersucht, ob während der Evolution von *Escherichia coli* Änderungen in deren Lebensstil mit verstärktem lateralen Gentransfer einher gingen. Es wurden 12 Lebensstil-Merkmale für 61 *E. coli*-/*Shigella*-Stämme betrachtet. Der nächste Abschnitt liefert eine Übersicht über die verschiedenen Lebensstile von *E. coli*.

10.1.1 Pathogenität der Stämme von Escherichia coli

Neben Archäen, Eukaryoten, anderen Enterobakterien und Streptokokken gehören auch verschiedene Stämme des Bakteriums *Escherichia coli* zur Normalflora im unteren Darmtrakt des Menschen und anderer Warmblüter. Die meisten Stämme von *Escherichia coli* sind apathogen, d. h., sie lösen keine Krankheiten aus, sondern pflegen eine kommensalische bzw. mutualistische Lebensweise (van Beneden, 1876). So versorgen die im Darm des Menschen lebenden *E. coli* ihren Wirt z.B. mit den Vitaminen K und B, die für die Entwicklung und ein intaktes Immunsystem notwendig sind (Munk und Dersch, 2008). Im Gegenzug sorgt der Wirt für eine stabile Lebensumgebung, in der es *E. coli* möglich ist, pathogene Bakterien zu verdrängen, was Infektionen vorbeugt. Apathogene *E. coli* sind auch an feuchten und warmen Hautregionen anzutreffen (Petkovsek et al., 2000). Viele *E. coli* können auch in anderen Habitaten als dem Darmtrakt überleben (siehe **Abb. 10.1**).

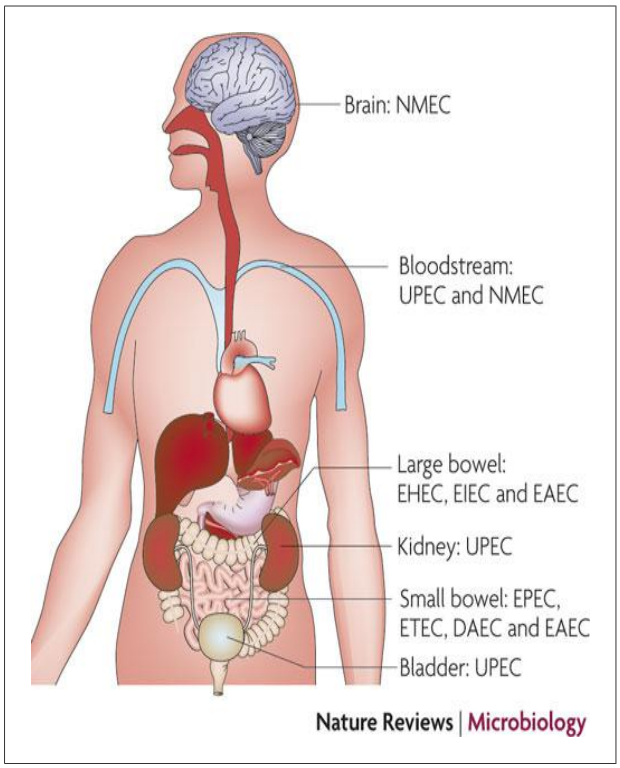


Abbildung 10.1*: Humanpathogene E. coli und ihre Gewebe (Fig. 1 aus Croxen and Finlay, 2010)*

10.1.2 Pathovare von Escherichia coli

Bestimmte Serotypen von *E. coli* sind pathogen und damit für Erkrankungen innerhalb und außerhalb des Darms verantwortlich (Kaper et al., 2004). Sie gehören nicht zur physiologischen Darmflora. Erstere werden als IPEC – intraintestinal pathogene *E. coli* -, letztere als ExPEC - extraintestinal pathogene *E. coli* - bezeichnet (Johnson und Russo, 2002).

Insbesondere darmpathogene *E. coli* (IPEC) sind für die hohe Zahl von Krankheits- und Todesfällen, nämlich ca. 160 Mio. Durchfallerkrankungen und rund 1 Mio. Todesfälle pro Jahr, verantwortlich (Hahn et al., 2008). So verursachten im Zeitraum Mai bis Juli 2011

enterohämorrhagische *E. coli* (EHEC) des Serotyps O104:H4 4000 Erkrankungsfälle mit dem hämolytisch-urämischen Syndrom (HUS) überwiegend in Norddeutschland, 53 Personen verstarben in Folge der Infektion (Appel et al., 2011). Laut Bundesinstitut für Risikobewertung (BfR) „handelte es sich um den bisher größten Krankheitsausbruch durch EHEC-Infektionen in Deutschland und bezogen auf die Anzahl der HUS-Fälle um den größten weltweit beschriebenen derartigen Ausbruch“.

Man unterscheidet sechs Pathovare bei den darmpathogenen *E. coli*: diffus adhärierende (DAEC), enteroaggregative (EAEC oder EaggEC), enteroinvasive (EIEC), enteropathogene (EPEC), enterotoxische (ETEC) und Verocytotoxin/Shigatoxin bildende *E. coli* (VTEC/STEC). Enterohämorrhagische *E.coli* (EHEC) sind eine Untergruppe der letzten Kategorie.

Obwohl in der Bevölkerung weniger wahrgenommen, stellen neben den darmpathogenen aber auch extraintestinal pathogene *E. coli* (ExPEC) eine ernsthafte Bedrohung für die Gesundheit dar (Johnson und Russo, 2002). Es sind dies überwiegend die Pathovare NMEC (Neugeborenenmeningitis-auslösende *E. coli*) und UPEC (uropathogene E. coli). **Tabelle 10.1** stellt den Pathovaren bei *E. coli* die von ihnen hervorgerufenen Krankheitsbilder gegenüber. Wie aus den Abkürzungen für die Pathovare ersichtlich ist, kolonisieren pathogene *E. coli* verschiedene Orte im menschlichen Körper (vgl. **Abb. 10.1**). Enteropathogene (EPEC), enterotoxische (ETEC) und diffus-adhärente (DAEC) *E. coli* kolonisieren den Dünndarm und verursachen Diarrhoe, während enterohämorrhagische (EHEC) und enteroinvasive (EIEC) *E. coli* für Krankheiten des Dickdarms verantwortlich sind. Enteroaggregative (EAEC) *E. coli* sind sowohl im Dünndarm als auch im Dickdarm anzutreffen und verursachen dort verschiedene Krankheiten. Uropathogene *E. coli* bevölkern die Harnwege und können in die Blase einwandern, wo sie eine Blasenentzündung verursachen. Wird die Infektion nicht behandelt, können uropathogene *E. coli* weiter in die Nieren wandern und dort eine Pyolenephritis (Nierenbeckenentzündung) verursachen. Sowohl die UPEC-Stämme als auch die Neugeborenenmeningitis auslösenden NMEC-Stämme können eine Blutvergiftung verursachen. NMEC kann die Blut-Hirn-Schranke überwinden und im zentralen Nervensystem eine Meningitis (Hirnhautentzündung) auslösen.

Pathotyp	Bedeutung	Klinik
AIEC	Adhärent-invasive E. coli	Vermutlich mit Morbus Crohn und Colitis ulcerosa assoziiert
APEC	Geflügelpathogene E. coli	Extraintestinale Krankheiten bei Geflügel
DAEC	Diffus-adhärente E. coli	Diarrhoe, extraintestinale Erkrankungen
EAEC/EAggEC	Enteroaggregative E. coli	Diarrhoe, extraintestinale Erkrankungen oder symptomlos
EHEC	Enterohämorrhagische E. coli	Diarrhoe, HUS (Hämolytisch-urämisches Syndrom)
EIEC	Enteroinvasive E. coli	Diarrhoe
EPEC	Enteropathogene E. coli	Diarrhoe, häufig bei Säuglingen
ETEC	Enterotoxische E. coli	Diarrhoe, Durchfall ähnlich wie bei Cholera
NMEC	Neonatale-Meningitis auslösende E. coli	Hirnhautentzündung bei Neugeborenen, Sepsis (falls im Blutstrom)
NTEC	Nekrotoxische E. coli	Diarrhoe,
UPEC	Uropathogene E. coli	Harnwegsinfektionen, Sepsis (falls im Blutstrom)
VTEC/STEC	Vero-/shigatoxin-bildende E. coli	Shigellose (Shigellenruhr), Diarrhoe

Tabelle 10.1*: Klinische Manifestation der E. coli-Pathovare*

Die Virulenzfaktoren und insbesondere die molekularen Pathogenitätsmechanismen der verschiedenen Pathovare werden nicht in dieser Arbeit beleuchtet. Für einen Einstieg in dieses Thema siehe z.B. die Reviews von Croxen und Finlay (Croxen und Finlay, 2010) sowie Kaper et al. (Kaper et al., 2004). Es folgt kurzer Abriß (ungekürzt nachzulesen bei Croxen und Finlay, 2010): *Escherichia coli* bedient sich verschiedener Virulenzstrategien, um Krankheiten beim Menschen zu verursachen. Weitgehend gemeinsam haben alle Pathovare außer EIEC, daß sie sich mittels

Fimbrien oder Pili an eine Wirtszelle heften. Danach unterlaufen sie die Prozesse der Wirtszelle, was sie häufig durch Sekretion von Proteinen bewerkstelligen. Durch die Manipulation von Signalwegen können *E. coli* gezielt Wirtszellen befallen und sie kolonisieren sowie Immunantworten des Wirtes umgehen. Am Ende steht der Ausbruch einer Krankheit (Bhavsar et al., 2007). Viele Virulenzfaktoren, die bei durch *E. coli* hervorgerufenen Krankheiten eine Rolle spielen, sind bekannt. Nicht für alle Pathovare bekannt und daher noch Gegenstand der Forschung sind die Interaktionen auf molekularer Ebene zwischen Wirtsproteinen und Virulenzfaktoren, die schließlich zur Ausprägung eines spezifischen Krankheitsbildes führen.

10.1.3 Mechanismen der Evolution pathogener Escherichia coli

Der Gewinn und der Verlust mobiler genetischer Elemente spielt eine entscheidende Rolle in der Modellierung der Genome pathogener Bakterien. Horizontaler Gentransfer (HGT) sorgt für eine schnelle Verteilung neuer Merkmale zwischen verschiedenen Organismen. Die Aneignung solcher Merkmale ist unter Umständen entscheidend für die Aufrechterhaltung der Fitness und sogar das Überleben eines Bakteriums, während es mit seinem Wirt ko-evolviert (Shames et al., 2009). Ausschließlich in den Chromosomen pathogener Bakterien oder in deren Plasmiden finden sich Verbände von Virulenzgenen, sog. Pathogenitätsinseln (PAIs – pathogenicity islands). Gewöhnlich werden PAIs von mobilen genetischen Elementen flankiert - Bakteriophagen, Insertionssequenzen oder Transposons – und in der Nachbarschaft von tRNA-Genen eingefügt. Viele der von *E. coli* bekannten Virulenzfaktoren befinden sich auf solchen PAIs, auf Plasmiden oder auch Prophagen, der latenten Form eines Bakteriophagen. Obwohl gewöhnlich defekt, können einige wenige Prophagen trotzdem infektiöse Partikel ausprägen (Asaldulghani et al., 2009). Durch lateralen Gentransfer erworbene Merkmale erlauben es dem empfangenden Bakterium, eine neue ökologische Nische zu bevölkern, und Selektionsdruck sorgt dann für Varianten, die diesem Selektionsdruck folgen. Die Evolution von Pathovaren erfolgt oft nicht ausschließlich entlang von Vererbungslinien. Die Virulenzfaktoren von EHEC beispielsweise wurden mehrfach unabhängig auf der *E. coli*-Phylogenie erworben (Ogura et al., 2009). Die Genome pathogener *E. coli* unterscheiden sich untereinander deutlich und sind bis zu 1 Mb größer als die Genome kommensaler *E. coli*, was auf den Erwerb oder den Verlust von Pathogenitätsinseln zurückzuführen ist. *Escherichia coli* verfügt in seinem Kerngenom[37] über ca. 2.200 Gene und in seinem Pangenom über 13.000 Gene. Obwohl die pathogenen *E. coli* für mehr als 5.000 Gene kodieren können, macht nur rund die Hälfte dieser Gene tatsächlich das Kerngenom aus, was maßgeblich für die Plastizität der Genome und deren genetische Vielfalt verantwortlich ist.

