ERKENNUNG UND ZEITLICHE EINORDNUNG VON HORIZONTALEN GENTRANSFERS IN RELATION ZUM LETZTEN GEMEINSAMEN VORFAHREN

Der Fakultät für Mathematik und Informatik
der Universität Leipzig eingereichte

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science
(B. Sc.)

im Fachgebiet Informatik vorgelegt von

Stefan Grote

geboren am 23.01.1999 in Halle (Saale), Deutschland

Leipzig, den 14. September 2021

AUTOR:
Stefan Grote

TITEL:
Erkennung und zeitliche Einordnung von horizontalen Gentransfers in Relation zum letzten gemeinsamen Vorfahren

INSTITUT:
Institut für Informatik
Fakultät für Mathematik und Informatik
Universität Leipzig

BETREUER:
Dr. Nicolas Wieseke

INHALTSVERZEICHNIS

1 EINLEITUNG

1.1 GEGENSTAND

Die Biologie lässt sich in viele verschiedene Bereiche unterteilen, die die unterschiedlichen Aspekte des Lebens auf einer wissenschaftlichen Ebene untersuchen. Einer dieser Bereiche ist die Evolutionsbiologie, die sich mit der Evolution von Lebewesen beschäftigt. Sie ist damit eng mit anderen Bereichen wie z.B. der Genetik verbunden, die sich mit der Vererbung von Genen befasst. Die Vererbung von Genen wird als Gentransfer bezeichnet. Dabei wird zwischen zwei Typen unterschieden: Der vertikale Gentransfer (VGT) beschreibt die Weitergabe von Erbmaterial eines Organismus an einen Nachkommen, z.B. durch Kreuzung. Der horizontale Gentransfer (HGT) tritt primär bei Bakterien auf (Ochman u. a., 2000; Keeling u. a., 2008). Hierbei kann ein Individuum Gene eines anderen Individuums einer potentiell anderen Spezies aufnehmen, es zu seinem eigenen Erbgut hinzufügen und bei der Fortpflanzung sogar reproduzieren. Das bedeutet unter anderem, dass sich genetische Anpassungen einer Spezies schnell auf andere übertragen können, was akute Gefahren und Entwicklungen mit sich ziehen kann (Lerminiaux u. a., 2019) und deswegen für die Evolutionsbiologie eine hohe Bedeutung hat. Die Begriffe vertikal und horizontal beziehen sich dabei auf die Richtung, in die sich das Gen in einem phylogenetischen Baum bewegt. Ein phylogenetischer Baum stellt den evolutionären und damit zeitlichen Verlauf von Spezies und/oder Genen dar. Eine der ältesten Darstellungen von evolutionären Bäumen wurde 1837 von Charles Darwin angefertigt (Abbildung 1.1), der die unterschiedlichen Ähnlichkeiten zwischen verschiedenen Spezies mit einer Abstammung in einer Baumstruktur erklärt hat. Insbesondere hob er dabei das Aussterben von Spezies hervor (Darwin, 1837).

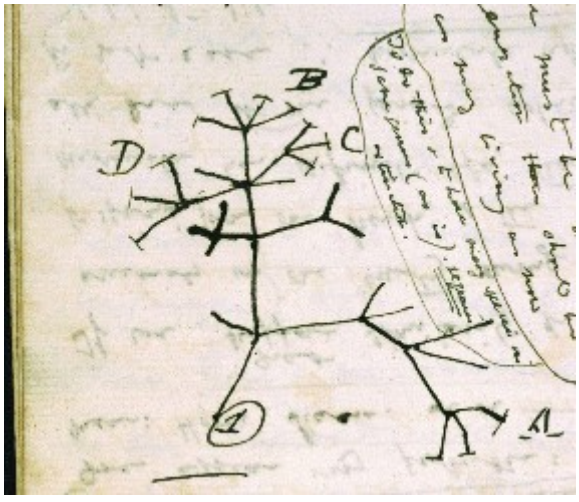

Abbildung 1.1: Darwins „Tree of Life“ (aus Darwin, 1837)

In dieser Arbeit wird ein Algorithmus entwickelt, der solche HGTs in Gendaten verschiedener Spezies anhand der Ähnlichkeit der Gene (evolutionäre Distanz) zueinander erkennt und für eine weitere Verarbeitung ausgibt.

1.2 PROBLEMATIK

HGTs sind im Vergleich zu VGTs zufällig und damit unberechenbar. Dadurch kann die Herkunft von Genen in Individuen, die aus HGTs herrühren, nicht immer direkt und eindeutig bestimmt werden. Dieser Umstand erschwert die Rekonstruktion von Gen- und Speziesbäumen anhand heutiger Gendaten in der Phylogenetik, da Genen Events wie Duplikationen und Aufteilungen möglicherweise falsch zugeordnet werden und der phylogenetische Baum somit fehlerhaft ist (Gophna u. a., 2005; Poptsova u. a., 2007).

1.3 MOTIVATION

Diese Arbeit bildet eine weitere Grundlage für zukünftige Forschungen, da sie zur Sicherheit und Verlässlichkeit von rekonstruierten phylogenetischen Bäumen beiträgt. Darüber hinaus werden weitere Rückschlüsse über die Herkunft von Genen möglich, die durch einen HGT entstanden sind. Unter anderem könnten dadurch ausgestorbene Spezies, die vielleicht auch noch gar nicht entdeckt wurden, aus heutigen Gendaten abgeleitet werden. Die Forschungen der Universität Leipzig bzw. des Institut für Informatik werden durch diese Arbeit aktiv unterstützt.

1.4 ZIELSETZUNG

Das Ziel dieser Arbeit ist die Entwicklung einer Anwendung, die nur mithilfe der Distanzen zwischen Genen verschiedener Spezies erkennen kann, welche dieser Gene bzw. Genpaare horizontal transferiert wurden. Diese erkannten Genpaare werden für eine mögliche weitere Verarbeitung ausgegeben.

1.5 AUFGABENSTELLUNG

1.5.1 Entwicklung eines Algorithmus

Bei der Entwicklung des Algorithmus liegt der Fokus auf folgenden Aspekten:

- Sind Performance-Fragen wie Laufzeit und Ressourcen für den Algorithmus wichtig?
- Unterstützt der Algorithmus Multithreading bzw. Parallelisierung, um die vorherige Aufgabe zu unterstützen?
- Wie gut ist der Algorithmus darin, HGTs zu erkennen und false-positives (fälschlicherweise als HGT eingestufte Genpaare) zu vermeiden?

1.5.2 Implementierung des Algorithmus

Bei der Implementierung sind folgende Aspekte zu beachten:

- Welche Programmiersprache (auch im Bezug auf Performance) ist geeignet?
- Welche Hilfsprogramme können verwendet werden, um aus Gendaten die Distanzen zu berechnen und in ein geeignetes Format zu portieren?
- Welche Variablen des Algorithmus bieten sich an, über Kommandozeilenparameter o.ä. für den Benutzer zur Individualisierung zugänglich gemacht zu werden?

1.6 AUFBAU DER ARBEIT

Die Arbeit gliedert sich in mehrere Abschnitte. Nach dieser Einleitung wird in Kapitel 2 näher auf die biologischen Hintergründe eingegangen und diese genauer erläutert. Dazu gehört ein Überblick über phylogenetische Grundlagen und Begriffe, phylogenetische Bäume und mögliche zeitliche Einordnungen von HGTs. Darüber hinaus werden bereits existierende Methoden zur Erkennung von HGTs vorgestellt.

Der in dieser Arbeit entwickelte Algorithmus wird in Kapitel 3 vorgestellt. Neben einem Beispiel und einer formalen Definition wird er anhand von Pseudocode genauer erklärt. Abschließend wird auf die Implementierung eingegangen, um die oben gestellten Fragen zu beantworten.

Die Ergebnisse des Algorithmus und deren Auswertung werden in Kapitel 4 analysiert und diskutiert.

Abschließend werden im letzten Kapitel 5 die wesentlichen Ergebnisse zusammengefasst und ein kritischer Ausblick bezüglich der weiteren Forschung gegeben.

2 GRUNDLAGEN DER PHYLOGENETIK

Die Phylogenetik ist ein Teilgebiet der Biosystematik und damit Teil der evolutionären Biologie. Sie untersucht die evolutionären Beziehungen von Organismen und Gruppen von Organismen bzw. Taxa anhand von vererbbaren Eigenschaften wie DNA-Sequenzen[1].

Der Begriff Taxa (sing. Taxon) kommt aus dem Bereich der Taxonomie und beschreibt eine Gruppe von Objekten die anhand von verschiedenen Kriterien und ähnlichen Eigenschaften zu dieser einen Klasse oder Kategorie gehören (Koschnik, 1992). Im Kontext der Biologie ist die Einordnung aller Lebewesen in die „Königreiche" wie Tiere, Pflanzen, Pilze, usw. (Ruggiero u. a., 2015) nur eine der gröbsten dieser Form. Weitere feinere Gruppierungen können z.B. anhand von Familien oder Spezies stattfinden.

In der numerischen Taxonomie bzw. der Phänetik werden solche Gruppierungen allein aufgrund von morphologischen Ähnlichkeiten vorgenommen (Ross, 1964). Diese Rückschlüsse waren bereits vor der Sequenzierung von Gendaten möglich, und ignorieren diese und andere evolutionäre Hinweise auch explizit. Im Gegensatz dazu nutzt die Phylogenetik auch genau diese Daten und Erkenntnisse. Aus diesem Grund sind die Ergebnisse der Phylogenetik meist genauer als die der Phänetik, da es nicht um mögliche Gruppierungen geht sondern um die wahrscheinlichste Gruppierung. Insbesondere liegt dabei der Fokus auf der Inferenz der evolutionären Hintergründe von heutigen Spezies, genauer der Rekonstruktion von phylogenetischen Bäumen (Heywood u. a., 1964). Diese werden in Kapitel 2.2 genauer erläutert.

Dabei nutzt die Phylogenetik algorithmische Ansätze, um diese Bäume zu rekonstruieren und andere Erkenntnisse zu erlangen. Sie gehört damit zum Bereich der Bioinformatik. Der in dieser Arbeit entwickelte Algorithmus zur Erkennung und Einordnung von HGTs behandelt also nur einen Teil dieser Rekonstruktion. In Kapitel 2.4 werden andere bereits existierende Methoden zur Erkennung von HGTs vorgestellt und anhand ihres Ansatzes gruppiert.

Um die Hintergründe dieser Arbeit zu verstehen, müssen zuerst die dafür relevanten Aspekte näher erläutert werden.

2.1 PHYLOGENETISCHE BEGRIFFE

In der Natur sind Mutationen ein ganz normales Phänomen. Eine Mutation ist eine spontane und permanente Veränderung des Erbguts einer Zelle, welche an Tochterzellen weitergegeben werden kann (Koschnik, 1997). Dabei sind natürliche Mutationen zuerst weder gut noch schlecht, sondern einfach ein essentieller Bestandteil der Evolution: Durch sie können sich Spezies weiterentwickeln und anpassen, wie eine reine Änderung des Verhaltens es niemals könnte. Diese Anpassungen können im Rahmen

1 https://www.biologyonline.com/dictionary/phylogenetics, abgerufen am 25.6.2021

der natürlichen Selektion in einer Spezies weitergegeben und verbreitet werden. Diese Selektion greift auch bei negativen Auswirkungen von Mutationen: Wenn z.B. lebenswichtige Funktionen durch eine Mutation eingeschränkt werden, wird diese Mutation früher oder später, je nach Ausmaß der Einschränkung, aussterben, da der Organismus eine geringere Lebenserwartung als andere Organismen derselben Art ohne diese Mutation hat.

Ein wichtiger Faktor ist dabei die Mutationsrate, die beschreibt wie häufig Mutationen in einem Gen oder Organismus in einem gewissen Zeitintervall vorkommen. Dieser Wert ist unter anderem abhängig von der Spezies und dem Alter des Organismus (Crow, 1997), und spielt für spätere Betrachtungen in dieser Arbeit eine wichtige Rolle.

Mutation ist dabei nur ein Oberbegriff. Bei den Genmutationen wird u.a. zwischen folgenden Arten bzw. Events unterschieden:

- Die häufigste Form der Mutation ist ein Einzelnukleotid-Polymorphismus (engl. „single-nucleotide polymorphism“, SNP). Dabei verändert sich nur ein einzelnes Basenpaar einer Sequenz. Bereits eine einzelne Mutation dieser Art kann sichtbare Auswirkungen haben. Komplexe Krankheiten wie z.B. Osteoporose treten meistens jedoch nur bei mehreren Mutationen auf (Singh u. a., 2011).

- Bei einer (Gen-)Duplikation wird ein Teil der DNA, welcher ein Gen enthält, dupliziert bzw. vervielfacht (Zhang, 2003). Die aus einer Duplikation enstandenen Gene werden als „paralogs“ bezeichnet. Sie haben danach meist verwandte, aber klar unterschiedliche Funktionen (Geiß u. a., 2018), weil genau dieselbe Funktion nicht noch einmal gebraucht wird und deswegen SNP-Mutationen oft häufiger an einem der Gene auftreten.
  Im folgenden Kapitel wird die Duplikation als Event an einem Baum gezeigt.

- Bei einer Deletion gehen Teile der DNA verloren. Das kann einzelne Basen, Gene oder ganze Chromosomen betreffen. Je nach Ausmaß werden eventuell fehlerhafte oder nicht funktionale Proteine gebildet, was unterschiedliche Folgen haben kann. Auch dieses Event wird im folgenden Kapitel noch einmal aufgegriffen.

- Für diese Arbeit nur indirekt relevant sind Insertions, weil sie im Bezug auf phylogenetische Bäume ähnliche Auswirkungen wie SNPs haben und deswegen auch nicht extra markiert werden. Bei einer Insertion werden ein oder mehrere Basenpaare zu einem Gen hinzugefügt. Die realen Auswirkungen einer solchen Mutation ähneln denen einer Deletion.

Der horizontale Gentransfer ist neben dem vertikalen Gentransfer die zweite Möglichkeit, wie Gene von einem Individuum auf ein anderes übertragen werden können. Dabei werden diese Gene also nicht durch Reproduktion bzw. Fortpflanzung weitergegeben. Dadurch ist er für Evolution als solche sehr wichtig, weil z.B. positive Effekte von Genmutationen nicht in einer Spezies bleiben sondern übertragen werden können. Es wird zwischen verschiedenen Arten des HGTs unterschieden (Ochman u. a., 2000; Gyles u. a., 2014):

- Als Konjugation wird die Übertragung von Genen zwischen zwei Zellen bezeichnet, wenn diese beiden Zellen direkten Kontakt haben. Die Spenderzelle stellt mit einem Pilus eine Verbindung bzw. Brücke zu der Empfängerzelle her. Über diese kann ein Plasmid in die Empfängerzelle übertragen werden, welches wiederum in ein Chromosom integriert werden kann.

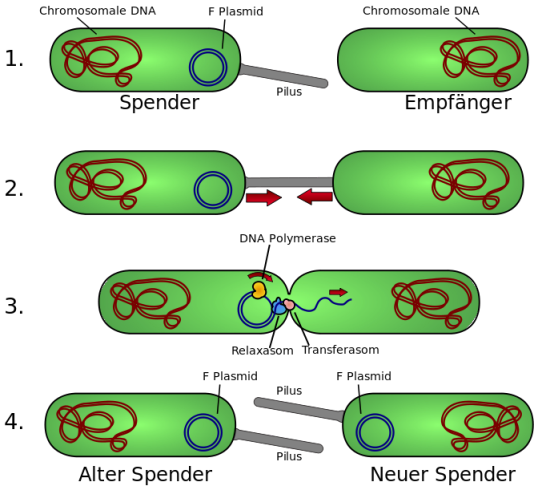

Abbildung 2.1: Ablauf eines Gentransfers bei einer Konjugation[2]

- Bei einer Transduktion wird Genmaterial durch ein Virus in eine Zelle injiziert, es ist also kein direkter Kontakt zwischen Spender- und Empfängerzelle notwendig. Dabei kann es sich sowohl um DNA des Virus handeln als auch um DNA von einer zuvor befallenen Zelle, welches das Virus zufällig von dieser aufgenommen hat.

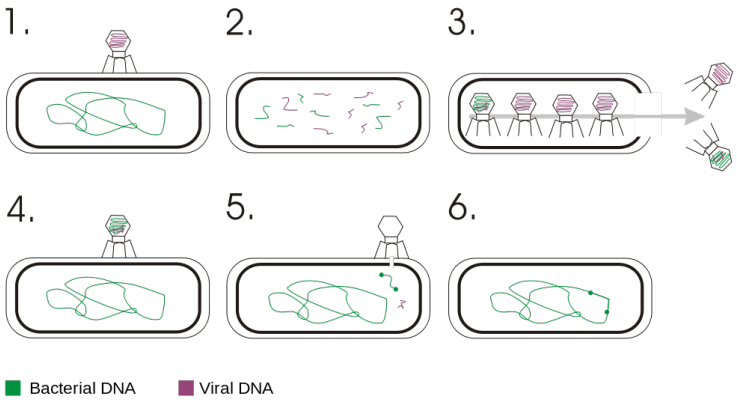

Abbildung 2.2: Ablauf eines Gentransfers bei einer Transduktion[3]
Die erste befallene Zelle produziert in 3. selber neue Viren, die teilweise noch Genmaterial der ersten Zelle enthalten. 4.-6. zeigt, wie eines dieser Viren eine neue Zelle befällt, und das Genmaterial der ersten Zelle eingebunden wird.

2 https://upload.wikimedia.org/wikipedia/commons/thumb/c/c2/Konjugation.svg/662px-Konjugation.svg.png, abgerufen am 13.09.2021

3 https://upload.wikimedia.org/wikipedia/commons/thumb/e/ee/Transduction_%28genetics%29en.svg/1024px-Transduction_%28genetics%29en.svg.png, abgerufen am 13.09.2021

- Die Transformation beschreibt die Aufnahme und Eingliederung von Genmaterial aus der Umgebung durch die Zellmembran in eine Zelle. Dafür relevant ist die Kompetenz einer Zelle, die natürlich nicht in allen Spezies vorhanden ist. Sie beschreibt, ob die Zelle Gene besitzt, deren Produkte wiederum eine Transformation durchführen können. Dieser Zustand kann bei manchen Spezies auch künstlich hervorgerufen werden (Johnston u. a., 2014).

Diese Arten von HGTs bzw. HGTs im Allgemeinen kommen primär bei Bakterien vor und beziehen sich damit auf Bakterienzellen, haben aber auch für Eukaryoten eine Bedeutung (Keeling u. a., 2008). Um die Relevanz von HGTs zu verdeutlichen, werden im folgenden jeweils ein Beispiel für Prokaryoten und danach für Eukaryoten näher behandelt.

Bakterien sind in der Lage, durch zufällige Mutationen eine Resistenz gegen Antibiotika aufzubauen, was wie bereits erläutert ein normaler natürlicher Vorgang ist. Die vermehrte Nutzung von Antibiotika im Kampf gegen bakterielle Krankheiten hat vor diesem Hintergrund aber einen entscheidenden Nachteil, der in den letzten Jahren mehr und mehr Aufmerksamkeit bekommen hat: Das verwendete Antibiotikum tötet nur die Bakterienstämme, die keine Resistenz entwickelt haben. Übrig bleiben nur die Bakterien, die eine Resistenz aufgebaut haben. Und sie sind damit auch die einzigen, die sich fortpflanzen können, und somit der ganze Bakterienstamm dann die gleiche Resistenz aufweist und in der Zukunft immun ist. HGTs kommen bei der weiteren Verbreitung dieser Resistenz ins Spiel: Die Resistenz ist bisher nur in einer Bakterienspezies vorhanden, kann aber durch einen HGT auch auf andere Bakterienspezies übertragen werden. Diese anderen Spezies sind dann ohne eine zufällige Mutation und ohne überhaupt schon einmal durch das Antibiotikum behandelt zu werden immun. Die Gefahr für solche Übertragungen mit (multi-)resistenten Keimen als Ergebnis ist dabei in Krankenhäusern am höchsten (Lerminiaux u. a., 2019; Kapley u. a., 2021).

