Karlsruhe Institute of Technology

Distanzbasierte Inferenz von Genbäumen unter Berücksichtigung ihres Speziesbaumes

Bachelorarbeit
von

Lukas Knirsch

An der Fakultät für Informatik
Institut für Theoretische Informatik

Erstgutachter:	Prof. Dr. Alexandros Stamatakis
Zweitgutachter:	TT-Prof. Dr. Thomas Bläsius
Betreuender Mitarbeiter:	Benoit Morel

Bearbeitungszeit: 1. Juni 2023 – 02. Oktober 2023

Selbstständigkeitserklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, 01. Oktober 2023

Zusammenfassung

In dieser Arbeit wird Spearfish, eine neue Methode zur distanzbasierten Inferenz von Genbäumen, entwickelt und getestet. Spearfish verwendet die paarweisen Distanzen der Gensequenzen, sowie die Distanzen der zugehörigen Spezies im Speziesbaum, in einem Clustering-Verfahren, um 10 Genbäume zu rekonstruieren. Der beste wird anschließend mithilfe eines statistischen Evaluierungsverfahrens ausgewählt.

Auf allen getesteten simulierten Datensätzen konnte gezeigt werden, dass die von Spearfish inferierten Bäume durchschnittlich eine Distanz von 0,213 zum echten Genbaum besitzen. Damit ist es 2,18-mal genauer als Methoden wie RAxML-NG, welche den Speziesbaum nicht berücksichtigen. Spearfish ist 25,85% ungenauer, aber 49,63% schneller als GeneRax, eine der führenden Methoden, die Genbäume mithilfe ihres Speziesbaumes korrigieren. So kann Spearfish verwendet werden, um Startbäume für GeneRax zu rekonstruieren oder bei goßen Datensätzen sogar zu ersetzen.

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einleitung

1.1. Motivation

Diese Arbeit beschäftigt sich mit der Rekonstruktion von Genbäumen (Abb. 2.2.2.2). Genbäume beschreiben und visualisieren die Evolutionsgeschichte einer Menge an Genen. Sie sind maßgeblich daran beteiligt, ein besseres Verständnis von Genen, Organismen und ihre Verwandtschaftsverhältnissen zu erlangen. Da Rekonstruktionsalgorithmen allerdings nur die Gensequenzen und nicht vollständige Genome zur Verfügung stehen, ist die korrekte Inferenz eines Genbaums schwer. Neuere Algorithmen wie GeneRax [1] berücksichtigen zusätzlich den Speziesbaum, da dieser zusätzliche hilfreiche Informationen enthält und können so sehr viel genauere Genbäume inferieren.

Alle bereits existierenden Methoden sind jedoch langsam und benötigen vor allem für sehr große Datensätze zu lange, um sinnvoll einsetzbar zu sein. Konträr dazu steigt weltweit die Menge der ermittelten DNA-Sequenzen aufgrund verbesserter und immer günstigerer Verfahren seit Jahren schnell an. Im Mai 2022 kostete das reine Sequenzieren einer Million Nukleinbasen, bei dem DNA aus einer Zelle extrahiert und die Abfolge der Basen digital ausgegeben wird, 0,006 Dollar und das Sequenzieren und Zusammensetzen eines gesamten menschlichen Genoms 525 Dollar [2]. Aufgrund dieser Entwicklung verdoppelt sich die Anzahl an Basen in GenBank, einer der größten Datenbanken für DNA- und Proteinsequenzen, ungefähr alle 18 Monate. In GenBank waren im Juni 2023 mehr als 1,9 Billionen Basen in über 240 Millionen Sequenzen enthalten [3]. Aufgrund dieser Entwicklung hin zu immer größeren Datenmengen, steht heutzutage nicht mehr nur die Genauigkeit der Softwaretools im Vordergrund. Damit alle verfügbaren Sequenzen genutzt werden können, werden bestehende Tools auf Geschwindigkeit optimiert und neue Methoden entwickelt, die eine schnellere Rekonstruktion erlauben.

Ein weiteres Problem der Bauminferenz ist der große Suchraum [4], auch *Tree-Space* genannt. Der Suchraum beschreibt die Menge aller möglichen Bäume (Formel 2.1), die mit höherer Anzahl an sequenzierten Genen exponentiell wächst [5]. Um den optimalen Baum zu finden, muss der Tree-Space durchsucht werden. Dafür werden unter anderem

statistische Verfahren angewendet, die jedoch mit höheren Kosten verbunden sind, da sie viel Zeit in Anspruch nehmen.

Im Rahmen dieser Arbeit wurde deshalb Spearfish (**SPE**cies tree **A**ware gene t**R**ee in**F**erence with d**IS**tance met**H**ods) entwickelt. Spearfish kann mithilfe der Distanzen im Speziesbaum und denen der Gensequenzen einen Genbaum inferieren. Der Speziesbaum der sequenzierten Spezies repräsentiert ihre evolutionäre Abstammung und kann, solange genug Sequenzen vorliegen, leichter als ein Genbaum rekonstruiert werden. Gleichzeitig verbessert eine Berücksichtigung des Speziesbaumes die Genauigkeit der Genbäume, wie Szöllősi et al. [6] in mehreren Softwaretools beobachtet haben. Im Vergleich zu zeichenbasierten Methoden benötigen distanzbasierte Methoden wie Spearfish weniger Zeit, um einen Genbaum zu inferieren. Dafür sind diese Bäume häufig ungenauer als die der zeichenbasierten Methden. Daraus ergeben sich für Spearfish zwei Anwendungsfälle. Zum einen können die von Spearfish inferierten Bäume als Startbäume für GeneRax verwendet werden und dadurch die Laufzeit von GeneRax verringern, da die Suche bereits von akkurateren Bäumen aus gestartet wird. Zum anderen kann Spearfish möglicherweise bei sehr großen Datensätzen, bei denen GeneRax zu langsam ist, eingesetzt werden und sehr viel genauere Genbäume inferieren, als andere Methoden, die bei diesen Datensätze bisher verwendet wurden.

1.2. Ziele dieser Arbeit

In dieser Arbeit soll auf bereits bestehende, schnelle, aber ungenaue Verfahren aufgebaut werden, um durch (geringe) Laufzeitverlängerungen sehr viel genauere Genbäume zu inferieren.

Die entwickelten Methoden werden am Ende mit simulierten Datensätzen evaluiert, wobei der Fokus auf der durchschnittlichen Genauigkeit liegt, mit der die Genbäume inferiert werden. Um die Ergebnisse einordnen zu können, werden die Verfahren mit Methoden, welche auf verschiedenen Ansätzen beruhen, verglichen. Während FastME [7] und RAxML-NG [8] nur die Gensequenzen verwenden, verwendet die zeichenbasierte Methode GeneRax auch den Speziesbaum. Zusätzlich wird auch auf den Laufzeitvorteil distanzbasierter Verfahren wie FastME oder Spearfish gegenüber den zeichenbasierten Verfahren eingegangen.

2. Grundlagen

In diesem Kapitel werden das notwendige Basiswissen, Fachbegriffe und Methoden erläutert. Zuerst werden grundlegende biologische Begriffe eingeführt. Als nächstes werden verschiedene phylogenetische Bäume eingeführt, sowie Operationen auf diesen. Danach werden bereits existierende Methoden zur Rekonstruktion von Genbäumen vorgestellt und klassifiziert.

2.1. Einführung in die Genetik

2.1.1. Definition von Spezies

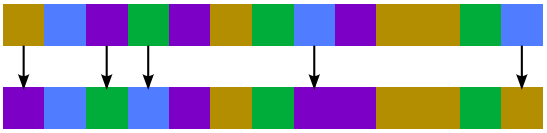
Obwohl der Begriff *Spezies* allgemein bekannt ist, gibt es keine einheitliche Definition. Häufig werden Lebewesen derselben Spezies zugeordnet, wenn sie sich in der Natur miteinander fortpflanzen können und ihre Nachkommen zeugungsfähig sind [9, 10]. Da diese Definition allerdings Bakterien nicht miteinschließt, gibt es noch die erweiterte Definition der Phylogenetik, welche sich speziell mit „Rekonstruktion der Stammesgeschichte“ [11, 12] beschäftigt. Ihre Definition besagt, dass eine Spezies ein „irreduzibles Cluster an Organismen [ist], welches eindeutige Unterschiede zu anderen solchen Clustern aufweist und in dem es ein (elterliches) Abstammungsmuster gibt“ [13]. Damit beinhaltet Cracraft's Definiton alle Lebewesen, die sich durch Paarung fortpflanzen, sowie alle Lebewesen, die sich durch Zellteilung vermehren, da auch hier eine Zelle als Vorfahr einer anderen definiert werden kann. In dieser Arbeit wird letztere Definition verwendet, da die vorgestellten Methoden in allen davon abgedeckten Fällen verwendet werden können.

2.1.2. Genom

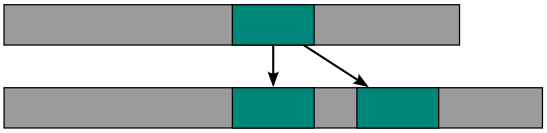
Alle Lebewesen bestehen aus mindestens einer Zelle, in der sich die Erbinformation des Organismus befindet. In allen Lebewesen wird die genetische Information in der Desoxyribonukleinsäure (DNS (engl. *DNA*)) gespeichert, die die typische Doppelhelixstruktur ausbildet. Die DNS wird aus Nukleotiden aufgebaut. Die Nukleotide bestehen aus Zucker, Phosphat und einer der vier Basen: *Adenin (A)*, *Guanin (G)*, *Cytosin (C)* und *Thymin*

ATCACACCAGTGTCTGCGTTCACAGCAGGCATCATCAGTAGCCTCCAGAGGC
CTCAGGTCCAGTCTCTAAAAATATCTCAGGAGGCTGCAGTGGCTGACCATTG
CCTTGGACCGCTCTTGGCAGTCGAAGAAGATTCTCCTGTCAGTTTGAGCTGG

Auszug aus dem menschlichen Genom, Chromosom 1 [14].



(a) Bildliche Darstellung der Substitution einzelner Nukleotide. Die verschiedenen Farben entsprechen den vier Basen.



(b) Schematische Darstellung einer Genduplikation, bei der das kopierte Gen an einer anderen Stelle im Genom eingebaut wird. In grau ist ein Chromosom zu sehen, während das kopierte Gen grün markiert ist.

Abbildung 2.1.: Beispiele molekularer Veränderungen.

(T). Bestimmte Nukleotidsequenzen, die *Gene*, codieren alle biologischen Information des Organismus, die von den Eltern vererbt werden. Gene werden als Zeichenketten wie in Abb. 2.1 dargestellt, wobei die Basennamen durch ihre Anfangsbuchstaben abgekürzt werden. Zusammen mit der *nicht-codierenden DNS* bilden sie das *Genom* eines Organismus.

2.1.3. Evolutionäre Prozesse

Bei einer Betrachtung der Evolutionsgeschichte mehrerer Spezies versucht man, Aussagen über die hypothetischen Vorfahren verwandter Spezies zu treffen. Der *jüngste gemeinsame Vorfahr (JGV)* [15] mehrerer Spezies ist dabei jene Art, ab der sich die Spezies unabhängig entwickelt haben. Das Prinzip des JGV kann genau so auch auf Gene übernommen werden. Der Begriff *Taxon*, der in der Biologie eine Gruppe an zusammengehörenden Teilen [10] bezeichnet, wird in dieser Arbeit speziell für eine einzelne Spezies oder ein Gen in einer Spezies verwendet, solange nicht anders erwähnt.

2.1.3.1. Evolution von Spezies

Spezies unterliegen dem Einfluss der Evolution. Zwei der wichtigsten Folgen der Evolution sind die *Artbildung* (engl. *speciation*) und das *Aussterben* (engl. *extinction*) einer Art [16]. Betrachtet man eine konkrete Art, so werden diese Folgen auch als *Evolutionsereignisse* bezeichnet. Bei einem Artbildungsereignis entstehen aus einer Art zwei neue Arten, während die alte Art (implizit) ausstirbt [10]. Ein Artaussterbeereignis ist durch den Tod des letzten Nachkommens einer Art definiert [17].

Molekulare Veränderungen der Spezies [18], genannt *Mutationen*, betreffen das Genmaterial und verändern so im Laufe der Zeit die DNA der Spezies [19]. Dabei werden einzelne Nukleinsäuren fälschlicherweise durch andere substituiert und nicht korrigiert, also ausgetauscht (Abb. 2.1a). Die Rate, mit der molekulare Veränderungen in der DNA auftreten, wird als *Mutationsgeschwindigkeit* bezeichnet. Je nach Position im Genom, sowie der beteiligten Nukleinsäuren, kann die Mutationsgeschwindigkeit variieren [20]. Bei ausreichend vielen Mutationen führen diese Veränderungen zu Artbildungs- sowie Artaussterbeereignissen.

2.1.3.2. Evolution von Genen

Betrachtet man nicht die Evolutionsgeschichte mehrerer Spezies, sondern nur die eines Gens, so bezeichnet man alle Nachkommen des Gens als *homolog* [21] zueinander. Ein homologes Genpaar kann in verschiedene Kategorien unterteilt werden, wobei für diese Arbeit besonders die Kategorien der orthologen und paralogen Genpaare von Bedeutung sind. Zwei Gene sind ortholog zueinander, wenn ihr JGV aufgrund eines Artbildungsereignisses kopiert wurde. Dahingegen werden Gene als *paralog* zueinander betrachtet, wenn es sich bei dem einen Gen um eine *Genduplikation* (engl. *duplication event*) des anderen Gens handelt [21]. Bei einer Genduplikation wird eine Kopie des Gens an einer anderen Stelle im Genom eingebaut [22], wie in Abb. 2.1b dargestellt.

Gene können zwar nicht aussterben, aber auch sie können verschwinden, indem sie bei Fortpflanzung nicht kopiert werden. Das zugehörige Ereignis wird als *Genverlust* (engl. *loss event*) [23] bezeichnet.

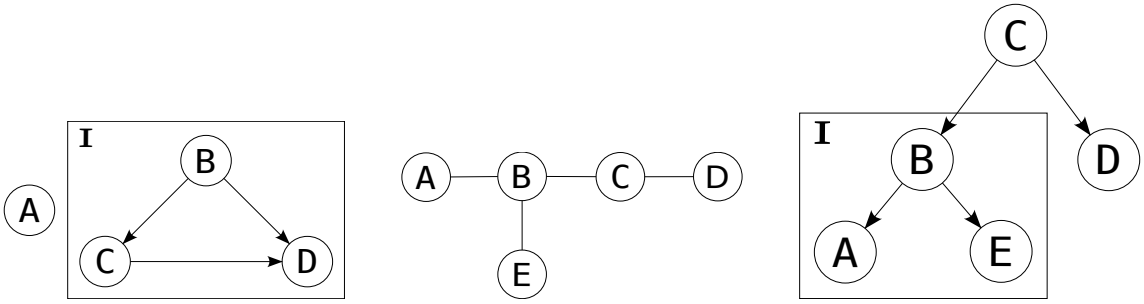
2.2. Bäume

2.2.1. Konzepte der Graphentheorie

Ein *einfacher Graph* (Abb. 2.2) ist das Paar $G = (V, E)$ der Knotenmenge V und der Kantenmenge $E \subseteq V^2$ [24]. Befindet sich ein Paar $e = (u, v) \in E$, so gibt es eine Kante von Knoten u zu v, der Graph ist also *gerichtet*. In Abb. 2.2c ist der gerichtete Graph $G = (V, E)$ mit der Knotenmenge $V = \{A, B, C, D\}$ und der Kantenmenge $E = \{(A, B), (A, D), (B, D)\}$ dargestellt, während Abb. 2.2b einen sogenannten *ungerichteten* Graphen zeigt. Ungerichtete Graphen sind Graphen, bei denen für jede Kante e auch ihre Rückkante (v,u) in E enthalten ist. Die zwei Kanten (u,v) und (v,u) können dann auch durch $e = \{u, v\}$ beschrieben werden.

Ein *Pfad* ist eine Abfolge von ungerichteten Kanten, wobei Kanten jeweils an dem Knoten beginnen, an dem die vorherige Kante endet [25]. Bei einem *zusammenhängenden* Graph existiert ein Pfad von jedem Knoten zu jedem anderen Knoten [24]. Somit ist der Graph in Abb. 2.2c nicht zusammenhängend, der umrandete Teilgraph I allerdings schon.

Ein *Baum* (Abb. 2.2b) ist in der Graphentheorie ein zusammenhängender Graph, der keine *Kreise* enthält [25]. Ein Kreis ist ein Pfad, dessen Start- und Endknoten gleich sind [24]. Die *Wurzel* eines gerichteten Baumes ist der Knoten, der keine eingehenden Kanten besitzt, von dem aus aber alle anderen Knoten erreichbar sind. Jeder Baum kann nur eine Wurzel haben. Bei einem ungerichteten Baum ist der Begriff der Wurzel nicht genauso sinnvoll, allerdings kann man einen ungerichteten Baum als einen gerichteten Baum interpretieren, indem man einen beliebigen Knoten als Wurzel definiert (Abb. 2.2c). Die Knoten mit keinen ausgehenden Kanten werden als *Blätter* bezeichnet, alle anderen Knoten inklusive der Wurzel als *innere* Knoten. Existiert eine Kante (u, v), so ist u der *Elternknoten* seines *Kindes* v [25]. Ein Baum, bei dem jeder Knoten exakt 0 oder 2 Kinder besitzt, ist ein *Binärbaum* [26]. Aufgrund der Baumeigenschaften ist jeder Teilgraph, der aus einem Knoten v und allen Knoten $U \subseteq V$ zu denen ein Pfad (v, u), $u \in U$ existiert, wieder ein Baum. Diesen bezeichnet man als den *Teilbaum* unter v [25].