Der enteroaggregative *E. coli*-Ausbruchstamm *O104:H4* ist ein Hybrid aus dem enteroaggregativen Stamm *O104:H4* und einem EHEC-Stamm (Karch, 2006). Die genaue Entstehung dieses Stamms ist noch nicht vollständig geklärt. Eine Theorie geht davon aus, daß ein afrikanischer EAEC-Stamm O104:H4 durch lateralen Gentransfer die Eigenschaft erworben hat, Shigatoxine zu produzieren (Tribe, 2011). Eine alternative Theorie besagt, daß es einen bisher unbekannten zur Shigatoxin-Produktion fähigen *O104:H4*-Stamm gibt, von dem sowohl der pathogene EHEC *O104:H4* als auch der apathogene EAEC *O104:H4* abstammen (Karch, 2001; Mellmann et al., 2011). Nicht nur der Gewinn von Genen, sondern auch ihr Verlust hat maßgeblich zur Ausprägung der verschiedenen Stämme pathogener *E. coli* beigetragen. So haben EIEC durch lateralen Gentransfer des Virulenzplasmids pINV die Fähigkeit erworben, invasiv zu sein (Maurelli et al., 2007). Eine Deletion in der Region ihres Genoms, das die Lysincarboxylase-Gene enthält, war verantwortlich dafür, daß EIEC ihre Fitness behielten und sich an einen intrazellulären Lebensstil anpaßten.

37 Gemeint ist die Menge der universellen Genfamilien aller *E. coli*-Stämme.

10.2 Material und Methoden

Die für die in diesem Kapitel vorgestellten Analysen benötigten Dateien befinden sich innerhalb der Ordnerstruktur im Projektordner *Lebensstil*. Die Daten im Ordner *Proteinfamilien_Ecoli_ ProteinOrtho* sind dort der Übersicht halber redundant abgelegt. Sie befinden sich auch im Ordner *Proteinfamilien_ProteinOrtho*. Jede Aussage zu *Salmonella* bezieht sich auf den von der NCBI-Seite heruntergeladenen Datensatz von 25 *Salmonellen*-Genomen (Stand September 2013), der der vorliegenden Arbeit zugrunde liegt. Jede Aussage zu *Escherichia coli* bezieht sich auf den von dort zum gleichen Zeitpunkt heruntergeladenen Datensatz von 61 *E. coli*-Genomen aus **Tabelle 8.1**.

10.2.1 Lebensstil-Merkmale

Es wurden Daten zu verschiedenen Facetten (vgl. **Tab 10.3**) des Lebensstils der 61 verwendeten *E. coli*-Stämme aus verschiedenen Quellen (siehe Dokument *Quellen.odt* im Ordner *Lebensstil*) gesammelt. Dabei wurde die von Skippington und Ragan (Skippington und Ragan, 2012) verwendete Lebensstil-Klassifikation nach dem Pathotyp übernommen und um die Gruppe „Humanpathogen“ erweitert. Die Merkmale wurden den Gruppen "IPEC", "ExPEC", "EHEC", "Human pathogen", "UPEC", "ETEC", "Commensal", "NMEC", "EAEC", "EPEC", "STEC" und "AIEC" zugeordnet.

Für einige Stämme ist nicht gesichert, ob sie als Human-Pathogen einzustufen sind. Das gilt für das Geflügel-Pathogen *E. coli APEC 01* (Dho-Moulin et al. 1999) sowie für die beiden *E. coli* KO11FL-Stämme. Für letztere ist unbekannt, ob sie nicht doch für andere Warmblüter pathogen sind. In dieser Arbeit werden alle drei Stämme als Nicht-Humanpathogene klassifiziert. Der gegen viele Antibiotika resistente Stamm *E. coli SMS-3-5* (Typ SECEC – Environmental *E. coli*) nimmt eine Sonderrolle ein, da er nach heutigem Kenntnisstand weder Pathogen noch Kommensale ist (Fricke et al., 2008). Alle anderen apathogenen *E. coli*-Stämme sind gleichzeitig auch Kommensalen. Die enterohämorrhagischen *E. coli* bilden eine Gruppe innerhalb der Shigatoxin/Verotoxin produzierenden Obergruppe STEC. Läßt man die *Shigellen* außer acht, sind die Klassen „EHEC“ und „VTEC/STEC“ deckungsgleich.

10.2.2 Rekonstruktion der anzestralen Lebensstile

Die anzestralen Lebensstil-Merkmalsmuster sind im Projektordner in der Datei *AnzestraleLebensstilMerkmale/Parsimony_count_of_events_per_site_per_branch.txt* abgelegt.

Anhand der bei den einzelnen *E. coli*-Stämmen im Datensatz an- oder abwesenden Merkmale und der im Newick-Format vorliegenden gewurzelten *E. coli*-Phylogenie (Datei *gemeinsameDateien/ ec_tree.tre*) wurde mit der GLOOME Gain Loss Mapping Engine in Version VR01.266 (Cohen und Pupko, 2010) das phyletische Muster der Merkmale für alle Äste der Phylogenie berechnet.

Kategorie	Variable	Belegung
Evolutionary model	Allow root freq to differ from stationary ones	no
	Loss only model (gain rate ~ 0)	no
	Correction for un-observable data	
	Minimum number of ones (1)	1
	Minimum number of zeros (0)	1
	Stochastic Mapping	no (nicht verwendet)
Results	Parsimony	yes
	Parsimony cost of gain	1
	Parsimony count of events per site	no (nicht verwendet)
	Parsimony count of events per branch	no (nicht verwendet)
	Parsimony count of events per site per branch	yes
	Additional features	no (nicht verwendet)

***Tabelle 10.2**: Verwendete GLOOME-Parameter*

GLOOME benötigt als Eingabe die binären Merkmalsmuster in Form eines multiplen Alignments im FASTA-Format: Zeilen korrespondieren mit Spezies und Spalten mit binären Mustern. Dazu wurden die Merkmale als 1 (Merkmal vorhanden) oder 0 (Merkmal nicht vorhanden) kodiert. Verwendet wurde der von GLOOME angebotene Maximum Parsimony-Ansatz unter Verwendung der in **Tabelle 10.2** dargestellten Parameterbelegung. Die Kostenmatrix belegt Gewinn und Verlust eines Merkmals mit gleichen Kosten. Die GLOOME-Ausgabedatei mit der Rekonstruktion des anzestralen Gengehalts auf der *E. coli*-Phylogenie findet sich im Ordner *AnzestralerGengehalt/ Parsimony_count_of_events_per_site_per_branch.txt*. Desweiteren produzierte GLOOME die Datei *Tree_with_inner_nodes_notation.ph*, die im übergeordneten Ordner abgelegt ist. Sie enthält die vom Benutzer als Eingabe spezifizierte Phylogenie im Newick-Format, jedoch um innere Knoten erweitert. Jeder Knotenname in dieser Datei enspricht dabei dem Namen eines Astes in obiger Datei. Der Knoten ist dabei stets der jüngere der beiden Knoten, die über diesen Ast verbunden werden.

10.2.3 Berechnung von Proteinfamilien

Das Perl-Programm *proteinOrtho* in Version 4.26 (Lechner et al., 2011) wurde verwendet, um orthologe Gruppen für die chromosomalen und plasmidischen Proteinsequenzen der *E. coli*-Stämme zu konstruieren. Dazu berechnete *proteinOrtho* mittels einer BLAST-Suche (Altschul et al., 1990) reziproke best Hits zwischen 287.840 Proteinen mit einer Sequenzidentität von mindestens 75% und einem *e-value* von 0.01. Um die Zusammensetzung der orthologen Gruppen bei Bedarf vergleichen zu können, wurde die gleiche Parameterbelegung verwendet wie zuvor bei der Herstellung von Proteinfamilien auf Grundlage von Syntenie und Sequenzähnlichkeit (**vgl. Abschnitt 8.2.2**).

Die von *proteinOrtho* produzierte Ausgabedatei *proteinOrtho_output_EC_75.txt* enthält 19.400 Proteinfamilien und befindet sich im Ordner *gemeinsameDateien/Proteinfamilien_ Ecoli_proteinOrtho*. Aus der Ausgabedatei wurde eine Matrix von GI-Nummern abgeleitet, deren Spalten Genome und deren Zeilen Proteinfamilien bzw. orthologe Gruppen von Proteinen repräsentieren. Ein „*“ anstelle einer GI-Nummer bedeutet, daß in der Familie kein Protein aus dem betroffenen Genom bzw. der betroffenen Spezies vorhanden ist. Die Matrix wurde in die Datei *gemeinsameDateien/gi_matrix_ec_75.txt* geschrieben. Die von *proteinOrtho* errechnete Konnektivität für die Proteinfamilien wurde dabei außer acht gelassen, alle Gruppen wurden in die

weitere Analyse übernommen. Dort, wo *proteinOrtho* eine Menge duplizierter Gene in verschiedenen Chromosomen ermittelte, wurde immer das Gen aus dem jeweils passenden Chromosom in die orthologe Gruppe übernommen.

10.2.4 Anzestraler Gengehalt

Analog zur Vorgehensweise für die Lebensstil-Merkmale wurden die Proteinfamilien mit dem Perl-Programm `getBinaryCodedOGs_FASTA.pl` für *E. coli* aus der Datei *proteinOrtho_output_EC_75.txt (s.o.)* in ein multiples FASTA-Alignment binärer Muster von Genanwesenheit (1) und -abwesenheit (0) umgeschrieben. Das Ergebnis ist in Datei *gemeinsameDateien/Proteinfamilien_Ecoli_proteinOrtho/ec_binaryCodedOG.fasta* abgelegt. Mit der gleichen GLOOME-Version wie zuvor wurden die anzestralen Gengehalte auf der *E. coli*-Phylogenie (*gemeinsameDateien/ec_tree.tre*) geschätzt. Die Parameterbelegung (siehe **Tab. 10.2**) wurde beibehalten.

Zwei Besonderheiten traten während der Berechnung auf: erstens stellte GLOOME bei der Analyse des Alignments fest, daß in der zweiten Spalte nur Einsen standen und demnach jedes *E. coli* das betreffende Protein besitzt. Daher änderte GLOOME den Parameter „Minimum number of zeros (0)“ auf 0. Zweitens änderte GLOOME die kürzesten Astlängen in der Phylogenie von 10^{-8} in 10^{-7} und begründete sein Vorgehen damit, daß die Äste zu kurz seien.

Die Gain-Loss-Ereignisse pro Ast und Gen, die das von GLOOME geschätzte anzestrale phyletische Muster ausbilden, sind in der Datei *gemeinsameDateien/AnzestralerGengehalt/Parsimony_count_of_events_per_site_per_branch.txt zu finden.* Es wurde ebenfalls wieder die Datei *Tree_with_inner_nodes_notation.ph* erzeugt. Die Datei ist identisch mit der Datei gleichen Namens im Ordner *AnzestraleLebensstilMerkmale*. Das ist auch erforderlich, da die Namen der inneren und äußeren Knoten auf der Phylogenie jeweils gleich sein müssen.

10.2.5 Korrelation von Lebensstil-Änderungen und Gen-Änderungen

Als Gradmesser dafür, wieviel höher die Intensität lateralen Gentransfers auf Ästen mit Änderungen im Lebensstil im Vergleich zu der auf Ästen ohne Änderungen im Lebensstil ist, wird jeweils der Quotient $\mathbf{Q}_{\mathbf{orig}}$ berechnet, der die Anzahl von Gen-Änderungen auf Ästen mit Lebensstil-Änderungen in Beziehung setzt zu der auf Ästen ohne Lebensstil-Änderungen. Ist er in einem konkreten Fall deutlich größer als 1, z.B. 2, kann das bedeuten, daß *E. coli* immer dann ungefähr doppelt so viele Gene ausgetauscht haben, wenn sie ihren Lebensstil geändert haben, wie zu Zeiten ohne solche Änderungen. Ob dieser Schluß zulässig ist, wird durch einen nachgeschalteten randomisierten Test geprüft. Dazu werden $\mathbf{Q}_{\mathbf{i}}$ für eine hohe Zahl von Permutationen der Lebensstil-Änderungen auf den Ästen berechnet und jeweils mit $\mathbf{Q}_{\mathbf{Orig}}$ verglichen. Es wird gezählt, für wieviele Permutationen $\mathbf{Q}_{\mathbf{i}}$ mindestens so groß wie $\mathbf{Q}_{\mathbf{Orig}}$ ist. Letzterer Wert geht in den *p*-Wert für den randomisierten Test ein.