Ein Beispiel für einen HGT zwischen Eukaryoten, der vor ca. 80 Jahren aufgetreten ist, wurde 2006 erkannt. Obwohl die anderen Gene der zwei beobachteten Pilzarten nur eine Ähnlichkeit von ungefähr 80% aufwiesen, wurde ein Genpaar mit einer Ähnlichkeit von 99,7% entdeckt. In diesem Fall wurde ein Virulenz-Genom übertragen, wodurch der durch die Empfängerspezies erzeugte Erreger, welcher Weizen befällt, signifikant verstärkt wurde (Friesen u. a., 2006). Durch HGTs können also auch existierende Krankheiten und Erreger verändert werden, was unerwartete Auswirkungen auf die Folgen und Behandlungen haben kann.

2.2 PHYLOGENETISCHE BÄUME

Phylogenetische Bäume sind das Ergebnis einer phylogenetischen Untersuchung. Sie stellen die Herkunft von Spezies und Genen dar und lassen Rückschlüsse auf die Evolution der betrachteten Elemente zu. Das gesamte heutige Leben auf der Erde könnte in einem einzelnen phylogenetischen Baum abgebildet werden, bei dem die Wurzel der „last universal common ancestor“ wäre (Weiss u. a., 2016). Von diesem einen Baum sind aber bisher nur Ausschnitte bekannt. Genau an dieser Stelle setzt der in dieser Arbeit entwickelte Algorithmus an: Er unterstützt die Rekonstruktion von phylogeneti-

schen Bäumen, indem er HGTs erkennt, die andernfalls die Rekonstruktion invalidieren könnten.

Da der Algorithmus unabhängig von einer gewählten Baumstruktur funktioniert, werden die für ihn relevanten Events nur an einem Beispielbaum erläutert. Abbildung 2.3 zeigt einen gewurzelten Genbaum, der in einen Speziesbaum eingebettet ist. Jede Kante stellt dabei den zeitlichen Verlauf eines Gens (mit verschiedenen Events) dar. Graue Ovale sind (ehemalige) Spezies, die sich in mehrere neue Spezies aufgeteilt haben. A, B, C und D sind die Spezies zu der sich die Wurzelspezies letztendlich aufgeteilt hat. a, a' sind also Gene von A, b von B usw.

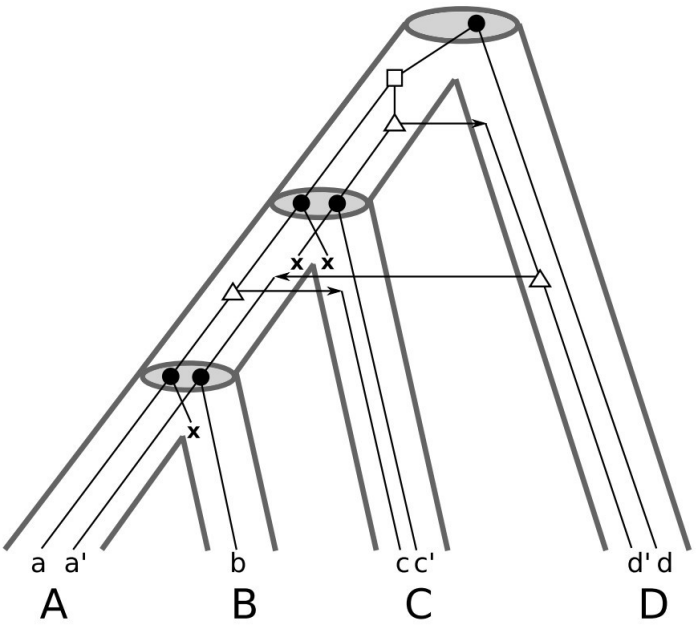

Abbildung 2.3: Phylogenetischer Baum (Geiß u. a., 2018)

Legende:
- • Speciation
- □ Duplikation
- × Deletion
- △ HGT

An jeder der Kanten kann jederzeit eins der in Kapitel 2.1 beschriebenen Mutation-Events auftreten, mit Ausnahme von SNPs. Sie kommen natürlich vor und verändern die einzelnen Gene, aber da die Mutationsrate verschieden ist, kann nie genau gesagt werden wann ein SNP aufgetreten ist. Das bedeutet auch, dass die „Höhe“ bzw. Einordnung von Events und Spezies in der Relation zu anderen Events nicht unbedingt korrekt sein muss. Die Ganggeschwindigkeit der molekularen Uhr, ein Begriff für die Abschätzung der Zeit seit der Aufteilung von Genen bzw. Spezies anhand von Mutationen, ist hier nicht konstant. In den meisten Modellannahmen so wie auch hier wird sie im Folgenden aber zusammen mit der Mutationsrate als konstant angenommen, da der genaue Zeitpunkt meist nicht wichtig, sondern die Orientierung schon ausreichend ist.

Eine Aufteilung einer Spezies ist immer mit einem Speciation-Event (•) der Gene in der Spezies verbunden. Dabei wird in beide neue Spezies eine Kopie des Genes weitergegeben, die jetzt unabhängig voneinander sind. Diese Genpaare werden als „orthologs“ bezeichnet. In den beiden neuen Spezies haben sie zuerst die gleiche und selbst nach

einiger Zeit (und damit Mutationen) immer noch eine ähnliche Funktion im Gegensatz zu „paralogs“, deren Funktionen meist auseinander gehen (Geiß u. a., 2018). Durch diese Mutationen erhöhen sich aber trotzdem die Unterschiede zwischen den beiden Genen, da die Gene nicht mehr von derselben Mutation betroffen sind. Das gleiche gilt auch für die Distanz zwischen den Genen, die für den Algorithmus (Kapitel 3.3) von essentieller Bedeutung ist.

Die Distanz zwischen zwei Genen beschreibt dabei und im Folgenden keine räumliche Distanz, sondern ist die genetische bzw. evolutionäre Distanz, die die Ähnlichkeit von zwei Genen charakterisiert. Je höher die Unterschiede zwischen zwei Genen bzw. Gensequenzen ist, desto höher ist auch die Distanz zwischen diesen Genen. Aus dieser Distanz lassen sich Rückschlüsse auf die evolutionäre Vergangenheit der Gene oder im größeren Kontext sogar Spezies ableiten: Wenn die Distanz gering ist, lässt das eine bis vor kurzem gemeinsame Vergangenheit schließen, eben bis z.B. ein Speciation-Event aufgetreten ist. Es gibt verschiedene Methoden zur Berechnung der genetischen Distanz. Eine der bekanntesten ist dabei die Nei Distanz (Nei, 1972).

Die Duplikation eines Gens wird durch ein □ mit einem Input Gen und zwei Output Genen (den „paralogs“) in den Bäumen dargestellt, die danach unabhängig voneinander sind. Eine Deletion bzw. der Verlust eines Genes wird durch ein $\times$ am Ende der Kante des Gens dargestellt. Ein horizontaler Gentransfer wird durch ein $\triangle$ markiert. Wie bei einer Duplikation gibt es dabei ein Input-Gen und zwei Output-Gene. Das Original des Gens verbleibt an der gleichen Position wie das Input Gen und ist vom HGT effektiv nicht betroffen. Es ist also dasselbe Gen wie das Input Gen. Der Transfer der Kopie wird durch einen Pfeil zum nächsten Event der Kopie in einer anderen Spezies angezeigt.

Da für die weiteren Betrachtungen in dieser Arbeit und insbesondere für den Ansatz des Algorithmus (Kapitel 3.1) die mathematischen Hintergründe von Bäumen und speziell phylogenetischen Bäumen nicht relevant sind, wird dafür an dieser Stelle auf Stadler u. a. (2017) sowie Geiß u. a. (2018) verwiesen.

2.3 ZEITLICHE EINORDNUNG VON HGTS

Wie bereits erwähnt sind HGTs zufällige Events, die auf Anhieb nicht genau zugeordnet werden können. Es ist aber möglich, HGTs anhand des Zeitpunktes der Aufteilung in Relation zu dem letzten gemeinsamen Vorfahren der beiden betroffenen Gene zu charakterisieren und sie dadurch korrekt erkennen zu können. HGTs können

- vor dem letzten gemeinsamen Vorfahren der Spezies der Gene („HÖHER“-Relation),
- zur gleichen Zeit wie die Aufteilung des letzten gemeinsamen Vorfahrens („GLEICH“-Relation) oder
- nach der Aufteilung des letzten gemeinsamen Vorfahrens („GERINGER“-Relation)

aufgetreten sein (Schaller u. a., 2021). Diese drei Möglichkeiten werden an entsprechenden Beispielen verdeutlicht. Die molekulare Uhr und die Mutationsrate werden jeweils als konstant angenommen.

Im Folgenden wird als „Best-Match“ das Gen einer anderen Spezies bezeichnet, welches dem aktuell betrachteten Gen am ähnlichsten bezüglich der Distanz ist. Zuerst wird die GERINGER-Relation anhand des Baumes in Abbildung 2.4 erläutert.

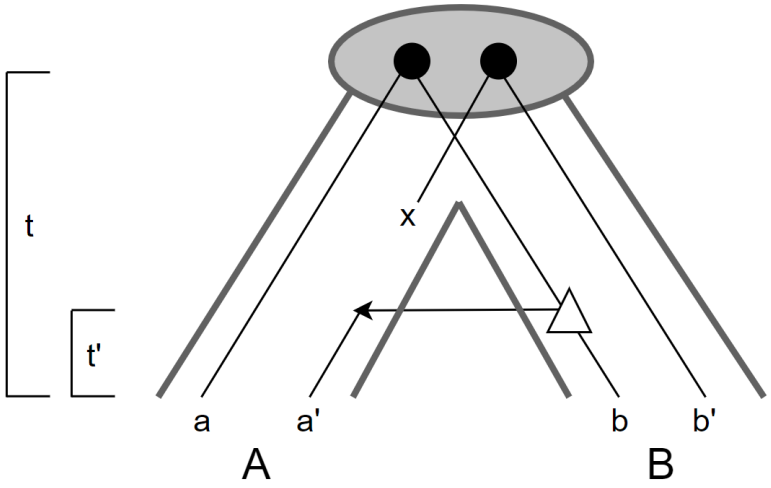

Abbildung 2.4: Beispiel: GERINGER-Relation. t und t' stehen für die vergangene Zeit im markierten Abschnitt. Weitere Notationen wie in 2.3

Der HGT von b nach a' in Abbildung 2.4 tritt zeitlich gesehen nach (genauer: nach $t - t'$) dem Aufteilen der Spezies und damit nach dem letzten gemeinsamen Vorfahren von A und B auf, welcher in diesem Fall gleichzeitig die Wurzel des Baumes ist. Das bedeutet, dass a' vor kurzem noch ein Gen von B bzw. genau b war, und deswegen jetzt die maximal möglichen Unterschiede und damit die Distanz zwischen a' und b nicht so groß sind wie etwa zwischen a und b: Die vergangene Zeit und damit die Anzahl an möglichen Mutationen zwischen a' und b ist mit $2 \cdot t'$ deutlich geringer als die Zeit $2 \cdot t$ zwischen a und b.

Dieser Umstand ist von essentieller Bedeutung, denn die geringere Distanz zwischen a' und b ist nicht nur speziell im Vergleich zu der Distanz zwischen a und b auffällig: Da HGTs immer noch die Ausnahme sind, ist die Distanz zwischen a' und b im Vergleich zu allen anderen Distanzen zwischen allen anderen Genpaaren (a_i, b_j), die nicht aus HGTs entstanden sind und aus A und B das Best-Match zueinander sind, deutlich geringer. Diese Unregelmäßigkeit kann nur aus den Distanzdaten erkannt und ausgewertet werden.

In Abbildung 2.3 würde z.B. bei einem Vergleich von a und c der unterste HGT als HGT der GERINGER-Relation erkannt werden. Der letzte gemeinsame Vorfahre von A und C ist die mittlere Spezies, und der HGT trat nach der Aufteilung dieses Vorfahren auf.

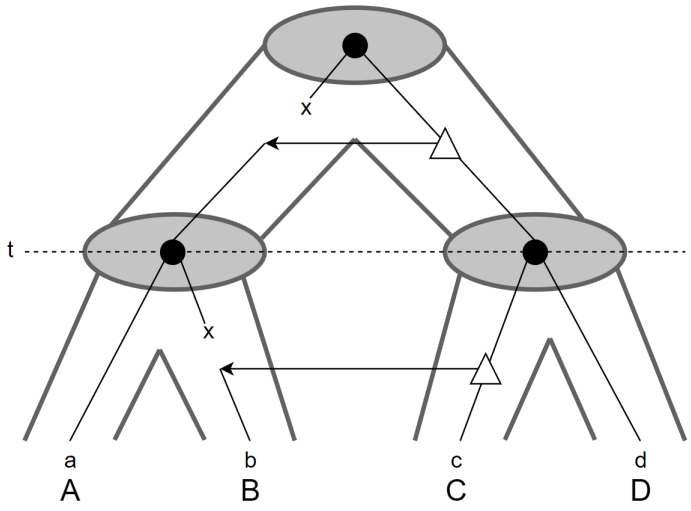

Abbildung 2.5: Beispiel: HÖHER-Relation. Notation wie in Abbildung 2.3

Anhand der HGTs in Abbildung 2.5 wird demonstriert, dass die Relation und Einordnung eines HGTs nicht nur abhängig von dem absoluten Zeitpunkt des HGTs sondern viel mehr von den beiden Genen abhängt, deren evolutionäre Vergangenheit untersucht wird.

Wenn die Vergangenheit von a und c betrachtet wird, ist zu erwarten, dass die Distanz zwischen den beiden Genen geringer ist als zwischen anderen, nicht eingezeichneten, Genen von A und C ohne HGT. Der erste HGT liegt zeitlich gesehen nach dem letzten gemeinsamen Vorfahren (der Wurzel), weswegen bei a im Gegensatz zu besagten anderen Genen von A weniger Zeit vergangen ist, in der sich die Distanz des jeweiligen Gens durch Mutationen zu c vergrößert haben könnte. Damit gehört dieser HGT im Bezug auf a und c zu der GERINGER-Relation.

Auf der anderen Seite kann aber auch die Vergangenheit von a und b untersucht werden. Der letzte gemeinsame Vorfahre von A und B ist das linke Kind der Wurzel. Der erste HGT ist zeitlich gesehen vor der Aufteilung dieser Spezies aufgetreten. Deswegen wird der HGT im Bezug auf a und b in die HÖHER-Relation eingeordnet. Zu erwarten ist außerdem, dass die Distanz zwischen a und b höher ist als zwischen anderen Genen aus A und B mit jeweils ihrem Best-Match. Das liegt daran, dass andere Gene aus A nur seit dem letzten gemeinsamen Vorfahren Zeit hatten um zu divergieren bzw. zu mutieren, während b diese Zeit und zusätzlich die Zeit seit dem ersten HGT hatte. Diese ungewöhnlich höhere Distanz von a zu b kann aus den Gendaten abgelesen und ausgewertet werden.

Der untere HGT stellt dabei nur die Verwandtschaft von a und b her, und ist für die Unterschiede der Distanz nicht verantwortlich. Zwischen b und c entsteht dadurch wieder eine GERINGER-Relation.

In Abbildung 2.3 ist z.B. der erste HGT im Bezug b und c' auf in die HÖHER-Relation einzuordnen, da der Zeitpunkt des HGTs wieder vor dem letzten gemeinsamen Vorfahren war. Im Gegensatz dazu würde dieser HGT bei der Betrachtung von c' und d' in die GERINGER-Relation eingeordnet, weil der letzte gemeinsame Vorfahre von C und D die Wurzel und damit vor dem HGT ist.

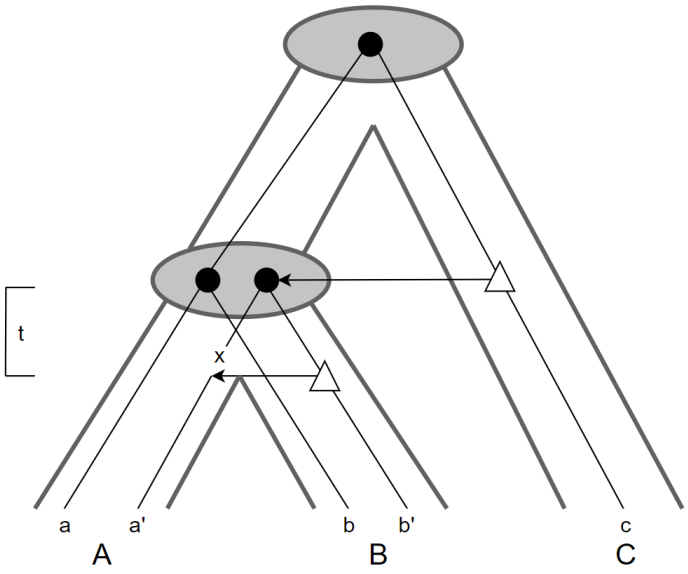

Abbildung 2.6: Beispiel: GLEICH-Relation. Notation wie in Abbildung 2.3

Die dritte und letzte Möglichkeit der Einordnung ist die GLEICH-Relation. Ein HGT bezüglich zwei Genen, welcher (fast) zur gleichen Zeit wie die Aufteilung des letzten gemeinsamen Vorfahrens dieser Gene aufgetreten ist, wird in diese Kategorie eingeordnet. Das „fast“ ist hier nur in Klammern, da es schwierig ist, den genauen Zeitpunkt eines HGTs zu bestimmen. In Abbildung 2.6 ist der HGT von c zum linken Kind der Wurzel relativ eindeutig einzuordnen: Die Kopie von c teilt sich bei der Aufteilung der Spezies mit auf und könnte vielleicht sogar als „Grund“ für die Unterscheidung der Spezies danach angesehen werden.

Schwieriger wird es bei der Einordnung des zweiten HGTs von b' nach a' unter der Annahme, dass die Zeit t minimal ist: Wenn der Baum bekannt ist, wäre es nach der Definition der GERINGER-Relation nur richtig, den HGT im Bezug auf b' und a' auch in die GERINGER-Relation einzuordnen. Aber in dieser Arbeit ist der Startpunkt der Untersuchung nicht der fertige korrekte Baum, sondern die reinen Distanzdaten zwischen den Genen. Und das bedeutet, dass die Einordnung nur anhand der Distanzen nicht sicher gewährleistet werden kann, weil der Unterschied zwischen einem HGT und einer Deletion (wie in Abbildung 2.6) im Gegensatz zu einer normalen Aufteilung („speciation“) des Gens nicht zu erkennen ist.

Selbst in dem Idealmodell mit konstanter Mutationsrate könnte ein solcher HGT nicht eindeutig bestimmt werden, da spätere Mutationen frühere Mutationen wieder umkehren können. Nur die maximal mögliche Anzahl der Unterschiede zwischen zwei Genen steigt proportional zu der vergangenen Zeit. Im Bezug auf Abbildung 2.6 könnte eine minimal größere Distanz zwischen a und b im Vergleich zu a' und b' also auch durch eine Mutation von a' während t erklärt werden, die eine frühere Mutation von a' während t wieder rückgängig gemacht hat. Die Distanz zu b' wäre dann nicht so groß wie wenn beide Mutationen eine dauerhafte Veränderung gebracht hätten.

Zusammengefasst ist die Einordnung von HGTs in die GLEICH-Relation prinzipiell möglich, aber für diese Arbeit aufgrund des Ansatzes mit dafür unzureichenden paarweisen Distanzvergleichen (Kapitel 3.1) nicht weiter relevant.

Im Verlauf dieser Arbeit werden verschiedene Formulierungen für die Einordnung von HGTs in eine Relation verwendet. Ein „HGT der HÖHER-Relation“ ist synonym für „es existieren zwei Gene, bei deren Betrachtung dieser HGT in die HÖHER-Relation eingeordnet wird“, da, wie gezeigt, die Einordnung eines HGTs immer abhängig von den betrachteten Genen ist. Aus Gründen der Übersichtlichkeit und Lesbarkeit wird auf die ausführliche Beschreibung an manchen Stellen verzichtet.