(a) Gerichteter Graph G. Teilgraph I ist zusammenhängend.

(b) Ungerichteter Baum, dessen Wurzel beliebig gewählt werden kann.

(c) Der Baum aus 2.2b, bei dem C als Wurzel gewählt wurde. $\{A, B, E\}(I)$ ist der Teilbaum unter B.

Abbildung 2.2.: Graphen und Bäume

2.2.2. Phylogenetische Bäume

Ein *phylogenetischer Baum*, auch als Stammbaum bezeichnet, stellt die evolutionären Beziehungen zwischen Genen, ganzen Spezies oder Proteinen graphisch dar. Er ist ein ungerichteter Baum, dessen Blätter die zu veranschaulichenden Einheiten repräsentieren. Die Kanten eines phylogenetischen Baumes, genannt *Äste*, repräsentieren die zeitlichen Abstände oder die Anzahl der aufgetretenen Mutationen. Innere Knoten repräsentieren hypothetische Vorfahren. Ein phylogenetischer Baum kann zwar auch gewurzelt sein, die meisten Verfahren berechnen allerdings ungewurzelte Bäume, da eine aussagekräftige Wurzel schwer zu definieren ist. Aufgrund der Annahme, bei einem Evolutionsereignis entstehen immer höchstens zwei Arten, sind phylogenetische Bäume meist binär. Unsicherheiten in der Inferenz können als Polytomie dargestellt werden, also einem Knoten mit mehr als zwei Kindern [10].

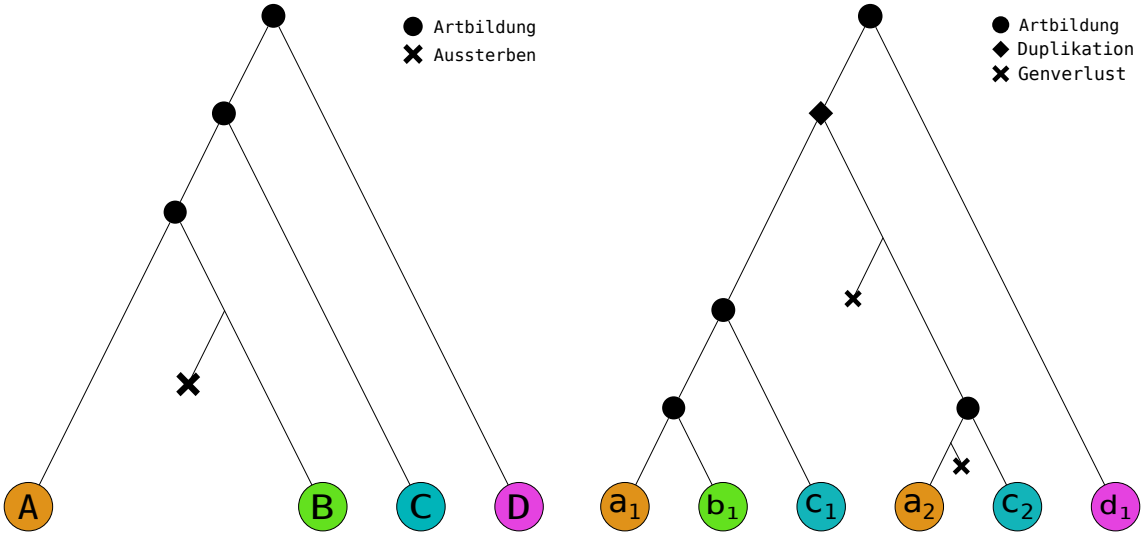
Die „Lage und Anordnung“ [27] der einzelnen Knoten wird durch die *Topologie* des Baumes beschrieben. Mit steigender Blattanzahl steigt die Anzahl der möglichen Topologien $\mathcal{T}$, also die Größe des Tree-Space, exponentiell. Ein ungerichteter Baum mit $n > 2$ Blättern besitzt $2n - 3$ Äste und deshalb existieren

$$|\mathcal{T}_n| = \prod_{i=3}^{n} (2i - 5) \tag{2.1}$$

viele Möglichkeiten, die Knoten anzuordnen. Formel 2.1 beschreibt die Anzahl der möglichen Topologien für einen gewurzelten Baum mit $n - 1$ Blättern [5].

2.2.2.1. Speziesbäume

Stellt man Spezies und ihre Evolutionshierarchie in einem phylogenetischen Baum dar, so spricht man von einem Arten- oder *Speziesbaum* (Abb. 2.3a). Innere Knoten bezeichnen immer eine Artenbildung [28], da kein Taxon des Speziesbaumes von einer Art ohne Nachkommen abstammen kann. Dieses Phänomen muss bei der Interpretation der Speziesbäume berücksichtigt werden, da Arten, die nicht *beobachtbar* sind, in keinem Baum auftauchen können. Beobachtbare Arten sind Arten, die entweder noch leben oder deren Genom aus



(a) Speziesbaum der Spezies A, B, C und D.

(b) Genbaum eines Gens der Spezies A, B, C und D, welches einmal dupliziert wurde.

Abbildung 2.3.: Zwei phylogenetische Bäume, links ein Speziesbaum und rechts ein zugehöriger Genbaum.

Fossilien gewonnen werden kann. Ein Beispiel für eine ausgestorbene, aber beobachtbare Art ist das Mammut [29], dessen Genom aus eingefrorenen Überresten rekonstruiert werden konnte. Der Speziesbaum in Abb. 2.3a stellt die evolutionäre Entwicklung der Spezies A, B, C und D dar. A und B sind am engsten miteinander verwandt, während die Art D sich schon früh von den anderen Arten abgespalten hat. Die ausgestorbene Spezies wurde eingezeichnet, obwohl sie eigentlich nicht sichtbar ist, um diese Eigenschaft zu verdeutlichen.

2.2.2.2. Genbäume

Genbäume (engl. *gene family tree*) sind phylogenetische Bäume, deren Blätter ein Gen einer Genfamilie darstellen (Abb. 2.3b). Genfamilien sind eine Menge aus homologen Genen [28]. Um die Evolution der Gene zu untersuchen, werden sie deshalb anhand ihrer Verwandtschaft in einen Genbaum eingeordnet. In Abb. 2.3b ist der Genbaum eines Gens der Spezies A, B, C und D aus dem Speziesbaum (Abb. 2.3a) dargestellt. Er ähnelt dem Speziesbaum, nach der Duplikation folgt die Evolution des Gens aber nicht mehr strikt der der Spezies. Am Ende enthalten A und C je zwei Kopien des Gens.

2.2.3. Vergleichsmethoden für Bäume

Um die Ähnlichkeit zweier Bäume bestimmen zu können, benötigt es einer Metrik $d : X \times X \to \mathbb{R}$ zwischen allen Elementen $x, y, z \in X$.

Mithilfe der *Robinson-Foulds (RF) Distanz* [30] kann die Ähnlichkeit zwischen zwei Bäumen berechnet werden. Die RF-Distanz d_{RF} zwischen zwei Bäumen T_1 und T_2 ist definiert als

$$d_{RF}(T_1, T_2) = B_1 \triangle B_2 = |B_1 \cup B_2| - |B_1 \cap B_2|. \tag{2.2}$$

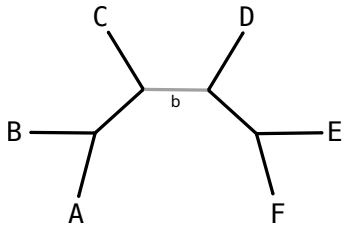


Abbildung 2.4.: Bipartition eines Baumes [31].

Die Menge aller *Bipartitionen* von Baum $i \in [\,1, 2\,]$ wird dabei mit B_i bezeichnet und definiert diesen eindeutig. Eine Bipartition $P := L_1 \,|\, L_2$ ist eine Unterteilung der Blätter in zwei disjunkte Teilmengen L_1 und L_2, die entweder durch L_1 und L_2 oder von dem trennenden Ast b zwischen L_1 und L_2 identifiziert wird. Bei einer *trivialen* Bipartition gilt $|\, L_1 \,| = 1$, zum Beispiel $P_1 = A|BCDEF$. Die in Abb. 2.4 gezeigte Bipartition wird durch Ast b identifiziert. Dadurch ergeben sich für $L_1 = \{A, B, C\}$ und $L_2 = \{D, E, F\}$. Die Bipartition b kann also auch durch $ABC|DEF$ dargestellt werden.

Da ein Baum mit n Blättern $2n - 3$ Äste hat, kann er nur $2n - 3$ Bipartitionen haben [30]. Für die RF-Distanz werden allerdings nur die nicht-trivialen Bipartitionen betrachtet. Von den ursprünglichen $2n - 3$ Bipartitionen verbleiben dadurch $2n - 3 - n = n - 3$ pro Baum.

Dies ist ein Nachteil für die RF-Distanz, da sie deshalb nur Werte im Bereich $d \in [\,0, 2(n-3)]$ annehmen kann (Formel 2.2). Da die Anzahl möglicher Bäume exponentiell wächst, gilt $|\, \mathcal{T}_n \,| \gg 2(n-3)$. Deshalb können große Bäumen schnell eine Distanz von 1 zueinander haben, obwohl sie nur $2(n-3)$ verschiedene Bipartitionen besitzen.

$$d_{rRF}(T_1, T_2) = \frac{B_1 \triangle B_2}{2(n-3)} \tag{2.3}$$

Die relative RF-Distanz *rRF* (Formel 2.3) bildet die Distanz zwischen zwei Bäumen auf den Wertebereich $[\,0, 1\,]$ ab und wird in dieser Arbeit meistens verwendet. Sie weißt dieselben Probleme wie die absolute RF-Distanz auf, kann aber besser verglichen werden, da sie die Distanz relativ zur Größe der Bäume angibt.

2.2.4. Markierungen auf Genbäumen

Gene durchlaufen unterschiedliche Evolutionsereignisse (Abs. 2.1.3.2) wie Duplikationen und Genverluste. Aus demselben Grund (Abs. 2.2.2.1) wie bei einem Speziesbaum, sind auch in einem Genbaum nur Artspaltungs- und Genduplikationsereignisse sichtbar. Das Verknüpfen der inneren Knoten mit der Information, ob sie ein Artspaltungs- oder Genduplikationsereigniss darstellen, wird als Markieren (engl. *Tagging*) bezeichnet. Dafür muss der Baum allerdings gewurzelt sein. Im Folgenden werden *Astral-Pro (APro)* [32] und *MAD* [33] vorgestellt, wobei APro einen Baum sowohl wurzeln als auch markieren kann, während MAD Bäume nur wurzelt.

2.2.4.1. ASTRAL-Pro

ASTRAL (Accurate Species TRee ALgorithm) [34] und die Weiterentwicklung *ASTRAL-Pro* (ASTRAL for PaRalogs and Orthologs) [32] sind zwei Methoden, um einen Speziesbaum

mithilfe vieler Genbäume zu inferieren. APro berücksichtigt dabei auch Duplikationen und markiert als ersten Schritt alle Genbäume. Dieser Schritt ist für diese Arbeit wichtig und wird im Folgenden mit dem Begriff „APro“ gleichgesetzt.

$$S(k_w) = \begin{cases} 0, & \text{falls } k_w \text{ Blatt, sonst} \\ S^l(k_w) + S^r(k_w) + & \begin{cases} 0, & \text{falls } \mathcal{C}^l(k_w) \cap \mathcal{C}^r(k_w) = \emptyset, \\ 1, & \text{falls } \mathcal{C}^l(k_w) = \mathcal{C}^r(k_w), \\ 2, & \text{falls } \bigoplus_{t \in \{l,r\}} (\mathcal{C}^t(k_w) \subsetneq \mathcal{C}(k_w)), \\ 3, & \text{sonst.} \end{cases} \end{cases} \tag{2.4}$$

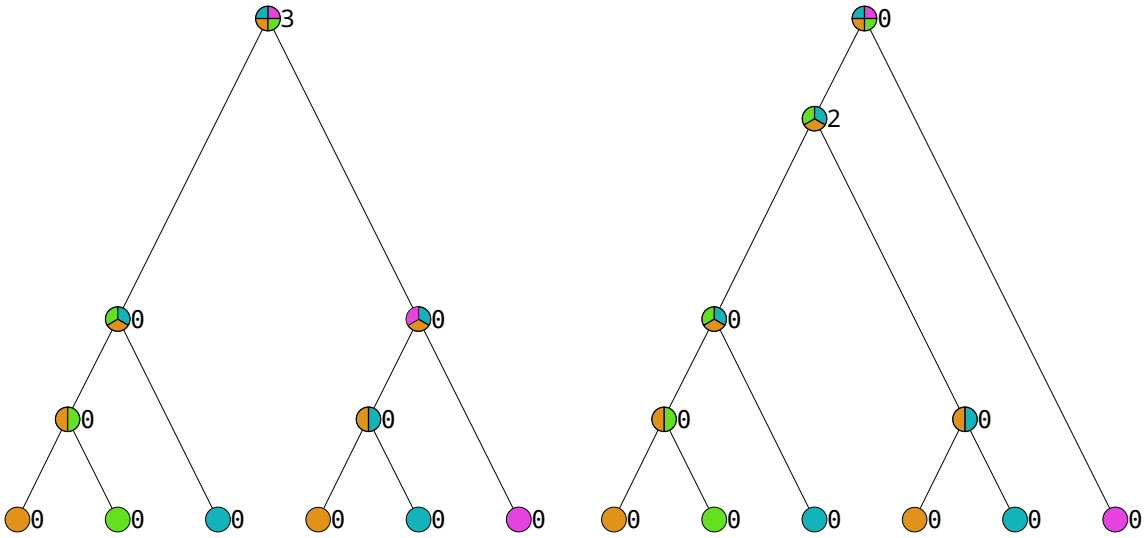
Zum Wurzeln eines Baumes wird der Baum von APro nacheinander an jedem Knoten gewurzelt und von dieser temporären Wurzel aus markiert. Der gesuchte Baum ist derjenige, bei dem die Wurzel den kleinsten *Score* S hat. Der Score $S(k_w)$ eines Knoten k in dem Baum mit der Wurzel w (Formel 2.4) kann rekursiv berechnet werden. Dabei ist $S^l(k)$ der zuvor berechnete Score des linken Kindes von k und $S^r(k)$ der des Rechten. Ein Knoten k wird als Duplikationsereignis markiert, wenn die Gene im linken Teilbaum von den gleichen Arten stammen, wie die des rechten Teilbaumes, also $S(k_w) \neq S^l(k_w) + S^r(k_w)$. Da aber auch Genverluste auftreten können, trifft nicht immer dieser Fall ein. APro unterscheidet deswegen zwischen „vollständiger“ Genduplikation (+1) und „teilweiser“ Genduplikation (+2; +3). Die Arten, von denen die Gene eines Baumes mit Wurzel k stammen, werden als *abgedeckt* bezeichnet und in der Menge $\mathcal{C}(k)$ beschrieben. Entsprechend sind $\mathcal{C}^l(k)$ und $\mathcal{C}^r(k)$ die abgedeckten Arten des linken beziehungsweise rechten Teilbaumes unter k.

In Abb. 2.5 sind zwei Bäume dargestellt, die APro nacheinander beim Wurzeln und Markieren des Genbaums aus Abb. 2.3b überprüft. Die Farben eines Knoten u repräsentieren die abgedeckten Arten $\mathcal{C}(u)$. Solange die zwei Kindknoten des Knotens u keine gemeinsame Farbe enthalten, gilt $S(u) = 0$. Für die Wurzel w des linken Baumes (Abb. 2.5a) überlappen sich die Farben, es liegt somit ein Duplikationsereignis vor. Allerdings taucht im linken Teilbaum die Farbe grün und im rechten Teilbaum die Farbe pink auf, $\mathcal{C}^l(w)$ und $\mathcal{C}^r(w)$ sind also nicht gleich. Da zusätzlich beide Kindknoten nicht alle vier Farben des Wurzelknotens enthalten, trifft auch der dritte Fall nicht zu. Somit wird der Score der Wurzel auf $S(w_w) = S^l(w_w) + S^r(w_w) + 3 = 3$ gesetzt. Im Gegensatz dazu hat die Wurzel t des rechten Teilbaums (Abb. 2.5b) den Score $S(t_t) = 0$, da die abgedeckten Arten seiner zwei Kindknoten keine gemeinsamen Arten (Farben) enthalten.

Da ein Baum mit n Blättern insgesamt $2n - 1$ Knoten hat, benötigt das Markieren eines Baumes $\mathcal{O}(n)$ Zeit. Wegen den $\mathcal{O}(n)$ möglichen Wurzeln markiert und wurzelt APro einen Ausgangsbaum in $\mathcal{O}(n^2)$ [32].

2.2.4.2. MAD

Im Gegensatz zu APro kann die *MAD*-Methode (Minimal Ancestral Deviation) [33] keine Duplikation erkennen, sondern nur eine aussagekräftige Wurzel für ungewurzelte Bäume definieren. Da APro allerdings zum Tagging einen gewurzelten Genbaum benötigt, kann dieser auch mit MAD gewurzelt werden und dieser Schritt in APro übersprungen werden. MAD versucht, Heterotachien zu berücksichtigen und eine qualitative Schätzung der besten



(a) Ein mit APro gewurzelter Baum mit den Scores an den Knoten.

(b) Der von APro ausgegebene Baum mit Scores, wenn nur die zwei Bäume in Abb. 2.5 geprüft werden würden.

Abbildung 2.5.: Zwei der Bäume, die APro iterativ überprüft.