Für diesen Schritt des Projekts wurde das Perl-Programm `compare_LS_vs_HGT.pl` entwickelt. Der einzige Kommandozeilen-Parameter gibt an, welches Lebensstil-Merkmal während des Programmlaufs untersucht werden soll. Für mehrere Merkmale sind demnach mehrere Programmläufe notwendig. Das Programm vergleicht die Verteilung der Gengewinne und -verluste mit der der Änderungen im Lebensstil auf der *E. coli*-Phylogenie. Die Phylogenie ist gewurzelt und besitzt damit eine Zeitachse; der Wurzelknoten ist dabei der jüngste gemeinsame Vorfahr aller endständigen Taxa. Ist im Folgenden die Rede von einem Vorgängerast, ist damit der ältere Ast

gemeint, der mit einem anderen Ast über einen Knoten verbunden ist. Der jüngere Ast wird dementsprechend als Nachfolgeast bezeichnet. Es folgt eine Beschreibung der Arbeitsschritte:

1. Die Phylogenie wird implizit in einer Datenstruktur vorgehalten, in der ein innerer oder äußerer Knoten im Baum jeweils auf seinen Vorgänger verweist. Diese Nachfolger-Vorgänger-Relationen wurden manuell anhand der GLOOME-Ausgabedateien erstellt. Die von GLOOME verwendete Bezeichnung der inneren und äußeren Knoten (siehe beide Dateien *Tree_with_inner_nodes_notation.ph)* wurde dabei beibehalten. Die Bezeichner der Äste werden im Programm gleichzeitig noch für den jüngeren der beiden Knoten an diesen Ästen verwendet.

2. Der Inhalt der Datei *gemeinsameDateien/gi_matrix_ec_75.txt* wird gelesen. Aus ihr geht die Einordnung der verschiedenen Proteine in orthologe Gruppen/Proteinfamilien hervor. Jede Zeile repräsentiert dabei eine Proteinfamilie. Zeile 1 entspricht Position 1 und damit Proteinfamilie 1 in der Datei *AnzestralerGengehalt/Parsimony_count_of_events_per_site_per_branch.txt*. Das Gleiche gilt auch für die identisch bezeichnete Datei im Ordner *AnzestraleLebensstilMerkmale*.

3. Für jede GI-Nummer wird die Annotation des jeweiligen Proteins gespeichert. Die Zuordnung wird aus der im Vorfeld konstruierten Datei *gemeinsameDateien/geneTable.txt* gelesen.

4. Das Perl-Programm liest die Daten zu Gengewinnen und -verlusten pro Ast aus der Datei *AnzestralerGengehalt/Parsimony_count_of_events_per_site_per_branch.txt* sowie die Daten zu Lebensstil-Änderungen aus der entsprechenden Datei gleichen Namens im Ordner *AnzestraleLebensstilMerkmale* ein.

In beiden Dateien wird der numerische Bezeichner „pos“ verwendet, der auf eine Spalte des jeweiligen für GLOOME verwendeten Eingabe-Alignments Bezug nimmt. Er entspricht im ersten Fall der Nummer der Proteinfamilie, im zweiten Fall einem der zwölf Lebensstil-Merkmale. Ein Gen oder Merkmal wird auf einem Ast entweder hinzugewonnen („gain“) oder verloren („loss“).

5. Die von GLOOME geschätzten Lebensstil-Änderungen und lateralen Gentransfers je Ast werden um den Zustand „retained“ („beibehalten“) ergänzt. Bezogen auf einen bestimmten Ast ist damit gemeint, daß ein bestimmtes Lebensstil-Merkmal bzw. Gen auf ihm weder gewonnen noch verloren wurde. Entweder ist es also auf ihm vorhanden und war schon auf dem Vorgängerast vorhanden, oder es ist nicht vorhanden und war auch auf dem Vorgängerast schon nicht vorhanden. Dieser dritte Zustand neben „gain“ und „loss“ hat den Zweck, diese in den Daten bereits vorhandene Mehrinformation zu nutzen.

6. Eine Zeichenkette $\mathbf{S_{orig}}$ aus den Zeichen „1“ für „gain“, „2“ für „loss“ oder „4“ für „retained“ wird gebildet. Ein Zeichen repräsentiert den Zustand des Lebensstil-Merkmals auf einem Ast. Die Reihenfolge der Zeichen wird von den alphabetisch aufsteigend sortierten Astnamen vorgegeben.

7. Auf den Ausgangsdaten wird der Quotient $\mathbf{Q_{Orig}}$ aus

$$\frac{\text{Summe aller Gen-Änderungen auf Ästen mit Lebensstil-Änderungen}}{\text{Anzahl Äste mit Lebensstil-Änderungen}}$$

und

$$\frac{\text{Summe aller Gen-Änderungen auf Ästen ohne Lebensstil-Änderungen}}{\text{Anzahl Äste ohne Lebensstil-Änderungen}}$$

berechnet. Ein Gen-Ereignis bzw. eine Gen-Änderung ist entweder ein Gengewinn oder ein Genverlust. Ist $\mathbf{Q_{Orig}} >= 2$, gab es auf den Ästen der *E. coli*-Phylogenie, auf denen sich der Lebensstil geändert hat, mindestens doppelt so viele laterale Gentransfers wie auf Ästen, auf denen

sich der Lebensstil nicht geändert hat.

8. Es muß noch untersucht werden, ob die Verteilung von Gen-Ereignissen und Lebensstil-Änderungen lediglich durch Zufall zustande gekommen ist. Dazu wurde der wie folgt ablaufende randomisierte Test konzipiert:

Das Programm berechnet *n* zufällige Permutationen $\mathbf{S_i}$ von $\mathbf{S_{orig}}$. Für jede Permutation $\mathbf{S_i}$ wird der Quotient $\mathbf{Q_i}$ berechnet. Die Variable *z* zählt die Anzahl $\mathbf{Q_i}$, für die gilt $\mathbf{Q_i} >= \mathbf{Q_{orig}}$.

9. Berechne die Wahrscheinlichkeit $p = (z+1) / (n + 1)$. Beispiel: Gab es bei 1.000 Durchäufen 11 Mal ein $\mathbf{Q_i} >= \mathbf{Q_{orig}}$, ist die Wahrscheinlichkeit nur ca. 1,1 %, daß die Ausgangsverteilungen der Lebensstil-Änderungen und Gen-Änderungen durch Zufall entstanden sind. Legt man ein Signifikanzniveau von 5 % zugrunde, ist das Ergebnis signifikant.

Das Programm legt die errechneten $\mathbf{Q_i}$, den Zähler von $\mathbf{Q_i}$ und den Nenner von $\mathbf{Q_i}$ zeilenweise in einer Datei *oddsRatios_ID*.txt* ab. Anstelle des Wildcards wird der Name des Lebensstil-Merkmals eingesetzt. Eine zweite Datei, *testResults_ID*.txt* wird angelegt, in der $\mathbf{Q_{Orig}}$ sowie Zähler und Nenner von $\mathbf{Q_{Orig}}$ gespeichert werden, daneben die Anzahl Äste, auf denen sich das Lebensstil-Merkmal geändert hat, die Zahl der $\mathbf{Q_i} >= \mathbf{Q_{Orig}}$ und *p* für diesen Test.

Das hier dargestellte Verfahren (Ordner *Experiment1*) existiert in einer zweiten Version in Form des Programms `compare_LS_vs_HGT_normiert.pl`. Hier wird die Zahl der Gen-Ereignisse pro Ast auf die Astlänge normiert (Ordner *Experiment2*). Zwei weitere Programmversionen wurden erstellt, die nicht ein einzelnes Lebensstil-Merkmal je Ast untersuchen, sondern die Summe aller Merkmalsänderungen je Ast.

Die Version, die die Zahl der Lebensstil-Veränderungen nicht auf die Astlänge normiert, heißt `compare_LS_vs_HGT_alleMerkmale.pl` *(*Ordner *Experiment3*), die andere Version heißt `compare_LS_vs_HGT_normiert_alleMerkmale.pl` (Ordner *Experiment4*).

Die Varianten des Programms `compare_LS_vs_HGT.pl` für die einzelnen Lebensstil-Merkmale verteilen im Randomisierungsschritt die Merkmale neu auf dem Baum, indem sie eine Permutation aller Lebensstil-Merkmale berechnen und sie über die Phylogenie legen. Dahingegen berechnen die Varianten, die die Lebensstil-Merkmale in Summe betrachten, eine Permutation der Äste und weisen diesen jeweils die Merkmale, die sich zusammen auf einem Ast befinden, zu.

10.3 Ergebnisse

Mit dem in *Korrelation von Lebensstil-Änderungen und Gen-Änderungen* dargestellten Protokoll wurden vier Experimente durchgeführt. Die Dateien für Experiment *i* befinden sich jeweils in einem Ordner *Experiment* mit angehängter Zahl *i*. Die Dateien *oddsRatio_ID*.txt*, *testsResults_ID*.txt* und *testsResults.txt* wurden jeweils mit der in diesem Ordner befindlichen Variante des Programms `compare_LS_vs_HGT.pl` erzeugt. Für den randomisierten Test wurde ein Signifikanzniveau von 5 % zugrundegelegt. Es wurden jeweils 100.000 Permutationen der Lebensstil-Änderungen erzeugt.

E. coli/Shigella-Genom	NCBI-Nummer	IPEC	ExPEC	EHEC	Hum.-Path.	UPEC	ETEC	Komm./Apath.	NMEC	EAEC	EPEC	VTEC/STEC	AIEC
IAI1 uid59377	NC_011741		x					x					
55989 uid59383	NC_011748	x			x					x			
ATCC 8739 uid58783	NC_010468							x					
HS	NC_009800							x					
K-12 substr MG1655 uid57779	NC_000913							x					
K-12 substr W3110 uid161931	NC_007779							x					
UMN026 uid62981	NC_011751		x		x								
APEC 01 uid58623	NC_008563		x										
S88 uid62979	NC_011742		x		x				x				
UTI89 uid58541	NC_007946		x		x	x							
ED1a uid59379	NC_011745							x					
536 uid58531	NC_008253		x		x	x							
CFT073 uid57915	NC_004431		x		x	x							
O127:H6 E2348 69 uid59343	NC_011601	x			x						x		
IAI39 uid59381	NC_011750		x		x								
SMS 3 5 uid58919	NC_010498		x					x					
042 uid161985	NC_017626	x			x					x			
ABU 83972_uid161975	NC_017631	x											
BL21 DE3 uid161947	NC_012971							x					
BL21 DE3 uid161949	NC_012892							x					
BW2952 uid59391	NC_012759							x					
B REL606 uid58803	NC_012967							x					
DH1 uid161951	NC_017625							x					
DH1 uid162051	NC_017638							x					
H10407 uid161993	NC_017633	x			x		x						
IHE3034 uid162007	NC_017628		x		x				x				
KO11FL uid162099	NC_017660							x					
KO11FL uid52593	NC_016902							x					
K-12 substr DH10B uid58979	NC_010473							x					
LF82 uid161965	NC_011993	x			x						x		x
NA114 uid162139	NC_017644		x		x	x							
O103:H2 uid41013	NC_013353	x		x	x							x	
O111:H 11128 uid41023	NC_013364	x		x	x							x	
O157:H7 EC4115 uid59091	NC_011353	x		x	x							x	
O157:H7 TW14359 uid59235	NC_013008	x		x	x							x	
O26:H11 uid41021	NC_013361	x		x	x							x	
O55:H7 CB9615 uid146655	NC_013941	x			x						x		
O55:H7 RM12579 uid162153	NC_017656	x			x						x		
O7:K1 CE10 uid162115	NC_017646		x		x				x				
O83:H1 NRG 857C uid161987	NC_017634	x			x						x		x
P12b (O15:H17) uid162061	NC_017663	x			x			x			x	x	
SE11 uid59425	NC_011415							x					
SE15 uid161939	NC_013654							x					
UM146 uid162043	NC_017632	x			x						x		x
UMNK88 uid161991	NC_017641	x			x		x						
W uid162011	NC_017635		x					x					
O157:H7 Xuzhou21 uid163995	NC_017906	x		x	x							x	
BL21 Gold DE3 pLysS AG uid59245	NC_012947							x					
D i14 uid162049	NC_017652		x		x	x							
D i2 uid162047	NC_017651		x		x	x							
O157:H7 EDL933 uid57831	NC_002655	x		x	x							x	
W uid162101	NC_017664		x					x					
Sh. boydii Sb227 uid58215	NC_007613	x			x							x	
Sh. boydii CDC3083 94 uid58415	NC_010658	x			x							x	
Sh. flexneri 2a 301 uid62907	NC_004741	x			x							x	
Sh. flexneri 2a 2457T uid57991	NC_004337	x			x							x	
Sh. flexneri 5 8401 uid58583	NC_008258	x			x							x	
Sh. dysenteriae Sd197 uid58213	NC_007606	x			x							x	
Sh. flexneri 2002017 uid159233	NC_017328	x			x							x	
Sh. Sonnei 53G uid84383	NC_016822	x			x							x	
Sh. sonnei Ss046 uid58217	NC_007384	x			x							x	

***Tabelle 10.3**: E. coli-Stämme und ihre Pathotypen*

10.3.1 Experiment 1 (siehe Tabelle 10.4)

- Lebensstil-Merkmale wurden getrennt betrachtet
- Die Summe der Gengewinne und -verluste je Ast wurde nicht auf die Astlänge normiert

Der für das Lebensstil-Merkmal ETEC errechnete Wert 1,783 für $\mathbf{Q_{Orig}}$ ist der höchste **Q**-Wert in diesem Experiment. Diesem Wert liegen allerdings lediglich Änderungen im Lebensstil auf nur zwei von 120 Ästen zugrunde. Der *p*-Wert 0,12, obwohl der niedrigste in diesem Experiment, ist daher nur beschränkt aussagekräftig und wie alle anderen *p*-Werte nichtsignifikant. Für die Merkmale IPEC, ExPEC, Human-Pathogen und Kommensale wurden Lebensstil-Änderungen auf genügend Ästen beobachtet. Für jeden der drei Merkmale wurde ein $\mathbf{Q_{Orig}}$ um 1,00 errechnet: 0,87, 0,94, 1,02 und 1,38. Die Anzahl von Gen-Ereignissen in den Ausgangsdaten entspricht der durchschnittlichen Zahl von Gen-Ereignissen in den randomisierten Daten, sowohl für die Äste mit Lebensstil-Änderungen als auch für Äste ohne Lebensstil-Änderung. Die Zahl der Gen-Ereignisse war pro Ast in allen Fällen ähnlich hoch (zwischen 322 und 373) außer für das Merkmal Kommensale (492). Die *p*-Werte 0,66, 0,54, 0,44 und 0,13 weisen die Ergebnisse als nichtsignifikant aus. Streng genommen müssen die *p*-Werte noch FDR-korrigiert werden. Dies wurde unterlassen, da alle Tests nichtsignifikant sind.