2.4 EXISTIERENDE METHODEN ZUR ERKENNUNG VON HGTS

Die Frage, wie HGTs erkannt werden können, stellt sich praktisch seit der ersten Beobachtung von HGTs im Jahre 1928 (Griffith, 1928). Die seitdem bis heute entwickelten Methoden lassen sich durch ihre Ansätze in zwei Kategorien gruppieren: parametrische und phylogenetische Methoden (Poptsova u. a., 2007; Ravenhall u. a., 2015).

Parametrische Methoden wurden vor den phylogenetischen Methoden entwickelt. Sie versuchen anhand der Eigenschaften eines einzelnen Genoms auszumachen, ob ein Gen horizontal transferiert wurde, da vor der Sequenzierung von Genen Vergleiche von Genen schwierig waren. Mögliche Eigenschaften reichen dabei u.a. von der Struktur der Gene (Worning u. a., 2000) bis hin zu dem GC-Gehalt, also wie viele Nukleinbasen des Gens die Basen Guanin und Cytosin sind (Daubin u. a., 2003). Diese Eigenschaften unterscheiden sich zwischen verschiedenen Spezies oft genug, um Unregelmäßigkeiten im Vergleich zu dem „Durchschnitt“ des Genoms in dieser Eigenschaft zu erkennen.

Der Vorteil ist dabei offensichtlich: das Genom der Originalspezies des horizontal transferierten Gens muss nicht bekannt sein, um einen Transfer trotzdem erkennen zu können. Auf der anderen Seite kann es aber durch Mutationen des horizontal transferierten Gens in der neuen Spezies zu Mutationen kommen, die eher der Art der neuen Spezies entsprechen. Das bedeutet, dass sich die Eigenschaften des Gens, die zur Erkennung seines Transfers genutzt werden, mit der Zeit denen der neuen Spezies anpassen, wodurch ein weit zurückliegender HGT möglicherweise nicht mehr erkannt werden kann (Ochman u. a., 1997).

Phylogenetische Methoden verfolgen einen anderen Ansatz: Sie verwenden insbesondere vorhandene Sequenzdaten verschiedener Genome, um aus dem Vergleich dieser Daten miteinander zu erkennen welche Gene möglicherweise durch einen HGT entstanden sind. Dadurch können auch zeitliche Einordnungen und damit das Modellieren der Evolution der betrachteten Spezies vorgenommen werden.

Hierbei gelten wieder gewisse Grenzen: wie im vorigen Kapitel 2.3 thematisiert können manche HGTs durch phylogenetische Methoden nicht immer korrekt erkannt werden. Es ist damit schwierig, verlässlich richtige Gen- und Speziesbäume zu erhalten, wodurch Annahmen, die aufgrund dieser Baumes getroffen werden, immer eine Unsicherheit enthalten. Dieser Umstand wird durch variable Mutationsraten noch verstärkt. Hinzu kommt, dass sich durch die Verwendung von mehreren Gensequenzen auch gleichzeitig auf genau diese Daten beschränkt wird. Im Gegensatz zu parametrischen Methoden können hier HGTs nicht richtig erkannt werden, wenn das Ursprungsgenom oder Empfängergenom nicht Teil der benutzten Sequenzdaten ist (Ravenhall u. a., 2015).

Um die Sicherheit von gefundenen Kandidaten für einen HGT zu erhöhen, bietet es sich also an, diese Methoden und andere Kriterien bis zu einem gewissen Grad zu kom-

binieren. Forschung zu diesen Themen kann z.B. bei Arvey u. a. (2009), Azad u. a. (2011) sowie Xiong u. a. (2012) nachgelesen werden. Zusammenfassend ist hervorzuheben, dass eine parametrische Methode allein oft zu ungenau ist und deswegen mehrere Parameter oder Ansätze gleichzeitig deutlich bessere Ergebnisse liefern. Wichtig ist dabei, dass jede Änderung der Kriterien gleichzeitig die Balance der false-positives (Genpaare, die als HGT erkannt werden, aber keine sind) und false-negatives (HGTs, die nicht erkannt werden) verändert: manche HGTs passen z.B. nur zu einem der gewählten Kriterien und werden dann fälschlicherweise ausgeschlossen (Poptsova u. a., 2007).

Für diese Arbeit sind primär nur die phylogenetischen Methoden relevant, da der später vorgestellte Algorithmus auch zu ihnen gehört. Phylogenetische Methoden können weiter in explizite und implizite Methoden unterteilt werden (Dessimoz u. a., 2008; Ravenhall u. a., 2015). In den folgenden Unterkapiteln werden die Ausführungen von Ravenhall u. a., 2015 für einen Überblick über bestehende und verwandte Methoden zur Erkennung von HGTs dargestellt.

2.4.1 Explizite Methoden

Explizite phylogenetische Methoden versuchen signifikante Unterschiede zwischen Gen- und Speziesbaum zu erkennen. Zum Beispiel kann die Vergangenheit von zwei Genen im Genbaum mit ihren Spezies im Speziesbaum verglichen werden: Unterschiede in der evolutionären Vergangenheit können auf einen HGT hindeuten: wenn etwa der letzte gemeinsame Vorfahre der Gene nur sehr kurz zurückliegt, der letzte gemeinsame Vorfahre der Spezies aber weiter in der Vergangenheit liegt, kann ein HGT vermutet werden. Zur Verdeutlichung dieses Ansatzes kann wieder Abbildung 2.3 genutzt werden: Der letzte gemeinsame Vorfahre der Spezies A und D ist die Wurzel des Baumes. Wenn die Gene a und d betrachtet werden, stimmt deren letzter gemeinsamer Vorfahre mit dem ihrer Spezies überein. Die Gene a' und d' hingegen sind aus einem HGT in der Vergangenheit von d' entstanden. Der Zeitpunkt der Aufteilung und somit der Zeitpunkt ihres letzten gemeinsamen Vorfahrens liegt somit deutlich vor dem letzten gemeinsamen Vorfahren ihrer jeweiligen Spezies, was auf einen HGT der GERINGER-Relation hindeutet.

Es gibt verschiedene Ansätze für die expiziten Methoden:

- Als Topologie Tests werden Methoden bezeichnet, die statistische Tests nutzen, um Gene bzw. Genfamilien zu erkennen, deren Topologie bzw. evolutionäre Vergangenheit sich schlecht in den Speziesbaum einordnen lassen. Das bestimmende Kriterium ist dabei das „Sequenzalignment“, also ob die Sequenz der Gene genug Ähnlichkeiten mit der erwarteten Sequenz aufweist, die anhand der Topologie der Gene abgeschätzt wird. Wenn die Unterschiede dabei zu groß sind, was bedeutet, dass die Topologie der Gene bzw. der Genfamilie nicht zu dem Speziesbaum passt, und diese Unterschiede auch nicht durch andere Events wie Duplikationen und Deletionen (siehe Abbildung 2.6 und Erklärung bezüglich a' und b') erklärt werden können, wird ein HGT vermutet.
  Ein Beispiel für diesen Ansatz ist der Approximately Unbiased Test (Shimodaira, 2002). Er kann für einen Baum berechnen, wie wahrscheinlich es ist, dass dieser Baum die tatsächliche Vergangenheit der zugrunde liegenden Gendaten beschreibt.

Jedem getesteten Baum wird dabei ein P-Wert zugewiesen, der genau diese Wahrscheinlichkeit beschreibt; je größer P für einen Baum ist, desto wahrscheinlicher ist es. Durch die Festlegung einer Grenze α ist es dann möglich, die Bäume auszuschließen, die unwahrscheinlich sind ($P < \alpha$), und gleichzeitig ein Set von mehreren Bäumen zu erhalten, die relativ wahrscheinlich sind ($P > \alpha$). Um HGTs mit diesem Test zu erkennen, wird die evolutionäre Vergangenheit der Spezies mit den vorhandenen Gendaten und einer gewählten Grenze α verglichen: Wenn der P-Wert des Speziesbaumes nicht größer als α ist bedeutet dies, dass sich die evolutionäre Vergangenheit der Spezies nicht besonders stark mit der Vergangenheit der betrachteten Gene ähnelt. Diese Unterschiede können durch HGTs erklärt werden (Shimodaira, 2002; Poptsova u. a., 2007).

- *Genome spectral approaches*-Methoden versuchen, durch die Betrachtung von Teilbäumen Rückschlüsse auf HGTs zu ermöglichen. Diese Teilbäume werden durch zwei Arten erzeugt:
  - Wenn eine Kante aus einem Baum entfernt wird, bleiben zwei nicht verbundene Teilbäume zurück. Diese beiden Teilbäume werden als Bipartition bezeichnet. Wenn diese Bipartitionen sowohl in dem Genbaum als auch dem Speziesbaum vorkommen, wird sie als kompatibel bezeichnet. Die Teilbäume und damit die evolutionäre Vergangenheit stimmt in beiden Bäumen überein.
  - Ein „(embedded) quartet“ ist ein Teilbaum eines Baumes, der vier Blätter hat. Diese Blätter können sowohl Blätter des Originalbaumes sein, als auch weitere Teilbäume. Wenn ein solches „quartet“ aus dem Speziesbaum auch im Genbaum vorkommt, wird es als kompatibel angenommen. Inkompatible quartets hingegen können auf HGTs hindeuten (Zhaxybayeva, 2006).

  Da es auch ohne HGTs zu Unterschieden in den Bäumen kommen kann, werden bei der Auswahl der Teilbäume nicht beliebige Aufteilungen vorgenommen, sondern eine entsprechende Bewertung mithilfe des Bootstrapping-Verfahrens oder der A-posteriori-Wahrscheinlichkeit durchgeführt (Alfaro, 2003; Douady, 2003; Poptsova u. a., 2007). Ausgewählte Aufteilungen werden als „strongly“ oder „well supported“ bezeichnet.

- Ein weiterer Ansatz ist das Entfernen von Teilbäumen an einer Stelle und dem Ansetzen dieses Teilbaumes an einer anderen Stelle (*Subtree pruning and regrafting*, SPR). Um damit HGTs zu erkennen, werden solche Operationen auf den Speziesbaum angewandt: Wenn der Genbaum und der Speziesbaum vorher gleich bzw. konsistent sind, können die Bäume nach einer solchen Operation nur inkonsistent, also nicht gleich, sein. Umgekehrt bedeutet das auch: wenn durch diese Operationen ein konsistenter Zustand erreicht wird, müssen die ursprünglichen Bäume inkonsistent gewesen sein, was wiederum auf HGTs hindeutet. Durch die Beobachtung, welche Operationen in einem konsistenten Zustand resultieren, können auch Aussagen über den Zeitpunkt und Richtung (welches Gen wurde wohin übertragen) des HGTs getroffen werden.
  Auch bei diesem Ansatz gibt es Kriterien, welche Operationen durchgeführt werden. SPR als Problem ist NP-schwer (Hickey, 2008), weswegen möglichst effiziente Verfahren benutzt werden müssen, um die Performance zu verbessern (Beiko, 2006; Abby, 2010).

- Modell-basierte Abgleichmethoden versuchen, dem Genbaum so Events zuzuordnen, damit er konsistent bezüglich des Speziesbaumes ist. Events sind dabei neben denen in Kapitel 2.1 beschriebenen Events wie HGTs, Duplikationen und Deletionen auch weitere wie „Incomplete lineage sorting“ und „Homologous recombination“. Verschiedene Modelle benutzen dabei verschiedene Events, um die Unterschiede zwischen den Bäumen zu erklären. Je mehr Events betrachtet werden, desto komplexer und vielfältiger werden auch die möglichen Lösungen, um den Genbaum konsistent zum Speziesbaum zu machen.

2.4.2 Implizite Methoden

Implizite phylogenetische Methoden untersuchen und vergleichen die Distanzen zwischen Genen und auch ihren Spezies. Der Hintergrund dieser Methoden wurde bereits in Kapitel 2.3 angesprochen: Durch HGTs der verschiedenen Relationen kommt es zu unerwarteten Distanzen, die bei der Betrachtung von den betroffenen Genen auffallen und einen HGT indizieren. Wie die expliziten Methoden sind auch die impliziten Methoden von zufälligen Entwicklungen und Mutationen betroffen, die zu Distanzen führen können, die dann fälschlicherweise als HGT interpretiert werden.

- Ein einfacher und naiver Weg, um HGTs zu erkennen, ist die Suche nach dem Best-Match in einer entfernt verwandten Spezies. Da die Best-Matches eines Gens eigentlich in nah verwandten Spezies zu vermuten sind, sind Treffer in entfernten Spezies auffällig. Eine Suche kann z.B. mit dem Tool BLAST (Altschul u. a., 1990) durchgeführt werden (Nelson u. a., 1999).
  Durch diesen Ansatz können nur HGTs der GERINGER-Relation erkannt werden, da nur sie im Gegensatz zu HGTs der HÖHER-Relation eine deutlich geringere Distanz haben. Das bedeutet, dass nur kürzliche HGTs identifiziert werden, und so viele HGTs übersehen werden. Hinzu kommt, dass das Top match nicht unbedingt auch das evolultionär ähnlichste Gen sein muss (Koski u. a., 2001).

- Eine weitere implizite Möglichkeit, HGTs zu erkennen, ist die Unterschiede zwischen Gen- und Speziesdistanzen zu untersuchen. Dazu werden vorher wieder Annahmen über die Gen- und Speziesbäume getroffen: Wenn nur Speciation-Events auftreten würden, würden die Gen- und Speziesbäume aller Gene und Spezies übereinstimmen. Zusätzlich wird, wie bereits erwähnt, eine konstante Molekulare Uhr bzw. Mutationsrate angenommen. In diesem Fall sollten sich die Distanzen zwischen den Genen und ihren Best-Matches proportional zu der Distanz zwischen den Spezies entwickeln. HGTs hingegen führen zu deutlich geringeren bzw. höheren Distanzen zwischen den Genen, die mit einem HGT verbunden sind, als zwischen Genen ohne HGT (Novichkov u. a., 2004).
  Ein Beispiel für diese Methode ist der Algorithmus DLIGHT (Dessimoz u. a., 2008). DLIGHT steht für „Distance Likelihood based Inference of Genes Horizontally Transfered“, und wie im Namen beschrieben werden Tests durchgeführt, die die Wahrscheinlichkeiten berechnen, ob für die betrachteten Gene ein HGT aufgetreten ist und ob kein HGT aufgetreten ist. Ist die Wahrscheinlichkeit, die für einen HGT spricht, deutlich höher, wird der HGT ausgegeben. Dabei können auch Aussagen über die Transferrichtung und die zeitliche Einordnung getroffen werden.

- Die Erkennung von HGTs durch Phylogenetische Profile geht auf die Forschung von Pellegrini, 1999 zurück. Proteine, die aufgrund einer gemeinsamen evolutionären Vergangenheit eine ähnliche Struktur besitzen und ähnliche Funktionen verkörpern, werden in Proteinfamilien zusammengefasst. Ein Phylogenetisches Profil ist ein Kette von Zahlen mit der Länge gleich der Anzahl der untersuchten Genome. Es beschreibt die Proteinstruktur der untersuchten Genome: Wenn ein Protein der gesuchten Proteinfamilie im n-ten Genom vorkommt, wird an der n-ten Stelle im Profil eine 1 geschrieben. Wenn keine Übereinstimmung gefunden wird, wird eine 0 geschrieben.
  HGTs werden durch die unerwartete Abwesenheit von Proteinen erkannt: Wenn ein Protein bzw. eine Proteinfamilie nur in einem Genom vorkommt, obwohl die phylogenetischen Profile bei der Betrachtung anderer Proteinfamilien durch ihre Ähnlichkeit eigentlich auf eine kürzliche Aufteilung und Verwandschaft hindeuten, ist dies ein Indikator dafür, dass das Protein und damit das Gen aus einem HGT entstanden ist und nur zu diesem einem Genom hinzugefügt wurde.
  Phylogenetische Profile eignen sich ebenfalls dafür, die evolutionäre Vergangenheit zu rekonstruieren (Csűrös u. a., 2009). Zum Beispiel können sehr ähnliche Proteine in einigen Genomen auf ein kürzliches Speciation-Event hindeuten, wodurch diese Genome näher verwandt als die, die diese sehr ähnlichen Proteine nicht haben.

3 METHODE

3.1 ANSATZ ZUR LÖSUNG

In dieser Arbeit wird ein impliziter Ansatz verfolgt, weil die Gen- bzw. Distanzdaten bereits verfügbar sind und deren Nutzung folglich sinnvoll ist. Die HGTs werden also anhand der berechneten Distanzen der Gene zueinander erkannt. Um das zu erreichen muss eine Metrik festgelegt werden die bestimmt, wann ein Genpaar bzw. eines der Gene aus einem HGT entstanden ist und nicht aus einem unbekannten anderen Event. Diese Metrik muss einen Ausgleich zwischen false-positives und false-negatives finden. False-positives sind hier Genpaare, die zwar eine geringe Distanz zueinander haben, aber tatsächlich nicht aus einem HGT entstanden sind. False-negatives beschreiben HGTs, die übersehen werden, weil die Distanz zwischen den Genen schon z.B. aufgrund eines länger zurückliegenden HGTs zu hoch sind.

Der Algorithmus wird nur paarweise Vergleiche von Distanzen durchführen. Deswegen beschränkt er sich auf die Erkennung von HGTs der HÖHER- und GERINGER-Relation, weil die Erkennung von GLEICH-Relationen nicht sicher gewährleistet werden kann (Kapitel 2.3). Auf der anderen Seite werden damit rechenintensive Baumoperationen und Berechnungen vermieden, was bei der Wahl der Programmiersprache eine gewisse Offenheit bietet. Darüber hinaus wird die Erkennung von HGTs nur in der Forschung verwendet, weswegen der Algorithmus nicht zeitkritisch ist, aber die Laufzeit trotzdem noch überschaubar sein sollte.

Ein weiterer wichtiger Punkt ist die Frage in welcher Form die Gendaten übergeben werden. Für den Algorithmus genügt die Annahme, dass auf alle Gene und jede Distanz zu allen anderen Genen ohne Mehraufwand zugegriffen werden kann, aber für die Implementierung eine angemessene Lösung gefunden werden muss. Um mögliche Kompatibilitätsprobleme zu vermeiden und Einheitlichkeit zu gewährleisten, ist die Benutzung von existierenden Möglichkeiten einer eigenen Implementierung zur Berechnung der Distanzen vorzuziehen.

Diese Aspekte dienen als Orientierung und Leitfaden für die Entwicklung und Implementierung des Algorithmus in diesem Kapitel. Um die Funktionsweise des Algorithmus besser darzustellen wird sie zuerst an einem Beispiel durchgeführt, dann verbal erläutert und zuletzt in Pseudocode umgesetzt.

NOTATION:

- Große Buchstaben $\{A, B, C, ...\} \in \mathcal{S}$ beschreiben verschiedene Spezies, wobei $\mathcal{S}$ das Set aller eingegebenen Spezies ist.
- Kleine Buchstaben mit Indizes $\{a_1, a_2, b_1, b_2, ...\}$ beschreiben einzelne Gene der entsprechenden Spezies (a_i ist ein Gen von A, b_j von B, ...)

Weitere Notationen werden bei Bedarf eingeführt.