Wurzel vorzunehmen [35]. Die ursprüngliche Implementierung von MAD [33] benötigt für einen Baum mit n Blättern $\mathcal{O}(n^3)$ Zeit [35]. Die effizientere Implementierung von Bryan und Charleston [35] benötigt nur $\mathcal{O}(n^2)$ Zeit, dafür $\mathcal{O}(n)$ zusätzlichen Platz. Sie wird in dieser Arbeit verwendet und im Folgenden vorgestellt.

Eine Position ρ im Baum kann entweder ein Knoten oder ein beliebiger Punkt auf einem Ast sein. MAD berechnet für jede Position einen *Ancestor Deviation Score* $S(\rho)$ (Formel 2.5), wobei dieser für die beste Wurzel minimiert werden muss. Dadurch kann, anders als bei APro, jeder beliebige Punkt im Baum als Wurzel ausgewählt werden.

$$S(\rho) = \sqrt{\frac{2}{n(n-1)} \sum_{x,y} \left(\frac{d_{x\rho} - d_{y\rho}}{d_{xy}} \right)^2}. \tag{2.5}$$

Dabei gibt d_{uv} von zwei Knoten u und v die Pfadlänge zwischen ihnen an, während x und y jeweils ein Blatt sind. Bryant und Charleston [35] zeigen, dass $S(\rho)$ zu minimieren äquivalent dazu ist, die strikt konvexe Funktion $G(\rho)$ (Formel 2.6) zu minimieren. Dadurch gibt es nur noch genau ein lokales Minimum, welches dadurch gleichzeitig auch das globale Minimum ist. Deshalb gibt es immer eine eindeutige Lösung, was das Problem vereinfacht. Mithilfe einer Tiefensuche und der Speicherung von $\mathcal{O}(n)$ Zwischenergebnissen kann deshalb in $\mathcal{O}(n^2)$ die optimale Wurzel bestimmt werden.

$$G(\rho) = \sum_{x,y} \left(\frac{d_{x\rho} - d_{y\rho}}{dxy} \right)^2. \tag{2.6}$$

2.3. Methoden zur Genbauminferenz

2.3.1. Distanzbasierte Ansätze

Distanzbasierte Methoden laufen in zwei Schritten ab. Zuerst wird eine Distanzmatrix D, welche die paarweisen Distanzen zwischen allen Sequenzen enthält, berechnet. Danach wird aus dieser Matrix ein Baum berechnet, der je nach verwendeter Methode unterschiedliche Eigenschaften besitzen kann.

2.3.1.1. Berechnung von Distanzmatrizen

Die n^2 paarweisen Distanzen von n zu vergleichenden Einheiten T unter der Metrik δ kann man in einer Distanzmatrix $(d_{ij}) = D \in \mathbb{R}^{n \times n}$ speichern. Für eine feste Reihenfolge $\sigma(T)$ der Elemente gilt $d_{ij} = \delta(\sigma_i(T), \sigma_j(T))$. Erfüllt die Metrik d die Voraussetzungen einer korrekten Distanz, so ist D eine symmetrische Matrix, deren Hauptdiagonalelemente $d_{ii} = 0$ sind.

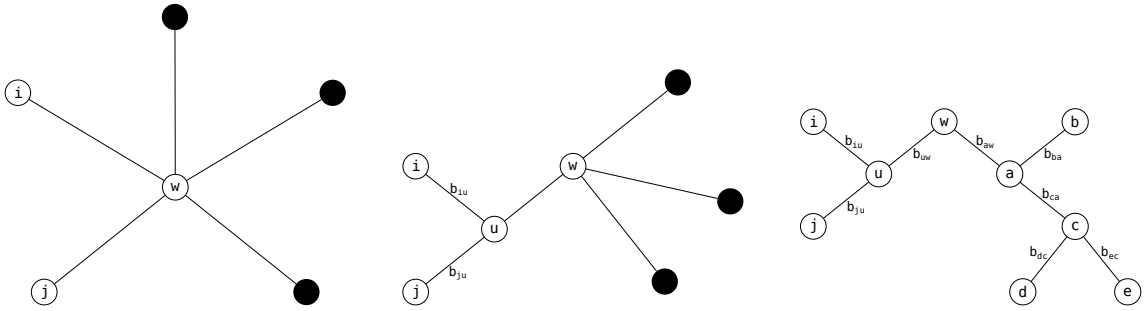
Für die Distanz zwischen zwei DNA-Sequenzen (im Folgenden *Strings* genannt) gibt es mehrere übliche Metriken. Die *Levenshtein-Distanz* misst die minimale Anzahl an Modifikationen eines einzelnen Zeichens - Einfügen, Löschen oder Umbenennen - um den einen String in den anderen umzuwandeln. Die *p-Distanz* ist das Verhältnis zwischen der Anzahl an unterschiedlichen Stellen der Strings und ihrer Länge [36]. Eine weitere Möglichkeit ist es, die Anzahl an Substitutionen (Mutationen) abzuschätzen, die seit der Aufspaltung in unabhängige Strings vergangen sind. Dafür benötigt man ein *Substitutionsmodell*, welches die Wahrscheinlichkeiten, mit denen die Nukleotide durch andere ersetzt werden, angibt [7].

2.3.1.2. Neighbor-Joining-Methode

Der *Neighbor-Joining-Algorithmus (NJ)* [37, 38] verknüpft n Elemente anhand einer Distanzmatrix zu einer binären Baumhierarchie in $\mathcal{O}(n^3)$. Er beruht auf *balanced minimal evolution (BME)* [39], das die Annahme trifft, der Baum mit der geringsten Summe der Astlängen ist am wahrscheinlichsten [40]. Solange eine Distanz (Abs. 2.3.1) auf der Elementmenge definiert wurde, ist NJ nicht auf DNA- oder Proteinsequenzen beschränkt, im Folgenden werden trotzdem Taxa als Eingabe angenommen.

NJ startet mit einem Graphen, bei dem alle Taxa T sternförmig zu einem zusätzlichen Knoten w in der „Mitte“ angeordnet sind und ihrem eigenen Cluster angehören (Abb. 2.6a). Iterativ werden jeweils die zwei Knoten $i, j \in V^{(n)} = T \cup \{w\}$, $i \neq j$ mit der geringsten Distanz gewählt und zu einem Teilbaum mit Wurzel $u \notin V^{(1)}$ zusammengefügt, der weiterhin mit dem Mittelknoten verbunden ist (Abb. 2.6b). Dafür wird zuerst eine neue Matrix $(a_{ij}) = A^{(1)}$ aus der ursprünglichen Distanzmatrix berechnet (Formel 2.7). Dabei gibt $r_t^{(1)} = \frac{1}{N-2} \sum_{k=1}^{N} d_{tk}^{(1)}$ mit $k \in V^{(1)}, N = |V^{(1)}|$ die Netto-Divergenz eines Taxon t zu allen anderen an.

$$a_{ij}^{(1)} = d_{ij}^{(1)} - (r_i^{(1)} + r_j^{(1)}). \tag{2.7}$$



(a) NJ starting tree. (b) NJ after the first iteration. (c) The finished neighbor-joined tree.

Abbildung 2.6.: Beispiel für den NJ-Algorithmus. Gefüllte Knoten werden in dem jeweils dargestelltem Schritt nicht benötogt.

Die Kantenlänge b_{tu}, $t \in \{i, j\}$ der zwei Taxa i und j mit der geringsten Distanz in A zu ihrem neuen Elternknoten $u \notin V^{(1)}$ ist durch Formel 2.8 gegeben.

$$b_{iu} = \frac{d_{ij}^{(1)} + r_i - r_j}{2},$$
$$b_{ju} = d_{i,j}^{(1)} - b_{iu}. \tag{2.8}$$

Dann werden die Distanzen aktualisiert, indem die Distanzen $d_{i,k}^{(1)}$ und $d_{j,k}^{(1)}$ mit $k \in V^{(1)}$ verworfen und neue Distanzen $d_{u,k}^{(2)}$ berechnet werden (Formel 2.9). Außerdem wird die Menge der aktiven Knoten aktualisiert ($V^{(2)} = V^{(1)} \setminus \{i, j\} \cup \{u\}$).

$$d_{uk}^{(1)} = \frac{d_{ik}^{(1)} + d_{jk}^{(1)} - d_{ij}^{(1)}}{2}. \tag{2.9}$$

Diese Abfolge wird ausgeführt, bis die letzten zwei Cluster zu einem zusammengefügt wurden (Abb. 2.6c), der Graph also ein Binärbaum geworden ist und $\| V^{(N-1)} \| = 1$ gilt. Der hochgestellte Index $^{(n)}$ bei r, M, D und V gibt dabei den Schritt n an, in dem diese Werte gültig sind.

2.3.1.3. FastME

FastME [7] ist eine distanzbasierte Methode, die zuerst eine Distanzmatrix mithilfe eines Substitutionsmodells berechnet. Danach wird ein abgewandelter NJ-Algorithmus angewandt und zum Schluss wird der entstehende Baum mithilfe von „Tree-Moves“ korrigiert [7, 41]. Tree-Moves verändern die Baumtopologie, indem zum Beispiel ganze Teilbäume miteinander getauscht werden (NNI) oder einzelne Äste entfernt und an einer anderen Stelle wieder eingefügt werden (SPR).

In FastME soll die Länge $l(T)$ (Formel 2.10) eines Baumes T minimiert werden, da auch FastME auf dem BME-Prinzip beruht. Dabei gibt δ_{ij} die (geschätzte) evolutionäre Distanz

zwischen den Taxa i und j an. Die Distanz d_ij gibt die Anzahl der Äste zwischen i und j an und nicht wie bei MAD die Summe der Länge dieser (Abs. 2.2.4.2).

$$l(T) = \sum_{i,j} 2^{1-d_{ij}} \delta_{ij} \tag{2.10}$$

Im Gegensatz zu NJ berechnet FastME allerdings nicht einen bestimmten Baum, sondern sucht einen optimalen Baum mithilfe von Tree-Moves. Ein Baum T', der durch NNI- oder SPR-Schritte aus T entstanden ist, wird akzeptiert, falls $c(T, T') < 0$ (Formel 2.11).

$$c(T, T') = l(T) - l(T') = \frac{1}{4}[(\delta^T_{AB} + \delta^T_{CD}) - (\delta^T_{AC} + \delta^T_{BD})]. \tag{2.11}$$

Dabei repräsentiert δ^T_{AB} die gewichtete Durchschnittsdistanz zwischen den Taxa der Teilbäume A und B. Besteht A aus n Teilbäumen und B aus m, so ist $\delta^T_{AB} = \frac{1}{n*m} \sum_{i=0}^{n} \sum_{j=0}^{m} \delta^T_{A_i B_j}$. Für zwei Teilbäume C und D, die jeweils nur noch aus einem einzelnen Taxon c beziehungsweise d bestehen, ist $\delta^T_{CD} = \delta_{cd}$ [41].

2.3.2. Zeichenbasierte Verfahren

Im Gegensatz zu distanzbasierten Methoden verwenden zeichenbasierte Methoden (engl. *character-based methods*) nicht die (paarweise) Distanz der Sequenzen, sondern optimieren einen Score, der direkt von den einzelnen Zeichen der Sequenzen abhängt [42] und mithilfe eines Substitutionsmodells (Abs. 2.3.1.1) gebildet wird. Die zwei wichtigsten zeichenbasierten Inferenzmethoden sind *Maximum Parsimony (MP)* und *Maximum Likelihood (ML)*.

MP ist eine Methode zur Rekonstruktion phylogenetischer Bäume, die versucht, den Baum zu finden, der die voliegenden Sequenzen mit der geringsten Anzahl an Mutationen erklärt [43]. Allerdings ist ein großes Problem von Maximum Parsimony das„*Long-Branch-Attraction*“-Phänomen [44]. Dabei können fehlerhafte Bäume berechnet werden. Treten in einem Baum, bei dem die Astlänge die Anzahl der aufgetretenen Mutation widerspiegelt, sowohl lange als auch kurze Äste auf, dann kann die Ähnlichkeit zweier Sequenzen überschätzt werden. Zwei Sequenzen mit langen Ästen können dann durch Zufall zu ähnlichen Sequenzen mutieren. Sie befinden sich in dem inferierten Baum deshalb nahe beieinander, obwohl die zugehörigen Taxa nicht eng miteinander verwandt sind. Die Sequenzen, die auf kurzen Ästen liegen, können im Gegenzug schon durch einige wenige Mutationen als sehr verschieden erscheinen und damit weiter voneinander entfernt in den rekonstruierten Baum eingebaut werden.

ML [45] ist eine statistische Methode, die in vielen Bereichen eingesetzt wird. In der Phylogenetik wird mit ihr und verschiedenen Evolutionsmodellen versucht, einen Baum zu finden, der die Daten am Besten erklärt (Formel 2.12). Die *phylogenetische Likelihood* L eines Baumes T, ist gerade die Wahrscheinlichkeit, mit der T zu den gegebenen *Sequenzalignment* A führt. Bei einem Sequenzalignment werden alle Sequenzen so aneinander angepasst, dass, wenn man sie untereinander schreiben würde, möglichst viele Spalten so wenig Basen wie möglich enthalten.

$$L(T \mid A) = \mathcal{P}(A \mid T). \tag{2.12}$$

Die von ML inferierten Bäume werden heutzutage im Vergleich zu den distanzbasierten Methoden oder MP als am akkuratesten angesehen [46]. Dies liegt zum Beispiel daran, dass der Likelihood-Score eines Baumes auch zulässt, dass eine einzelne Nukleotidbase mehrmals auf einem einzigen Ast mutiert. ML wird zum Beispiel in RAxML-NG [8] eingesetzt.

2.3.2.1. Berücksichtigung des Speziesbaums

Die phylogenetische Likelihood L aus Formel 2.12 berücksichtigt nur die alignierten Sequenzen und nicht den Speziesbaum. Die *Joint Likelihood*

$$\tilde{L}(T,S \mid A) = L(T \mid A)\mathcal{P}(T \mid S)$$

dagegen ist das Produkt der Wahrscheinlichkeit, dass der Speziesbaum unter einem gegebenem Substitutionsmodell zum Genbaum T führt und der phylogenetischen Likelihood [1]. Dadurch können beim Auftreten von Duplikationen oder Verlusten meist genauere Genbäume inferiert werden als bei andere Methoden. Die fehlenden Informationen durch zu ähnliche Sequenzen können mithilfe des Speziesbaums ergänzt werden und so eine Evolutionsgeschichte abbilden, die näher an der der zugehörigen Spezies ist.

MP, ML und FastME inferieren einen Baum, indem sie den Tree-Space nach dem besten Baum durchsuchen. FastME benutzt verschiedene Heuristiken [7], um den Tree-Space schneller zu durchsuchen. Dahingegen benötigt ein einzelner Parsimony- oder Likelihood-Score bereits $\mathcal{O}(|Zeichen| * |Taxa|)$ Zeit. Da Zeichenbasierte Methoden zusätzlich auch auf größeren Datenmengen arbeiten und nicht nur auf n^2 vielen Distanzen, ist ihre Laufzeit asymptotisch bedeutend größer als die Distanzbasierter Methoden. Dafür ist der inferierte Baum meist genauer [42].

2.4. Verwandte Arbeiten

RAxML-NG, FastME und IQ-Tree [47] können Genbäume mithilfe der alignierten Gensequenzen inferieren. Durch die geringe Länge einer Gensequenz im Vergleich zur Länge eines ganzen Genoms, wie es bei der Rekonstruktion eines Speziesbaumes verwendet werden kann, besitzen all diese Methoden für eine genaue Genbaumrekonstruktion oft zu wenig Information [48] und sind deshalb für die Inferenz von Genbäumen nicht so gut geeignet wie Verfahren, welche die Genbäume mithilfe ihres Speziesbaumes korrigieren [1].

Andere Arbeiten beschäftigen sich nicht mit ML sondern mit MP. So kann ecceTERA [49] nicht nur Genbäume in einen Speziesbaum eingliedern, sondern auch mithilfe der Joint-Likelihood den Speziesbaum berücksichtigende Genbäume inferieren.

PHYLDOG [50] und GSR [51] sind zwei probabilistische Modelle, die auf unterschiedliche Weisen Genbäume unter Berücksichtigung ihres Speziesbaumes inferieren. PHYLDOG kann dabei mehrere Genbäume und den Speziesbaum gleichzeitig aus den DNA-Sequenzen ermitteln. Beide Arbeiten bestätigen, dass der Speziesbaum wertvolle Informationen für die Rekonstuktion eines Genbaumes enthält und Tools wie RAxML-NG oder PhyML [52], welches auch auf ML beruht, ungenauere Ergebnisse erzielen, da der Speziesbaum nicht berücksichtigt wird [6].

Ein weiteres wichtiges Softwaretool ist ALE [53]. Es nimmt als Eingabe einen Speziesbaum sowie eine Menge an Genbäumen und ihren Wahrscheinlichkeiten, die mithilfe eines bayesischen Schätzers wie MrBayes [54] aus den Gensequenzen rekonstruiert wurden. Danach werden die einzelnen Genbäume mithilfe der Joint Likelihood unter Berücksichtigung ihres Speziesbaumes zusammengefügt. ALE wurde beispielsweise in [55] verwendet, um die von ausgestorbenen Spezies gebildeten Proteine rekonstruieren zu können.

GeneRax [1] basiert wie RAxML-NG [8] auch auf ML, aber verwendet die modifizierte Joint-Likelihood (Abs. 2.3.2). Der größte Vorteil gegenüber RAxML-NG liegt allerdings in der Verwendung des Speziesbaumes. GeneRax ist eines der führenden Korrektur- und Rekonstruktionsmethoden für Genbäume.

3. Spearfish

3.1. Arbeitsweise von Spearfish

Im Verlauf dieser Arbeit wurde *Spearfish* entwickelt. Spearfish steht für „**SPE**cies tree **A**ware gene t**R**ee in**F**erence with d**IS**tance met**H**ods“ (dt. *„Distanzbasierte Inferenz von Genbäumen unter Berücksichtigung ihres Speziesbaumes“*). Spearfish ist ein Programm, welches mehrere Teilschritte und -programme ineinander vereint. Im Folgenden soll nun ein Überblick über die einzelnen Schritte gegeben werden.