Lebensstil-Merkmal	Q_{Orig}	Anzahl Äste mit Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Orig.)	Mittlere Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Random.)	Anzahl Äste ohne Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Orig.)	Mittlere Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Random.)	Anzahl $Q_i >= Q_{orig}$	p-Wert
IPEC	0,869	12	3865	4352	108	40034	39547	66024	0,66
ExPEC	0,945	10	3472	3630	110	40427	40269	54407	0,54
EHEC	0,753	2	553	724	118	43346	43175	57495	0,57
Human-Pathogen	1,024	11	4111	3987	109	39788	39912	43943	0,44
UPEC	1,017	5	1860	1815	115	42039	42084	43843	0,44
ETEC	1,783	2	1288	726	118	42611	43173	11761	0,12
Kommensale	1,385	9	4431	3263	111	39468	40636	12769	0,13
NMEC	1,443	3	1566	1088	117	42333	42811	19180	0,19
EAEC	1,259	2	917	726	118	42982	43173	30569	0,31
EPEC	0,918	5	1685	1817	115	42214	42082	53420	0,53
VTEC/STEC	0,811	6	1797	2180	114	42102	41719	66152	0,66
AIEC	1,215	2	886	723	118	43013	43176	31859	0,32

Tabelle 10.4*: Einzelne Lebensstil-Merkmale, nicht normiert, 100.000 Permutationen*

10.3.2 Experiment 2 (siehe Tabelle 10.5)

- Lebensstil-Merkmale wurden getrennt betrachtet
- Die Summe der Gengewinne und -verluste je Ast wurde auf die Astlänge normiert

Da die Astlängen in der verwendeten *E. coli*-Phylogenie sehr stark schwanken, ist die vorgenommene Normierung der Gen-Ereignisse auf die Astlängen sinnvoll. Durch die Normierung haben sich die $\mathbf{Q_{Orig}}$ deutlich verändert. Wie zuvor wurde das höchste $\mathbf{Q_{Orig}}$, hier 6,900, für das Merkmal *ETEC* berechnet. Der randomisierte Test für dieses Merkmal war mit $p = 0{,}04$ signifikant, basiert jedoch nur auf Lebensstil-Änderungen auf 2 von 120 Ästen. Alle anderen Ergebnisse sind nichtsignifikant. Insbesondere das Ergebnis für das Merkmal *Kommensale*, $\mathbf{Q_{Orig}} = 0{,}081$, $p = 0{,}9$, fällt völlig anders aus als in Experiment 1. Die *p*-Werte für dieses Experiment müßten ebenfalls FDR-korrigiert werden. Das würde den Test für *ETEC* voraussichtlich ebenfalls nichtsignifikant machen.

Lebensstil-Merkmal	Q_{Orig}	Anzahl Äste mit Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Orig.)	Mittlere Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Random.)	Anzahl Äste ohne Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Random.)	Mittlere Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Random.)	Anzahl $Q_i >= Q_{orig}$	p-Wert
IPEC	1,690	12	3865	4356	108	40034	39543	19903	0,20
ExPEC	1,550	10	3624	3624	110	40427	40275	25202	0,25
EHEC	0,002	2	553	727	118	43346	43172	73882	0,74
Human-Pathogen	0,026	11	4111	3990	109	39788	39909	96020	0,96
UPEC	0,005	5	1860	1813	115	42039	42086	85138	0,85
ETEC	6,900	2	1288	725	118	42611	43174	4216	0,04
Kommensale	0,081	9	4431	3264	111	39468	40635	90001	0,90
NMEC	0,027	3	1566	1086	117	42333	42813	59990	0,60
EAEC	4,034	2	917	725	118	42982	43174	10857	0,11
EPEC	2,440	5	1685	1814	115	42214	42085	16808	0,17
VTEC/STEC	0,080	6	1797	2176	114	42102	41723	80492	0,80
AIEC	5,958	2	886	727	118	43013	43172	6901	0,07

***Tabelle 10.5**: Einzelne Lebensstil-Merkmale, normiert, 100.000 Permutationen*

10.3.3 Experiment 3 (siehe Tabelle 10.6)

- Lebensstil-Merkmale wurden kumuliert betrachtet
- Die Summe der Gengewinne und -verluste je Ast wurde nicht auf die Astlänge normiert

Gegenüber den Experimenten 1 und 2 wurden zwei Änderungen vorgenommen: zum einen werden in den Experimenten 3 und 4 alle 12 Lebensstil-Merkmale zusammen betrachtet, indem alle Veränderungen der Lebensstil-Merkmale auf jedem Ast aufsummiert werden. Dabei spielt die „Richtung“, d. h. Gewinn oder Verlust, keine Rolle. Zum anderen wurden die Lebensstil-Merkmale für den randomisierten Test nach einer anderen Strategie neu verteilt. Es wurde zunächst eine Permutation der Äste erzeugt, danach wurden die Lebensstil-Merkmale auf dem ehemaligen Ast dem neuen zugewiesen. Das bedeutet, daß die Merkmale nicht mehr individuell von einem beliebigen Ast auf einen anderen verteilt werden, sondern nur zusammen auf einen anderen Ast. Das Randomisierungsverfahren lieferte kein $\mathbf{Q_i}$, das mindestens so groß ist wie $\mathbf{Q_{orig}}$. Der p -Wert für den Test ist damit kleiner oder gleich (0+1)/(100.000+1) ≈ 0,00001. Der Test ist demnach signifikant.

10.3.4 Experiment 4 (siehe Tabelle 10.6)

- Lebensstil-Merkmale wurden kumuliert betrachtet
- Die Summe der Gengewinne und -verluste je Ast wurde auf die Astlänge normiert

Für dieses Experiment wurden die Veränderungen aller 12 Lebensstil-Merkmale pro Ast aufsummiert, und deren Summe wurde auf die Astlänge normiert. Während des Randomisierungsschritts wurden die Lebensstil-Merkmale auf die gleiche Weise wie in Experiment 3 auf die Äste verteilt. Für 1.845 von 100.000 Permutationen wurde ein $\mathbf{Q_i}$ errechnet, das mindestens genauso groß wie $\mathbf{Q_{orig}}$ = 0,180 ist. Der p-Wert 0,0180 weist den Test als signifikant aus.

Die Ergebnisse der Experimente 3 und 4 (vgl. **Tab. 10.6**) erlauben den Schluß, daß *Escherichia coli* während ihrer Evolution immer dann mehr lateralen Gentransfer betrieben haben, wenn sich ihr Lebensstil geändert hat.

Q_{Orig}	Anzahl Äste mit Lebensstil-Änderungen (Orig.)	Anzahl Äste ohne Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Orig.)	Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Orig.)	Durchschn. Anzahl HGT auf Ästen mit Lebensstil-Änderungen (Random.)	Durchschn. Anzahl HGT auf Ästen ohne Lebensstil-Änderungen (Random.)	Anzahl $Q_i >= Q_{orig}$	p-Wert	Auf Astlänge normiert
2,549	41	79	25002	18897	15013	28886	0	0,00001	Nein (Exp. 3)
2,765					15002	28897	1845	0,01800	Ja (Exp. 4)

Tabelle 10.6*: Lebensstil-Merkmale kumuliert, 100.000 Permutationen*

10.3.5 Zusammenhang zwischen der Zahl der Gen-Änderungen und Astlängen

Das Diagramm aus **Abbildung 10.2** stellt den Astlängen die Zahl der Gen-Ereignisse auf den jeweiligen Ästen gegenüber. Grundlage für das Diagramm sind die Daten in der Datei *numHGT_vs_branchLengths.txt* im Unterordner *Rohdaten/lifestyleProjekt.* Der Spearman-Korrelationskoeffizient für die Größen „Astlängen“ und „Zahl der Gen-Ereignisse“ ist 0,254, was auf eine schwache Korrelation hinweist.

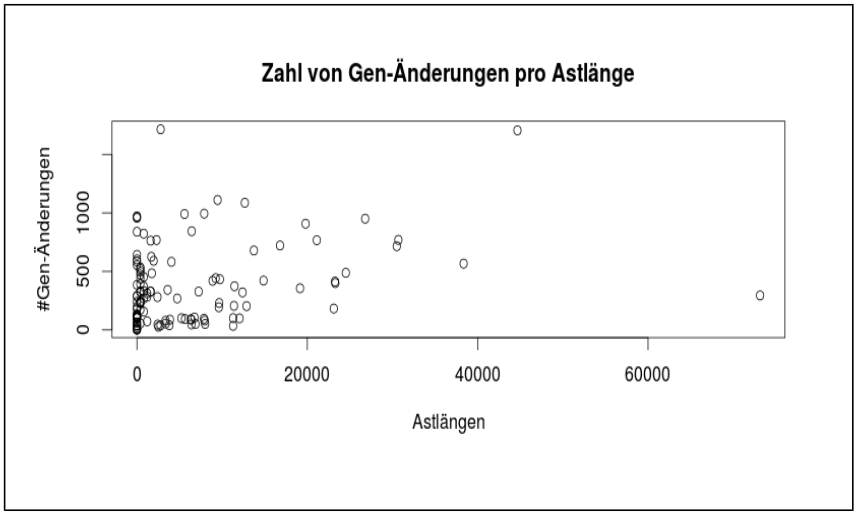


Abbildung 10.2*: Anzahl HGT-Ereignisse und Astlängen*

10.3.6 Zahl der HGT-Ereignisse auf Ästen mit und ohne Lebensstil-Änderungen

Für die in **Abschnitt 10.3.5** verwendeten Daten wurde die Zahl der Gen-Änderungen auf Ästen mit Lebensstil-Änderungen denen auf Ästen ohne Lebensstil-Änderungen gegenübergestellt. Das entsprechende, in R erstellte Diagramm ist in **Abbildung 10.3** dargestellt. Mit hoher Wahrscheinlichkeit sind die Gen-Ereignisse weder auf Ästen mit Lebensstil-Änderungen (*p*-Wert 0,012) noch auf Ästen ohne Lebensstil-Änderungen (*p*-Wert $5{,}565 \cdot 10^{-7}$) normalverteilt, wie die entsprechenden, in R durchgeführten Shapiro-Tests ergeben haben. Die Verteilungen wurden daher mit dem Kolmogorov-Smirnov-Test auf Übereinstimmung getestet. Aus dem *p*-Wert $9{,}04 \cdot 10^{-7}$ für den Test geht hervor, daß die verglichenen Stichproben mit sehr hoher Wahrscheinlichkeit nicht aus der gleichen Grundgesamtheit stammen. Die Verteilung der Gen-Änderungen auf Ästen mit

Lebensstil-Änderungen stimmt demnach nicht mit der der Gen-Änderungen auf Ästen ohne Lebensstil-Änderungen überein. Dieses Ergebnis unterstützt die eingangs aufgestellte Hypothese in der Hinsicht, als daß diese besagt, daß lateraler Gentransfer in der Evolution von *E. coli* nicht mit gleichmäßiger Rate geschehen ist.