3.2 ERKLÄRUNG AN EINEM BEISPIEL

Seien A, B und C drei Spezies mit jeweils drei verschiedenen Genen. Die Distanzen zwischen den Genen dieser Spezies wird durch die folgende Matrix gegeben. Um zusätzliche Darstellungsformen zu vermeiden, folgt sie dem Schema des erst in Kapitel 3.5.1 vorgestellten Nexus-Formats. Für den Algorithmus nicht relevante Distanzwerte (Distanzen zwischen den Genen derselben Spezies und die 0-Diagonale von Genen zu sich selbst) werden durch „`------`“ dargestellt. Die Werte in dieser Matrix dienen nur der Verdeutlichung und Darstellung, wie der Algorithmus HGTs erkennt und orientieren sich nicht an realen Daten.

```
BEGIN ALLDISTANCES;
    distances
        name=example triangle=both
            A/00001 ------ ------ ------ 5.4000 5.6000 5.7000 9.1000 9.3000 9.6000
            A/00002 ------ ------ ------ 5.6000 6.3000 5.8000 9.2000 1.7000 9.4000
            A/00003 ------ ------ ------ 5.5000 5.7000 5.1000 9.8000 9.4000 9.1000
            B/00001 5.4000 5.6000 5.5000 ------ ------ ------ 7.2000 7.5000 7.7000
            B/00002 5.6000 6.3000 5.7000 ------ ------ ------ 7.4000 7.1000 7.6000
            B/00003 5.7000 5.8000 5.1000 ------ ------ ------ 7.8000 7.7000 7.3000
            C/00001 9.1000 9.2000 9.8000 7.2000 7.4000 7.8000 ------ ------ ------
            C/00002 9.3000 1.7000 9.4000 7.5000 7.1000 7.7000 ------ ------ ------
            C/00003 9.6000 9.4000 9.1000 7.7000 7.6000 7.3000 ------ ------ ------
            ;
```

Abbildung 3.1: Matrix mit Beispielwerten

Die Gene aller Spezies besitzen also jeweils eine Distanz zueinander. Die Distanzen zwischen den Genen von A und B liegen dabei im Bereich von `5.1` bis `6.3`, und zwischen B und C im Bereich von `7.1` bis `7.8`. Die Gene von A und C haben mit der Ausnahme von `A/00002` zu `C/00002` eine Distanz von `9.1` bis `9.8` zueinander.

Aus dem Vergleich der Distanzen zwischen den Genen der Spezies zueinander kann man ablesen, dass die Ähnlichkeit von A und B größer ist als die Ähnlichkeit von A und C sowie B und C. A und B sind damit enger verwandt, was auf einen Speziesbaum ähnlich zu dem in Kapitel 2.2 hindeutet (ohne D und unabhängig von den HGTs). Das heißt: Sucht man in anderen Spezies das ähnlichste Gen für ein beliebiges a_i (das „Best Match“), findet man es am ehesten in B und nicht in C. Dieser Umstand wird vom Algorithmus überprüft und bei einer Verletzung dieser Annahme genauer untersucht, ob es sich um einen HGT gehandelt haben könnte.

In diesem Beispiel wird diese Annahme durch `A/00002` zu `C/00002` verletzt: Obwohl alle anderen Gene a_i zu c_j eine Distanz von $\geq$`9.1` besitzen, haben diese beiden Gene nur eine Distanz von `1.7`.

Der Algorithmus vergleicht für jedes Gen x aus jeder Spezies X die Reihenfolge der Spezies und sortiert sie entsprechend welche Spezies das nächst-ähnliche Gen zu x besitzt. Für `A/00001` und `A/00003` ist diese Reihenfolge jeweils $\{B, C\}$, für `A/00002` aber $\{C, B\}$[1]. Diese Überprüfung findet anhand der Verteilung der Positionen von C im Bezug

1 Best Matches für `A/00001`: {`B/00001`, `C/00001`}, für `A/00002`: {`C/00002`, `B/00002`}, für `A/00003`: {`B/00003`, `C/00003`}

auf andere Gene von X statt. In diesem Beispiel wäre die Verteilung von C im Bezug auf `A/00002` $[0, 2]$[2], weil sowohl für `A/00001` als auch `A/00003` B das ähnlichste und C nur das zweit-ähnlichste Gen besitzt.

Ist die Position von C signifikant weiter vorne als in allen anderen Reihenfolgen von allen anderen a_i bzw. der Verteilung, würde der Algorithmus einen möglichen HGT erkennen. In diesem Fall ist die Position von C in der Reihenfolge von `A/00002` 0, was signifikant (siehe letzer Absatz) kleiner ist als die sonstigen Positionen von C ($[0, 2]$). Der mögliche HGT muss aber noch durch die Prüfung von `C/00002` zu `A/00002` bestätigt werden: nur wenn für beide gilt, dass das Best Match der anderen Spezies deutlich näher liegt als erwartet wird der HGT ausgegeben.

Wichtig ist also, dass der Algorithmus nicht nur eine geringe Distanz liest und es als HGT versteht, sondern vielmehr versucht, diese geringere Distanz in Relation mit der erwarteten Position zu setzen. Der genaue Distanzwert verliert an Bedeutung sobald die Reihenfolge der Spezies berechnet wurde.

Wie oben vermerkt ist das Beispiel entsprechend beschränkt: die Position von C im Bezug auf `A/00002` ist hier in der Verteilung nur um eine Stelle nach vorne verschoben, und ob das bereits als „signifikant“ bezeichnet werden sollte ist diskussionwürdig. Hätten die Gene von B und C eine ähnliche durchschnittliche Distanz zu den Genen von A, wären solche unregelmäßigen Vertauschungen aufgrund der zufälligen Mutationsrate sogar normal. Deswegen gibt es genau die Einschränkung, dass die Position signifikant weiter vorne sein muss, die dann aber entsprechend auch erst bei Daten mit mehr als nur drei Spezies wirklich Bedeutung erlangt. Dies wäre z.B. der Fall wenn die Matrix durch beliebig viele weitere Spezies erweitert wird, von denen alle Gene zu den Genen von A eine Distanz zwischen `1.8` und `9.0` haben. Bei fünf hinzugefügten Spezies, die diese Bedingung erfüllen, würde die Position 0 von C im Bezug auf `A/00002` dann mit der Verteilung von C mit $[0, 0, 0, 0, 0, 0, 2]$ schon eher intuitiv als „signifikant“ kleiner bezeichnet werden.

3.3 FORMALE DEFINITION DES ALGORITHMUS

In diesem Kapitel werden der Algorithmus und seine Abschnitte formal erläutert. Es wird wieder angenommen, dass alle Distanzen zwischen allen Genen bekannt sind. Für ein besseres Verständnis werden die verschiedenen Schritte mit dem Pseudocode in Kapitel 3.4 verknüpft.

3.3.1 Best-Match-Distanzen für jedes Gen

Zuerst muss für jedes Gen a_i jeder Spezies $A \in \mathcal{S}$ eine geordnete Liste $\mathcal{L}_{a_i}$ von Spezies erstellt werden. Diese Liste wird aufsteigend anhand der Best-Match-Distanzen von a_i zu jeder Spezies sortiert, wobei Best-Match-Distanz hier einfach die kürzeste Distanz zwischen a_i und allen b_j mit $B \in \mathcal{S} \setminus A$ ist.

Wenn es mehrere Spezies gibt, zu denen a_i dieselbe Best-Match-Distanz hat, werden diese Spezies auch als gleich weit entfernt behandelt. Es kann also z.B. mehr als eine Spezies geben, die ein Gen besitzt welches a_i am ähnlichsten ist, und diese Spezies

2 C ist null Mal an Position 0 und zwei Mal an Position 1

werden dann alle an die erste Stelle bzw. Position gesetzt. Da der restliche Algorithmus und insbesondere die spätere Verteilung der Spezies von diesem Aspekt aber nicht betroffen sind, wird er, auch der Übersichtlichkeit wegen, nicht mehr explizit behandelt. Das ist nur möglich, weil die Verteilungen einer Spezies unabhängig von den Positionen der anderen Spezies berechnet wird. Ob zwei Spezies „am nächsten" sind ist für die Verteilung irrelevant.

Diese Liste $\mathcal{L}_{a_i}$ beschreibt also eine Reihenfolge von Spezies, welche das von der Distanz her ähnlichste Gen zu a_i besitzen und dient später als Ausgangspunkt, um die Distanzen anderer Gene zu a_i in einen Kontext einordnen zu können. Diese Listen werden für alle Gene von A erstellt, bevor mit dem nächsten Schritt begonnen wird. Die Listen der Gene anderer Spezies werden für die folgenden Vergleiche von a_i nicht benötigt.

Im Pseudocode wird dieser Schritt bis Zeile 6 realisiert. Die verwendete Methode `berechneBestMatchReihenfolge` berechnet dabei genau $\mathcal{L}_{a_i}$, wobei a_i das als Parameter übergebene Gen ist.

3.3.2 Vergleich mit anderen Genen

Jedes Gen a_i jeder Spezies $A \in \mathcal{S}$ wird jetzt mit allen Genen b_j jeder Spezies $B \in \mathcal{S} \setminus A$ verglichen. Dabei werden diese b_j nacheinander anhand der Distanz zu a_i überprüft, beginnend bei dem mit der kleinsten Distanz (dem „Best Match", notiert b_1).

Der Vergleich findet mit Hilfe der Liste $\mathcal{L}_{a_i}$ statt: die Position von B in $\mathcal{L}_{a_i}$ wird entsprechend der Distanz zwischen a_i und b_j neu berechnet. Da $\mathcal{L}_{a_i}$ auf den Best-Match-Distanzen basiert, wird sich die Position von B bei dem ersten Vergleich nicht verändern, weil b_j bzw. b_1 das Best-Match ist. Bei den folgenden Vergleichen kann B in dieser Liste nur nach hinten rutschen, da die Distanz von a_i zu b_j und damit B im Vergleich zu den anderen Spezies in $\mathcal{L}_{a_i}$ nur schlechter werden kann. Diese neue temporäre Liste wird als $\mathcal{L}'_{a_i}(b_j)$ notiert und nur für den Vergleich von a_i mit b_j benötigt.

Um diese Liste $\mathcal{L}'_{a_i}(b_j)$ auswerten zu können, wird noch ein Vergleich über die durchschnittliche Entfernung von B und A benötigt. Als Vergleich bietet sich hier die Verteilung der Positionen von B im Bezug auf A an. Diese Verteilung kann aus den Listen $\mathcal{L}_{a_1}$ bis $\mathcal{L}_{a_n}$ (ohne $\mathcal{L}_{a_i}$) gebildet werden, wobei n die Anzahl aller Gene von A ist. Aus diesen Listen werden die Spezies ignoriert, die nicht in $\mathcal{L}_{a_i}$ vorkommen, da sie für diese Verteilung bezüglich (a_i, b_j) nicht relevant sind. Diese Verteilung wird als $\mathcal{D}_{a_i}(B)$ notiert.

Die Berechnung der Verteilung findet von Zeile 8 bis 19 statt.

Entscheidend ist hierbei dass, wenn b_j als Kandidat verworfen wird, auch alle nachfolgenden b_k mit $k > j$ übersprungen werden können. Die Position von B in $\mathcal{L}'_{a_i}(b_k)$ kann nur gleich oder schlechter sein als in $\mathcal{L}'_{a_i}(b_j)$, weil nacheinander aufsteigend nach der Distanz verglichen wird, vgl. erster Absatz. Anders ausgedrückt: Wenn b_j nach der gewählten Metrik nicht aus einem HGT entstanden ist, ist es für alle b_k noch unwahrscheinlicher.

Die beschriebene Abbruchbedingung wird durch `break;` in Zeile 42 umgesetzt.

3.3.3 Auswertung

Die Berechnung eines Grenzwertes zur Einschätzung der Position findet von Zeile 21 bis 32 statt. Die Position von B in $\mathcal{L}'_{a_i}(b_j)$ kann jetzt mit der Verteilung $\mathcal{D}_{a_i}(B)$ bzw. dem daraus berechneten Grenzwert verglichen werden:

- Eine ähnliche Position deutet darauf hin, dass die Distanz zwischen a_i und b_j den anderen Distanzen von beliebigen a_x zu ihren Best-Matches aus B ähnelt. Daher ist anzunehmen, dass die gleiche Zeit seit dem letzten gemeinsamen Vorfahren vergangen ist, und es deswegen zu einer ähnlichen Anzahl von Mutationen und dadurch Vergrößerungen der Distanz bei a_i und a_x gekommen ist.
  a_i und b_j wird als Kandidatenpaar verworfen, und es kann mit a_{i+1} fortgefahren werden (siehe letzter Absatz Kapitel 3.3.2).

- Eine signifikant kleinere bzw. frühere Position von B in $\mathcal{L}'_{a_i}(b_j)$ als in $\mathcal{D}_{a_i}(B)$ deutet auf einen HGT in einer GERINGER-Relation hin: Die Unterschiede zwischen a_i und b_j sind deutlich geringer als zwischen anderen beliebigen a_x zu ihren Best-Matches aus B, was auf deutlich weniger Mutationen und damit Vergrößerungen der Distanz hinweist, was wiederum primär nur aus einer kürzeren Zeit seit dem letzten gemeinsamen Vorfahren herrühren kann. Es ist also wahrscheinlich, dass a_i erst nach diesem Vorfahren zu dem Genpool von A hinzugefügt wurde, und möglicherweise das Original oder die Kopie von b_j ist.
  (a_i, b_j) wird als Kandidatenpaar gespeichert und es kann mit (a_i, b_{j+1}) fortgefahren werden, denn b_{j+1} könnte auch noch als Kandidat in Frage kommen.

- Eine deutlich spätere Position ist ebenfalls auffällig, und kann vom Algorithmus auf eine ähnliche Weise erkannt werden: Anstatt nur zu überprüfen ob die Position von B signifikant kleiner ist (wie im Pseudocode Kapitel 3.4), kann gleichzeitig überprüft werden ob die Position signifikant größer ist

Wenn a_i und b_j sich gegenseitig als Kandidatenpaar erkennen, d.h. sowohl (a_i, b_j) als auch (b_j, a_i) wurden als Kandidatenpaar gespeichert, wird a_i mit b_j am Ende als erkannter HGT der GERINGER-Relation zusammen mit möglichen anderen HGTs zurückgegeben.

Dieser Schritt findet von Zeile 34 bis 37 statt. Wie im folgenden Kapitel beschrieben, wird auf die Darstellung für die HÖHER-Relation verzichtet.

3.4 PSEUDOCODE DES ALGORITHMUS

Der nachfolgende Pseudocode ist an Java angelehnt, da auch die finale Implementierung in Java realisiert wurde (Kapitel 3.5.2). Es wird angenommen, dass auf alle Distanzen zwischen allen Genen zugegriffen werden kann.

Aus Gründen der Übersichtlichkeit ist die Erkennung von HGTs der HÖHER-Relation im Pseudocode nicht extra vermerkt. Sie funktioniert auf die gleiche Weise wie die Erkennung der GERINGER-Relation, nur dass nicht im kleinsten Teil der Verteilung gesucht wird sondern im größten Teil. Dabei wird der gleiche Prozentwert benutzt.

Algorithm 1 Verteilungsalgorithmus

```
Require: Set 𝒮 aller Spezies
for all Spezies X aus 𝒮 do
  for all Gen x aus X do
    for all Spezies Y aus 𝒮 \ X do
      for all Gen y aus Y, aufsteigend sortiert anhand der Distanz von y zu x do
        List<Spezies> reihenfolge = berechneBestMatchReihenfolge(x);
        int positionVonKandidatenspezies = reihenfolge.getPositionVon(Y);

        int[] verteilung; //... von den Positionen, die Y im Bezug auf X
        int anzahlWerteInVerteilung = 0;
        for all Gen x' aus X do
          List<Spezies> reihenfolgeTemp = berechneBestMatchReihenfolge(x');
          for int i = 0; i < reihenfolgeTemp.size(); i++ do
            if reihenfolgeTemp.get(i) == Y then
              verteilung[i]++;
              anzahlWerteInVerteilung++;
              break;
            end if
          end for
        end for

        //Berechne, wann eine Position „signifikant“ kleiner ist
        //Als Beispiel: Wenn die Position in den kleinsten 5% der Verteilung liegt
        Double prozent = 0.05;
        int grenzwert = 0;
        int anzahlGefundenerWerte = 0;
        for int i = 0; i < verteilung.length; i++ do
          anzahlGefundenerWerte += verteilung[i];
          if anzahlGefundenerWerte ≥ anzahlWerteInVerteilung * prozent then
            grenzwert = i;
            break;
          end if
        end for

        if positionVonKandidatenspezies < grenzwert then
          //Möglicher HGT zwischen x und y erkannt. Wenn dieser auch
          //ausgehend von y bestätigt wird, werden x und y als HGT-Paar
          //der GERINGER-Relation ausgegeben.
        else
          //Da die Gene von Y aufsteigend überprüft werden muss, wenn y_i kein
          //HGT ist, y_{i+1} auch nicht mehr überprüft werden, da y_i schon näher
          // an x war und nicht erkannt wurde.
          break;
        end if
      end for
    end for
  end for
end for
```

Mit dem Pseudocode können außerdem zwei Fragen aus Kapitel 1.5.1 beantwortet werden: Es ergibt sich eine exponentielle Laufzeit von mindestens $O(n^5)$, da keiner der Größen in der FOR-Schleifen linear ist. In der Praxis wird diese Laufzeit aber nie erreicht, da durch die verschiedenen `breaks` die Loops vorzeitig beendet werden. Es werden z.B. nie alle Gene aus Y betrachtet (außer Y hat nur ein Gen).

Außerdem ist Multithreading bzw. Parallelisierung möglich: Jede Iteration des Codeblockes 5 bis 43 läuft unabhängig voneinander. Deswegen ist es theoretisch möglich, mit genügend Prozessoren eine Laufzeit von O(n) zu erreichen. Tatsächlich könnten die restlichen FOR-Schleifen ebenfalls noch parallel gelöst, wodurch sich eine Laufzeit von O(1) ergeben würde. Diese Überlegungen sind natürlich nicht praxistauglich: Bei einem entsprechend großen Datensatz wäre ein durchschnittliches System bereits durch die Parallelisierung der ersten FOR-Schleife ausgelastet. Hinzu würde dann noch der Overhead der verschiedenen Threads kommen: Der eigentliche Gewinn durch die `breaks` geht verloren, da sofort alle Gene aus Y überprüft werden und es insgesamt mehr berechnet wird.

3.5 IMPLEMENTIERUNG

In diesem Kapitel wird die Implementierung des in Kapitel 3.3 beschriebenen Algorithmus näher erläutert. Dafür werden zuerst die in Kapitel 1.5.2 aufgestellten Fragen beantwortet, bevor auf die wichtigsten Aspekte des Codes eingegangen wird.

3.5.1 SWP 2018/19 und das Nexus Dateiformat

An der Universität Leipzig findet jährlich im Wintersemester das Modul Softwaretechnik-Praktikum (SWP)[3] für Informatikstudenten im Bachelor statt. Ziel dieses Moduls ist das erste Sammeln von Erfahrungen im Bereich der Softwareentwicklung in einem Team sowie die Organisation von größeren Projekten. Dabei wird jede Gruppe von einem Betreuer unterstützt.

Im Wintersemester 2018/19 bearbeitete die Gruppe nw18a das Thema „Rekonstruktion phylogenetischer Bäume“. Das Endprodukt dieser Gruppe[4] war eine umfassende Softwaresuite zum Einlesen und Verarbeiten von Gendaten, programmiert in Java. Sie kann u.a. zufällig von dem Tool ALFSim[5] (Dalquen u. a., 2012) generierte Gendaten einlesen und die verarbeiteten Daten in das Nexus-Dateiformat (Maddison u. a., 1997) exportieren. Relevant für diese Arbeit ist dabei primär der Export des `ALLDISTANCES`-Blocks mithilfe von `alfextract.jar`[6]:

3 `http://bis.informatik.uni-leipzig.de/de/Lehre/BAMA/SWP`, abgerufen am 25.6.2021
4 `http://pcai042.informatik.uni-leipzig.de/swp/SWP-19/nw18a/`, abgerufen am 25.6.2021
5 `http://www.alfsim.org/`, abgerufen am 25.6.2021
6 `alfextract.jar -i path/to/alfsimResultFolder --all -o output.nex`

```
BEGIN ALLDISTANCES;
    distances
        name=exampletree triangle=both
            SE001/00001 0.0000 2.5000 8.5000
            SE002/00001 2.5000 0.0000 5.0000
            SE003/00001 8.5000 5.0000 0.0000
            ;
    distances
        name=secondtree triangle=both
            ...
```

`SE00X` ist hier der Name einer Spezies. Die Zahl hinter dem Speziesnamen nach dem Backslash ist eine eindeutige ID des Gens dieser Spezies. Danach folgen die Matrix- bzw. Distanzwerte: `SE001/00001` hat zu `SE002/00001` eine Distanz von 2.5, `SE001/00001` zu `SE003/00001` eine Distanz von 8.5, usw.