Spearfish erfordert die Bereitstellung des Speziesbaumes, seiner Astlängen, welche die durchschnittliche Anzahl an Substitutionen pro Base angeben, und aller zugehörigen Gensequenzen als Eingabe. In einem ersten Schritt erfolgt eine Vorverarbeitung dieser Eingabedaten, bei der diese in zwei separate Distanzmatrizen umgerechnet werden (Abs. 3.1.1), die Speziesbaummatrix $(s_{ij}) = S$ und die Gensequenzmatrix $(g_{ij}) = G$. Anschließend wird ein vorläufiger Startbaum (Abs. 3.1.2) konstruiert, der als Ausgangspunkt für die spätere Korrektur dient. Dafür werden die Knoten dieses Startbaumes analysiert, um Duplikations- oder Artbilungsereignisse zu identifizieren und zu markieren. Abhängig von den Markierungen des Startbaumes, werden spezifische Einträge in G ausgewählt (Abs. 3.1.3). Diese werden mithilfe der entsprechenden skalierten Einträge in S korrigiert (Abs. 3.1.4). Das Ergebnis dieser Korrekturen ist eine modifizierte Distanzmatrix G', die im Anschluss an FastME [7] übergeben wird, um den Genbaum zu berechnen.

Da in den Experimenten (Kap. 4) kein einzelner, optimaler Skalierungsfaktor erkennbar wurde, der immer zu den akkuratesten inferierten Bäumen führt, werden mehrere Genbäume unter Verwendung verschiedener Skalierungsfaktoren für S berechnet. Um den endgültigen Genbaum zu ermitteln, evaluiert (Abs. 3.1.5) Spearfish alle erzeugten Bäume mit GeneRax [1] und wählt den mit dem höchsten Joint-Likelihood-Score aus.

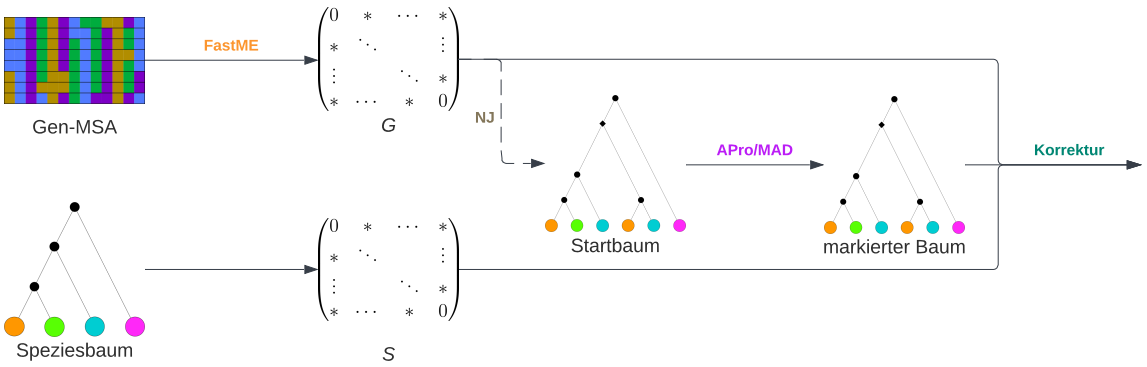


Abbildung 3.1.: Spearfish (Teil 1): Berechnung von G aus dem Multiplen Sequenzalignment (MSA, oben), S (unten) und der zu korrigierenden Elemente.

3.1.1. Berechnung der Distanzmatrizen

Die Distanzmatrix S der Pfadlänge im Speziesbaum, gibt ein Maß für den Verwandschaftsgrad der Spezies. Da Spearfish Astlängen nicht selber schätzen kann, müssen diese vorgegeben werden.

Die Distanzmatrix G wird mithilfe von FastME berechnet (Abb 3.1). FastME verwendet entweder die reine p-Distanz oder ein Modell für die Mutationsraten der einzelnen Basen (Abs. 2.3.1.1), um die paarweisen Distanzen der DNA-Sequenzen zu berechnen.

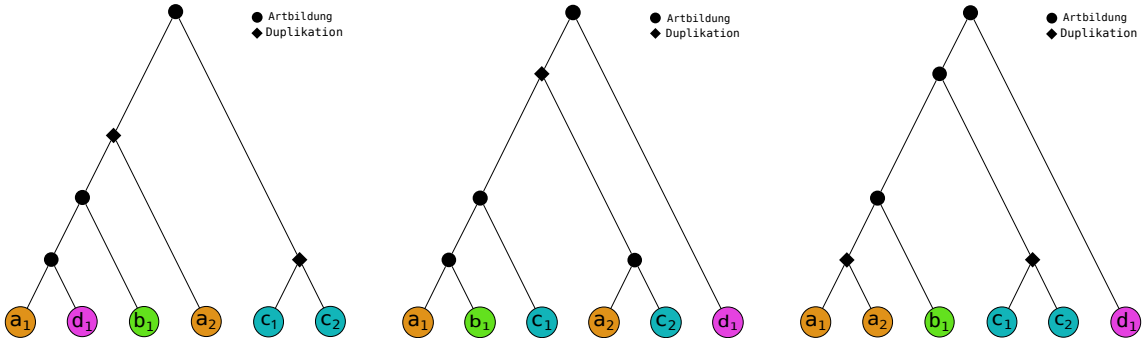
3.1.2. Berechnung des Startbaumes

Der Startbaum T_{st} dient als Referenz für APro oder MAD. Er wird mithilfe von NJ aus G berechnet, wobei jede Zeile der Matrix einem Gen der Genfamilie entspricht und somit ein Blatt pro Zeile angelegt wird. Zusätzlich zu der Länge der Äste wird für jedes Blatt die zugehörige Spezies hinterlegt, da diese Information von APro benötigt wird. Auch dieser Schritt von Spearfish ist in Abb. 3.1 dargestellt.

Da die Gensequenzmatrix für die Rekonstruktion des Startbaums nicht korrigiert wird, hätte FastME auch verwendet werden können, um nicht nur die Matrix G, sondern direkt den Baum T_{st}, zu berechen. Dagegen steht allerdings, dass Spearfish nicht nur die Rekonstruktion eines unkorrigierten Startbaumes unterstützt, dies wird in Abs. 3.2 weiter ausgeführt. Außerdem zeigen die Experimente, dass FastME nicht alle getesteten Astlängen unterstützt (Abs. 4.2.3). Durch einen eigenen NJ-Algorithmus (Abs. 2.3.1.2) kann Spearfish so in mehr Fällen als FastME verwendet werden. Dadurch benötigt dieser Schritt $\mathcal{O}(|Gentaxa|^3)$ lange.

3.1.3. Auswahl der zu korrigierenden Elemente

Alle Duplikationsereignis werden im Speziesbaum nicht dargestellt. Deshalb kann die Verwandtschaft zweier paraloger Gene nicht mithilfe des Speziesbaumes rekonstruiert werden. Die zwei Gene können zwar unterschiedlichen Spezies angehören, deren Artbildungsereignis im Speziesbaum sichtbar ist, allerdings kann der Verwandtschaftsgrad der beiden Spezies stark von dem der zwei Gene abweichen. Ein Beispiel dafür ist in Abb. 3.2c gegeben. Der



(a) **Unterkorrektur.** Sehr ähnliche Sequenzen (wenige Mutationen) resultieren in einer von vielen Topologien, die nicht die Verwandschaft der Spezies berücksichtigen.

(b) **Echter Genbaum** aus Abb. 2.3b.

(c) **Überkorrektur.** Alle Genduplikationen einer Spezies werden zusammengruppiert, da der Speziesbaum zu stark berücksichtigt wird.

Abbildung 3.2.: Probleme der Über- und Unterkorrektur.

dargestellte Genbaum ist das Ergebnis einer extremen Überkorrektur durch den Speziesbaum. Die Gene a_1 und a_2 der Art A, sowie die Gene der Art C sind im echten Genbaum aus Abb. 3.2b jeweils weit voneinander entfernt. Da a_1 und a_2 allerdings in der selben Art vorkommen, ist die Distanz ihrer Arten im Speziesbaum 0. Wird das Verwandschaftsverhältnis der vier Gene nun mithilfe der Distanzen aus dem Speziesbaum korrigiert, kommt es zur Überkorrektur, wenn der Speziesbaum zu stark gewichtet wird. Alle Gene einer Art werden zuerst zusammengruppiert, wodurch sich Gen- und Speziesbaum stark ähneln.

Im Gegensatz dazu stehen Sequenzen, deren JGV ein Artneubildungsereignis war und die wenige Mutationen erfahren haben. Dadurch ist ihre Verwandschaft nur anhand der Sequenz schwer zu rekonstruieren. In diesen Teilbäumen bildet der Speziesbaum eine gute Grundlage für den Genbaum und erhöht die Genauigkeit, mit der die Verwandtschaft rekonstruiert werden kann. Die Abb. 3.2a zeigt hierzu ein Gegenbeispiel. Angenommen, alle Gene sind ähnlich mutiert und das Gen a_1 und d_1 haben sich, obwohl ihre Arten sich schon lange aufgetrennt haben, in dieselbe Richtung entwickelt. Die Distanz zwischen den Gensequenzen von a_1 und d_1 ist somit kleiner als die aller anderen. Ohne Berücksichtigung des Speziesbaumes werden sie deshalb im Genbaum zusammengruppiert. Das Gleiche passiert auch mit den zwei Genen der Art C. Der daraus entstehende Genbaum kann, wie in diesem Fall, stark vom echten Genbaum abweichen.

Eine Korrektur, die sehr geringe Distanzen in S (Überkorrektur), sowie sehr geringe Distanzen in G (Unterkorrektur), zu stark berücksichtigt, erhöht die Rate der falsch rekonstruierten Verwandschaften und somit die Ungenauigkeit des inferierten Baumes. Existiert ein mit Genduplikation und Artneubildungen markierter Baum, so sollten deshalb nur die Genpaare mit einem Artneubildungsereignis als JGV korrigiert werden. In Spearfish kann der JGV zweier Knoten in $\mathcal{O}(|Baumhöhe|)$ berechnet werden. Um diese Auswahl in Spearfish treffen zu können, wird der Startbaum benötigt. Dieser kann entweder mit MAD oder APro in $\mathcal{O}(|Taxa|^2)$ markiert werden oder unmarkiert gelassen werden (Abb. 3.1).

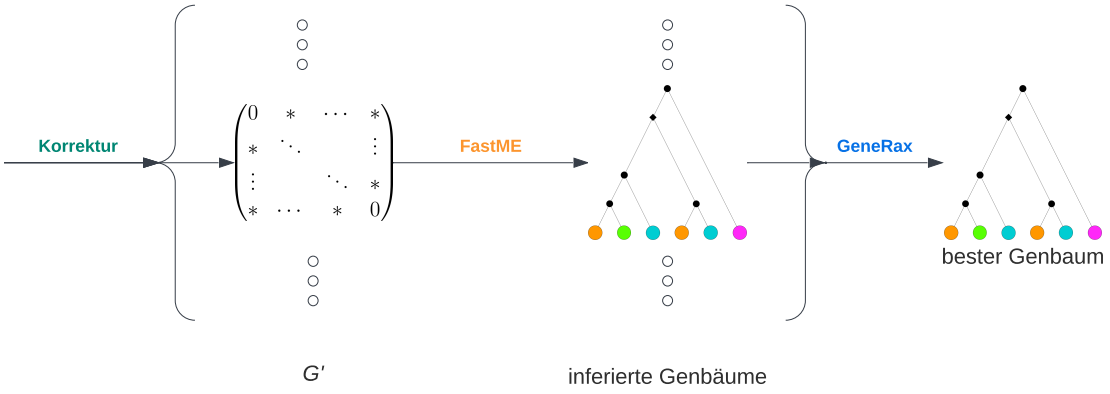


Abbildung 3.3.: Spearfish (Teil 2): Korrektur der Gensequenzdistanzmatrix G und Rekonstruktion mit darauf folgender Evaluatation des besten Baumes.

Ein unmarkierter Baum ist gleichbedeutend mit einem Baum, bei dem alle Knoten als Artneubildungsereigniss markiert sind.

3.1.4. Korrektur

Die Korrektur der Distanzmatrix geschieht elementweise. Dabei wird ein Eintrag g_{ij} nur dann korrigiert, wenn der JGV der zwei Gene i und j im vorherigen Schritt als Artneubildungsereignis markiert wurde. Wurde der Startbaum nicht markiert, so werden nun alle Matrixeinträge korrigiert. Die neue Distanz $g'_{\lambda,ij}$ in der korrigierten Distanzmatrix $(g'_{\lambda,ij}) = G'_\lambda$, wird mit Formel 3.1 berechnet. Dabei sind $\tilde{i}$ und $\tilde{j}$ die Indizes der Spezies in S, denen die Gene i und j angehören. Diese Zuordnung wird gleichzeitig mit der Berechnung der Matrizen S und G erstellt. Der Faktor λ gehört zu einer Menge an Skalaren Λ, die die Skalierung der Speziesmatrix angeben. Für jedes $\lambda \in \Lambda$ wird eine eigene Matrix G'_λ berechnet, wobei wiederum FastME Für jede dieser Matrizen einen Genbaum T_λ inferiert (Abb. 3.3). Da die elementweise Korrektur einer Matrix mit n Gentaxa $\mathcal{O}(n^2)$ Zeit kostet und die Rekonstruktion eines Baumes mit FastME $\mathcal{O}(kn^2)$ [7], wobei k die Anzahl an NNI- und SPR-Zügen ist, benötigt der Korrekturschritt von Spearfish $\mathcal{O}(|\Lambda|kn^2)$ Zeit.

$$g'_{\lambda,ij} = \begin{cases} g_{ij} + \lambda * s_{\tilde{i}\tilde{j}}, & \text{falls JGV von } i \text{ und } j \text{ Artbildungsereignis,} \\ g_{ij}, & \text{sonst.} \end{cases} \tag{3.1}$$

3.1.5. Evaluation

Da die Menge aller von FastME inferierten Genbäume T_λ nach dem letzten Schritt mehr als einen Baum enthält, Spearfish allerdings nur einen Baum am Ende ausgeben soll, müssen die Bäume noch evaluiert werden. Dafür wird mithilfe von GeneRax der Joint-Likelihood-Score (Abs. 2.3.2) für jeden inferierten Baum berechnet. Dabei besitzt ein Baum mit einem höheren Score eine höhere Wahrscheinlichkeit, dem echten Baum ähnlicher zu sein, als ein Baum mit einem niedriegerem Joint-Likelihood-Score. Deshalb sucht Spearfish den Baum mit dem maximalen Score und gibt diesen am Ende aus (Abb. 3.3). Dieser Schritt benötigt $\mathcal{O}((|Zeichen| + |Speziestaxa|) * |Gentaxa|)$ pro Baum.

3.1.6. Laufzeitanalyse

Für n Gentaxa, m Speziestaxa, k FastME-Iterationen, einer maximalen Sequenzlänge von l und einer Startbaumhöhe von h benötigt Spearfish maximal

$$\begin{aligned} &\mathcal{O}(n^3 + n^2 + n^2h + n^2 + |\Lambda|kn^2 + (l+m)n) \\ &= \mathcal{O}(n^3 + |\Lambda|(kn^2 + (l+m)n)) \end{aligned} \tag{3.2}$$

viel Zeit. Ohne den Startbaum zu berechnen und zu markieren, kann die asymptotische Laufzeit auf $\mathcal{O}(|\Lambda|(kn^2 + (l+m)n))$ verkürzt werden.

3.2. Anpassungsmöglichkeiten

Abgesehen von den verwendeten Substitutionsmodellen zur Berechnung und dem Speziesbaum kann Spearfish auf mehrere Arten modifiziert werden, die in den folgenden Experimenten nicht getestet wurden. Diese sind entweder

- die Angabe eines bereits existierendes Startbaum für die Auswahl der Korrekturpaare,
- eine Änderung des Skalierungsfaktors zur Berechnung der Distanzmatrix für den Startbaum oder
- eine Änderung der Skalierungsfaktoren für die Korrektur.

Durch die Angabe eines Startbaumes kann die Methode iterativ ausgeführt werden, wobei die Ergebnisse des vorherigen Laufs in den nächsten einfließen. Die Änderung der Skalierungsfaktoren geschieht frei nach Ermessen, wobei der Abstand der Korrekturfaktoren enger oder weiter je nach gewünschter Laufzeit gewählt werden sollten.

4. Experimente & Ergebnisse

4.1. Experimenteller Aufbau

Für die Experimente wurde das „Haswell-Cluster“ des Heidelberg Institute for Theoretical Studies (HITS) verwendet. Jede getestete Methode lief jeweils auf einer Node. Eine Node besteht aus jeweils zwei Intel Haswell CPUs (E5-2630v3) mit einer Taktrate von 2.40GHz und 64GB RAM. Jede CPU besitzt acht physische Kerne und unterstützt HyperThreading, wodurch theoretisch 32 Threads parallel auf jeder Node ausgeführt werden können. Für die Experimente wurde HyperThreading jedoch nicht verwendet, wodurch jede Methode 16 Threads zur Verfügung standen, allerdings wurden alle Tools mithilfe von MPI parallelisiert.