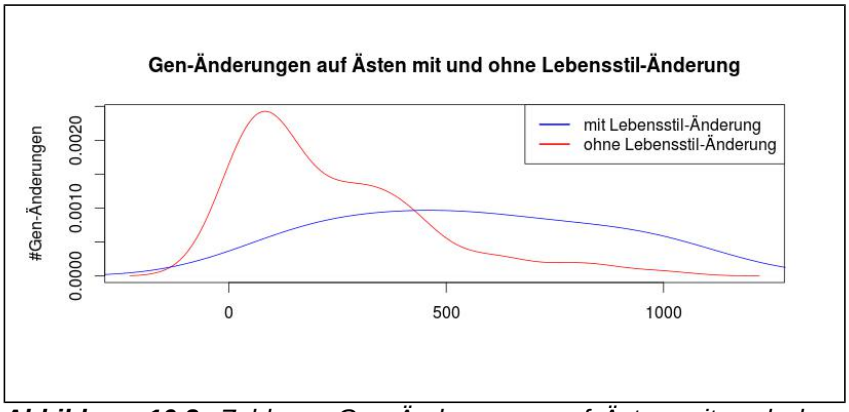


Abbildung 10.3: *Zahl von Gen-Änderungen auf Ästen mit und ohne Lebensstil-Änderung*

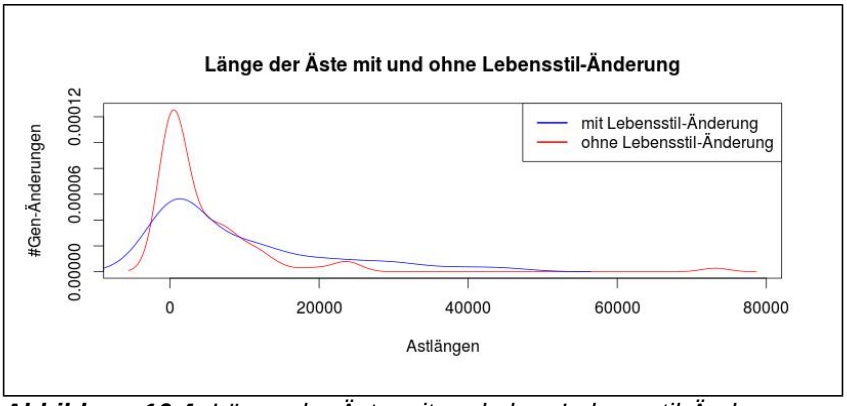


Abbildung 10.4: *Länge der Äste mit und ohne Lebensstil-Änderung*

Die Verteilungen für die Längen der Äste mit und ohne Lebensstil-Änderungen (siehe **Abb. 10.4**) haben einen ähnlichen Mittelwert nahe Null. Die Ergebnisse aus diesem Abschnitt lassen vermuten, daß die Zahl von Gen-Änderungen auf der E. coli-Phylogenie unabhängig von den Astlängen ist.

10.4 Diskussion

In diesem Projekt wurde die Hypothese untersucht, daß *Escherichia coli/Shigella* im Laufe der Evolution immer dann verstärkt horizontalen Gentransfer betrieben haben, wenn sie ihren Lebensstil geändert haben. Diese Idee erscheint plausibel: zum einen geht die Forschung davon aus, daß lateraler Gentransfer in der Prokaryotenevolution ein sehr häufiges, praktisch gewöhnliches Phänomen ist, zum anderen gibt es Hinweise darauf, daß Bakterien sich schnell an neue Umgebungen anpassen, indem sie Fähigkeiten durch den Gewinn oder Verlust von Genen

hinzugewinnen bzw. modifizieren. Es wurde ein Verfahren vorgestellt, mit dem untersucht wurde, ob Änderungen in zwölf unterschiedlichen Lebensstilen bei *E. coli* tatsächlich mit verstärktem lateralen Gentransfer einhergingen und ob dieses Ergebnis vertrauenswürdig ist.

Letztendlich konnte die eingangs aufgestellte Hypothese gestützt werden. Das gelang, als in den Experimenten 3 und 4 alle Änderungen in den Lebensstil-Merkmalen kumuliert betrachtet wurden. Diese Praxis ist statistisch zulässig, da lediglich zwischen den zwei Klassen „keine Änderung im Lebensstil“ und „mindestens eine Änderung im Lebensstil“ unterschieden werden muß. Insbesondere spielt es dabei keine Rolle, ob die Lebensstil-Merkmale teilweise voneinander abhängig sind. Die statistischen Untersuchungen aus den **Abschnitten 10.3.5** und **10.3.6** unterstützen ebenfalls die Hypothese. So scheint die Zahl der Gen-Änderungen auf Ästen der Phylogenie nicht von den Astlängen abzuhängen. Die Verteilung der Gen-Änderungen auf Ästen mit Lebenssstil-Änderung weicht deutlich von der auf Ästen ohne Lebensstil-Änderung ab, und die Verteilungen der Längen von Ästen mit und ohne Lebensstil-Änderungen haben einen ähnlichen Mittelwert.

Wieso waren keine Korrelationen zwischen den einzelnen Lebensstil-Merkmalen und der Zahl von Gen-Veränderungen auf der Phylogenie festzustellen? Für viele Lebensstil-Merkmale war die Grundvoraussetzung einer erhöhten Zahl von Gen-Änderungen bei gleichzeitiger Änderung im Lebensstil auf der *E. coli*-Phylogenie gar nicht erst erfüllt. Weiterhin war eine verläßliche Aussage für bestimmte Lebensstil-Merkmale oft von vornherein unmöglich, da Änderungen auf nur wenigen Ästen stattgefunden hatten.

Mit welchen Maßnahmen könnte das Verfahren verbessert werden?

- *Gain- und Loss*-Ereignisse wurden nicht differenziert betrachtet. Sie sollten auch getrennt ausgewertet werden. Die These, daß ein bestimmtes Gen auf der Phylogenie gewonnen wurde, immer wenn ein Vorfahr von *E. coli* pathogen geworden ist, kann mit dem jetzigen Ansatz beispielsweise nicht erfaßt werden.
- Die Ergebnisse hängen sehr stark von dem Verfahren ab, mit dem die anzestralen Gengehalte und Lebensstil-Merkmale rekonstruiert wurden. In diesem Projekt wurde das von GLOOME angebotene Verfahren benutzt, das auf Maximum Parsimony (MP) basiert. MP verteilt Änderungen auf der Phylogenie nach dem Gedanken maximaler Sparsamkeit. Das führt dazu, daß man auf einem Ast niemals mehr als eine Änderung pro Merkmal antrifft. Diesen Sachverhalt kann man so interpretieren, daß die Astlänge für die Verteilung der Merkmale wenigstens in der Praxis keine Rolle spielt. Obwohl hier MP in einem seiner typischen Anwendungsbereiche benutzt worden ist, weist es bestimmte Schwächen auf (Swofford et al., 2001). Es erscheint lohnenswert, ebenfalls das von GLOOME offerierte „Stochastic Mapping“-Verfahren auszuprobieren, das auf Maximum-Likelihood basiert. In vom Verfasser durchgeführten Versuchen kam es allerdings vor, daß das Verfahren *Gain*- und *Loss*-Ereignisse für das gleiche Merkmal mehrfach auf einem Ast rekonstruierte. Gegen die Benutzung eines ML-Ansatzes im Rahmen der hier dargestellten Analysen spricht auch, daß ML ein explizites Modell für die Verteilung der Ereignisse auf der Phylogenie benutzt, das von der Astlänge abhängig ist. Die eingangs aufgestellte These erfordert es jedoch, daß verstärkter lateraler Gentransfer bei *E. coli* ausschließlich von Änderungen im Lebensstil abhängig ist und ausdrücklich nicht von der Astlänge.
- In den Experimenten 10.2 und 10.4 wurde die Summe der HGT-Ereignisse jeweils auf die Astlänge normiert. Da es Hinweise darauf gibt, daß die Zahl der HGT-Ereignisse nur sehr schwach von der Astlänge abhängig ist (siehe **Abschnitt 10.3.5**), bringt dieses Vorgehen wahrscheinlich keinen Erkenntnisgewinn.
- Man könnte im Randomisierungsschritt die Lebensstil-Änderungen mit einer zur Astlänge

proportionalen Wahrscheinlichkeit neu auf die Äste verteilen. Auf kurzen Ästen fänden dann weniger Lebensstil-Änderungen statt als auf langen, was das Ergebnis des Tests im Vergleich zur Ausgangssituation wahrscheinlich deutlich verändern würde. Diese Handlungsweise fußt aber auf der Annahme, daß in längeren evolutionären Zeiträumen bei *E. coli* auch mehr Änderungen im Lebensstil erfolgt sind. Diese neuerliche Normierung auf die Astlängen bringt möglicherweise keinen zusätzlichen Nutzen, da die Zahl der Genveränderungen bereits auf die Astlänge normiert worden ist. Die Normierung wurde vorgenommen, da es möglich ist, daß für Änderungen im Lebensstil und Veränderungen im Gengehalt eine scheinbare Korrelation beobachtet wird, diese Größen jedoch tatsächlich nur jede für sich mit der Astlänge korrelieren.

11 Ausblick

Es war ein Teilziel dieser Arbeit, robuste Referenzphylogenien für möglichst viele vollständig sequenzierte *Escherichia coli/Shigellen*- und *Salmonellen*-Stämme zu inferieren. Das ist gelungen. Für 25 *Salmonellen* wurde eine Phylogenie berechnet, deren Äste durchweg sehr gut statistisch unterstützt werden. Für 61 von 63 vollständig sequenzierten *E. coli* wurde erstmalig eine Phylogenie erzeugt, deren Äste bis auf eine Ausnahme ebenfalls durchweg hohe Konfidenzwerte aufweisen. Zusätzlich spiegelt sie die für *E. coli* definierten phylogenetischen Gruppen wider.

Die über Maximum-Likelihood und ein Bayesianisches Modell inferierten *Salmonellen*-Phylogenien sind deckungsgleich, die entsprechenden *E. coli*-Phylogenien unterscheiden sich nur durch eine Multifurkation in einem Teilbaum. Phylogenien, die mit anderen Methoden erzeugt wurden, weisen einen großen symmetrischen Abstand zu den Referenzphylogenien auf. Der Autor hat das Abstandsmaß *ssRF* für Bäume implementiert, das eine Fortentwicklung des symmetrischen Abstands ist. Es attestiert subjektiv als ähnlich wahrgenommenen Bäumen tatsächlich einen kurzen Abstand, indem es nur signifikante Äste berücksichtigt. Selbst bei Vergleichen mit diesem Abstandsmaß bleiben die Unterschiede zwischen den Phylogenien teilweise beträchtlich. Bei den angesprochenen Methoden handelt es sich einerseits um einen Ansatz, der hochkonservierte Genfamilien benutzt („iTol“), und andererseits um einen Ansatz („Idee von Sergei Maslov“), der den Prozentsatz rekombinanter Regionen zwischen zwei Genomen als Abstandsmaß benutzt. Mit ihnen war die Hoffnung verbunden, daß sie noch besser aufgelöste bzw. robustere Bäume liefern würden, waren sie doch so ausgelegt worden, daß sie unempfindlich gegenüber lateralem Gentransfer sind. Diese Hoffnung hat sich nicht erfüllt. Die „iTol“-Methode hat für die sehr nah verwandten *E. coli*- und *Salmonellen*-Stämme statistisch schlecht unterstützte Bäume geliefert. Zwar wurde dazu das von den Urhebern des „iTol“-Baumes veröffentlichte Verfahren nicht völlig nachgebaut. Der Verfasser ist allerdings der Ansicht, daß dies das Ergebnis nicht beeinflußt hat. Die von Sergei Maslov vorgeschlagene Methode hat in diesem Sinne nur etwas bessere Bäume produziert. Es bleibt abzuwarten, ob sich der Bootstrap-Support der „Maslov-Bäume“ insgesamt verbessert, wenn – wie in der Diskussion zu **Kap. 8** vorgeschlagen – die rekombinanten Regionen zwischen unterschiedlichen Genfamilien nicht mehr in die Abstandsberechnung einfließen.

Hauptgegenstand dieser Arbeit waren zwei In-silico-Analysen, die die inferierten Phylogenien benutzt haben, um Zusammenhänge zwischen anzestralem Gengehalt und Lebensstil-Änderungen bzw. der Fähigkeit, essentielle, in der Umwelt vorhandene Nährstoffe zu metabolisieren, zu untersuchen. Im Rahmen der ersten Analyse gelang es, die Ausgangshypothese zu unterstützen, nach der Änderungen im Lebensstil von *E. coli* mit verstärktem lateralen Gentransfer einhergingen. Dieses Ergebnis deutet die Möglichkeit an, daß Prokaryoten Fähigkeiten, die für das Überleben in einer neuen Lebensumgebung notwendig sind, durch die Aufnahme geeigneter Gene erwerben.