Die Tags `name` und `triangle` sind in diesem Kontext nicht gebrauchte Zusatzinformationen, die das Erstellen und Einlesen des Blockes beeinflussen. `name` ist dabei der Name eines Baumes, während `triangle` definiert, welche Hälften der Matrix in dem Block gespeichert sind. Mögliche Werte sind hierbei `upper`, `lower` und `both`, und beschreiben jeweils die Hälften ober- bzw. unterhalb der `0.0000`-Diagonale. Da die Werte an dieser Diagonale gespiegelt sind, kann man bei der Erstellung der Nexus-Datei mit `alfextract.jar` darauf verzichten, beide Hälften zu speichern.

Dieser Block enthält somit alle Distanzen, die der vorgestellte Algorithmus benötigt, wobei verschiedene Bäume getrennt voneinander gelistet sind. Für den Algorithmus bedeutet das, dass Gene verschiedener Bäume eine Distanz von ∞ bzw. einen definierten maximalen Wert zueinander haben.

Darüber hinaus wurde im SWP-Projekt die Klasse `NexusReader.java` erstellt, welche Nexus-Dateien allgemein und insbesondere diese `ALLDISTANCES`-Blöcke einliest und in eine geeignete Struktur in Java überträgt. Da sowohl der `ALLDISTANCES`-Block als auch dieser Reader von nw18a entwickelt wurden, bietet es sich an beide Strukturen in eine eigene Implementierung zu übernehmen.

3.5.2 Grundlagen

An erster Stelle steht die Entscheidung über die verwendete Programmiersprache(n). Eine objektorientierte Sprache erscheint hier sinnvoll, da mit Spezies und Genen immer wieder auf Attribute der Instanzen dieser Klassen zugegriffen und sie verglichen werden. Die Entscheidung fiel aus mehreren Gründen auf Java:

- Wie in Kapitel 3.5.1 beschrieben ist das SWP-Projekt in Java programmiert. Um die Datenstruktur und insbesondere `NexusReader.java` zu übernehmen und zusätzliche ausführbare Dateien zu vermeiden, muss also Java gewählt werden.
- Java ist plattformunabhängig. Das fertige Programm funktioniert damit auf beliebigen Betriebssystemen und muss nicht neu kompiliert werden.

- Der Algorithmus benötigt keine grafische Oberfläche, da während der Ausführung keine Eingaben getätigt werden müssen. Optionale Ausgaben können auch nur auf der Konsole erscheinen.
- Es werden keine externen Libraries benötigt, da alle Strukturen und Operationen des Algorithmus allein mit Java realisierbar sind. Eine grafische Oberfläche würde hier sogar hinderlich sein da z.B. JavaFX nicht mehr im Standard-JDK enthalten ist.
- Da auf rechenintensive Berechnungen verzichtet wird, ist die schlechtere Performance von Java im Vergleich zu z.B. C++ vernachlässigbar. Trotzdem wurde ein grundlegendes Multithreading verwendet, da es sich auf der einen Seite anbietet und auf der anderen Seite eine deutliche Verbesserung der Laufzeit ermöglicht[7].

Das normale Java liefert also schon alle für den Algorithmus benötigte Komponenten. Darüber hinaus gibt es bei der Version praktisch keine Einschränkungen: Verwendet wurde OpenJDK Version 1.8.0_282 und ausgeführt werden kann es mit JRE Version 1.8.0_281[8], wobei eine Aufwärtskompatibilität gegeben ist.

Um den Algorithmus zu individualisieren, sind Konsolenparameter beim Start ausreichend. Eine Config-Datei ist wegen der relativ geringen Anzahl an möglichen Konfigurationen nicht erforderlich. Wenn der Nutzer Parameter wiederverwenden möchte, ist dies auch über ein eigenes Start-Skript möglich.

3.5.3 Datenstruktur

Obwohl der Algorithmus keine Datenbank oder ähnliche persistente Speicherungen benötigt, ist es doch sinnvoll, intern datenbankähnliche Strukturen zu verwenden und klare Parent-Child-Konzepte zu etablieren.

SPECIES.JAVA

Diese Klasse dient primär der Übersichtlichkeit der Implementierung. Ihr einziges eigenes Attribut `name` könnte auch als Attribut zu `Gene.java` hinzugefügt werden. Dann könnten Instanzen davon immer noch anhand dieses Namens in „gedachte“ Spezies gruppiert werden. Um sich aber weiter an das Paradigma der objektorientierten Programmierung zu halten und den Code intuitiver zu machen, bleibt diese Klasse als „Container“ für Gene erhalten.

Die Klasse hat folgende Attribute:

- `String name`: Name dieser Spezies
- `List<Gene> genes`: Liste von allen Genen die in der Nexus-Datei für diese Spezies gefunden wurden

7 Der Zeitgewinn hängt von der Struktur der Daten und der Leistungsfähigkeit der CPU ab. Beobachtet wurde u.a. eine mehr als dreifache Geschwindigkeit bei 100% CPU Auslastung gegenüber <20% Auslastung ohne Multithreading. Verwendet wurde eine Nexus Datei mit 36 Spezies und knapp 4000 Genen.

8 `https://www.oracle.com/java/technologies/javase-downloads.html`, abgerufen am 25.6.2021

GENE.JAVA

Diese Klasse repräsentiert ein einzelnes Gen und enthält alle für den Algorithmus bzw. den Verlgeichsprozess wichtige Distanzwerte. Um die Verteilungen $\mathcal{D}$ zu berechnen, kann über `parent.genes` auf die Liste $\mathcal{L}$ der anderen Gene der gleichen Spezies zugegriffen werden.

Das bedeutet außerdem, dass Distanzwerte doppelt gespeichert werden: Die Distanz zwischen einem Gen a_i und b_j ist in beiden Objekten in der `distances` Map hinterlegt.

Die Klasse hat folgende Attribute:

- `Long id`: Eindeutige ID des Gens innerhalb seiner Spezies
- `Species parent`: Spezies, zu der dieses Gen gehört
- `Map<Species, Double> bestMatchDistances`: Map in der die Best-Match Distanzen zu allen anderen Spezies gespeichert sind. Wird zu Beginn des Algorithmus gefüllt und zur Erstellung der Liste $\mathcal{L}$ (Kapitel 3.3.1) genutzt
- `List<Gene> potentialCandidateGenes`: Liste von Genen, die für dieses Gen potentielle Kandidaten für einen HGT sind
- `Map<Gene, Double> distances`: Map, die alle Distanzen von diesem Gen zu allen anderen Genen enthält

DISTANCESPECIESLISTTUPLE.JAVA

Die Listen $\mathcal{L}$ haben diese Klasse als Element. Da entsprechend Kapitel 3.3.1 mehrere Spezies die gleiche Distanz zum aktuellen Gen haben können, können sie mit diesem Tupel an der gleichen Position in $\mathcal{L}$ stehen und verfälschen die Verteilung nicht.

Die Klasse hat folgende Attribute:

- `double distance`: Distanz, die alle Spezies in `species` zum aktuell betrachteten Gen haben
- `List<Species> species`

GENEDISTANCES.JAVA

Eine Instanz dieser Klasse wird vom `NexusReader.java` für jeden Baum erstellt, den er beim Einlesen einer Nexus Datei findet. Die Attribute sind dann eine direkte Repräsentation der Daten in der Nexus-Datei.

Die in der Matrix enthaltenen Werte werden dann genutzt, um das `distance` Attribut der `Gene.java` Instanzen korrekt mit allen Distanzen zu befüllen.

Die Klasse hat folgende Attribute:

- `List<Species> species`: Alle Spezies, die in diesem Baum vorkommen
- `List<Gene> genes`: Alle Gene, die in diesem Baum vorkommen
- `Matrix<Double> distances`: Matrix aller Distanzwerte aus diesem Baum
- `String treeName`: Name des Baums

3.5.4 Aspekte der finalen Implementierung

Der Quellcode ist Open-Source und kann auf Github (Kapitel 4) eingesehen und beliebig weiter verwendet werden. Im gleichen Repository befindet sich auch die ausführbare JAR-Datei und somit das Ergebnis der Implementierung. Auf sie sowie den Code wird hier nicht genauer eingegangen, da der Quellcode durch Kapitel 3.4 schon ausreichend beschrieben wird und die JAR-Datei nur die kompilierte Version des Quellcodes ist. Daher werden in diesem Kapitel nur noch kurz die Eigenschaften thematisiert, die nicht direkt mit dem Code zu tun haben.

KOMMANDOZEILENPARAMETER

- Hilfe `[-h] [-help]` erklärt diese Parameter
- Konsolenoutput. Default: Aktueller Fortschritt wird ausgegeben.
  - `[-s] [-silent]`: Kein Konsolenoutput (außer Fehlermeldungen)
  - `[-v] [-verbose]`: Alle potentiellen Kandidaten sowie alle letztendlich bestätigten HGTs werden ausgegeben
- Input Datei im Nexus-Dateiformat `[-i filename]`. Kein Defaultwert, es muss explizit eine Datei angegeben werden!
  Möglich sind hier der Dateiname, wenn sie im gleichen Verzeichnis wie die JAR liegt, sowie relative und absolute Pfade.
- Output Datei `[-o output]`. Default: output.nex
- Prozentsatz, in welchem Bereich der Verteilung $\mathcal{D}$ das Match sein soll. Default: In den kleinsten 0.05 bzw. 5%. `[-p 0.05]`
- Gesuchte Relation `[-higher] [-lesser]`. Default: Beide Relationen werden gesucht. Beschränkt die Suche nach HGTs auf die spezifizierte Relation
- Multithreading `[-dm] [-disableMultithreading]`. Deaktiviert Multithreading, wodurch die CPU Auslastung verringert aber dafür die Laufzeit erhöht wird

RESTRIKTIONEN DES INPUTS

Die Input-Datei wird im Nexus-Formati mit einem `ALLDISTANCES`-Block erwartet. Dabei kann nur genau eine Input-Datei angegeben werden. Der Algorithmus ist zwar in der Lage, eine beliebige Anzahl von Bäumen einzulesen (innerhalb eines `ALLDISTANCES`-Block), aber auf die Möglichkeit, mehrere Nexus Dateien mit mehreren solcher Blöcke als Input angeben zu können, wurde aus Gründen der Konsistenz der Daten verzichtet.

Um mehrere Dateien eingeben zu können, müssen sie zuerst in eine einzelne Nexus-Datei zusammengeführt werden.

NEXUSREADER.JAVA

Anstelle des Original `NexusReader.java` aus dem SWP-Projekt wurde eine gekürzte Version `NexusReaderDistances.java` verwendet, welche nur `ALLDISTANCES`-Blöcke einlesen kann und den restlichen Inhalt der Nexus-Datei überspringt. Darüber hinaus wurden die Attribute der Klasse leicht modifiziert, um besser für die hier verwendete Datenstruktur verwendet werden zu können.

OUTPUT-FORMAT

Der Output der Applikation ist wieder eine Nexus-Datei. Sie enthält einen Block `BEGIN HGTRELATIONS`, in dem für jeden Genbaum eine Matrix gespeichert wird. Diese Matrix stellt die HGT-Relationen zwischen den Genen dieses Baumes dar: -1 steht für einen HGT der GERINGER-Relation, 1 für einen HGT der HÖHER-Relation und eine 0 für keinen erkannten HGT.

Der folgende Block dient der Verdeutlichung, wie eine Output-Datei aussehen könnte:

```
#NEXUS
BEGIN HGTRELATIONS;
[matrices each describing the type of horizontal gene transfer between genes]
    [LESSER=-1, EQUALORDEFAULT=0, GREATER=1]
    hgtrelation
        name=G1 triangle=both
            SE001/00001  0  0  0  0  0  0
            SE002/00001  0  0 -1  0  0  1
            SE002/00002  0 -1  0  0  0  0
            SE003/00001  0  0  0  0  0  0
            SE003/00002  0  0  0  0  0  0
            SE004/00001  0  1  0  0  0  0
            ;
    hgtrelation
        name=G2 triangle=both
            ...
END;
```

4 ANALYSE DER ERGEBNISSE

4.1 METHODIK

Das Ergebnis dieser Arbeit ist der implementierte Algorithmus zur Erkennung und Einordnung von HGTs. Der Quellcode sowie eine ausführbare JAR-Datei befinden sich in einem Repository unter

`https://github.com/Varauk/HorizontalGeneTransferDetection`.

Im Folgenden wird die Fähigkeit des Algorithmus, HGTs korrekt zu erkennen, anhand von zwei Beispieldatensätzen untersucht (siehe Kapitel 1.5.1). Diese Beispieldatensätze, die im oben genannten Repository unter `src/resources` zu finden sind, bestehen aus den Nexus-Dateien `out10.nex` und `out1000.nex`, die zufällige, mit ALFSim generierte, Gendaten mit 10 bzw. 1000 Genen pro Spezies enthalten. Verwendet wurde dafür die im gleichen Verzeichnis liegende Parameter-Datei `alf-params.drw`.

Da diese Nexus-Dateien die tatsächlichen Distanzen zwischen den Genen enthalten, diese aber in der Praxis nicht bekannt sind, wäre eine Auswertung nur aufgrund dieser Werte nicht aussagekräftig. Um reale Umstände zu simulieren wird ein weiteres Ergebnis des SWP-Projektes (Kapitel 3.5.1) verwendet: Das Skript `AlfToDist.sh` berechnet mithilfe von `distmat`[1] aus dem EMBOSS Softwarepaket[2] (Rice u. a., 2000) aus den von ALFSim generierten FASTA-Dateien[3] die Ähnlichkeit von allen Genen zueinander. Diese Ähnlichkeiten bzw. Distanzen werden dann als Nexus-Datei ausgegeben, die den gleichen `ALLDISTANCES`-Block und damit das gleiche Format wie die von `alfextract.jar` generierte Nexus-Datei hat. Sie kann damit ebenfalls als Input für diesen Algorithmus bzw. dessen Implementierung genutzt werden. Diese Dateien liegen ebenfalls im `src/resources` Verzeichnis, namentlich `dists10.nex` und `dists1000.nex`.

Wie gut der Algorithmus darin ist HGTs korrekt zu erkennen, kann durch einen Vergleich der Liste der erkannten HGTs mit dem in der jeweiligen `out`-Datei enthaltenen `RELATIONS`-Blocks bewertet werden. Dieser Block enthält unter anderem für jeden Genbaum eine Matrix[4], die die tatsächlichen HGT-Relationstypen zwischen den Genen zueinander enthält. Dabei wird die gleiche Notation wie in der Output-Datei dieses Algorithmus (Kapitel 3.5.4) benutzt: -1 steht für die GERINGER-Relation, 0 für die GLEICH-Relation und 1 für die HÖHER-Relation. Dadurch ergibt sich eine einfache, nachvollziehbare Vergleichbarkeit der Ergebnisse.

Beobachtet werden insbesondere drei Werte:

- Anzahl der erkannten und korrekt eingeordneten HGTs

1 `http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/distmat.html`, abgerufen am 25.6.2021
2 `http://emboss.sourceforge.net/`, abgerufen am 25.6.2021
3 `https://zhanglab.dcmb.med.umich.edu/FASTA/`, abgerufen am 25.6.2021
4 `type=RCB name=baumname triangle=both`

- Anzahl der erkannten, aber falsch eingeordneten HGTs (false-positives)
  Dieser Wert in Relation mit der Anzahl der korrekt eingeordneten HGTs ist signifikant für die Bewertung des Algorithmus. Obwohl es aufgrund der Zufälligkeit von Mutationen und Evolution unwahrscheinlich ist keine HGTs falsch einzuordnen, sollte er so gering wie möglich sein. Andernfalls wäre die Verwendung dieses Algorithmus nicht empfehlenswert.

- Anzahl der nicht erkannten HGTs (false-negatives)
  Da der verwendete Ansatz HGTs nur anhand von starken Abweichungen von der erwarteten Distanz zwischen Genen erkennt ist es wahrscheinlich, dass die Anzahl der false-negatives deutlich höher ist als die der erkannten HGTs. HGTs der HÖHER- und GERINGER-Relation, deren Zeitpunkte sich der GLEICH-Relation annähren, sind in den meisten Fällen nicht in z.B. den kleinsten oder größten 5% der Verteilung. Im Gegensatz zu den false-positives ist eine hohe Anzahl hier aber unkritisch, da wenige, aber dafür sichere, Ergebnisse immer noch hilfreich sind.

Um mehrere Vergleiche durchzuführen und gleichzeitig die Auswirkung des Parameters `p` zu verdeutlichen, werden die Ergebnisse mit variierenden Prozentwerten berechnet. Die Anzahl der gefundenen HGTs und die Anzahl der nicht erkannten HGTs addiert sich bei einer `out`-Datei nicht zwangsläufig mit dem Wert derselben `out`-Datei bei anderen `p` Werten oder der zugehörigen `dists`-Datei, weil die Anzahl von richtig erkannten GLEICH-Relationen unterschiedlich sein kann. Da dieser Algorithmus nur HÖHER- und GERINGER-Relationen erkennt ist eine Übereinstimmung bei der GLEICH-Relation nur der Standard-Fall, wenn weder HÖHER noch GERINGER erkannt wurde. Dadurch gehören sie in keine der eben beschriebenen Kategorien.

Der Übersichtlichkeit halber werden die vergleichbaren Prozentwerte anhand des Graphen in Abbildung 4.1 ausgewertet; die zugrunde liegenden Tabellen befinden sich im Anhang A (Tabellen A.1 - A.11).

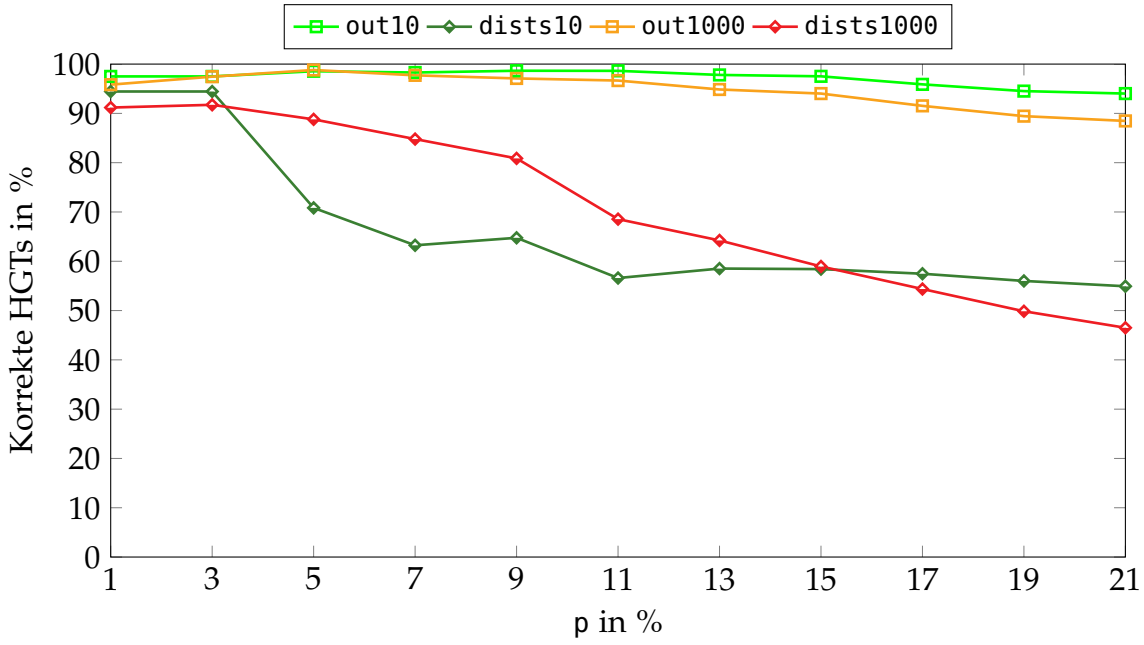

Abbildung 4.1: Anteil der korrekt eingeordneten HGTs in Abhängigkeit von `p`

Um ein detaillierteres Bild über die Korrektheit der Ergebnisse des Algorithmus zu bekommen, können die erkannten HGTs auch aufgeteilt in GERINGER- und HÖHER-Relation untersucht werden. In Abbildung 4.2 wird der Anteil der korrekt erkannten HGTs im Datensatz mit 1000 Genen pro Spezies in die Relationen aufgeteilt, um daran eventuelle Unterschiede zu erkennen. Dieser Datensatz wurde gewählt, da sich durch die erhöhte Anzahl an Genen und Spezies die Kurven eher dem Erwartungswert einer zufälligen Generierung annähern.