4.1.1. Getestete Methoden

Die in Spearfish verwendeten Teilschritte können nach Belieben variiert werden. In dieser Arbeit wurden mehrere Varianten getestet, wobei besonderer Wert auf die verschiedenen Algorithmen zur Elementauswahl gelegt wurde. Zur Berechnung der Gensequenz-Distanzmatrizen wird in jeder der getesteten Varianten die p-Distanz (Abs. 2.3.1.1) verwendet und somit der Einfluss eines Substitutionsmodells nicht getestet. Der Startbaum, der verwendet wird, um Genpaare als paralog oder ortholog zu markieren, wird immer mithilfe meiner eigenen Implementierung des NJ-Algorithmus (Abs. 2.3.1.2) aus der unkorrigierten Gensequenzmatrix inferiert. Auch hier wurde der Einfluss eines anderen Startbaumes, zum Beispiel der inferierte Baum aus einem vorherigen Lauf, nicht getestet, obwohl Spearfish dies unterstützt. Im Folgenden wird eine spezifische Variation von Spearfish über den String $tagTI_{|\Lambda|}$ definiert, der nur die variierten Methoden spezifiziert.

Dabei gibt *tag* den Markierungsalgorithmus an und kann einen der drei Werte $apro, mad$ oder *all* annehmen. Wird der Startbaum mit dem APro-Algorithmus (Abs. 2.2.4.1) gewurzelt und markiert, so gilt $tag = apro$. Wird hingegen MAD (Abs. 2.2.4.2) verwendet, um den Baum zu wurzeln und danach APro, um den gewurzelten Baum zu markieren, so ergibt sich *mad*. Der Algorithmus *all* bezeichnet die Spearfish-Variante, bei welcher der Startbaum nicht markiert wird und alle Genpaare korrigiert werden.

Parameter	Wert
Artneubildungsrate	5e−9
Artaussterberate	4,9e−9
Anzahl Spezies	25
Anzahl Genbäume pro Speziesbaum	100
Astlängenfaktor	1
Genduplikationsrate (Ereignisse/Generation)	4,9e−10
Genverlustrate	4,9e−10
Gentransferrate	0
Genkonversionsrate	0
Effektive Populationsgröße	10
Generationen	$\mathcal{LN}(21{,}25; 0{,}2)$
Globale Substitutionsrate	$\mathcal{LN}(-21{,}0; 0{,}1)$
Spezies-spezifische Heterogenitäts-Gammastruktur	$\mathcal{LN}(1{,}4; 1)$
Gen-spezifische Heterogenitäts-Gammastruktur	$\mathcal{LN}(1{,}551533; 0{,}6931472)$
Genbaumast-spezifische Gammastruktur	$\mathcal{LN}(1{,}5; 1)$

Tabelle 4.1.: Verwendete Ausgangsparameter für SimPhy.

Der Inferierungsalgorithmus TI der korrigierten Genbäume ist entweder FM für FastME (Abs. 2.3.1.3) oder NJ für die NJ-Implementierung von Spearfish, die auch für die Inferenz des Startbaumes genutzt wird. Die Rekonstruktion mit NJ wurde zu Beginn der Experimente getestet. Da die dabei inferierten Bäume allerdings weniger akkurat als die von FM waren und sich die gleichen Muster abbildeten, wurde NJ im weiteren Verlauf der Experimente nicht weiter getestet.

Da in den Experimenten Spearfish die Distanzmatrix zuerst mit 80 und später mit 10 Skalierungsfaktoren (Abs. 3.1.4) getestet wurde, gibt $|\Lambda|$ die Anzahl der Faktoren an, um zwischen den zwei ansonsten identischen Varianten unterscheiden zu können. Die ersten 80 Faktoren wurden anhand der Ergebnisse aus ersten Testläufen festgelegt und befinden sich in dem Intervall $[0, 100]$. Mit größer werdenden Werten wurde der Abstand zwischen ihnen vergrößert.

Damit definiert zum Beispiel $madFM_{80}$ die Variante, bei welcher der unkorrigierte Startbaum mithilfe des MAD-Algorithmus gewurzelt wird und deren 80 korrigierte Matrizen von FastME zu Bäumen rekonstruiert werden.

Die verschiedenen Variationen von Spearfish wurden nicht nur mit sich selbst, sondern auch mit den inferierten Bäumen von GeneRax, RAxML-NG und FastME, verglichen.

4.1.2. Simulierte Datensätze

Die entwickelten Methoden wurden auf mehreren simulierten Datensätzen getestet und mit anderen Softwaretools verglichen. Ein Datensatz besteht aus einer Menge an Variationen einiger weniger Parameter. Jede dieser Variationen beinhaltet je 100 Simulationen, die mit denselben Parametern generiert wurden. Am Beispiel des DUPLOS-Datensatzes, in dem die Genduplikations- und die Genverlustrate 5 unterschiedliche Werte annehmen kann (Tab. 4.2), bedeutet das, dass dieser aus $5 * 100 = 500$ einzelnen Simulationen besteht. Die Daten wurden mithilfe der Simulationsprogramme SimPhy [56] und INDELible [57]

Datensatz	Parameter	Wert
SPECIES	Anzahl an Spezies	15; **25**; 50; 75; 100
BRALEN	Astlängenfaktor	0,0; 0,5; **1**; 2; 3
DUPLOS	Faktor der Genduplikationsrate Faktor der Genverlustrate	0,0; 0,5; **1**; 2; 3

Tabelle 4.2.: Variierte Parameter je Datensatz für SimPhy. Fettgedruckte Werte sind die Parameter des BASE-Datensatzes.

Parameter	Wert
DNA-Sequenzlänge	50; **100**; 250; 1000
Basengrundfrequenzen	Dirichlet (A = 36, C = 26, G = 28, T = 32)
Basenübergangsraten	Dirichlet (TG = 5, TC = 16, TA = 3, GC = 6, GA = 15, CA = 5)

Tabelle 4.3.: Verwendete INDELible-Simulationsparameter, die DNA-Sequenzlänge wird im Datensatz SITES variiert. Der fettgedruckte Wert ist der Parameter des BASE-Datensatzes.

generiert. Die gewählten Parameter basieren auf denen der Experimente, mit welchen SpeciesRax [58] getestet wurde. SimPhy erstellt zuerst Spezies- und Genbäume mit den Parametern aus den Tabellen 4.1 und 4.2, woraufhin INDELible dann DNA-Sequenzen auf diesen Genbäumen abhängig von den Parametern in Tab. 4.3 simuliert.

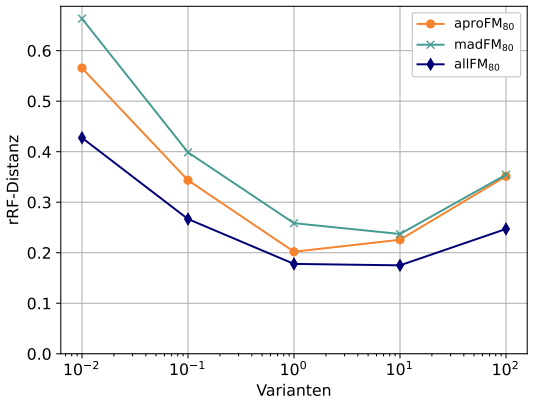
Da SimPhy die simulierten Daten nicht validiert, können Genfamilien auftreten, bei denen zu wenige Informationen vorliegen, um einen Genbaum sinnvoll rekonstruieren zu können. Deshalb werden alle Genfamilien mit weniger als vier unterschiedlichen Gensequenzen von allen Berechnungen ausgeschlossen. Über alle 1600 simulierten Datensätze hinweg, werden so ungefähr fünf Prozent aller Familien nicht berücksichtigt.

4.2. Ergebnisse auf den simulierten Datensätzen

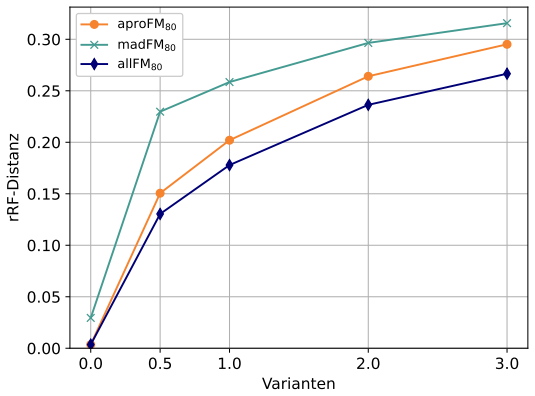
Im Folgenden werden die Ergebnisse auf den einzelnen Datensätzen näher besprochen. Dabei liegt der Fokus hauptsächlich auf der Genauigkeit, mit der die einzelnen Methoden einen Genbaum inferieren. Dabei wird jeweils die durchschnittliche rRF-Distanz (Abs. 2.2.3) zwischen dem echten Genbaum und aller inferierten Bäume der Methode je Datensatz-Variation berechnet. Um zu beurteilen, inwieweit es im Laufe dieser Arbeit gelungen ist, mithilfe von distanzbasierten Methoden unter Berücksichtigung des Speziesbaumes eine akkurate und gleichzeitig schnelle Methode zu entwickeln, werden zusätzlich die Laufzeiten der einzelnen Methoden auf dem Cluster miteinander verglichen.

4.2.1. Die Standardparameter

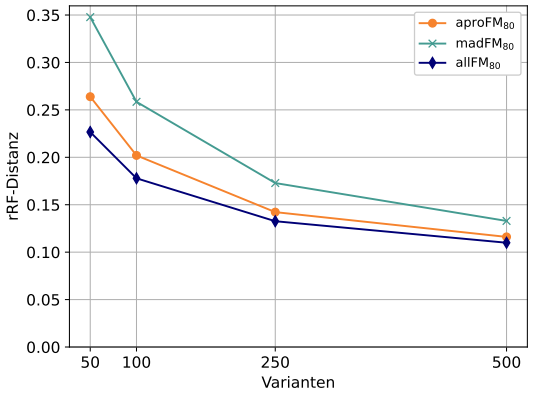
Jeder Datensatz basiert auf den selben Standardparametern. Diese werden je nach Experiment verändert, wobei immer nur ein Parameter je Datensatz variiert wird. Eine Außnahme bildet der DUPLOS-Datensatz, da die Genverlust- und Genduplikationsrate immer gemeinsam verändert werden. Die Standardwerte sind in den beiden Tabellen 4.2 und 4.3 gegeben und sind, bei einer Liste von Werten, jeweils die fettgedruckten. Da die



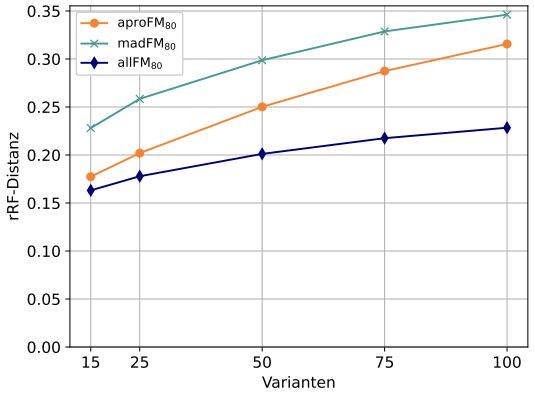
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)



(c) Variierte Sequenzlängen (SITES)



(d) Variierte Anzahl an Spezies (SPECIES)

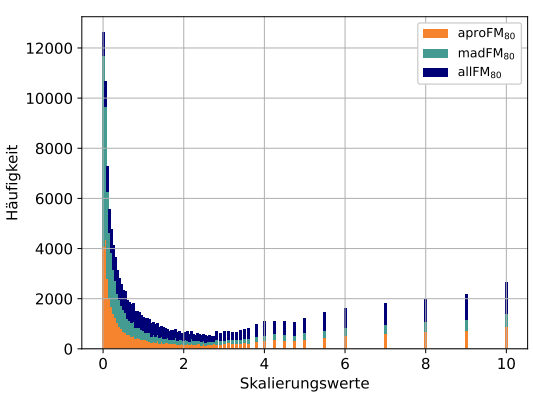
Abbildung 4.1.: Durchschnittliche rRF-Distanzen zwischen dem inferierten und dem echten Genbaum.

Faktoren der Genduplikations- und Genverlustraten standardmäßig auf eins gesetzt sind, finden in jedem Datensatz Duplikationen und Verluste statt.

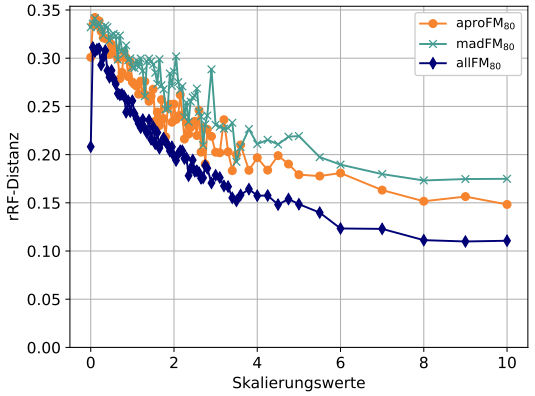
4.2.2. Einfluss des Markierungsalgorithmus auf die Genauigkeit

Als erstes wurden drei Variationen von Spearfish getestet, bei denen jeweils der Markierungsalgorithmus verändert wurde. Dafür wurde jeweils derselbe Startbaum verwendet und anschließend 80 verschiedene korrigierte Distanzmatrizen berechnet. Die durchschnittliche Distanzvon $allFM_{80}$ über alle Datensätze beträgt dabei 0,197 und 0,240 beziehungsweise 0,285 für $aproFM_{80}$ und $madFM_{80}$ respektive.

Im Datensatz BRALEN, bei dem der Faktor der Astlängen und somit die Anzahl der aufgetretenen Mutationen verändert wurde, ermöglichen eine mittlere Anzahl an Mutationen bei einem Astlängenfaktor von eins und zehn die beste Rekonstruktion mit einer durchschnittlichen Distanz von 0,176 für $allFM_{80}$, 0,214 für den APro- und 0,248 für den MAD-Markierungsalgorithmus (Abb. 4.1a). Die Ergebnisse des DUPLOS- und des SPECIES-Datensatzes (Abb. 4.1b und 4.1d) sind wenig überraschend, da mit steigender Komplexität der Eingaben durch erhöhte Duplikations- und Verlustrate oder einer größeren Anzahl an Arten die Genbauminferenz schwieriger wird. Ein ähnliches Bild zeigt sich auch



(a) Anzahl der Bäume mit dem Skalierungsfaktor λ, die von GeneRax ausgewählt wurden.



(b) Durchschnittliche rRF-Distanz der Bäume aus Abb. 4.2a zum echten Genbaum.

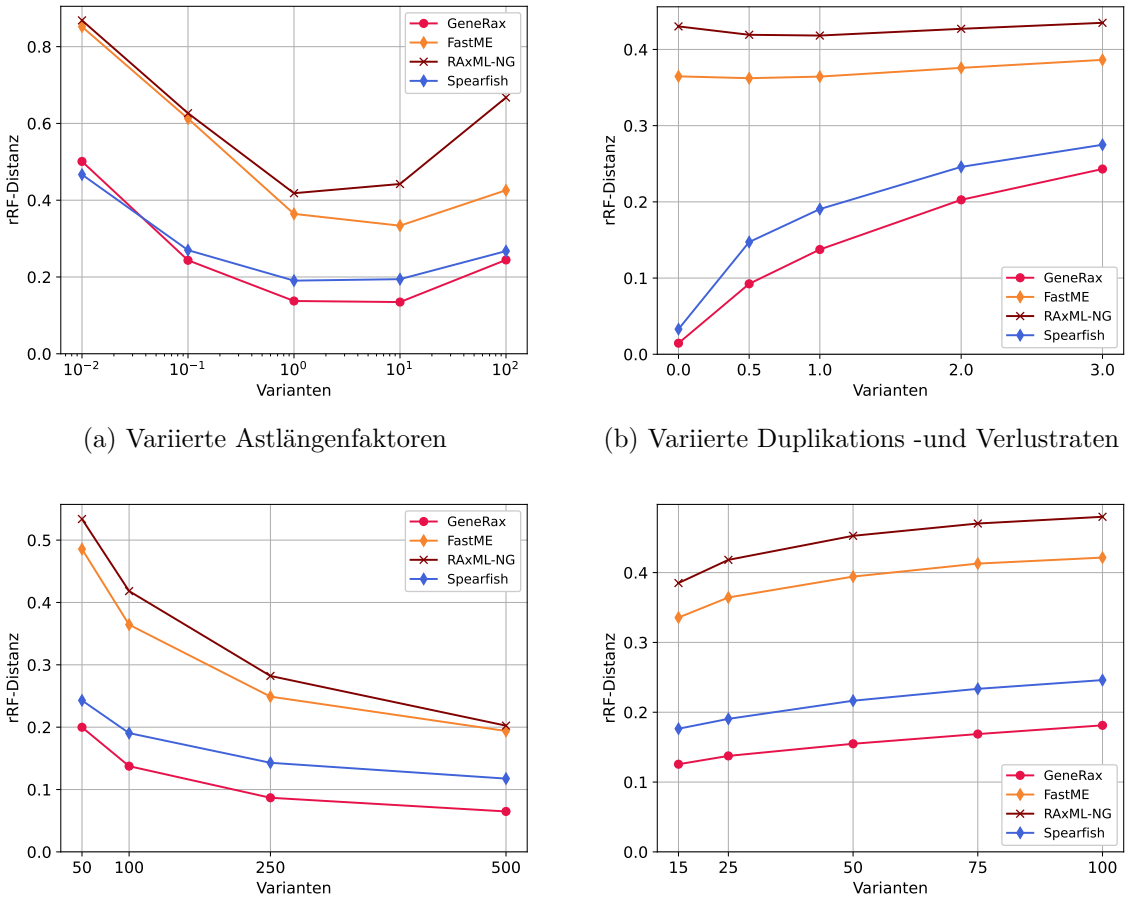
Abbildung 4.2.: Genauere Einsicht in die Verteilung der Skalierungsfaktoren der Distanzmatrizen im SPECIES-Datensatz.

durch eine Verlängerung der DNA-Sequenzen in Abb. 4.1c. Dadurch sind diese tendenziell verschiedener voneinander, wodurch sich die Distanzen ähnlicher Sequenzen stärker von denen anderer DNA-Sequenzen unterscheiden.