Im zweiten Experiment wurden mathematisch signifikante Paarungen von Genen und Nährstoffen identifiziert, von denen eine größere Zahl dem Autor tatsächlich biologisch sinnvoll erscheint, wenngleich sie nur einen geringen Prozentsatz aller betrachteten Paarungen ausmachen. Als sehr positiv zu bewerten ist, daß dabei signifikante Paarungen aus Transporter und Nährstoff ermittelt wurden, die nicht explizit in den Eingangsdaten vorhanden waren. Die entwickelte Methode vermag es also allem Anschein nach, neue Ergebnisse zu liefern. Der Verfasser hält die Resultate dieses Experiments insgesamt für vielversprechend, stellen sie doch in Aussicht, daß *E. coli* und *Salmonellen* im Laufe ihrer Evolution Gene nicht mit gleichbleibender Rate horizontal aufgenommen bzw. abgegeben haben, sondern dann, wenn die jeweiligen Genprodukte tatsächlich benötigt bzw. nicht mehr benötigt worden sind.

Der Verfasser geht ferner davon aus, daß die von ihm rekonstruierten Referenzphylogenien die

wahren Phylogenien von *E. coli* und *Salmonellen* hinreichend gut annähern. Beide Analysen sollten unter Verwendung der mit den anderen Methoden rekonstruierten Phylogenien wiederholt werden. So kann ausgeschlossen werden, daß etwaige Trends in den Ergebnissen, die robust gegenüber Perturbationen in der Phylogenie sind, unentdeckt bleiben.

Leider haftet allen Rekonstruktionen von Stammesgeschichten, gleichgültig, ob als Baum oder als Netz dargestellt, gleichermaßen der Makel fehlender Überprüfbarkeit an. Mit der stetig anwachsenden Zahl von Computersimulationen, die auf solchen Rekonstruktionen aufbauen, steigt jedoch auch das Vertrauen in die Rekonstruktionen, wann immer es gelingt, die Ergebnisse in-vivo zu verifizieren. Würden zukünftige Studien ergeben, daß die Gestalt metabolischer Netzwerke von Prokaryoten maßgeblich durch lateralen Gentransfer geformt wird, und wird man schließlich die molekularen Mechanismen kennen, welche die Vorfahren von *E. coli* und *Salmonellen* zu Pathogenen gemacht haben, wäre dies ein entscheidender Schritt für die medizinische und molekularbiologische Forschung.

Literaturverzeichnis

Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19.6 (1974): 716-23.

Altschul, S. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215.3 (1990): 403-10.

Altschul, S. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research* 25.17 (1997): 3389-402.

Appel, B. et al. EHEC-Ausbruch 2011. Bundesamt Für Risikobewertung (2011). http://www.bfr.bund.de/de/ehec_ausbruch_2011-128212.html. Letzter Zugriff: 6/2014.

Asadulghani, M., Y. Ogura, T. Ooka, T. Itoh, A. Sawaguchi, A. Iguchi, K. Nakayama und T. Hayashi. The Defective Prophage Pool of Escherichia Coli O157: Prophage–Prophage Interactions Potentiate Horizontal Transfer of Virulence Determinants. Ed. Howard Ochman. *PLoS Pathogens* 5.5 (2009): E1000408.

Asano, T., J. Jansson, K. Sadakane, R. Uehara und G. Valiente. Faster Computation of the Robinson–Foulds Distance between Phylogenetic Networks. *Information Sciences* 197 (2012): 77-90.

Atkinson, H. J., J. H. Morris, T. E. Ferrin und P. C. Babbitt. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. Ed. I. King Jordan. *PLoS ONE* 4.2 (2009): E4345.

Bhavsar, A. P., J. A. Guttman und B. B. Finlay. Manipulation of Host-cell Pathways by Bacterial Pathogens. *Nature* 449.7164 (2007): 827-34.

Van Beneden, P. J. Die Schmarotzer Des Thierreichs Mit 83 Abbildungen in Holzschnitt. Leipzig: F. A. Brockhaus, 1876.

Benjamini, Y. und Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57.1 (1995):289–300.

Berry, V. und O. Gascuel. On the Interpretation of Bootstrap Trees: Appropriate Threshold of Clade Selection and Induced Gain. *Molecular Biology and Evolution* 13.7 (1996): 999-1011.

Billera, Louis J., S. P. Holmes und K. Vogtmann. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics* 27.4 (2001): 733-67.

Blattner, F. R. The Complete Genome Sequence of Escherichia Coli K-12. *Science* 277.5331 (1997): 1453-462.

Bogdanovicz, D., K. Giaro, B. Wrobel. TreeCmp: Comparison of Trees in Polynomial Time. *Evolutionary Bioinformatics* 8 (2012):475-487.

Borenstein, E., M. Kupiec, M. W. Feldman und E. Ruppin. Large-scale Reconstruction and Phylogenetic Analysis of Metabolic Environments. *Proceedings of the National Academy of Sciences* 105.38 (2008): 14482-4487.

Bouckaert, R, J. Heled, D. Kühnert, T. G. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. und A. J. Drummond. BEAST2: A software platform for Bayesian evolutionary analysis. PLOS Computational Biology 10.4 (2014):e1003537.

Braasch, B. Horizontaler Gentransfer in Metabolischen Netzwerken von E. coli Stämmen. Univ. Masterarbeit, Heinrich-Heine-Universität Düsseldorf, 2012.

Breton, C. Structures and Mechanisms of Glycosyltransferases. *Glycobiology* 16.2 (2005): 29R-37R.

Burger, W. und M. J. Burge. Digitale Bildverarbeitung. Springer Verlag, Heidelberg, Berlin; 2005.

Cardona, G., M. Llabres, F. Rossello und G. Valiente. Metrics for Phylogenetic Networks I: Generalizations of the Robinson-Foulds Metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6.1 (2009): 46-61.

Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* 17.4 (2000): 540-52.

Caugant, D. A., B. R. Levin, I. Orskov, F. Orskov, C. Svanborg Eden und R. K. Selander. Genetic diversity in relation to serotype in Escherichia coli. Infect. Immun. 49 (1985):407-413.

Chaudhuri, R. R. und I. R. Henderson. The Evolution of the Escherichia Coli Phylogeny. *Infection, Genetics and Evolution* 12.2 (2012): 214-26.

Clermont, O., S. Bonacorsi und E. Bingen. Rapid and simple determination of the Escherichia coliphylogenetic group. *Appl. Environ. Microbiol.* 66 (2000):4555-4558.

Cohen, O. und T. Pupko. Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* 27.3 (2010): 703-13.

Croxen, M. A. und B. B. Finlay. Molecular Mechanisms of Escherichia Coli Pathogenicity. *Nature Reviews Microbiology* (2009): 1811–1817.

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

Dagan, T., Y. Artzy-Randrup und W. Martin. Modular Networks and Cumulative Impact of Lateral Transfer in Prokaryote Genome Evolution. *Proceedings of the National Academy of Sciences* 105.29 (2008): 10039-0044.

Darriba, D., G. L. Taboada, R. Doallo und D. Posada. JModelTest 2: More Models, New Heuristics and Parallel Computing. *Nature Methods* 9.8 (2012).

Dellen, R. Rekonstruktion und vergleichende Analyse des Nährstoffbedarfs von Bakterien aus Genomdaten. Univ. Masterarbeit, Heinrich-Heine-Universität Düsseldorf, 2010.

Desjardins, P., B. Picard, B. Kaltenböck, J. Elion und E. Denamurl. Sex in Escherichia Coli Does Not Disrupt the Clonal Structure of the Population: Evidence from Random Amplified Polymorphic DNA and Restriction-fragment-length Polymorphism. *Journal of Molecular Evolution* 41.4 (1995): 440-48.

Dho-Moulin, M. und J. M. Fairbrother. Avian pathogenic Escherichia coli (APEC). *Vet Res.* 30.2-3 (1999):299-316.

Dixit, P. D., T. Y. Pang, F. W. Studier, S. Maslov. Quantifying vertical and horizontal evolutionary dynamics of the basic genome of E. coli. (2013, zur Veröffentlichung eingereichtes Manuskript).

Doolittle, W. F. und E. Bapteste. Inaugural Article: Pattern Pluralism and the Tree of Life Hypothesis. *Proceedings of the National Academy of Sciences* 104.7 (2007): 2043-049.

Drummond, A. J. und A. Rambaut. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolutionary Biology* 7.1 (2007): 214

Edwards, J. S., R. U. Ibarra und B. Ø. Palsson. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology* 19.2 (2001):125-30.

Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7.1 (1979): 1-26.

Efron, B. Bootstrap Confidence Levels for Phylogenetic Trees. *Proceedings of the National Academy of Sciences* 93.14 (1996): 7085-090.

Eizirik, E. Molecular Dating and Biogeography of the Early Placental Mammal Radiation. *Journal of Heredity* 92.2 (2001): 212-19.

Elena, S. F., T. S. Whittam, C. L. Winkworth, M. A. Riley und R. E. Lenski. Genomic divergence of Escherichia coli strains: evidence for horizontal transfer and variation in mutation rates. *Int Microbiol* 8.4 (2005):271-278.

Elzanowski, A. A. und J. Ostell. The Genetic Codes. NCBI, 2013. http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG11. Letzter Zugriff: 10/2013.

Enright, A. J. An Efficient Algorithm for Large-scale Detection of Protein Families. *Nucleic Acids Research* 30.7 (2002): 1575-584.

Escherich, T. Über die Bakterien des Milchkothes. *Aerztliches Intelligenzblatt 32 (1884): 243.*

Eßer, C. Der Einfluß lateralen Gentransfers auf die Evolution prokaryotischer Genome am Beispiel von Alpha-Proteobakterien und Stämmen von Escherichia coli. Univ. Diss., Heinrich-Heine-Universität Düsseldorf. 2010.

Estabrook, G. F., F. R. McMorris und C. A. Meacham. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Zoology* 34.2 (1985): 193-200.

Felsenstein, J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology* 22.3 (1973): 240.

Felsenstein, J. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution 17.6 (1981): 368-76.*

Felsenstein, J. Inferring Phylogenies. Sunderland, MA: Sinauer Associates, 2004.

Felsenstein, J. Molecular Sequence Programs. The University of Washington, 2008. http://evolution.genetics.washington.edu/phylip/doc/sequence.html. Letzter Zugriff: 7/2014.

Fisher, R. A. On the interpretation of $\chi 2$ from contingency tablesund the calculation of P. *Journal of the Royal Statistical Society* 85.1 (1922): 87–94.

Fricke, W. F., M. S. Wright, A. H. Lindell, D. M. Harkins, C. Baker-Austin, J. Ravel und R. Stepanauskas. "nsights into the Environmental Resistance Gene Pool from the Genome Sequence of the Multidrug-Resistant Environmental Isolate Escherichia Coli SMS-3-5. *Journal of Bacteriology* 190.20 (2008): 6779-794.

Friedrich, A.W., H. Karch. Infektiologie des Gastrointestinaltraktes Teil IV; Infektionen mit enteropathogenen Escherichia coli. ISBN 978-3-540-413592. Springer Verlag Heidelberg, 2006.

Fritzemeier, J. Automatisierte Rekonstruktion und Analyse metabolischer Netzwerke. Univ. Masterarbeit, Heinrich-Heine-Universität Düsseldorf, 2012.

Frost, L. S., R. Leplae, A. O. Summers und A. Toussaint. Mobile Genetic Elements: The Agents of Open Source Evolution. *Nature Reviews Microbiology* 3.9 (2005): 722-32.

Gadagkar, S. R., M. S. Rosenberg und S. Kumar. Inferring Species Phylogenies from Multiple Genes: Concatenated Sequence Tree versus Consensus Gene Tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 304B.1 (2005).

Gelius-Dietrich, G. Der Einfluß der Hydrogenosomen des anaeroben Pilzes Neocallimastix frontalis – Proteomanalyse und EST-Sequenzierung. Univ. Diss., Heinrich-Heine-Universität Düsseldorf, 2008.

Gorecki, P. und O. Eulenstein. ISBRA'12 Proceedings of the 8th International conference on Bioinformatics Research and Applications (2012):115-126. Springer Verlag, Berlin, Heidelberg.

Greiner, R., U. Konietzny und K. D. Jany. Purification and Characterization of Two Phytases from Escherichia Coli. *Archives of Biochemistry and Biophysics* 303.1 (1993): 107-13.