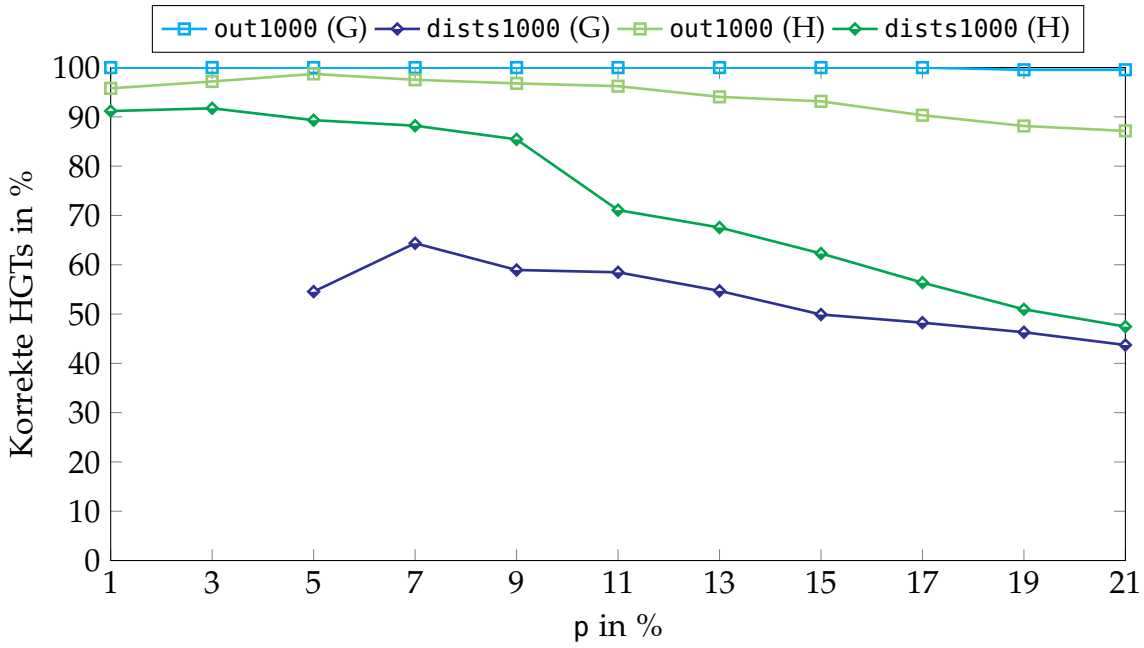

Abbildung 4.2: Vergleich von korrekt erkannten HGTs der GERINGER- und HÖHER-Relation in Abhängigkeit von p. Mit (G) markierte Kurven basieren auf HGTs der GERINGER-Relation, mit (H) markierte HGTs auf denen der HÖHER-Relation

Die Auswirkungen der Kurven in Abbildung 4.2 auf die zusammengefassten Kurven in Abbildung 4.1 ist abhängig von der absoluten Anzahl der korrekt erkannten HGTs der verschiedenen Relationen; dieses Verhältnis wird im folgenden Kapitel genauer untersucht.

Auch im restlichen Kapitel werden Werte für p immer in Prozent angegeben; der Kommandozeilenparameter erwartet die Dezimaldarstellung.

4.2 AUSWERTUNG

Aus den in Abbildung 4.1 vorgestellten Testdaten lassen sich folgende zentrale Erkenntnisse und Ergebnisse ableiten:

- Eine Erhöhung von p führt meistens zu mehr false-positives.

  Dieses Verhalten ist grundsätzlich erwartet, da der Ansatz des Algorithmus genau darauf abzielt die extremen Abweichungen durch HGTs zu erkennen. Je größer p ist desto höher ist die Wahrscheinlichkeit, dass eine durch zufällige Mutationen entstandene ungewöhnlich geringe bzw. hohe Distanz fälschlicherweise als HGT

erkannt wird.
An manchen Stellen steigt der Anteil der korrekt erkannten HGTs. Das tritt sowohl bei kleinen Werten für `p` als auch bei `p` = 11 für `dists10` auf. Diese Abweichungen entstehen durch entsprechend generierte HGTs: Kleinere Schwankungen sind durch die Zufälligkeit der Generierung normal, die generelle Verschlechterung des Verhältnis der erkannten HGTs wird dadurch aber nur wenig beeinflusst.

- Der Anteil von korrekt erkannten HGTs ist in den Dateien mit den tatsächlichen Distanzen deutlich höher und stabiler.

  Während die Kurven von `out10` und `out1000` beide konstant gute Ergebnisse repräsentieren und erst bei ungefähr `p`=11 merklich schlechter werden, brechen die Kurven von `dists10` und `dists1000` schon deutlich früher bei `p`=3 ein. Darüber hinaus kommen starke Veränderungen nur bei den `dists`-Dateien vor: Die größte Änderung beträgt bei den `dists`-Dateien 23,61% (`dists10` bei der Erhöhung von `p`=3 auf 5), wohingegen bei den `out`-Dateien die größte Änderung lediglich 2,49% (`out1000` bei der Erhöhung von `p`=15 auf 17) beträgt.
  Dass der Algorithmus mit den tatsächlichen Distanzwerten bessere Ergebnisse liefert als mit simulierten Praxiswerten hat zwei primäre Gründe:

  – Bei den tatsächlichen Werten gibt es weniger zufällige Abweichungen, wie sie bei den Praxiswerten durch deren Berechnung bzw. Abschätzung entstehen. Die zentrale Annahme des Algorithmus wird dadurch seltener durch von der Erwartung abweichenden Distanzwerten verletzt. Dies ist sozusagen der „naive“ Grund, warum in den meisten Fällen bei der Auswertung von Daten tatsächliche (Ideal-)Werte meistens bessere Ergebnisse liefern als beobachtete Werte.

  – Die von ALFSim generierten Distanzwerte verschiedener Gene gleichen sich oft, sogar Genbaum-übergreifend. Zum Beispiel kommt der Distanzwert 164.864800 in `out10` in den Bäumen G1-G8 jeweils immer häufiger als 25 Mal vor. Wie in Kapitel 3.3.1 erwähnt werden Gene bzw. Spezies und deren Best-Match Gen, die dieselbe Distanz zu einem Referenzgen haben, auch als gleich weit entfernt behandelt. Die daraus resultierende Gruppierung auf Positionen der Verteilung kann zu folgendem Phänomen führen:
    Betrachtet wird ein Gen a. Wenn z.B. der berechnete Grenzwert (siehe Kapitel 3.4) 1 ist, würde der Algorithmus nur das erste bzw. ähnlichste Gen zu a in der Verteilung betrachten. Wenn zwei andere Gene b und c mit a in einer GERINGER-Relation sind und die gleiche Distanz zu a haben, können sie gleichzeitig als HGT erkannt werden. In den mit `distmat` berechneten Werten ist dies unwahrscheinlich: Die Distanz von b und c zu a ist vermutlich nicht gleich, wodurch nur das ähnlichste Gen als HGT erkannt wird. Das andere Gen hingegen, welches vielleicht eine nur minimal größere Distanz hat, wird bei einem entsprechenden Grenzwert verworfen. Dadurch sinkt die Anzahl der korrekt erkannten HGTs. Mit `p`=15 übersteigt die Anzahl der gefundenen HGTs bei `dists1000` sogar die bei `out1000`, was aber nur aufgrund der falschen Erkennung von HGTs passiert.

- Die Kurven der `out`-Dateien und der `dists`-Dateien ähneln sich untereinander.

  Die Ähnlichkeit der Kurvenverläufe der `out`-Dateien und der `dists`-Dateien bestätigt, dass die relative Anzahl an korrekt erkannten HGTs unabhängig von der Anzahl der Gene und der Spezies ist. Dadurch kann (bei ähnlichen Parameterwerten) das gleiche Ergebnis für beliebige andere Simulationen abgeschätzt werden. Bei den `out`-Dateien trat von `p`=1 bis 21 insgesamt eine Verschlechterung von $-3,48\%$ bei `out10` und $-7,36\%$ bei `out1000` auf. Bei den `dists`-Dateien betragen diese Veränderungen $-39,51\%$ und $-44,67\%$ bei `dists10` bzw. `dists1000`.
  Starke Einbrüche, wie bei `dists10` bei der Erhöhung von `p`=3 auf 5, sind in der Form bei keiner anderen Kurve zu beobachten. Bei diesem neuen Wert wurden sechs neue falsche HGTs erkannt, ohne dass neue richtige HGTs dazu kamen. Derartige „unglückliche“ Ereignisse in der Generierung werden in `dists1000` durch die deutlich höhere Anzahl von Genen ausgeglichen, wodurch sich die Generierung dem Erwartungswert annähert. Das erklärt den konstanteren Verlauf der Kurve von `dists1000`.

- Die Grundannahme des Algorithmus über die Verteilung ist nicht immer korrekt.

  Die Kernidee des Algorithmus ist, dass Gene, die in einer GERINGER-Relation sind, auch eine geringere Distanz zueinander haben. Das gleiche gilt umgekeht für Gene, die in einer HÖHER-Relation sind (siehe Kapitel 2.3). Diese Annahme stimmt zwar in den meisten Fällen (was durch die positiven Ergebnisse bei niedrigen Werten für `p` bestätigt wird), aber nicht immer. Ein Beispiel dafür ist die Relation zwischen den Genen `SE001/00052` und `SE008/00014` im Genbaum G3 (Abbildung 4.4: Laut der zu diesem Baum gehörenden Matrix im `RELATIONS`-Block sind diese Gene aus einem HGT der GERINGER-Relation entstanden. Der Algorithmus hingegen erkennt keine GERINGER-Relation, sondern sogar eine HÖHER-Relation. Die Distanz zwischen den beiden Genen ist mit 150.5902[5] die dritthöchste (und nur viertniedrigste) Distanz, die Gene von `SE008` zu `SE001` haben. Dementsprechend ist es nicht überraschend, dass keine GERINGER-Relation erkannt wurde und hier sogar die Kriterien für eine HÖHER-Relation erfüllt sind.
  Die folgenden Bäume wurden mithilfe des Tree Viewers des ETE Toolkits[6] erzeugt. Rote Einsen markieren dabei Speciation- bzw. Aufteilungs-Events. Der angegebene Maßstab steht für die Distanz zwischen den jeweiligen Objekten. Obwohl die Distanz zwischen Spezies hier nicht weiter relevant ist, und die Distanz zwischen den Genen besser aus `out10.nex` erfasst werden kann, ist dadurch trotzdem ein schneller Überblick über die verschiedenen Distanzen möglich. Die benutzte Baumcodierung befindet sich in `out10.nex` im `TREES`-Block. Abbildung 4.3 zeigt den Speziesbaum dieses Datensatzes.

5 150.5902 ist die tatsächliche Distanz zwischen den Genen. Der HGT wird eigentlich nur bei der Auswertung von `dists10` erkannt, aber da nur die auf den tatsächlichen Werten basierenden Bäume zur Verfügung stehen, wird das Problem an diesen erläutert. Auch in `dists10` weicht die Distanz nicht deutlich ab.

6 `http://etetoolkit.org/treeview/`

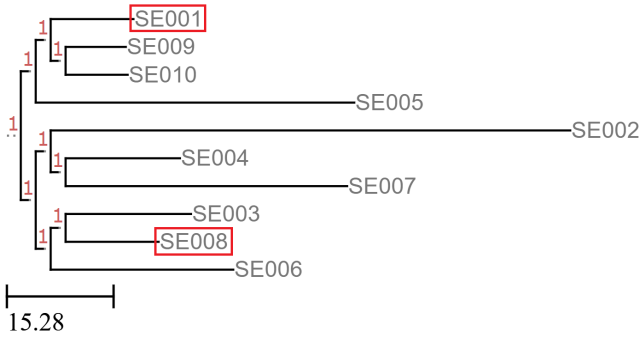

Abbildung 4.3: Speziesbaum des Datensatzes mit zehn Spezies

Der letzte gemeinsame Vorfahre von `SE001` und `SE008` ist die Wurzel des Baums. Es ist somit nicht möglich, dass ein HGT zwischen Genen dieser Spezies aufgetreten ist, der nach der Definition in Kapitel 2.3 in der HÖHER-Relation eingeordnet wird, da jeder HGT nur nach dem letzten gemeinsamen Vorfahren aufgetreten sein kann. Abbildung 4.4 zeigt den Genbaum G3.

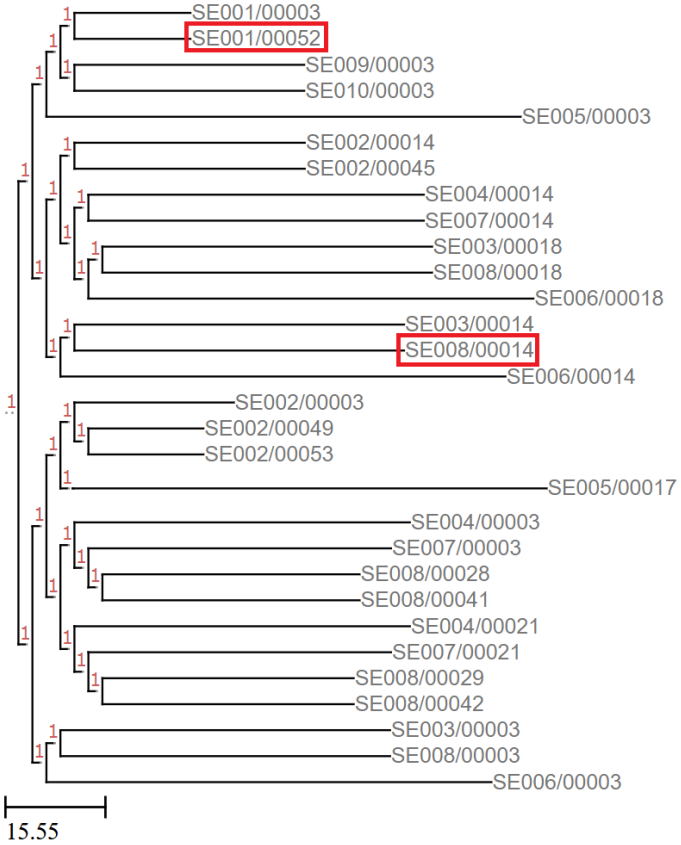

Abbildung 4.4: Genbaum G3 aus `out10.nex`

Der letzte gemeinsame Vorfahre der Gene `SE001/00052` und `SE008/00014` ist wie im Genbaum erkennbar also nicht die Wurzel, sondern eines der Kinder der Wurzel. Die Einordnung in die GERINGER-Relation ist also korrekt, trotz der relativ großen Distanz. Diese Umstände führen u.a. zu false-positives.

- In Relation zu den richtig erkannten HGTs gibt es sehr viele false-negatives.

  Wie zuvor erwähnt sind false-negatives ein erwartetes Problem, welches eine Empfehlung zur Nutzung dieses Algorithmus als zusätzliches Mittel zur Erkennung von HGTs zwar nicht beeinflusst, aber maßgeblich die alleinige Nutzung dieses Algorithmus zur Erkennung von HGTs bestimmt: Die Anzahl der false-negatives ist in fast allen Tabellen aus Anhang A ein Vielfaches der Anzahl der gefundenen HGTs und der korrekt erkannten HGTs. Selbst bei Werten für `p` wie 21, die nicht zu empfehlen sind, da für `dists1000` bereits mehr als die Hälfte der gefundenen HGTs false-positives sind, sind es immer noch fast zwei Mal so viele false-negatives wie gefundene HGTs. Ein großer Teil aller HGTs bleibt also unerkannt.
  Der Hauptgrund dafür ist, wie in Kapitel 4.1 angesprochen, dass nur HGTs im kleinsten bzw. größten Bereich der Verteilung erkannt werden. Im Bezug auf z.B. die GERINGER-Relation sind das also HGTs die erst sehr kurz zurückliegen; sobald die Zeit seit dem HGT (t' in Abbildung 2.4) so groß wird, dass zufällige Mutationen zu große Unterschiede zwischen der Kopie und dem Ursprungsgen verursachen können, werden diese HGTs nicht mehr sicher erkannt. Der verwendete Ansatz ist dementsprechend beschränkt, was sich auch so in den Testdaten widerspiegelt.

- Die sichersten Ergebnisse liefert ein Wert kleiner gleich 3 für `p`.

  Obwohl die Auswertung auf den tatsächlichen Werten eigentlich vielversprechend ist, ist letztendlich doch die Performance auf den (simulierten) Praxiswerten entscheidend. Und hier ist offensichtlich, dass Werte größer als 3 schnell zu einem stark erhöhten Anteil von false-positives führen, die wiederum auf den Ergebnissen aufbauende Arbeiten invalidieren. Da selbst mit minimalen Werten für `p` keine 100% richtige Erkennung stattfindet, muss sowieso ein Trade-Off bezüglich Anzahl der erkannten HGTs und Korrektheit dieser HGTs eingegangen werden. Im Folgenden wird sich primär auf die Auswertung des Datensatzes mit 1000 Genen pro Spezies bezogen, da es dabei wahrscheinlicher ist, dass sich dessen Werte denen einer zufälligen Generierung und damit dem Erwartungswert annähern.
  Im Bezug auf möglichst viele korrekt erkannte HGTs scheint dieser Trade-Off am besten bei `p`=3 zu sein. Größere Werte führen zwar zu mehr erkannten HGTs, aber auch zu mehr false-positives. Letztendlich liegt die Entscheidung beim Nutzer, wie sicher er seine Ergebnisse haben möchte. Werte größer als 3 für `p` sind also nur bedingt zu empfehlen.

Zusätzlich zu diesen „allgemeinen“ Erkenntnissen können noch speziellere Ergebnisse aus Abbildung 4.2 und den zugrunde liegenden Daten präzisiert werden:

- Für die GERINGER-Relation werden bessere Resultate bei den tatsächlichen Werten gefunden, aber schlechtere bei den simulierten Praxiswerten.

  Grund für diese Unterschiede ist die Simulation der Distanzwerte durch ALFSim, die wie bereits erwähnt häufig mehrfach verwendet werden (siehe Kapitel 4.2, zweiter Stichpunkt), . Beispielhaft wird G8 im Datensatz mit zehn Spezies untersucht (siehe Anhang B). Der Umstand wird am kleineren Datensatz ebenso deutlich, und die Tabellen in Anhang B.2 bleiben dadurch überschaubar.

Alle Gene, die in der GERINGER-Relation zueinander stehen, haben auch eine geringere Distanz zueinander als die anderen Distanzen der Gene zu anderen Genen. Tatsächlich sind die Distanzwerte ansich schon besonders: 58.117 (zwischen `SE006/00008` und `SE006/00047`) kommt im gesamten Datensatz nur hier vor, während 94.1424 (zwischen `SE007/00008` und `SE008/00034`) zwar auch in anderen Genbäumen vorkommt, aber auch dort immer für einen HGT der GERINGER-Relation steht. Im Bezug auf diese beiden Spezies ist es auch die kleinste vorkommende Distanz zwischen ihren Genen (vgl. Anhang B.2). Das gleiche gilt für 104.345 (zwischen `SE004/00008` und `SE008/00034`), diese Distanz kommt aber auch an anderen Stellen vor. Die Distanz zwischen `SE007/00008` und `SE008/00034` in `dists10` beträgt 69.25, die sich aber in das obere Drittel aller Distanzen zwischen `SE007` und `SE008` einordnet, und mit `p=2` sogar als HGT der HÖHER-Relation erkannt wird. Aus den Tabellen in Anhang B.2 ist auch erkennbar, dass es noch einen weiteren derartigen Außreißer gibt: `SE007/00019` mit `SE008/00026` in G4 haben eine Distanz von 85.04 (bzw. 148.8088) zueinander, und werden schon bei `p=9` als HGT der HÖHER-Relation erkannt. In beiden Fällen ist auffällig, dass eine vorher relativ geringe Distanz zu einer relativ hohen Distanz geworden ist. Dies kommt auch in der anderen Richtung vor. Die Distanz von `SE007/00021` mit `SE008/00014` beträgt in `out10` 164.8648, was die größte Distanz zwischen Genen von `SE007` und `SE008` ist. In `dists10` ist die berechnete Distanz nur noch 26.02, und damit sogar die zweit-kleinste Distanz (vgl. Anhang B.2).
Wichtig ist also, wie oft diese Unterschiede vom Algorithmus auch falsch eingeordnet werden. Dies wird durch die folgende Tabelle 4.1 präzisiert.