Im Vergleich sind die Bäume, bei denen jeder Eintrag ihrer Distanzmatrix korrigiert wurde, in jedem Datensatz genauer. Insbesondere bei einer höhreren Anzahl an Arten werden um einiges bessere Ergebnisse erzielt. Während bei 15 Arten die *apro*-Variante eine nur 8,75% schlechtere Durchschnittsdistanz besitzt, so sind die inferierten Bäume bei 100 Arten ungefähr 38,19% ungenauer.

Die Korrektur von 80 Distanzmatrizen, die Rekonstruktion der Genbäume mithilfe dieser und insbesondere die nachfolgende Evaluierung aller 80 Bäume mithilfe von GeneRax benötigt durchschnittlich 10min55s. Im Vergleich dazu benötigt GeneRax zur Rekonstruktion eines genaueren Baumes nur 2min22s. Da die Laufzeit von Spearfish mit 80 Skalierungsfaktoren somit zu groß ist, wurden von diesen 80 Faktoren zehn Stück ausgewählt. Die Auswahl wurde anhand der zwei Diagramme in Abb. 4.2 getroffen, die für jeden Datensatz ähnlich aussehen. Beide Diagramme wurden auf einen maximalen Skalierungsfaktor von zehn begrenzt, um die Lesbarkeit zu erhalten.

In Abb. 4.2a wird kumulativ die Anzahl der Bäume dargestellt, deren Distanzmatrix aus der Korrektur mit dem Faktor λ entstanden ist und die den besten Joint-Likelihood-Score in ihrer Familie besitzen. Dabei ist eine starke Tendenz zu sehr kleinen Faktoren sichtbar. Die Bäume, welche aus den mit den fünf kleinsten Faktoren $\{0; 0{,}05; 0{,}1; 0{,}15; 0{,}2\}$ korrigierten Matrizen inferiert werden, bilden 28% von den ungefähr $|Varianten| * |Familien| * |Datensätze| = 40 - 50.000$ möglichen Bäumen. Die sehr großen Faktoren wie zehn, 50 oder 100 werden zwar nicht ganz so häufig (2628, 5512 und 8503) ausgewählt, aber auch in diese Richtung wächst die Anzahl der gewählten Bäume. Gegen diese Verteilung stehen die durchschnittlichen Distanzen all der Bäume, deren Häufigkeitsverteilung in Abb. 4.2a gegeben ist. Je höher der Skalierungsfaktor der Speziesmatrix, desto geringer ist die durchschnittliche Distanz vom echten Genbaum zu allen ausgewählten Bäumen, die mithilfe



(a) Variierte Astlängenfaktoren

(b) Variierte Duplikations -und Verlustraten

(c) Variierte Sequenzlängen

(d) Variierte Anzahl an Spezies

Abbildung 4.3.: Durchschnittliche rRF-Distanzen zwischen dem inferierten und dem echten Genbaum. Spearfish verwendet $allFM_{10}$.

dieses Faktors inferiert wurden, wie in Abb. 4.2b abgebildet ist. Für den Faktor 10 bedeutet dies eine 2.25-al bessere Durchschnittsdistanz von 0,145 im Vergleich zum Faktor 0,05.

Um sowohl die hohe Rate an kleinen Faktoren, die durchschnittlich besseren Bäume, sowie alle Bereiche abzudecken, damit Unter- und Überkorrekturen (Abs. 3.1.4) bestmöglich vermieden werden können, wurden deshalb zehn Faktoren ausgewählt, die eine Tendenz zu kleinen Werten haben, aber auch aus mittleren und großen Faktoren bestehen. Diese zehn Faktoren werden im Folgenden verwendet.

4.2.3. Vergleich der Genauigkeit der Verfahren

Nach dem Vergleich des Einflusses der verschiedenen Markierungsalgorithmen auf die Genauigkeit der inferierten Bäume, wird Spearfish nun mit GeneRax, RAxML-NG und FastME verglichen, drei bereits existierenden Methoden. GeneRax ist eine der führenden Methoden zur Genbaumkorrektur sowie -rekonstruktion. Da $allFM_{80}$ in jedem Datensatz die besten Ergebnis gegenüber $aproFM_{80}$ und $madFM_{80}$ erzielt, werden im Folgenden nur Bäume von Spearfish berücksichtigt, bei denen jeder Eintrag der Distanzmatrizen korrigiert wurde. Zusätzlich werden die Skalierungsfaktoren auf die in Abs. 4.2.2 beschriebenen zehn Stück begrenzt, um die Auswirkungen auf Genauigkeit und Laufzeit erkennen zu können.

Im Vergleich zu der durchschnittlichen rRF-Distanz 0,197 der Bäume von $allFM_{80}$ zum echten Genbaum, verschlechtert diese sich bei einer Reduktion der Skalierungsfaktoren, wodurch $allFM_{10}$ eine durchschnittliche Distanz von 0,212 erreicht. Dagegen stehen die durchschnittlichen Distanzen der anderen drei Methoden. Während GeneRax mit 0,169 die genauesten Bäume inferiert, liegt die Distanz der FastME-Bäume bei 0,403 und der von RAxML-NG inferierten Bäume bei 0,463.

Die von FastME und die von Spearfish inferierten Bäume liegen minimal 0,076 auseinander und können bei sehr kleinen Astlängen (Abb. 4.3a), keinen Duplikationen und Genverlusten (Abb. 4.3b) oder kurzen Gensequenzen (Abb. 4.3c) bis zu 0,385 auseinander liegen.

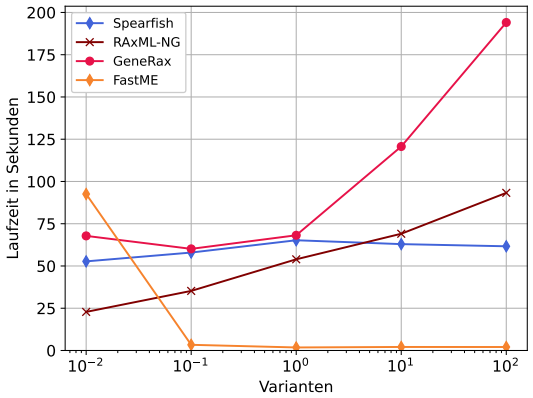
Bei einem differenzierten Blick auf die einzelnen Datensätze ergibt sich ein ähnliches Bild wie beim Vergleich der Markierungsalgorithmen. Bei sehr wenigen oder sehr vielen Mutationen rekonstruieren alle getesteten Methoden bis zu 3,7-mal ungenauere Bäume als bei einer mittleren Anzahl an Mutationen (Abb. 4.3a). Bei erhöhter Komplexität durch mehr Arten im SPEZIES-Datensatz oder mehr Evolutionsereignisse im DUPLOS-Satz sinkt die durchschnittliche Genauigkeit. Während die rRF-Distanz der Bäume von Spearfish und GeneRax bei 0,033 beziehungsweise 0,015 liegt, wenn keine Duplikationen auftreten, so steigt die durchschnittliche Distanz auf 0,275 beziehungsweise 0,243 bei einer Duplikationsrate von 1,47e−9 (Tab. 4.1). Dagegen bleibt die durchschnittliche Distanz bei FastME sowie RAxML-NG ungefähr bei jeweils 0,371 und 0,426.

4.2.4. Laufzeiten der Verfahren

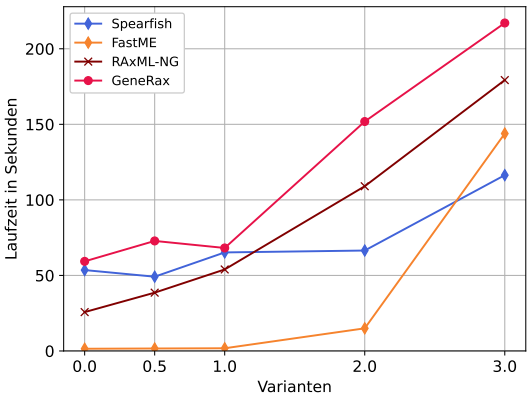
Um die Nützlichkeit von Spearfish beurteilen zu können, muss nicht nur die Genauigkeit der Inferenz evaluiert werden, sondern auch die Geschwindigkeit, mit der dies geschieht. Dafür wurde jeweils die Zeit gemessen, die eine Methode benötigt, um 100 Genbäume einer Variation eines Datensatzes zu inferieren. Bei der Berechnung der durchschnittlichen Laufzeit einer Variation, also dem Durchschnitt von 100 Durchläufen, wurde jeweils die längste sowie die kürzeste Zeit heraus gefiltert.

Mit durchschnittlich 10min55s Laufzeit von Spearfish mit $|\Lambda| = 80$, sind alle anderen getesteten Methoden sehr viel schneller. Durch die Verringerung der Baumanzahl von Spearfish um den Faktor acht verkürzt sich die durchschnittliche Laufzeit für $allFM_{10}$ auf 1min11s. Dies ist eine Geschwindigkeitszunahme um den Faktor neun. Dabei muss allerdings beachtet werden, dass die obige durchschnittliche Laufzeit mit 80 Faktoren auch die benötigte Zeit für das Wurzeln und Markieren des Startbaums mit APro und MAD berücksichtigt, wodurch die Zeit der ALL-Methode etwas überschätzt wird. Aufgeschlüsselt auf die einzelnen Schritte ergibt sich, dass die Umrechnung der Sequenzen bei $allFM_{10}$ 8,32% der Gesamtlaufzeit benötigt, die Korrektur 3,92%, die Rekonstruktion der zehn Bäume 18,50% und die anschließende Evaluierung durch GeneRax 69,25%. GeneRax benötigt durchschnittlich 2min22s, RAxML-NG 1min45s und FastME rekonstruiert Genbäume in durchschnittlich 19s.

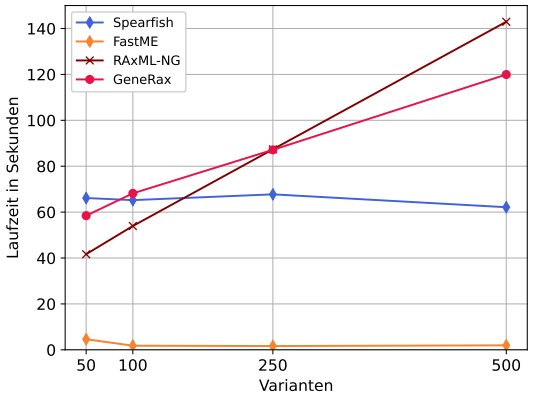
Die Verhalten der Laufzeiten der vier getesteten Methoden GeneRax, RAxML-NG, FastME und $allFM_{10}$ auf den vier Datensätzen lässt sich in zwei Gruppen einteilen. Die erste



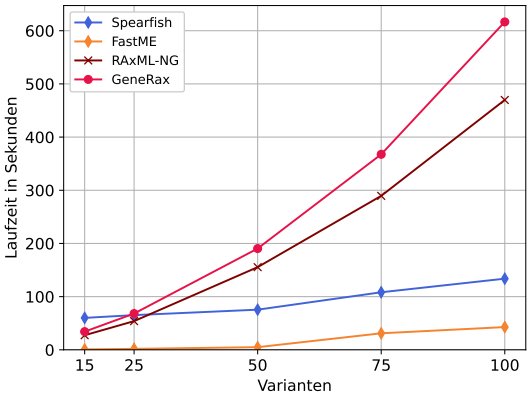
(a) Variierte Astlängenfaktoren



(b) Variierte Duplikations -und Verlustraten



(c) Variierte Sequenzlängen



(d) Variierte Anzahl an Spezies

Abbildung 4.4.: Laufzeiten der Methoden. Spearfish verwendet $allFM_{10}$.

Gruppe besteht aus den Datensätzen BRALEN und SITES, die zweite aus DUPLOS und SPECIES.

Die Parameter von BRALEN und SITES erhöhen im Vergleich zum BASE-Datensatz nicht die Anzahl an Arten oder Genen, die in den jeweiligen Bäumen dargestellt werden. Beide variierten Parameter wirken sich nur auf die Gensequenzen direkt aus. Da die Berechnung der Gensequenz-Distanzmatrizen allerdings nur 8,92% der Laufzeit von Spearfish ausmacht, wirken sich die veränderten Längen nicht stark auf die Laufzeit aus, welche durchschnittlich bei 1min02s liegt. Da auch FastME eine distanzbasierte Methode ist und nach dem selben Prinzip arbeitet, bleibt die Laufzeit von FastME annährend konstant bei ungefähr 12s. Für sehr kleine Astlängen benötigt FastME allerdings 1min33 und ist in diesem Fall langsamer als alle anderen Methoden. Im Gegensatz zu den zwei distanzbasierten Methoden steigt die Laufzeit von GeneRax und RAxML-NG mit erhöhter Sequenzlänge, durch die direkte Abhängig der Laufzeit von dieser. Auch bei vermehrten Mutationen benötigen die beiden zeichenbasierten Methoden mehr Zeit zur Inferenz, da sie direkt auf den Sequenzen arbeiten und somit bei mehr Mutationen mehr Berechnungen durchführen müssen. GeneRax ist tendenziell etwas langsamer als RAxML-NG mit jeweils einer durchschnittlichen Laufzeit von 1min34s und 1min07s.

Im Gegensatz zu BRALEN und SITES stehen die Datensätze SPECIES und DUPLOS. Die Laufzeit von FastME und somit auch jene von Spearfish ist direkt abhänging von der Anzahl der Gene. Da eine höhere Anzahl an Arten zu einer größeren Anzahl an Genen führt, wirkt sich dies auch auf die Laufzeit aus. Während FastME bei 25 Arten unter 2s und Spearfish 1min05s benötigt, benötigt FastME für viermal so viele Arten fast 24 Mal so lange (43s) und Spearfish ungefähr doppelt so lang (2min14s). GeneRax und RAxML-NG benötigen hingegen jeweils neunmal beziehungsweise 8,7-mal so lange für 100 Arten wie für 25. Beim DUPLOS-Datensatz benötigen die zeichenbasierten Methoden 3min37s und sind insbesondere im Vergleich zu Spearfish durchschnittlich nur noch 1,4-mal so langsam. Während FastME in nahezu allen Experimenten die schnelleste Methode ist, benötigt Spearfish für die Bäume des DUPLOS-Datensatzes weniger Zeit. Eine Möglichkeit dafür ist, dass die korrigierten Distanzmatrizen von Spearfish mehr Informationen enthalten und FastME so die Suche nach einem optimalen Baum erleichtert.

In allen vier Datensätzen ist Spearfish für kleine Werte meist etwas langsamer wie GeneRax. Bei einer Sequenzlänge von 50 Basen inferiert GeneRax die Genbäume beispielsweise 8s schneller. Je größer die Werte werden, desto schneller wird Spearfish im Vergleich zu GeneRax oder RAxML-NG. Bei einem Astlängenfaktor von 100 benötigt Spearfish 2min12s weniger als GeneRax und 32s weniger als RAxML-NG, während dieser Unterschied bei 100 Spezies auf 8min03s und 5min36s respektive wächst. FastME ist durchschnittlich 3,8-mal schneller als Spearfish.

5. Zusammenfassung & Ausblick

5.1. Diskussion

Zu Beginn dieser Arbeit wurden zwei Ziele festgelegt. Zum einen sollte eine akkurate, zum anderen eine schnelle Methode entwickelt werden, um Genbäume mithilfe ihres Speziesbaums zu inferieren. Dafür wurde die neue Methode Spearfish entwickelt. Spearfish ist eine distanzbasierte Methode, die zusätzlich zu den Gensequenzen die Distanzen zwischen den einzelnen Spezies im Speziesbaum berücksichtigt. Dadurch kann die Genauigkeit der inferierten Bäume erhöht werden, womit diese durchschnittlich ungefähr nur 25% ungenauer als GeneRax ist. Allerdings werden die Bäume auch durchschnittlich 4,6-mal schneller rekonstruiert, weitere Verbesserungen werden in Abs. 5.2 forgeschlagen.

Interessanterweise erzielt die Spearfish-Variante $allFM$ bessere Ergebnisse als die Varianten $aproFM$ und $madFM$, bei denen ein markierter Startbaum verwendet wird, um eine Überkorrektur durch den Speziesbaum zu vermeiden. Diese Beobachtung widerspricht den Erwartungen. Eine Erklärung hierfür könnte sein, dass der Startbaum in den meisten Fällen zu ungenau ist und dadurch die Duplikations- und Artneubildungsereignisse falsch zugeordnet werden, wodurch wiederum die falschen Genpaardistanzen korrigiert werden.

Im Laufe der Experimente wurde außerdem die Anzahl der Skalierungsfaktoren $|\Lambda|$ verkleinert. Dadurch konnte die Laufzeit stark verkürzt werden, die Genauigkeit blieb aber weitestgehend erhalten. Deshalb werden in Spearfish standardmäßig zehn Faktoren verwendet.

5.2. Ausblick

Die Experimente (Kap. 4) haben gezeigt, dass die Evaluierung der inferierten Bäume mithilfe von GeneRax über 60% der gesamten Laufzeit von Spearfish ausmacht. Eine naheliegende Verbesserung von Spearfish ist deshalb, diesen Schritt zu beschleunigen. Eine Möglichkeit dafür wäre, die Modellparameter wie die Astlänge oder die Duplikations- und Verlustrate nicht zu optimieren.