Greiner, R. N.-G. Carlsson und M. Larsson Alminger. Stereospecificity of Myo-inositol Hexakisphosphate Dephosphorylation by a Phytate-degrading Enzyme of Escherichia Coli. *Journal of Biotechnology* 84.1 (2000): 53-62.

Guindon, Stéphane und Olivier Gascuel. A Simple, Fastund Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology* 52.5 (2003): 696-704.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk und O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59.3 (2010): 307-21.

Hahn, H. Medizinische Mikrobiologie Und Infektiologie. Heidelberg: Springer, 2008. 241f.

Hanakahi, L. Binding of Inositol Phosphate to DNA-PK and Stimulation of Double-Strand Break Repair. *Cell* 102.6 (2000): 721-29.

Hedges, S. B., J. E. Blair, M. L. Venture und J. L. Shoe. *A molecular time scale of eukaryote evolution and the rise of complex multicellular life*. BMC Evol Biol 4 (2004):2.

Hein, J. Reconstructing Evolution of Sequences Subject to Recombination Using Parsimony. *Mathematical Biosciences* 98.2 (1990): 185-200.

Hemm, Matthew R., Brian J. Paul, Thomas D. Schneider, Gisela Storz und Kenneth E. Rudd. Small Membrane Proteins Found by Comparative Genomics and Ribosome Binding Site Models. *Molecular Microbiology* 70.6 (2008): 1487-501.

Herzer, P. J., S. Inouye, M. Inouye und T. S. Whittam. Phylogenetic distribution if branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli. *J. Bacteriol.* 172 (1990):6175-6181.

Hughes, Joseph. TreeRipper web application: towards a fully automated optical tree recognition software. *BMC Bioinformatics* 12 (2011):178.

Huson, D. H. SplitsTree: Analyzing and Visualizing Evolutionary Data. *Bioinformatics* 14.1 (1998): 68-73.

Huson, D. H. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23.2 (2005): 254-67.

Johnson, J. R. und T. A. Russo. Extraintestinal Pathogenic Escherichia Coli : The Other Bad E Coli. *Journal of Laboratory and Clinical Medicine* 139.3 (2002): 155-62.

Kanehisa, M. The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* 32.90001 (2004): 277D-80.

Kaper, James B., James P. Nataro und H. L. T. Mobley. Pathogenic Escherichia Coli. *Nature Reviews Microbiology* 2.2 (2004): 123-40.

Karch, Helge. EHEC O104:H4 Und Die Folgen. *BIOspektrum* 17.6 (2011): 616-20.

Katoh, K. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research* 30.14 (2002): 3059-066.

Katoh, K. MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment. *Nucleic Acids Research* 33.2 (2005): 511-18.

Katoh, K. und H. Toh. Parallelization of the MAFFT Multiple Sequence Alignment Program. *Bioinformatics* 26.15 (2010): 1899-900.

Kaur, Jasmine und S. K. Jain. Role of Antigens and Virulence Factors of Salmonella Enterica Serovar Typhi in Its Pathogenesis. *Microbiological Research* 167.4 (2012): 199-210.

Keseler, I. M., A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martinez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muniz-Rascado, Q. Ong, S. Paley, I. Schroder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen und P. D. Karp. EcoCyc: Fusing Model Organism Databases with Systems Biology. *Nucleic Acids Research* 41.D1 (2012): D605-612.

Kloesges, T., O. Popa, W. Martin und T. Dagan. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. *Molecular Biology and Evolution* 28.2 (2011): 1057-074.

Kunin, V. The Net of Life: Reconstructing the Microbial Phylogenetic Network. *Genome Research* 15.7 (2005): 954-59.

Kupczok, A., A. von Haeseler und S. Klaere. An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees. *Journal of Computational Biology* 15.6 (2008): 577-91.

Kupczok, A. Postprocessing phylogenies, univ. Diss., Universität Wien, 2010.

Kuhner, Mary K. Coalescent Genealogy Samplers: Windows into Population History. *Trends in Ecology & Evolution* 24.2 (2009): 86-93.

Lan, Ruiting und Peter R. Reeves. Escherichia Coli in Disguise: Molecular Origins of Shigella. *Microbes and Infection* 4.11 (2002): 1125-132.

Laubach, T. und A. von Haeseler. TreeSnatcher: Coding Trees from Images. *Bioinformatics* 23.24 (2007): 3384-385.

Laubach, T., A. von Haeseler und M. J. Lercher. TreeSnatcher Plus: Capturing Phylogenetic Trees from Images. *BMC Bioinformatics* 13.1 (2012): 110.

Lawrence, J. G. Und A. C. Retchless. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods Mol Biol.* 532 (2009):29-53.

Lee, J., S. R. Hiibel, K. F. Reardon, T. K. Wood. Identification of stress-related proteins in Escherichia coli using the pollutant cisdichloroethylene. *J Appl Microbiol.* 108.6 (2010):2088-102.

Leebens-Mack, J., T. Vision, E. Brenner, J. E. Bowers, S. Cannon, M. J. Clement, C. W. Cunningham, C. de Pamphilis, R. de Salle, J. J. Doyle, J. A. Eisen, X. Gu, J. Harshman, R. K. Jansen, E. A. Kellogg, E. V. Koonin, B. D. Mishler, H. Philippe, J. C. Pires, Y. L. L. Qiu, S. Y. Rhee, K. Sjölander, D. E. Soltis, P. S. Soltis, D. W. Stevenson, K. Wall, T. Warnow und C. Zmasek. Taking the first steps towards a standard for reporting on phylogenies: Minimum information about a phylogenetic analysis (MIAPA). *J Integr Biol* 10 (2006):231-237.

Lechner, M., S. Findeiß, L. Steiner, M. Marz, P. F. Stadler und S. J. Prohaska. Proteinortho: Detection of (Co-)orthologs in Large-scale Analysis. *BMC Bioinformatics* 12.1 (2011): 124.

Letunic, I. und Bork P. Interactive Tree Of Life (iTOL): An Online Tool for Phylogenetic Tree Display and Annotation. *Bioinformatics* 23.1 (2006): 127-28.

Letunic, I. und Bork P. Interactive Tree Of Life V2: Online Annotation and Display of Phylogenetic Trees Made Easy. *Nucleic Acids Research* 39.Web Server (2011): W475-478.

Lin, Y., R. Vaibhav und B. M. E. Moret. A Metric for Phylogenetic Trees Based on Matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2011): n. Pag.

Lipman, D. und W. Pearson. Rapid and Sensitive Protein Similarity Searches. *Science* 227.4693 (1985): 1435-441.

Maddison, W. P. und D. R. Maddison. *Mesquite: A Modular System for Evolutionary Analysis*. N.p., 2011. http://mesquiteproject.org/mesquite/mesquite.html. Letzter Zugriff: 6/2014.

Maurelli, A. T. Black Holes, Antivirulence Genesund Gene Inactivation in the Evolution of Bacterial Pathogens. *FEMS Microbiology Letters* 267.1 (2007): 1-8.

McCloskey, D., B. Ø. Palsson und A. M. Feist. Basic and Applied Uses of Genome-scale Metabolic Network Reconstructions of Escherichia Coli. *Molecular Systems Biology* 9 (2013):661.

McNeil, L. K., Aziz R. K. In silico Reconstruction of the Metabolic and Pathogenic Potential of Bacterial Genomes Using Subsystems. *Genome Dynamics* 6 (2009): 21-34.

Mellmann, A., D. Harmsen und C. A. Cummings. Prospective genomic characterisation of the German enterohaemorhagic Escherischia coli O104:H4 outbreak by next generation sequencing technologies. *PloS One* 6 (2011):e22751.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller und E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21.6 (1953): 1087.

Minin, V. N. und M. A. Suchard. Counting Labeled Transitions in Continuous-time Markov Models of Evolution. *Journal of Mathematical Biology* 56.3 (2007): 391-412.

Monk, J. M., P. Charusanti, R. K. Aziz, J. A. Lerman, N. Premyodhin, J. D. Orth, A. M. Feist und B. Ø. Palsson. Genome-scale Metabolic Reconstructions of Multiple Escherichia Coli Strains Highlight Strain-specific Adaptations to Nutritional Environments. *Proceedings of the National Academy of Sciences* 110.50 (2013): 20338-0343.

Munk, K. und P. Dersch. Mikrobiologie: 43 *Tabellen*. Stuttgart: Thieme, 2008.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryderk und S. J. O'Brien. Molecular phylogenetics and the origins of placental mammals. Nature 409 (2001):614-618.

Nielsen, R. Mapping Mutations on Phylogenies. *Systematic Biology* 51.5 (2002): 729-39.

Norman, A., L. H. Hansen und S. J. Sorensen. Conjugative Plasmids: Vessels of the Communal Gene Pool. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1527 (2009): 2275-289.

Needleman, S. B. und C. D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* 48.3 (1970): 443-53.

Ochman, H. Evidence for Clonal Population Structure in Escherichia Coli. *Proceedings of the National Academy of Sciences* 81.1 (1984a): 198-201.

Ochman, H. und R. K. Selander. Standard reference strains of Escherichia coli from natural populations. J. Bacteriol. 157 (1984b):690-693.

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

Ogura, Y., T. Ooka, A. Iguchi, H. Toh, M. Asadulghani, K. Oshima, T. Kodama, H. Abe, K. Nakayama, K. Kurokawa, T. Tobe, M. Hattori und T. Hayashi. Comparative Genomics Reveal the Mechanism of the Parallel Evolution of O157 and Non-O157 Enterohemorrhagic Escherichia Coli. *Proceedings of the National Academy of Sciences* 106.42 (2009): 17939-7944.

Orskov, I., F. Orskov, B. Jann und K. Jann. Serology, chemistryund genetics of O and K antigens of Escherichia coli. *Bacteriol. Rev* 41 (1977):667-710.

Ostell, J. M. Integrated access to heterogeneous biomedical data from NCBI. *IEEE Eng. Med. Biol.* 14 (1995):730–736.

Ostell, J.M. Nucleic Acid and Protein Analysis: A Practical Approach. Oxford: IRL Press (1996): 31-43.

Pál, Csaba, Balázs Papp und Martin J. Lercher. Adaptive Evolution of Bacterial Metabolic Networks by Horizontal Gene Transfer. *Nature Genetics* 37.12 (2005): 1372-375.

Pál, C., B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver und L. D. Hurst. Chance and Necessity in the Evolution of Minimal Metabolic Networks. *Nature* 440.7084 (2006): 667-70.

Pattengale, N. D., E. J. Gottlieb und B. M. E. Moret. Efficiently Computing the Robinson-Foulds Metric. *Journal of Computational Biology* 14.6 (2007): 724-35.

Petkovsek, Z., K. Elersic, M. Gubina, D. Zgur-Bertok und M. S. Erjavec. Virulence Potential of Escherichia Coli Isolates from Skin and Soft Tissue Infections. *Journal of Clinical Microbiology* 48.9 (2010): 3462-463.

Pol, D. Empirical Problems of the Hierarchical Likelihood Ratio Test for Model Selection. *Systematic Biology* 53.6 (2004): 949-62.

Pol, D. und I. H. Escapa. Unstable taxa in cladistic analysis: identification and the assessment of relevant characters. *Cladistics* 25 (2009): 515–527.

Poolman, B., J. Knol, C. van der Does, P. J. F. Henderson, W. Liang, G. Leblanc, T. Pourcher und I. Mus-Veteau. Cation and Sugar Selectivity Determinants in a Novel Family of Transport Proteins. *Molecular Microbiology* 19.5 (1996): 911-22.

Popa, O., E. Hazkani-Covo, G. Landan, W. Martin und T. Dagan. Directed Networks Reveal Genomic Barriers and DNA Repair Bypasses to Lateral Gene Transfer among Prokaryotes. *Genome Research* 21.4 (2011): 599-609.

Porwollik, S. Evolutionary Genomics of Salmonella: Gene Acquisitions Revealed by Microarray Analysis. *Proceedings of the National Academy of Sciences* 99.13 (2002): 8956-961.

Posada, D. und T. Buckley. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology* 53.5 (2004): 793-808.

Puigbo, P., S. Garcia-Vallve und J. O. McInerney. TOPD/FMTS: A New Software to Compare Phylogenetic Trees. *Bioinformatics* 23.12 (2007): 1556-558.

Pupo, G. M., L. Ruiting, P. R. Reeves und P. R. Baverstock. Population Genetics of Escherichia Coli in a Natural Population of Native Australian Rats. *Environmental Microbiology* 2.6 (2000): 594-610.

Rambaut, A. TreeThief. TreeThief: a tool for manual phylogenetic tree entry. http://microbe.bio.indiana.edu:7131/soft/iubionew/molbio/evolution/phylo/TreeThief/main.html, 2000. Letzter Zugriff: 10/2013.