Datensatz	p in %	(G) als (H) eingestuft	(H) als (G) eingestuft
`out10`	1	0	0
`dists10`	1	16	1
`out1000`	1	31	0
`dists1000`	1	54	4
`out10`	15	1	0
`dists10`	15	30	1
`out1000`	15	119	0
`dists1000`	15	398	78

Tabelle 4.1: Aufteilung der falsch eingeordneten HGTs. (G) steht für HGTs der GERINGER-Relation und (H) für die der HÖHER-Relation.

In allen Fällen werden deutlich mehr HGTs der GERINGER-Relation fälschlicherweise als HGT der HÖHER-Relation erkannt als anders herum. Im Bezug auf die Kurven in Abbildung 4.2 bedeutet dies, dass `out1000` (G) von dieser Fehlerquelle nicht betroffen ist und deswegen bis `p=19` keine false-positives entstehen, was bei `out1000` (H) nicht der Fall ist. Hier gibt es keine „perfekten“ Ergebnisse und immer ein paar false-positives.
Aus der Tabelle kann noch eine weitere Erkenntnis abgeleitet werden, die sich auf die Kurven der `dists`-Daten bezieht: Offensichtlich sind HGTs der GERINGER-Relation eher davon betroffen, fälschlicherweise in die falsche Relation eingeordnet

zu werden. Daraus lässt sich schlussfolgern, dass es bei HGTs der GERINGER-Relation eher zu Entwicklungen kommt, die dafür sorgen, dass sie falsch eingeordnet werden, womit auch die schlechteren Ergebnisse der korrekten Erkennung von HGTs der GERINGER-Relation begründet werden können.

- Es werden deutlich mehr HGTs der HÖHER-Relation als der GERINGER-Relation erkannt.

  Zum einen existieren einfach mehr HGTs der HÖHER-Relation als der GERINGER-Relation, die gefunden werden könnten: Insgesamt gibt es im Datensatz 25926 HGTs der HÖHER-Relation, und nur 19436 HGTs der GERINGER-Relation. Daraus ergibt sich ein Verhältnis von $\frac{19436}{25926} \approx \frac{3}{4}$. Die Unterschiede in den Ergebnissen sind aber noch deutlicher: Bei keinem verwendeten Wert für `p` bei sowohl der Betrachtung der `out`-Datei als auch bei der `dists`-Datei wird überhaupt das Verhältnis $\frac{1}{3}$ erreicht.
  Hier kann wieder die Schlussfolgerung aus Tabelle 4.1 im vorherigen Absatz angewendet werden: Wenn HGTs der GERINGER-Relation eher davon betroffen sind, in die falsche Relation eingeordnet zu werden, dann sind sie auch eher bzw. erst recht davon betroffen, nicht mehr nur in ihre Relation eingeordnet zu werden, egal ob die stattdessen als GLEICH oder als HÖHER erkannt wurden. Anders ausgedrückt: Dass ein HGT in die entgegengesetzte Relation eingeordnet wird ist der Extremfall. Deutlich häufiger wird er nur in die GLEICH-Relation eingeordnet. Die genauen Häufigkeiten werden in Tabelle 4.2 dargestellt.

Datensatz	`p` in %	(G) als (E) eingestuft	(H) als (E) eingestuft	Verhältnis
`out1000`	1	18682	20140	$\frac{18682}{20140} \approx 0,92 > \frac{3}{4}$
`dists1000`	1	18902	22974	$\frac{18902}{22974} \approx 0,82 > \frac{3}{4}$
`out1000`	15	18180	18683	$\frac{18180}{18683} \approx 0,97 > \frac{3}{4}$
`dists1000`	15	17687	21275	$\frac{17687}{21275} \approx 0,83 > \frac{3}{4}$

Tabelle 4.2: Aufteilung der falsch eingeordneten HGTs. (G) steht für HGTs der GERINGER-Relation, (H) für die der HÖHER-Relation und (E) für die GLEICH-Relation.

  Obwohl die absoluten Werte in der vierten Spalte immer größer sind, wird durch die Verhältnis-Spalte trotzdem deutlich, dass HGTs der GERINGER-Relation deutlich häufiger in eine andere Relation eingeordnet werden als HGTs der HÖHER-Relation: Die Verhältnisse sind in allen Fällen größer als $\frac{3}{4}$, welches das erwartete Verhältnis unter der Annahme ist, dass die falsche Einordnung bei HGTs der GERINGER- und der HÖHER-Relation gleich häufig vorkommt. Dementsprechend kommt es zu deutlich mehr false-negatives und schlechteren Ergebnissen, wenn nur nach HGTs der GERINGER-Relation gesucht wird.

- Aus diesen Erkenntnissen lässt sich die vorherige Aussage über geeignete Werte für `p` präzisieren: Wenn nur nach HÖHER-Relationen gesucht wird, liefern auch höhere Werte relativ sichere Ergebnisse.

  Entsprechend der Auswertung von Abbildung 4.1 wurden letztendlich nur sehr geringe Werte für `p` als anwendbar empfohlen. Aus der Auswertung von Abbildung 4.2 ist aber zu erkennen, dass eine Verwendung von verschiedenen Werten

für p in Abhängigkeit von der gesuchten Relation insgesamt zu mehr erkannten HGTs führt, während die Anzahl von false-positives immer noch gering bleibt. Die geringste Anzahl von false-positives wird offensichtlich erreicht, wenn HGTs der GERINGER-Relation komplett ignoriert werden (p=0), da diese für alle Werte von p schlechtere Ergebnisse liefern.
Bei der Erkennung von HGTs der HÖHER-Relation liefern nach Abbildung 4.2 Werte bis p=9 fast dasselbe Verhältnis an korrekt erkannten HGTs. Anzumerken ist dabei, dass der Einbruch der Kurve von p=9 zu 11 an einer entsprechenden Generierung der Gendistanzen liegt; im Datensatz mit zehn Genen pro Spezies kommt ein ähnlicher Einbruch bei diesen Werten nicht vor. In realen Daten sind solche Einbrüche aber ebenfalls möglich. Eine klare Vorgabe, welche Werte für p für welche Relation „am besten" sind, ist deswegen nicht möglich. Aus den vorgestellten Ergebnissen lässt sich aber ableiten, dass für die HGTs der HÖHER-Relation ein größerer Wert für p als für HGTs der GERINGER-Relation gewählt werden kann.

Die Erkennung der HÖHER-Relation ist aber nicht nur im Bezug auf das Verhältnis von true-positives und false-positives besser. Abbildung 4.5 zeigt die Entwicklung von true-positives in Abhängigkeit von p, wobei die gleichen Tabellen (Anhang A) wie bisher verwendet werden.

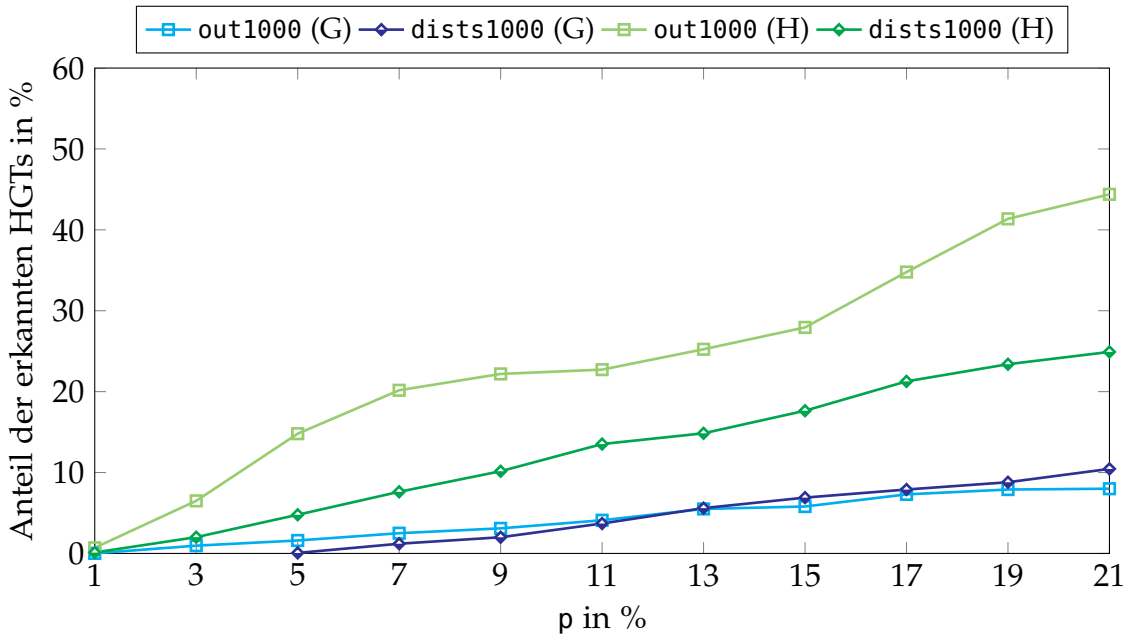

Abbildung 4.5: Anteil der korrekt erkannten HGTs einer Relation zu der Gesamtanzahl der HGTs dieser Relation in Abhängigkeit von p. Insgesamt gibt es 19.436 HGTs der GERINGER-Relation und 25.926 HGTs der HÖHER-Relation. Mit (G) markierte Kurven basieren auf HGTs der GERINGER-Relation, mit (H) markierte HGTs auf denen der HÖHER-Relation.

Wieder zeigen die Kurven der HÖHER-Relation deutlich bessere Ergebnisse als die der GERINGER-Relation. Zum Beispiel werden für p=9 in beiden Dateien mehr als doppelt so viele HGTs der HÖHER-Relation korrekt erkannt als der GERINGER-Relation, sowohl absolut als auch im Bezug auf die Anzahl der vorhandenen HGTs

der Relationen. Trotzdem wird durch Abbildung 4.5 noch einmal verdeutlicht, wie begrenzt die Erkennung von HGTs mit diesem Ansatz tatsächlich ist. Für `p=9` in `dists1000` werden nur ca. 10% der HGTs der HÖHER-Relation erkannt, für die GERINGER-Relation sind es nur 2%. Ein Großteil der vorhandenen HGTs bleibt damit unerkannt.

5 FAZIT UND AUSBLICK

In dieser Arbeit wurde ein Ansatz zur Erkennung von horizontalen Gentransfers (HGT) anhand der Distanzen zwischen Genen entwickelt. HGTs sind neben dem vertikalen Gentransfer bzw. der klassischen Fortpflanzung die zweite Möglichkeit, wie Gene von einem Individuum auf ein anderes übertragen werden können. Die Unberechenbarkeit der Auswirkungen trifft auf die in Kapitel 2.1 beschriebene Zufälligkeit der HGTs. Um sie zu erkennen, sind in der Vergangenheit bereits verschiedene Ideen vorgestellt worden. Neben expliziten Methoden, die die phylogenetischen Bäume und damit die evolutionäre Vergangenheit der Gene darstellen und daraus Informationen über HGTs ableiten, gibt es auch implizite Methoden, die die evolutionäre Distanz, also die Unterschiede zwischen Genen in verschiedenen Spezies, untersuchen. Zu dieser Art gehört auch der hier entwickelte Algorithmus: Nur mithilfe von den Distanzdaten können HGTs mit einer gewissen Wahrscheinlichkeit korrekt erkannt werden.

Um die Ergebnisse zu charakterisieren, wurde eine Einteilung von HGTs in verschiedene Relationen vorgestellt, namentlich die HÖHER-Relation, die GLEICH-Relation und die GERINGER-Relation. Die verschiedenen Relationen beziehen sich auf die Zeitpunkte des Auftretens des HGTs und der Aufteilung der Spezies der Gene, die an dem HGT beteiligt waren. Zum Beispiel kann ein HGT nach der Aufteilung einer Spezies zwischen denen daraus entstandenen neuen Spezies auftreten (siehe Abbildung 2.4). Es ist zu erwarten, dass sich die am HGT beteiligten Gene mehr ähneln, also eine geringere Distanz zueinander haben, als andere Gene aus den beiden Spezies. Dies resultiert aus der geringeren Zeit, die diese beiden Gene hatten, um durch zufällige Mutationen zu divergieren. Der HGT wird im Bezug auf diese beiden Gene damit in die GERINGER-Relation eingeordnet. Die Einordnung in die anderen Relationen funktioniert analog. Diese Einteilung ermöglicht es, die Bedeutung der erkannten HGTs für weiterführende Projekte zu präzisieren.

Die Überprüfung des entwickelten Ansatzes fand mithilfe einer Java-Implementation und zufälligen Gendaten statt. Diese Gendaten wurden von dem Tool ALFSim[1] (Dalquen u. a., 2012) generiert, welches anhand verschiedener Parameter zufällige Genbäume erzeugt. Die genauen Details können in Kapitel 3.5.1 und 4.1 nachgelesen werden.

Das Ergebnis dieser Auswertung war: während der Ansatz grundsätzlich funktioniert, kommt es trotzdem zu einem hohen Anteil an falsch eingeordneten HGTs. Durch die Veränderung des Parameters `p` wird eine Abwägung zwischen insgesamt erkannten HGTs und davon korrekt erkannten HGTs getroffen. Zum Beispiel werden für `p=9` im `dists`-Datensatz nur 3.727 HGTs von 45.362 vorhandenen HGTs erkannt. Von diesen erkannten HGTs sind $\approx$ 81% korrekte HGTs und $\approx$ 19% false-positives, also auffällige Distanzen, die von dem gewählten Ansatz als HGT eingestuft werden aber tatsächlich

1 `http://www.alfsim.org/`, abgerufen am 25.6.2021

nicht aus einem HGT entstanden sind. Ob und wie diese false-positives behandelt werden, muss in möglichen auf dieser Arbeit aufbauenden Projekten entwickelt werden.

Überraschenderweise kam es bei der Erkennung der verschiedenen Relationen zu deutliche Unterschieden. Die HÖHER-Relation wird insgesamt besser erkannt, sowohl im Bezug auf das Verhältnis von true-positives zu false-positives, als auch der Anteil der erkannten true-positives von allen vorhandenen HGTs dieser Relation. Aus diesem Grund ist es sinnvoll, die gefundenen HGTs der verschiedenen Relationen mit einem unterschiedlichen Bias zu begegnen. Wie stark die Auswirkungen auf reale Daten sind, lässt sich in Anbetracht der fehlenden Möglichkeiten zur sicheren Verifikation der gefundenen HGTs nicht eindeutig bestimmen. Hier bietet sich aber ein Vergleich zu anderen Methoden zur Erkennung von HGTs an, um gefundene Ergebnisse anhand anderer Kriterien zu bestätigen.

Ein Grund für die bessere Erkennung der HÖHER-Relation könnte deren Eigenschaft sein, dass sie auch „über“ dem betrachteten Genbaum auftreten können. Ein Gen könnte also von einer Spezies, die gar nicht betrachtet wird, zu einer im Datensatz vorhandenen Spezies übertragen werden. Die Aufteilung der betrachteten Spezies findet dann innerhalb des Baumes statt, die ursprüngliche Aufteilung der Gene aber bereits außerhalb des Baumes, sozusagen „über“ der Wurzel. Zur Verdeutlichung kann Abbildung 2.5 betrachtet werden: Sei das linke Kind der Wurzel der betrachtete Teilbaum. Die Relation zwischen a und b wird weiterhin als HÖHER-Relation erkannt, obwohl der relevante HGT außerhalb des betrachteten Baumes stattgefunden hat. Unter der Annahme, dass der verwendete Baum selbst vollständig ist, also z.B. das linke Kind der Wurzel nicht noch ein weiteres Kind besitzt, können so nur HGTs der HÖHER-Relation auftreten. Für HGTs der GERINGER-Relation müsste der Zeitpunkt der Aufteilung der Spezies bekannt sein.

Zuletzt stellt sich die Frage, was vielleicht noch aus den Distanzdaten abgeleitet werden kann. Die alleinige Verwendung von paarweisen Vergleichen, wie sie hier vorgestellt wurde, scheint nicht vorbehaltlos für die Erkennung von HGTs geeignet zu sein. Zum einen sind die gefundenen Ergebnisse immer von der beschriebenen Unsicherheit geprägt, zum anderen werden nur relativ wenig Informationen über HGTs extrahiert. Die zeitliche Einordnung in die Relationen ist zwar bereits ein hilfreicher Schritt, um HGTs besser zu verstehen und in weiterführenden Projekten zu verwenden, aber es gibt noch weitere wichtige Aspekte. Beispiele dafür sind eine genauere zeitliche Einordnung bzw. Reihenfolge der HGTs innerhalb der Relationen, die Unterscheidung zwischen Spender- und Empfänger-Genom und schlussendlich die Erkennung der GLEICH-Relation. Diese kann nur mit paarweisen Vergleichen nicht von einem normalen Speciation-Event unterschieden werden.

Ob und wie diese genaueren Informationen abgeleitet werden, muss in weiteren Arbeiten ermittelt werden. Eine mögliche Verbesserung könnte durch die Erweiterung von paarweisen Vergleichen zu Vergleichen, die drei Gene gleichzeitig betrachten, gewonnen werden, da diese das komplexe Gesamtbild der Evolution besser erfassen können.

LITERATUR

Abby, Sophie S. (2010). „Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests.“ In: *BMC Bioinformatics* 11. URL: https://pubmed.ncbi.nlm.nih.gov/20550700/.

Alfaro, Michael E. (2003). „Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence.“ In: *Molecular biology and evolution* 20(2). URL: https://pubmed.ncbi.nlm.nih.gov/12598693/.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers und David J. Lipman (1990). „Basic local alignment search tool“. In: *Journal of Molecular Biology* 215.3, S. 403–410. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

Arvey, Aaron J, Rajeev K. Azad, Alpan Raval und Jeffrey G. Lawrence (2009). „Detection of genomic islands via segmental genome heterogeneity.“ In: *Nucleic Acids Res* 37. URL: https://pubmed.ncbi.nlm.nih.gov/19589805/.

Azad, Rajeev K. und Jeffrey G. Lawrence (2011). „Towards more robust methods of alien gene detection“. In: *Nucleic Acids Res* 39, e56. URL: https://pubmed.ncbi.nlm.nih.gov/21297116.

Beiko, Robert G. (2006). „Phylogenetic identification of lateral genetic transfer events.“ In: *BMC evolutionary biology* 6. URL: https://pubmed.ncbi.nlm.nih.gov/16472400/.

Crow, James F. (1997). „The high spontaneous mutation rate: Is it a health risk?“ In: *Proceedings of the National Academy of Sciences* 94.16, S. 8380–8386. ISSN: 0027-8424. DOI: 10.1073/pnas.94.16.8380. eprint: https://www.pnas.org/content/94/16/8380.full.pdf. URL: https://www.pnas.org/content/94/16/8380.

Csűrös, Miklós und István Miklós (Juni 2009). „Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model“. In: *Molecular Biology and Evolution* 26.9, S. 2087–2095. ISSN: 0737-4038. DOI: 10.1093/molbev/msp123. eprint: https://academic.oup.com/mbe/article-pdf/26/9/2087/13643390/msp123.pdf. URL: https://doi.org/10.1093/molbev/msp123.

Dalquen, Daniel A., Maria Anisimova, Gaston H. Gonnet und Christophe Dessimoz (2012). „ALF–a simulation framework for genome evolution“. In: *Molecular biology and evolution* 29. URL: https://pubmed.ncbi.nlm.nih.gov/22160766/.

Darwin, Charles (1837). *Notebook B*.