Weitere Verbesserungsmöglichkeiten sind eine bessere Wahl des Startbaums und das Entwickeln neuer Formeln, die eine bessere Gewichtung des Speziesbaums zulassen. Außerdem wurde nicht getestet, welchen Einfluss die verschiedenen Berechnungsmodelle von FastME zur Berechnung der Distanzmatrizen haben. Zusätzlich kann auch die Wahl der Skalierungsfaktoren weiter verbessert und optimiert werden.

Beachten werden sollte, dass der Abstand zwischen zwei Spezies im Speziesbaum in den Experimenten als die Pfadlänge zwischen den Spezies im echten Speziesbaum definiert wurde 3.1.1. Da dieser nur bei Simulationen bekannt ist, müssen die Distanzen auf eine andere Art berechnet werden. Eine Möglichkeit wäre, den Speziesbaum mithilfe von RAxML-NG oder FastME zu berechnen.

Außerdem kann Spearfish weiterentwickelt werden, um *horizontale Gentransfers (HGTs)* sowie *Incomplete Lineage Sorting (ILS)* zu berücksichtigen und so noch genauere Bäume zu rekonstruieren. HGTs gehören wie Duplikationen und Verlusten zu den Evolutionsereignissen, die in einem Genbaum wichtig sind, während ILS das Produkt von Aufspaltungen ist, die noch nicht komplett vollzogen worden sind.

Literatur

[1] Benoit Morel u. a. „GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss“. In: *Molecular Biology and Evolution* 37.9 (1. Sep. 2020), S. 2763–2774. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa141.

[2] Kris A. Wetterstrand. *DNA Sequencing Costs: Data*. Genome.gov. 16. Mai 2023. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (besucht am 19. 07. 2023).

[3] *GenBank and WGS Statistics*. National Center for Biotechnology Information (NCBI). 15. Juni 2023. URL: https://www.ncbi.nlm.nih.gov/genbank/statistics/ (besucht am 19. 07. 2023).

[4] L. L. Cavalli-Sforza und A. W. F. Edwards. „Phylogenetic analysis. Models and estimation procedures“. In: *American Journal of Human Genetics* 19.3 (Mai 1967), S. 233–257. ISSN: 0002-9297.

[5] Joseph Felsenstein. „The Number of Evolutionary Trees“. In: *Systematic Biology* 27.1 (1. März 1978), S. 27–33. ISSN: 1063-5157. DOI: 10.2307/2412810.

[6] Gergely J. Szöllősi u. a. „The Inference of Gene Trees with Species Trees“. In: *Systematic Biology* 64.1 (1. Jan. 2015), e42–e62. ISSN: 1076-836X, 1063-5157. DOI: 10.1093/sysbio/syu048.

[7] Vincent Lefort, Richard Desper und Olivier Gascuel. „FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program“. In: *Molecular Biology and Evolution* 32.10 (1. Okt. 2015), S. 2798–2800. ISSN: 0737-4038. DOI: 10.1093/molbev/msv150.

[8] Alexey M Kozlov u. a. „RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference“. In: *Bioinformatics* 35.21 (1. Nov. 2019), S. 4453–4455. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz305.

[9] Theodosius Dobzhansky. *Genetics and the origin of species*. 3. ed., rev., 2. pr. Bd. 11. Columbia biological series ; 11. New York [u.a.]: Columbia Univ. Press, 1953. 364 S.

[10] Johann Wolfgang Wägele. *Grundlagen der phylogenetischen Systematik*. 2., überarb. Aufl. München: Pfeil, 2001. 320 S. ISBN: 978-3-931516-93-2.

[11] *Phylogenetik*. URL: https://www.spektrum.de/lexikon/biologie/phylogenetik/51550 (besucht am 13. 09. 2023).

[12] *Phylogenie*. URL: https://www.spektrum.de/lexikon/biologie/phylogenie/51553 (besucht am 13. 09. 2023).

[13] Joel Cracraft. „Species Concepts and Speciation Analysis“. In: *Current Ornithology*. Hrsg. von Richard F. Johnston. Current Ornithology. New York, NY: Springer US, 1983, S. 159–187. ISBN: 978-1-4615-6781-3. DOI: 10.1007/978-1-4615-6781-3_6.

[14] Genome Reference Consortium. *Homo sapiens chromosome 1, GRCh38 reference primary assembly*. Version Number: 2. 20. Dez. 2013. URL: http://www.ncbi.nlm.nih.gov/nuccore/CM000663.2 (besucht am 09. 09. 2023).

[15] Michael A. Bender u. a. „Finding least common ancestors in directed acyclic graphs“. In: *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. SODA '01. USA: Society for Industrial und Applied Mathematics, 9. Jan. 2001, S. 845–854. ISBN: 978-0-89871-490-6. DOI: 10.5555/365411.365795.

[16] Charles Darwin. *On the Origin of Species by Means of Natural Selection (or Preservation of Favoured Races in the Struggle for Life)*. 1866.

[17] JV Chamary. „Aussterben“. In: *50 Schlüsselideen Biologie*. Hrsg. von JV Chamary. Berlin, Heidelberg: Springer, 2016, S. 196–199. ISBN: 978-3-662-48381-7. DOI: 10.1007/978-3-662-48381-7_49.

[18] Werner Buselmaier. „Mutation, Selektion und der Takt der molekularen Uhr“. In: *Der Gen-Kultur-Konflikt*. Berlin: Springer, 2016, S. 27–31. ISBN: 978-3-662-49395-3. DOI: 10.1007/978-3-662-49395-3_6.

[19] Roderic D. M. Page und Edward C. Holmes. *Molecular evolution: a phylogenetic approach*. Oxford ; Malden, MA: Blackwell Science, 1998. 346 S. ISBN: 978-0-86542-889-8.

[20] James F. Crow. „The high spontaneous mutation rate: Is it a health risk?“ In: *Proceedings of the National Academy of Sciences of the United States of America* 94.16 (5. Aug. 1997), S. 8380–8386. ISSN: 0027-8424. DOI: 10.1073/pnas.94.16.8380.

[21] Eugene V. Koonin. „Orthologs, Paralogs, and Evolutionary Genomics 1“. In: (1. Dez. 2005). DOI: 10.1146/annurev.genet.39.073003.114725.

[22] Michael Lynch. „Gene Duplication and Evolution“. In: *Science* 297.5583 (9. Aug. 2002). Publisher: American Association for the Advancement of Science, S. 945–947. DOI: 10.1126/science.1075472.

[23] E. J. Vallender. „4.07 - Molecular Evolution and Phenotypic Change“. In: *Evolution of Nervous Systems (Second Edition)*. Hrsg. von Jon H. Kaas. Second Edition. DOI: 10.1016/B978-0-12-804042-3.00108-1. Oxford: Academic Press, 2017, S. 101–119. ISBN: 978-0-12-804096-6.

[24] Reinhard Diestel. *Graphentheorie*. 5. Auflage. Lehrbuch. Berlin, Heidelberg: Springer Spektrum, 2017. ISBN: 978-3-662-53633-9.

[25] Kurth Mehlhorn und Peter Sanders. *Algorithms and Data Structures*. Berlin, Heidelberg: Springer, 2008. ISBN: 978-3-540-77977-3. DOI: 10.1007/978-3-540-77978-0.

[26] Donald Ervin Knuth. *The art of computer programming*. Addison-Wesley series in computer science and information processing. Reading, Mass.: Addison-Wesley, 1969. ISBN: 0-201-48541-9.

[27] *Duden | Topologie | Rechtschreibung, Bedeutung, Definition, Herkunft*. URL: https://www.duden.de/rechtschreibung/Topologie (besucht am 13. 09. 2023).

[28] Luay Nakhleh, Derek Ruths und Hideki Innan. „Gene Trees, Species Trees, and Species Networks“. In: *Meta-analysis and Combining Information in Genetics and Genomics* (7. Juli 2009), S. 275–293. ISSN: 978-1-58488-522-1. DOI: 10.1201/9781420010626.ch17.

[29] Hendrik N. Poinar u. a. „Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA“. In: *Science* 311.5759 (20. Jan. 2006). Publisher: American Association for the Advancement of Science, S. 392–394. DOI: 10.1126/science.1123360.

[30] D. F. Robinson und L. R. Foulds. „Comparison of phylogenetic trees“. In: *Mathematical Biosciences* 53.1 (1. Feb. 1981), S. 131–147. ISSN: 0025-5564. DOI: 10.1016/0025-5564(81)90043-2.

[31] Alexandros Stamatakis. *The Exelixis Lab*. URL: https://cme.h-its.org/exelixis/web/teaching/slides.html (besucht am 11. 09. 2023).

[32] Chao Zhang u. a. „ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy“. In: *Molecular Biology and Evolution* 37.11 (1. Nov. 2020). Hrsg. von Jeffrey Thorne, S. 3292–3307. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msaa139.

[33] Fernando Domingues Kümmel Tria, Giddy Landan und Tal Dagan. „Phylogenetic rooting using minimal ancestor deviation“. In: *Nature Ecology & Evolution* 1.7 (19. Juni 2017), S. 0193. ISSN: 2397-334X. DOI: 10.1038/s41559-017-0193.

[34] S. Mirarab u. a. „ASTRAL: genome-scale coalescent-based species tree estimation“. In: *Bioinformatics* 30.17 (1. Sep. 2014), S. i541–i548. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu462.

[35] David Bryant und Michael Charleston. *MAD roots for large trees*. 7. Nov. 2018. DOI: 10.48550/arXiv.1811.03174. arXiv: 1811.03174[q-bio].

[36] Masatoshi Nei und Sudhir Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, USA, 2000.

[37] N Saitou und M Nei. „The neighbor-joining method: a new method for reconstructing phylogenetic trees.“ In: *Molecular Biology and Evolution* 4.4 (1. Juli 1987), S. 406–425. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040454.

[38] J A Studier und K J Keppler. „A note on the neighbor-joining algorithm of Saitou and Nei.“ In: *Molecular Biology and Evolution* 5.6 (1. Juli 1988), S. 729–731. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040527.

[39] Andrey Rzhetsky und Masatoshi Nei. „A simple method for estimating and testing minimum-evolution trees“. In: *Mol Biol Evol* 9.5 (1992), S. 945–967. ISSN: 1537-1719. DOI: 10.1093/oxfordjournals.molbev.a040771.

[40] A Rzhetsky und M Nei. „Theoretical foundation of the minimum-evolution method of phylogenetic inference.“ In: *Molecular Biology and Evolution* 10.5 (1. Sep. 1993), S. 1073–1095. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040056.

[41] Richard Desper und Olivier Gascuel. „Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting“. In: *Molecular Biology and Evolution* 21.3 (1. März 2004), S. 587–598. ISSN: 0737-4038. DOI: 10.1093/molbev/msh049.

[42] Geetika Munjal, Madasu Hanmandlu und Sangeet Srivastava. „Phylogenetics Algorithms and Applications“. In: *Ambient Communications and Computer Systems*. Hrsg. von Yu-Chen Hu u. a. Advances in Intelligent Systems and Computing. Singapore: Springer, 2019, S. 187–194. ISBN: 9789811359347. DOI: 10.1007/978-981-13-5934-7_17.

[43] Walter M. Fitch. „Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology“. In: *Systematic Biology* 20.4 (1. Dez. 1971), S. 406–416. ISSN: 1063-5157. DOI: 10.1093/sysbio/20.4.406.

[44] Johannes Bergsten. „A review of long-branch attraction“. In: *Cladistics* 21.2 (2005), S. 163–193. ISSN: 1096-0031. DOI: 10.1111/j.1096-0031.2005.00059.x.

[45] R. A. Fisher und Edward John Russell. „On the mathematical foundations of theoretical statistics“. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594 (Jan. 1997). Publisher: Royal Society, S. 309–368. DOI: 10.1098/rsta.1922.0009.

[46] T Heath Ogden und Michael S Rosenberg. „Multiple Sequence Alignment Accuracy and Phylogenetic Inference“. In: *Systematic Biology* 55.2 (1. Apr. 2006), S. 314–328. ISSN: 1063-5157. DOI: 10.1080/10635150500541730.

[47] Lam-Tung Nguyen u. a. „IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies“. In: *Molecular Biology and Evolution* 32.1 (1. Jan. 2015), S. 268–274. ISSN: 0737-4038. DOI: 10.1093/molbev/msu300.

[48] Julia Haag u. a. „From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses“. In: *Molecular Biology and Evolution* 39.12 (1. Dez. 2022), msac254. ISSN: 1537-1719. DOI: 10.1093/molbev/msac254.

[49] Edwin Jacox u. a. „ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony“. In: *Bioinformatics* 32.13 (1. Juli 2016), S. 2056–2058. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw105.

[50] Bastien Boussau u. a. „Genome-scale coestimation of species and gene trees“. In: *Genome Research* 23.2 (1. Feb. 2013). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, S. 323–330. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.141978.112.

[51] Örjan Åkerborg u. a. „Simultaneous Bayesian gene tree reconstruction and reconciliation analysis“. In: *Proceedings of the National Academy of Sciences* 106.14 (7. Apr. 2009). Publisher: Proceedings of the National Academy of Sciences, S. 5714–5719. DOI: 10.1073/pnas.0806251106.

[52] Stéphane Guindon und Olivier Gascuel. „A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood“. In: *Systematic Biology* 52.5 (1. Okt. 2003), S. 696–704. ISSN: 1063-5157. DOI: 10.1080/10635150390235520.

[53] Gergely J. Szöllősi u. a. „Efficient Exploration of the Space of Reconciled Gene Trees“. In: *Systematic Biology* 62.6 (1. Nov. 2013), S. 901–912. ISSN: 1063-5157. DOI: 10.1093/sysbio/syt054.

[54] Fredrik Ronquist u. a. „MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space“. In: *Systematic Biology* 61.3 (1. Mai 2012), S. 539–542. ISSN: 1063-5157. DOI: 10.1093/sysbio/sys029.

[55] Mathieu Groussin u. a. „Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees“. In: *Molecular Biology and Evolution* 32.1 (1. Jan. 2015), S. 13–22. ISSN: 0737-4038. DOI: 10.1093/molbev/msu305.

[56] Diego Mallo, Leonardo De Oliveira Martins und David Posada. „SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees“. In: *Systematic Biology* 65.2 (1. März 2016), S. 334–344. ISSN: 1063-5157. DOI: 10.1093/sysbio/syv082.

[57] William Fletcher und Ziheng Yang. „INDELible: A Flexible Simulator of Biological Sequence Evolution“. In: *Molecular Biology and Evolution* 26.8 (1. Aug. 2009), S. 1879–1888. ISSN: 0737-4038. DOI: 10.1093/molbev/msp098.

[58] Benoit Morel u. a. „SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss“. In: *Molecular Biology and Evolution* 39.2 (Jan. 2022), msab365. ISSN: 1537-1719. DOI: 10.1093/molbev/msab365. eprint: https://academic.oup.com/mbe/article-pdf/39/2/msab365/42426267/msab365.pdf.

[59] Aaron J. Mussig. *PhyloDM*. Version 3.0.0. Juni 2023. DOI: 10.5281/zenodo.3998716.

[60] *p-ranav/argparse: Argument Parser for Modern C++*. GitHub. URL: https://github.com/p-ranav/argparse (besucht am 27. 09. 2023).

[61] *BenoitMorel/MPIScheduler*. GitHub. URL: https://github.com/BenoitMorel/MPIScheduler (besucht am 30. 09. 2023).

[62] Jaime Huerta-Cepas, François Serra und Peer Bork. „ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data“. In: *Molecular Biology and Evolution* 33.6 (1. Juni 2016), S. 1635–1638. ISSN: 0737-4038. DOI: 10.1093/molbev/msw046.

[63] Peter J. A. Cock u. a. „Biopython: freely available Python tools for computational molecular biology and bioinformatics“. In: *Bioinformatics* 25.11 (1. Juni 2009), S. 1422–1423. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp163.

Anhang

A. Quellcodes

A.1. Spearfish

Spearfish inklusive der Pipeline, die zum Testen verwendet wurde, ist unter `https://github.com/knirschl/ba-code.git` verfügbar. Ein eigenständiges Repository für Spearfish ist zum aktuellen Zeitpunkt nicht vorhanden.

A.2. Verwendete Software

Die folgenden Softwaretools werden von Spearfish verwendet:

- PhyloDM [59]
- FastME [7]
- argparse [60]
- MADroot [35]
- GeneRax [1]
- MPIScheduler [61]
- Zusätzlich dazu wurden der NJ-Algorithmus [37] und APro [32] selbst implementiert.