Rambaut, A. FigTree. The University of Edinburgh. Letzter Zugriff: 29. Mai 2014. http://tree.bio.ed.ac.uk/software/figtree/, 2012. Letzter Zugriff: 6/2014.

Reed, J. L. und B. Ø. Palsson. Thirteen Years of Building Constraint-Based In Silico Models of Escherichia Coli. *Journal of Bacteriology* 185.9 (2003): 2692-699.

Reeves, P. R., L. Bin, Z. Zhou, D. Li, D. Guo, Y. Ren, C. Clabots, R. Lan, J. R. Johnson und L. Wang. Rates of Mutation and Host Transmission for an Escherichia Coli Clone over 3 Years. Ed. Baochuan Lin. *PLoS ONE* 6.10 (2011): E26907.

Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander und T. S. Whittam. Parallel evolution of virulence in pathogenic Escherichia coli. *Nature* 406 (2000):56-67.

Rice, P., I. Longden und A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16.6 (2000): 276-77.

Robinson, D. F. und L. R. Foulds. Comparison of weighted labelled trees. *Lecture Notes in Mathematics* 748 (1979):119-126. Springer Verlag, Berlin.

Robinson, D. F. und L. R. Foulds. Comparison of Phylogenetic Trees. *Mathematical Biosciences* 53.1-2 (1981): 131-47.

Rokas, A., B. L. Williams, N. King und S. B. Carroll. Genome-scale Approaches to Resolving Incongruence in Molecular Phylogenies. *Nature* 425.6960 (2003): 798-804.

Ronquist, F. und J. P. Huelsenbeck. MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models. *Bioinformatics* 19.12 (2003): 1572-574.

Rosenberg, M. S. Heterogeneity of Nucleotide Frequencies Among Evolutionary Lineages and Phylogenetic Inference. *Molecular Biology and Evolution* 20.4 (2003): 610-21.

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

Schönknecht, G., W.-H. Chen, C. M. Ternes, G. G. Barbier, R. P. Shrestha, M. Stanke, A. Bräutigam, B. J. Baker, J. F. Banfield, R. M. Garavito, K. Carr, C. Wilkerson, S. A. Rensing, D. Gagneul, N. E. Dickenson, C. Oesterhelt, M. J. Lercher und A. P. M. Weber. Gene Transfer from Bacteria and Archaea Facilitated Evolution of an Extremophilic Eukaryote. *Science* 339.6124 (2013): 1207-210.

Schulenburg, J. GOCR. Sourceforge, 2013. GOCR: open-source character recognition. jocr.sourceforge.net. Letzter Zugriff: 6/2014.

Selander, R. K., D. A. Caugant und T. S. Whittam. Genetic structure and variation in natural populations of *Escherichia coli*. In: Neidhardt F. C., J. L. Ingraham, K. B. Low, B. Magasanik, M-Schaechter, H. E. Umbarger, Herausgeber. Escherichia coli and Salmonella typhimurium: cellular and molecular biology. Washington, D. C.: American Society for Microbiology (1987):1625–1648.

Shames, S. R., S. D. Auweter und B. B. Finlay. Co-evolution and Exploitation of Host Cell Signaling Pathways by Bacterial Pathogens. *The International Journal of Biochemistry & Cell Biology* 41.2 (2009): 380-89.

Shi, J., Y. Zhang, H. Luo und J. Tang. Using Jackknife to Assess the Quality of Gene Order Phylogenies. *BMC Bioinformatics* 11.1 (2010): 168.

Sims, G. E. und S.-H. Kim. Whole-genome Phylogeny of Escherichia Coli/Shigella Group by Feature Frequency Profiles (FFPs). *Proceedings of the National Academy of Sciences* 108.20 (2011): 8329-334.

Skippington, E. und M. A. Ragan. Phylogeny Rather than Ecology or Lifestyle Biases the Construction of Escherichia Coli-Shigella Genetic Exchange Communities. *Open Biology* 2.9 (2012): 120112.

Smillie, C. S., M. B. Smith, J. Friedman, O. X. Cordero, L. A. David und E. J. Alm. Ecology Drives a Global Network of Gene Exchange Connecting the Human Microbiome. *Nature* 480.7376 (2011): 241-44.

Sneath, A. und R. R. Sokal. Numerical Taxonomy: The Principle and Practice of Numerical Classification. W. H. Freeman and Co., San Francisco.

Sokal, R. R. und C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38 (1958):1409–1438.

Stajich, J. und E. Birney. The Bioperl Project. *ACM SIGBIO Newsletter* 20.2 (2000): 13-14.

Stamatakis, A. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22.21 (2006): 2688-690.

Stamatakis, A.: RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30.9 (2014):1312-1313.

Sukumaran, J. und M. T. Holder. DendroPy: A Python Library for Phylogenetic Computing. *Bioinformatics* 26.12 (2010): 1569-571.

Susko, E. Using minimum bootstrap support for splits to construct confidence regions for trees. *Evol. Bioinform. Online* 2(2007):137-151.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis und J. S. Rogers. Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods. *Systematic Biology* 50.4 (2001): 525-39.

Talavera, G. und J. Castresana. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology* 56.4 (2007): 564-77.

Thomas, C. M. und K. M. Nielsen. Mechanisms Of and Barriers To Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology* 3.9 (2005): 711-21.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. El Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguénec, M. Lescat, S. Mangenot, V. Martinez-Jéhanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. Saint Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Médigue, E. P. C. Rocha und E. Denamur. Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths. Ed. Josep Casadesús. *PLoS Genetics* 5.1 (2009): E1000344.

Ungewickell, E. Inositol Hexakisphosphate Binds to Clathrin Assembly Protein 3 (AP-3/AP180) and Inhibits Clathrin Cage Assembly in Vitro. *Journal of Biological Chemistry* 270.1 (1995): 214-17.

Van Dongen, S. Graph Clustering by Flow Simulation. Univ. Diss., Universität Utrecht, 2000.

Vanorsdel, C. E., S. Bhatt, R. J. Allen, E. P. Brenner, J. J. Hobson, A. Jamil, B. M. Haynes, A. M. Genson und M. R. Hemm. The Escherichia Coli CydX Protein Is a Member of the CydAB Cytochrome Bd Oxidase Complex and Is Required for Cytochrome Bd Oxidase Activity. *Journal of Bacteriology* 195.16 (2013): 3640-650.

Varma, A. und B. Ø. Palsson. Parametric sensitivity of stoichiometric flux balance models applied to wild-type Escherichia coli metabolism. *Biotechnol Bioeng.* 45.1 (1995):69-79.

Vieira, G., V. Sabarly, P.-Y. Bourguignon, M. Durot, F. Le Fevre, D. Mornico, D. Vallenet, O. Bouvet, E. Denamur, V. Schachter und C. Medigue. Core and Panmetabolism in Escherichia Coli. *Journal of Bacteriology* 193.6 (2011): 1461-472.

Wang, Y., Y. Xiangsheng, D. Guangda und X. Fangsen. Overexpression of PhyA and AppA Genes Improves Soil Organic Phosphorus Utilisation and Seed Phytase Activity in Brassica Napus. Ed. Tianzhen Zhang. PLoS ONE 8.4 (2013): E60801.

Waterman, M. S. und T. F. Smith. *On the Similarity of Dendrograms. Journal of Theoretical Biology 73.4 (1978): 789-800.*

Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Orskov, I. Orskov und R. A. Wilson. Clonal relationships among Escherichia coli strains that cause hemorrhagic colitis and infantile diarrhea. *Infect. Immun.* 61 (1993):1619-1629.

Wicherts, J. M., D. Borsboom, J. Kats und D. Molenaar. The poor availability of psychological research data for reanalysis. *American Psychologist* 61 (2006):726-728.

Winfield, M. D. und E. A. Groisman. Role of Nonhost Environments in the Lifestyles of Salmonella and Escherichia Coli. *Applied and Environmental Microbiology 69.7 (2003): 3687-694.*

Wolf, Y. I. Coelomata and Not Ecdysozoa: Evidence From Genome-Wide Phylogenetic Analysis. *Genome Research* 14.1 (2003): 29-36.

Wyss, M., R. Brugger, A. Kronenberger, R. Remy, R. Fimbel, G. Oesterhelt, M. Lehmann und A. P. van Loon. Biochemical characterization of fungal phytases (myo-inositol hexakisphosphate phosphohydrolases): catalytic properties. *Appl Environ Microbiol* 65.2 (1999):367-73.

Yang, L., L. Jelsbak, R. L. Marvig, S. Damkiaer, C. T. Workman, M. H. Rau, S. K. Hansen, A. Folkesson, H. K. Johansen, O. Ciofu, N. Hoiby, M. O. A. Sommer und S. Molin. Evolutionary Dynamics of Bacteria in a Human Host Environment. *Proceedings of the National Academy of Sciences* 108.18 (2011): 7481-486.

Danksagung

Ich danke Prof. Dr. Martin J. Lercher für die Möglichkeit, eine Dissertation an seinem Lehrstuhl zu verfassen. Wir haben beide nicht damit gerechnet, daß die vorliegende, sehr „biologielastige“ Arbeit dabei herauskommen würde, enthielt der anfangs erstellte Plan doch sehr viel mehr „angewandte Informatik“. Im Ergebnis freue ich mich aber darüber, daß ich dadurch die Chance erhalten habe, hochgradig interdisziplinär zu arbeiten und mir unbekanntes Territorium erforschen zu können.

Die stets kameradschaftliche und immer motivierende Atmosphäre am Lehrstuhl für Bioinformatik tat ihr übriges, um mich immer wieder zur Fortführung meiner Arbeit anzuspornen. Die erhellenden Diskussionen mit meinen Kollegen, insbesondere Christian Eßer, Jonathan Fritzemeier und Ulrich Wittelsbürger, haben mich sehr oft weiter gebracht, wenn ich mich gedanklich oder emotional in einer Sackgasse befand. Aber auch unsere anderen (ehemaligen und aktuellen) Kollegen haben dazu beigetragen, daß wir alle immer gern am Lehrstuhl waren – auch außerhalb der Kernarbeitszeit. Ohne besondere Reihenfolge sind dies Janina Maß, David Heckmann, Deya Alzoubi, Na Gao, Jodie Napp, Sabine Thuß, Daniel Hartleb, Abdelmonem Desouki, Nina Knipprath, Gabriel Gelius-Dietrich, Ingo Paulsen, Dominic Mainz, Indra Mainz, Jochen Kohl, Esther Sundermann, Jan Wolfertz, Bastian Pfeifer, Simone Linz, Michael Roßkopf, Benjamin Braasch, Rafael Dellen, Karim Benyaa, Trupti Gohel, Anas Amro, Abla Jaber, Kollegen und Freunde von anderen Lehrstühlen, unter ihnen Ingo Weigelt, Michael Jastram, Jens Bendisposto und Ivaylo Dobrikov, sowie diverse Bachelor- und Masterarbeits-Studenten, deren Namen mir leider entfallen sind.

Ich danke unserem „guten Geist“ Anja Walge für ihre humorvolle und unkomplizierte Hilfe, mit der sie uns allen stets beigestanden und uns gegen den bürokratischen Wahnsinn abgeschottet hat. Mein Dank geht auch an die anderen Sekretärinnen des Instituts für Informatik, die uns immer unkompliziert geholfen haben: Angela Rennwanz, Sabine Freese, Claudia Kiometzis u. a. Ohne sie würde der Laden nicht laufen.

Danke sagen möchte ich auch den anderen Professoren am Institut für Informatik. Sie haben es geschafft, mich durch die Qualität ihrer Lehre immer wieder neu für die verschiedensten Themen der Informatik zu begeistern.

Ich danke den einschlägigen Produzenten von Kaffeebohnen und Kaffeepulver für den endlosen Nachschub an koffeinhaltigen Heißgetränken. Ein Leben ohne Kaffee ist möglich, aber sinnlos.

Vor allem danke ich meiner Lebensgefährtin Annika Hoinkes und unserem Sohn Noah für ihre schier endlose Geduld und ihren Beistand. Wie oft mußten sie auf mich verzichten, als ich wieder einmal die „jetzt aber wirklich letzte“ Berechnung abschloß. Wie oft habe ich zu Hause für schlechte Laune gesorgt. Einmal mußten sie sogar ohne mich in den Urlaub fahren!

Besonderer Dank gebührt meinen Eltern. Sie haben mir stets mit Rat und Tat zur Seite gestanden und mich finanziell unterstützt. Ohne sie wäre ich heute nicht da, wo ich bin.

Erklärung

Ich versichere an Eides Statt, daß die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. 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

_____

(Thomas Laubach)