Die folgenden Softwaretools werden im Laufe der Experimente *außerhalb* von Spearfish verwendet:

- SimPhy [56]
- INDELible [57]
- FastME [7]
- GeneRax [1]

- RAxML-NG [8]
- MPIScheduler [61]
- PhyloDM [59]
- ETE3-Toolkit [62]
- BioPython [63]

B. Weitere Ergebnisse

B.1. Vergleich zwischen NJ und FastME als Inferierungsalgorithmus in Spearfish

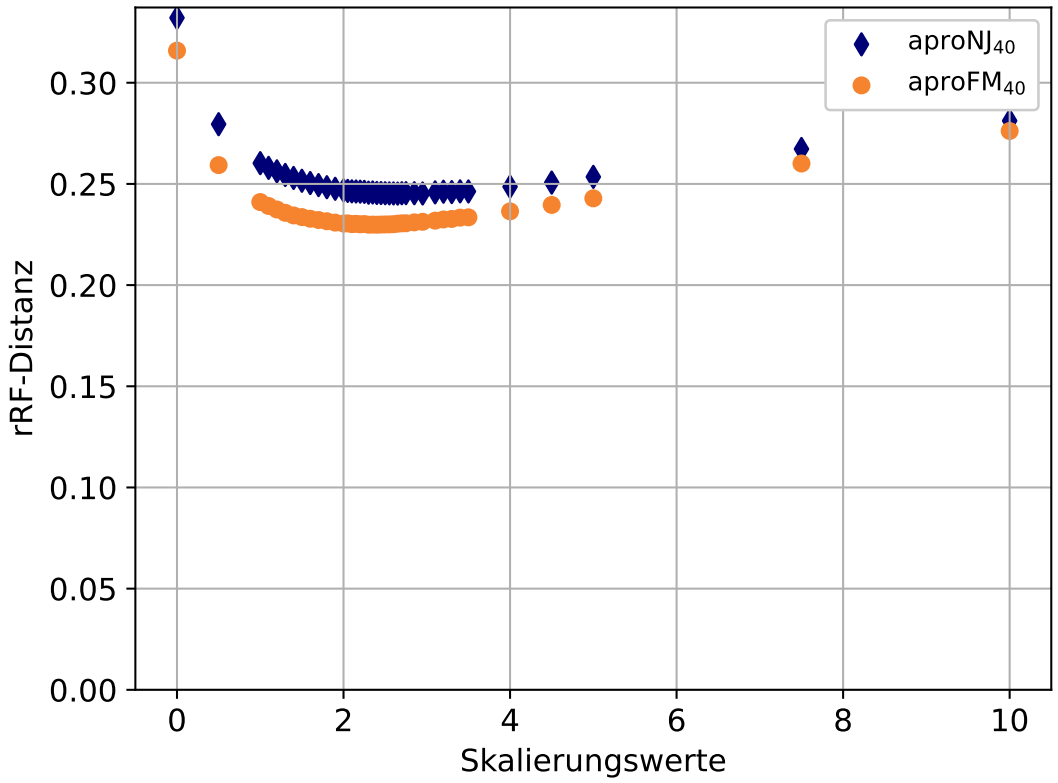
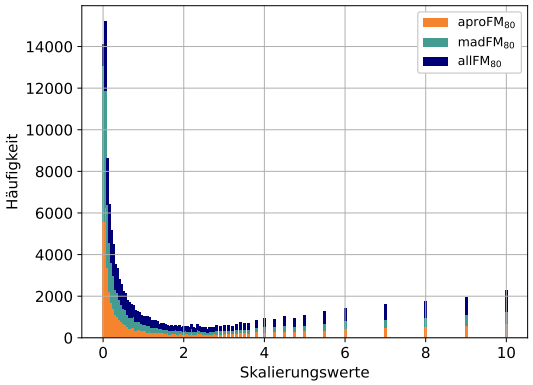
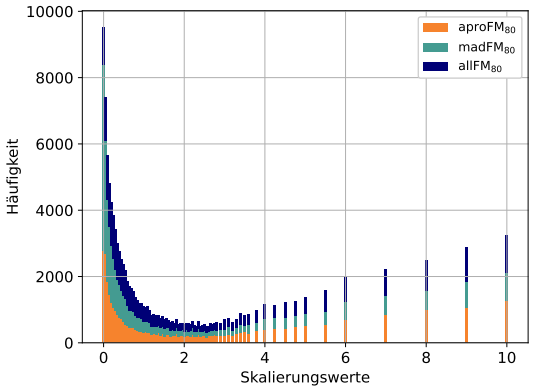


Abbildung B.1.: Durchschnittliche rRF-Distanzen zwischen dem inferierten und dem echten Genbaum in einem Experiment mit den Standardparametern.

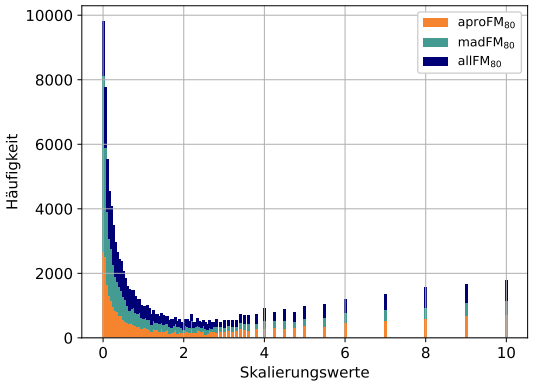
B.2. Einfluss der Markierungsalgorithmen



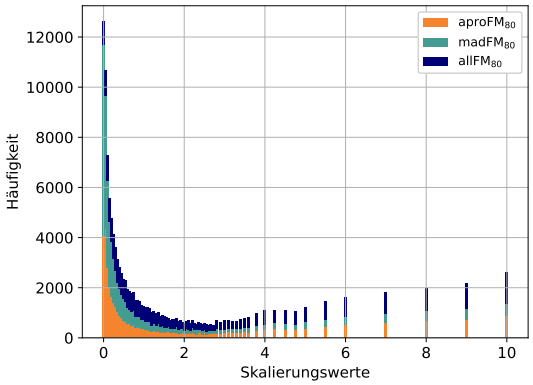
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)

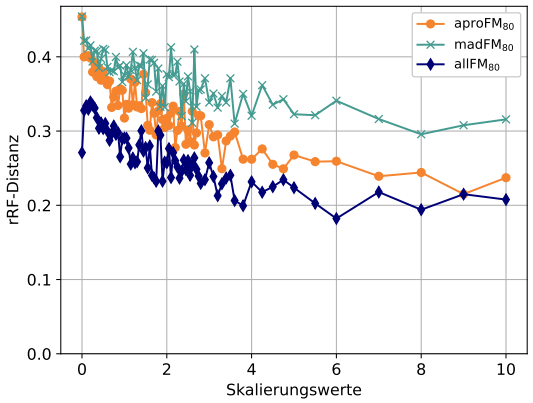


(c) Variierte Sequenzlängen (SITES)

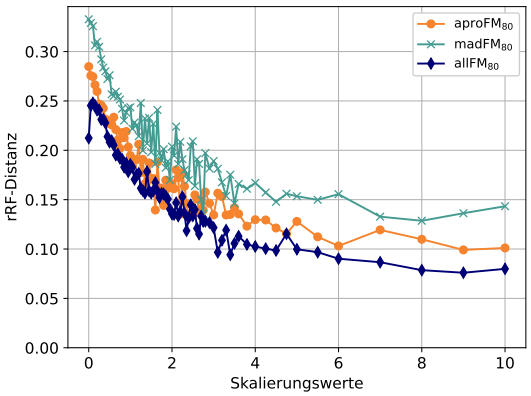


(d) Variierte Anzahl an Spezies (SPECIES)

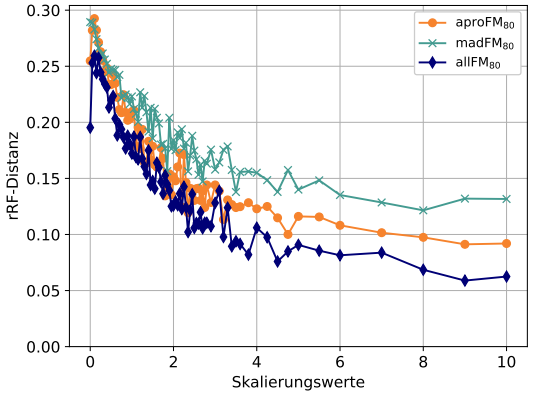
Abbildung B.2.: Absolute Anzahl an Bäumen, die aus einer Matrix rekonstruiert wurden, welche mit dem jeweiligen Faktor berechnet wurde.



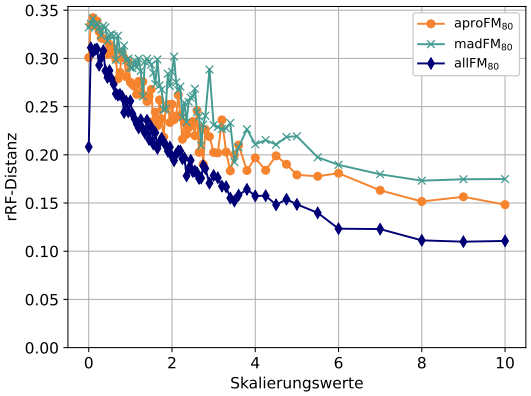
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)

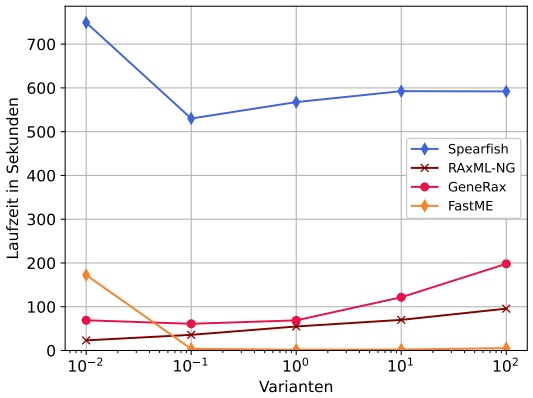


(c) Variierte Sequenzlängen (SITES)

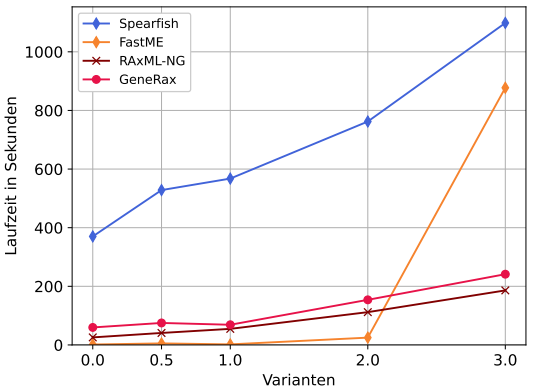


(d) Variierte Anzahl an Spezies (SPECIES)

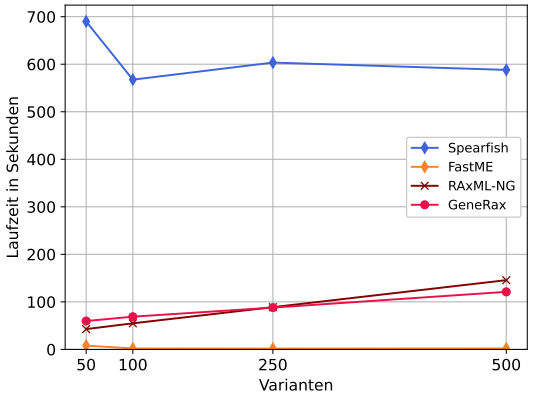
Abbildung B.3.: Durchschnittliche rRF-Distanzen zwischen dem inferierten und dem echten Genbaum aller ausgewählten Bäume (siehe Abb. B.2.



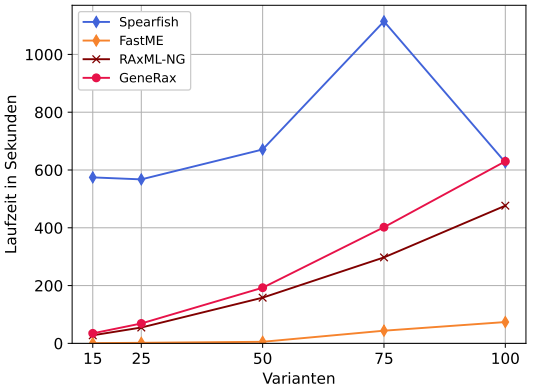
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)



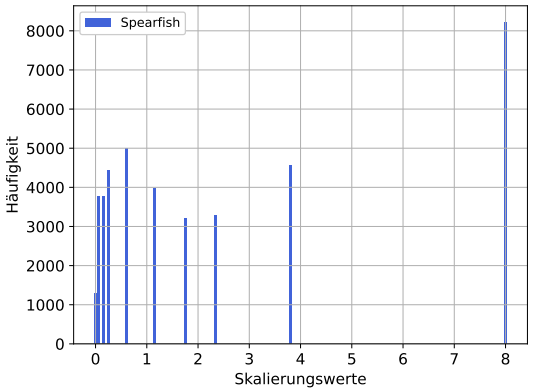
(c) Variierte Sequenzlängen (SITES)



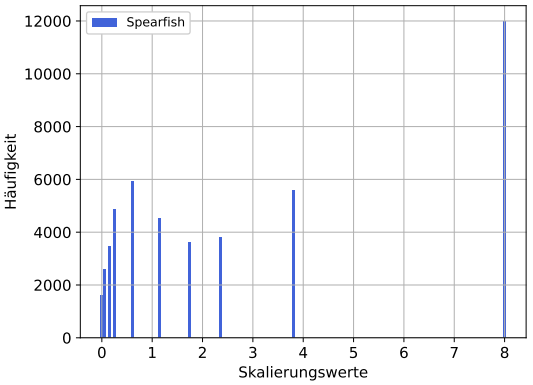
(d) Variierte Anzahl an Spezies (SPECIES)

Abbildung B.4.: Durchschnittliche Laufzeit von Spearfish mit 80 Skalierungsfaktoren und je einer der drei Variationen $aproFM$, $madFM$, $allFM$ gegenüber den anderen Verfahren.

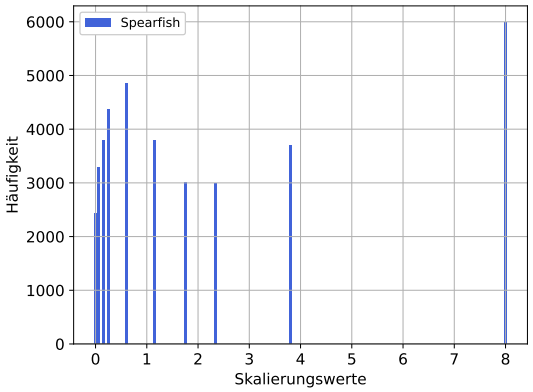
B.3. Vergleich der Verfahren



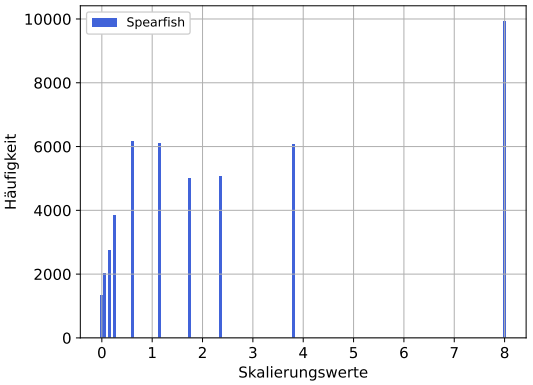
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)

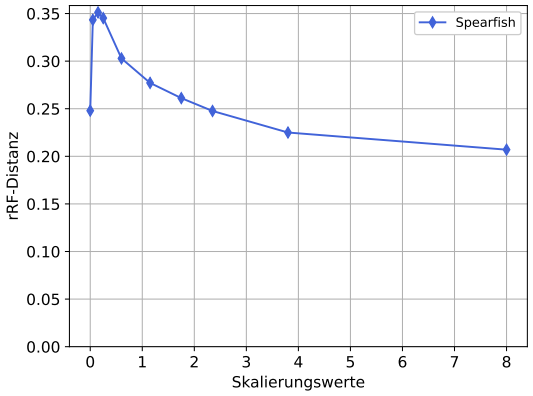


(c) Variierte Sequenzlängen (SITES)

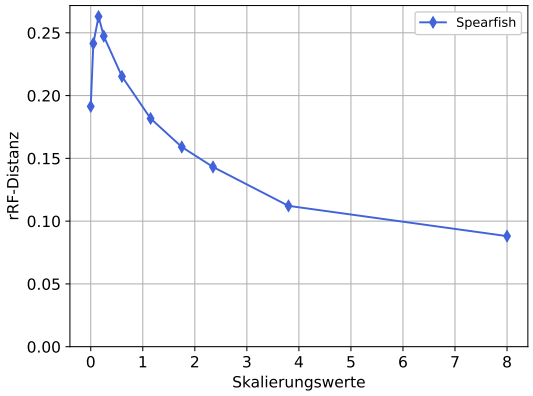


(d) Variierte Anzahl an Spezies (SPECIES)

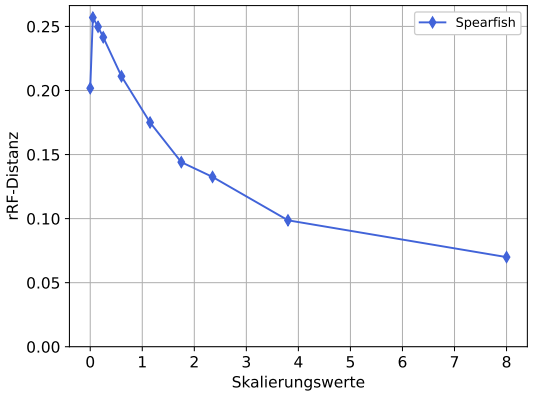
Abbildung B.5.: Absolute Anzahl an Bäumen, die aus einer Matrix rekonstruiert wurden, welche mit dem jeweiligen Faktor berechnet wurde.



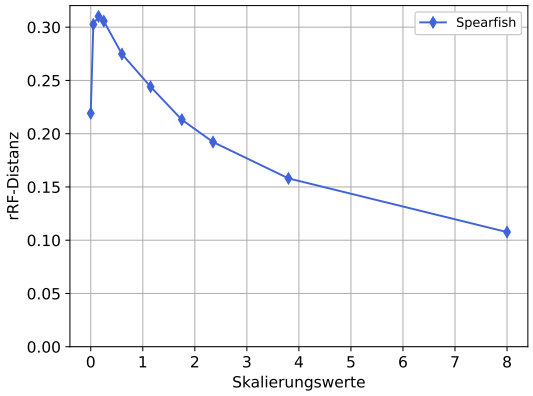
(a) Variierte Astlängenfaktoren (BRALEN)



(b) Variierte Duplikations -und Verlustraten (DUPLOS)



(c) Variierte Sequenzlängen (SITES)



(d) Variierte Anzahl an Spezies (SPECIES)

Abbildung B.6.: Durchschnittliche rRF-Distanzen zwischen dem inferierten und dem echten Genbaum aller ausgewählten Bäume (siehe Abb. B.5.