Daubin, Vincent, Emmanuelle Lerat und Guy Perrière (2003). „The source of laterally transferred genes in bacterial genomes“. In: *Genome Biology* 4, R57. URL: https://doi.org/10.1186/gb-2003-4-9-r57.

Dessimoz, C., D. Margadant und G.H. Gonnet (2008). „DLIGHT – Lateral Gene Transfer Detection Using Pairwise Evolutionary Distances in a Statistical Framework“. In: *Lecture Notes in Computer Science* 4955. URL: https://doi.org/10.1007/978-3-540-78839-3_27.

Douady, Christophe J. (2003). „Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability.“ In: *Molecular biology and evolution* 20(2). URL: https://pubmed.ncbi.nlm.nih.gov/12598692/.

Friesen, Timothy L., Eva H. Stukenbrock, Zhaohui Liu, Steven Meinhardt, Hua Ling, Justin D. Faris, Jack B. Rasmussen, Peter S. Solomon, Bruce A. McDonald und Richard P. Oliver (2006). „Emergence of a new disease as a result of interspecific virulence gene transfer“. In: *Nature Genetics* 38, S. 953–956. URL: https://doi.org/10.1038/ng1839.

Geiß, Manuela, John Anders, Peter F. Stadler, Nicolas Wieseke und Marc Hellmuth (2018). „Reconstructing gene trees from Fitch's xenology relation“. In: *Journal of Mathematical Biology* 77, S. 1459–1491. URL: https://doi.org/10.1007/s00285-018-1260-8.

Gophna, Uri, W. Ford Doolittle und Robert L. Charlebois (2005). „Weighted genome trees: refinements and applications.“ In: *Journal of bacteriology* 187(4). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC545607/.

Griffith, F (1928). „The Significance of Pneumococcal Types“. In: *The Journal of hygiene* 27, S. 113–159. URL: https://pubmed.ncbi.nlm.nih.gov/20474956.

Gyles, C. und P. Boerlin (2014). „Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease“. In: *Veterinary Pathology* 51.2. PMID: 24318976, S. 328–340. DOI: 10.1177/0300985813511131. eprint: https://doi.org/10.1177/0300985813511131. URL: https://doi.org/10.1177/0300985813511131.

Heywood, V. H. und J. McNeill (1964). „Phenetic and Phylogenetic Classification“. In: *Nature* 203, S. 1220–1224. URL: https://doi.org/10.1038/2031220a0.

Hickey, Glenn (2008). „SPR distance computation for unrooted trees.“ In: *Evolutionary bioinformatics online* 4. URL: https://pubmed.ncbi.nlm.nih.gov/19204804/.

Johnston, Calum, Bernard Martin, Gwennaele Fichant, Patrice Polard und Jean-Pierre Claverys (2014). „Bacterial transformation: distribution, shared mechanisms and divergent control“. In: *Nature Reviews Microbiology* 12, S. 181–196. URL: https://doi.org/10.1038/nrmicro3199.

Kapley, Atya und Shailendra Yadav (2021). „Antibiotic resistance: Global health crisis and metagenomics.“ In: *Biotechnology reports (Amst)*. URL: https://pubmed.ncbi.nlm.nih.gov/33732632/.

Keeling, Patrick J. und Jeffrey D. Palmer (2008). „Horizontal gene transfer in eukaryotic evolution“. In: *Nature Reviews Genetics* 9, S. 605–618. URL: https://doi.org/10.1038/nrg2386.

Koschnik, Wolfgang J. (1992). *Standardwörterbuch für die Sozialwissenschaften*. 2. Aufl. München, Deutschland: Saur Verlag. ISBN: 3598110804.

– (1997). *Molekulare Genetik*. 7. Aufl. Stuttgart, Deutschland: Thieme. ISBN: 3134770075.

Koski, L. B. und G. B. Golding (2001). „The closest BLAST hit is often not the nearest neighbor.“ In: *Journal of molecular evolution* 52(6). URL: https://pubmed.ncbi.nlm.nih.gov/11443357/.

Lerminiaux, Nicole A. und Andrew D. S. Cameron (2019). „Horizontal transfer of antibiotic resistance genes in clinical environments“. In: *Can J Microbiol.* 91. URL: https://pubmed.ncbi.nlm.nih.gov/30248271/.

Maddison, David R., David L. Swofford und Wayne P. Maddison (1997). „Nexus: An Extensible File Format for Systematic Information“. In: *Systematic Biology* 46. URL: https://academic.oup.com/sysbio/article/46/4/590/1629695.

Nei, Masatoshi (1972). „Genetic Distance between Populations“. In: *The American Naturalist* 106(949). URL: https://doi.org/10.1086/282771.

Nelson, Karen E. u. a. (1999). „Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.“ In: *Nature* 399. URL: https://pubmed.ncbi.nlm.nih.gov/10360571/.

Novichkov, Pavel S., Marina V. Omelchenko, Mikhail S. Gelfand, Andrei A. Mironov, Yuri I. Wolf und Eugene V. Koonin (2004). „Genome-wide molecular clock and horizontal gene transfer in bacterial evolution.“ In: *Journal of bacteriology* 186(19). URL: https://pubmed.ncbi.nlm.nih.gov/15375139/.

Ochman, Howard und Jeffrey G. Lawrence (1997). „Amelioration of Bacterial Genomes: Rates of Change and Exchange“. In: *Journal of Molecular Evolution* 44, S. 383–397. URL: https://doi.org/10.1007/PL00006158.

Ochman, Howard, Jeffrey G. Lawrence und Eduardo A. Groisman (2000). „Lateral gene transfer and the nature of bacterial innovation“. In: *Nature* 405. URL: https://ui.adsabs.harvard.edu/abs/2000Natur.405..299O/abstract.

Pellegrini, M. (1999). „Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.“ In: *Proceedings of the National Academy of Sciences of the United States of America* 96(8). URL: https://pubmed.ncbi.nlm.nih.gov/10200254/.

Poptsova, Maria S. und J. Peter Gogarten (2007). „The power of phylogenetic approaches to detect horizontally transferred genes“. In: *BMC Evol Biol* 7, S. 45. URL: https://pubmed.ncbi.nlm.nih.gov/17376230.

Ravenhall, Matt, Nives Škunca, Florent Lassalle und Christophe Dessimoz (2015). „Inferring Horizontal Gene Transfer“. In: *PLoS Comput Biol* 11(5). URL: https://doi.org/10.1371/journal.pcbi.1004095.

Rice, Peter, Ian Longden und Alan Bleasby (2000). „EMBOSS: The European Molecular Biology Open Software Suite“. In: *Trends in Genetics* 16.6, S. 276–277. ISSN: 0168-9525. DOI: https://doi.org/10.1016/S0168-9525(00)02024-2. URL: https://www.sciencedirect.com/science/article/pii/S0168952500020242.

Ross, Herbert H. (1964). „Principles of Numerical Taxonomy“. In: *Systematic Biology* 13.1-4, S. 106–108. ISSN: 1063-5157. DOI: 10.2307/sysbio/13.1-4.106. eprint: https://academic.oup.com/sysbio/article-pdf/13/1-4/106/4827553/13-1-4-106.pdf. URL: https://doi.org/10.2307/sysbio/13.1-4.106.

Ruggiero, Michael A., Dennis P. Gordon, Thomas M. Orrell, Nicolas Bailly, Thierry Bourgoin, Richard C. Brusca, Thomas Cavalier-Smith, Michael D. Guiry und Paul M. Kirk (2015). „A Higher Level Classification of All Living Organisms“. In: *PLOS ONE* 10.4, S. 1–60. DOI: 10.1371/journal.pone.0119248. URL: https://doi.org/10.1371/journal.pone.0119248.

Schaller, David, Manuel Lafond, Peter F. Stadler, Nicolas Wieseke und Marc Hellmuth (2021). „Indirect Identification of Horizontal Gene Transfer“. In: *Journal of Mathematical Biology*. URL: https://doi.org/10.1007/s00285-021-01631-0.

Shimodaira, Hidetoshi (2002). „An approximately unbiased test of phylogenetic tree selection.“ In: *Systematic biology* 51(3). URL: https://pubmed.ncbi.nlm.nih.gov/12079646/.

Singh, Monica, Puneetpal Singh, Pawan Kumar Juneja, Surinder Singh und Taranpal Kaur (2011). „SNP–SNP interactions within APOE gene influence plasma lipids in

postmenopausal osteoporosis“. In: *Rheumatology International* 31, S. 421–423. URL: https://doi.org/10.1007/s00296-010-1449-7.

Stadler, Peter F., Nicolas Wieseke und Marc Hellmuth (2017). „The mathematics of xenology: di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations“. In: *Journal of Mathematical Biology* 75, S. 199–237. URL: https://doi.org/10.1007/s00285-016-1084-3.

Weiss, Madeline C., Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin (2016). „The physiology and habitat of the last universal common ancestor“. In: *Nature Microbiology* 1, S. 16116. URL: https://doi.org/10.1038/nmicrobiol.2016.116.

Worning, P., L. J. Jensen, K. E. Nelson, S. Brunak und D. W. Ussery (2000). „Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima“. In: *Nucleic Acids Res* 28, S. 706–709. URL: https://pubmed.ncbi.nlm.nih.gov/10637321.

Xiong, Dapeng, Fen Xiao, Li Liu, Kai Hu, Yanping Tan, Shunmin He und Xieping Gao (2012). „Towards a better detection of horizontally transferred genes by combining unusual properties effectively“. In: *PLoS One* 7, e43126. URL: https://pubmed.ncbi.nlm.nih.gov/22905214.

Zhang, Jianzhi (2003). „Evolution by gene duplication: an update“. In: *Trends in Ecology & Evolution* 18, S. 292–298. eprint: http://www.umich.edu/~zhanglab/publications/2003/Zhang_2003_TIG_18_292.pdf. URL: https://doi.org/10.1016/S0169-5347(03)00033-8.

Zhaxybayeva, Olga (2006). „Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events.“ In: *Genome research* 16(9). URL: https://pubmed.ncbi.nlm.nih.gov/16899658/.

APPENDIX

A TABELLEN DER GRAPHEN

Die folgenden Tabellen beschreiben die Ergebnisse des Programms für verschiedene Werte für p mit den verschiedenen Input-Dateien. Die Tabellen sind die Grundlagen für die Graphen in Kapitel 4.

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	80	78 (97.5%)	2 (2.5%)	959
dists10	18	17 (94.44%)	1 (5.56%)	1.020
out1000	192	184 (95.83%)	8 (4.17%)	45.178
dists1000	34	31 (91.18%)	3 (8.82%)	45.331

Tabelle A.1: Auswertung mit $p = 0,01$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	80	78 (97.5%)	2 (2.5%)	959
dists10	18	17 (94.44%)	1 (5.56%)	1.020
out1000	1.925	1.876 (97,45%)	49 (2,55%)	43.486
dists1000	569	522 (91,74%)	47 (8,26%)	44.840

Tabelle A.2: Auswertung mit $p = 0,03$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	138	136 (98.55%)	2 (1.45%)	901
dists10	24	17 (70.83%)	7 (29.17%)	1.020
out1000	4.180	4.130 (98,8%)	50 (1,2%)	41.232
dists1000	1.409	1.251 (88,79%)	158 (11,21%)	44.111

Tabelle A.3: Auswertung mit $p = 0,05$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	175	172 (98.29%)	3 (1.71%)	865
dists10	68	43 (63.24%)	25 (36.76%)	994
out1000	5.843	5.710 (97,72%)	133 (2,28%)	39.652
dists1000	2.616	2.218 (84,79%)	398 (15,21%)	43.144

Tabelle A.4: Auswertung mit $p = 0,07$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	222	219 (98.65%)	3 (1.35%)	818
dists10	105	68 (64.76%)	37 (35.24%)	969
out1000	6.555	6.364 (97,09%)	191 (2,91%)	38.998
dists1000	3.727	3.014 (80,87%)	713 (19,13%)	42.348

Tabelle A.5: Auswertung mit $p = 0,09$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	291	287 (98.63%)	4 (1.37%)	750
dists10	159	90 (56.6%)	69 (43.4%)	947
out1000	6.919	6.688 (96,66%)	231 (3,34%)	38.674
dists1000	6.169	4.229 (68,55%)	1.940 (31,45%)	41.133

Tabelle A.6: Auswertung mit $p = 0,11$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	318	311 (97.8%)	7 (2.2%)	726
dists10	176	103 (58.52%)	73 (41.48%)	934
out1000	8.023	7.610 (94,85%)	413 (5,15%)	37.752
dists1000	7.686	4,937 (64.23%)	2.749 (35,77%)	40.425

Tabelle A.7: Auswertung mit $p = 0,13$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	322	314 (97.52%)	8 (2.48%)	723
dists10	214	125 (58.41%)	89 (41.59%)	912
out1000	8.913	8.380 (94,02%)	533 (5,98%)	36.982
dists1000	10.049	5.924 (58,95%)	4.125 (41,05%)	39.438

Tabelle A.8: Auswertung mit $p = 0,15$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	341	327 (95.89%)	14 (4.11%)	710
dists10	261	150 (57.47%)	111 (42.53%)	887
out1000	11.397	10.432 (91,53%)	965 (8,47%)	34.930
dists1000	12.988	7.061 (54,37%)	5.927 (45,63%)	38.301

Tabelle A.9: Auswertung mit $p = 0,17$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	365	345 (94.52%)	20 (5.48%)	692
dists10	291	163 (56.01%)	128 (43.99%)	874
out1000	13.713	12.265 (89,44%)	1.448 (10,56%)	33.097
dists1000	15.607	7.781 (49,86%)	7.826 (50,14%)	37.581

Tabelle A.10: Auswertung mit $p = 0,19$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
out10	368	346 (94.02%)	22 (5.98%)	691
dists10	335	184 (54.93%)	151 (45.07%)	853
out1000	14.773	13.070 (88,47%)	1.703 (11,53%)	32.292
dists1000	18.254	8.490 (46,51%)	9.764 (53,49%)	36.872

Tabelle A.11: Auswertung mit $p = 0,21$

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	2	2 (100%)	0 (0%)	45.360
`dists1000` (G)	0	—-	—-	45.362
`out1000` (H)	190	182 (95,79%)	8 (4,21%)	45.180
`dists1000` (H)	34	31 (91,18%)	3 (8,82%)	45.331

Tabelle A.12: Auswertung mit $p = 0,01$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	186	186 (100%)	0 (0%)	45.176
`dists1000` (G)	0	—-	—-	45.362
`out1000` (H)	1.739	1.690 (97,18%)	49 (2,82%)	43.672
`dists1000` (H)	569	522 (91,74%)	47 (8,26%)	44.840

Tabelle A.13: Auswertung mit $p = 0,03$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	304	304 (100%)	0 (0%)	45.058
`dists1000` (G)	22	12 (54,55%)	10 (45,45%)	45.350
`out1000` (H)	3.876	3.826 (98,71%)	50 (1,29%)	41.536
`dists1000` (H)	1.387	1.239 (89,33%)	148 (10,67%)	44.123

Tabelle A.14: Auswertung mit $p = 0,05$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	480	480 (100%)	0 (0%)	44.882
`dists1000` (G)	376	242 (64,36%)	134 (35,64%)	45.120
`out1000` (H)	5.363	5.230 (97,52%)	133 (2,48%)	40.132
`dists1000` (H)	2.240	1.976 (88,21%)	264 (11,79%)	43.386

Tabelle A.15: Auswertung mit $p = 0,07$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	609	609 (100%)	0 (0%)	44.753
`dists1000` (G)	643	379 (58,94%)	264 (41,06%)	44.983
`out1000` (H)	5.946	5.755 (96,79%)	191 (3,21%)	39.607
`dists1000` (H)	3.084	2.635 (85,44%)	449 (14,56%)	42.727

Tabelle A.16: Auswertung mit $p = 0,09$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	797	797 (100%)	0 (0%)	44.565
`dists1000` (G)	1.240	725 (58,47%)	515 (41,53%)	44.637
`out1000` (H)	6.122	5.891 (96,23%)	231 (3,77%)	39.471
`dists1000` (H)	4.929	3.504 (71,09%)	1.425 (28,91%)	41.858

Tabelle A.17: Auswertung mit $p = 0,11$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	1.065	1.065 (100%)	0 (0%)	44.297
`dists1000` (G)	1.989	1.088 (54,7%)	901 (45,3%)	44.274
`out1000` (H)	6.958	6.545 (94,06%)	413 (5,94%)	38.817
`dists1000` (H)	5.697	3.849 (67,56%)	1.848 (32,44%)	41.513

Tabelle A.18: Auswertung mit $p = 0,13$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	1.137	1.137 (100%)	0 (0%)	44.225
`dists1000` (G)	2.707	1.351 (49,91%)	1.356 (50,09%)	44.011
`out1000` (H)	7.776	7.243 (93,15%)	533 (6,85%)	38.119
`dists1000` (H)	7.342	4.573 (62,29%)	2.769 (37,71%)	40.789

Tabelle A.19: Auswertung mit $p = 0,15$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	1.415	1.415 (100%)	0 (0%)	43.947
`dists1000` (G)	3.204	1.546 (48,25%)	1.658 (51,75%)	43.816
`out1000` (H)	9.982	9.017 (90,33%)	965 (9,67%)	36.345
`dists1000` (H)	9.784	5.515 (56,37%)	4.269 (43,63%)	39.847

Tabelle A.20: Auswertung mit $p = 0,17$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	1.549	1.542 (99,55%)	7 (0,45%)	43.820
`dists1000` (G)	3.708	1.717 (46,31%)	1.991 (53,69%)	43.645
`out1000` (H)	12.164	10.723 (88,15%)	1.441 (11,85%)	34.639
`dists1000` (H)	11.899	6.064 (50,96%)	5.835 (49,04%)	39.298

Tabelle A.21: Auswertung mit $p = 0,19$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

	gefundene HGTs	korrekte HGTs	falsche HGTs	nicht erkannte HGTs
`out1000` (G)	1.568	1.561 (99,55%)	7 (0,45%)	43.801
`dists1000` (G)	4.648	2.032 (43,72%)	2.616 (56,28%)	43.330
`out1000` (H)	13.205	11.509 (87,16%)	1.696 (12,84%)	33.853
`dists1000` (H)	13.606	6.458 (47,46%)	7.148 (52,54%)	38.904

Tabelle A.22: Auswertung mit $p = 0,21$. Zeilen mit (G) beinhalten nur HGTs der GERINGER-Relation überprüft, Zeilen mit (H) die der HÖHER-Relation

B AUSSCHNITTE AUS DEM DATENSATZ MIT ZEHN SPEZIES

B.1 NEXUS DATEIEN

In diesem Kapitel wird der Aufbau der verwendeten Testdateien (und damit des benötigten Inputs) anhand von Ausschnitten verdeutlicht.

out10.nex

```
BEGIN ALLDISTANCES;
...
  distances
    name=G8 triangle=both
      SE001/00008 0.000000 77.538400 77.538400 147.524000 164.864800 164.864800 164.864800
          164.864800 164.864800 164.864800 164.864800 164.864800 164.864800
      SE009/00008 77.538400 0.000000 71.476200 147.524000 164.864800 164.864800 164.864800
          164.864800 164.864800 164.864800 164.864800 164.864800 164.864800
      SE010/00008 77.538400 71.476200 0.000000 147.524000 164.864800 164.864800 164.864800
          164.864800 164.864800 164.864800 164.864800 164.864800 164.864800
      SE005/00008 147.524000 147.524000 147.524000 0.000000 164.864800 164.864800 164.864800
          164.864800 164.864800 164.864800 164.864800 164.864800 164.864800
      SE004/00008 164.864800 164.864800 164.864800 164.864800 0.000000 104.345000 104.345000
          150.578600 150.578600 150.578600 150.578600 150.578600 150.578600
      SE007/00008 164.864800 164.864800 164.864800 164.864800 104.345000 0.000000 94.142400
          150.578600 150.578600 150.578600 150.578600 150.578600 150.578600
      SE008/00034 164.864800 164.864800 164.864800 164.864800 104.345000 94.142400 0.000000
          150.578600 150.578600 150.578600 150.578600 150.578600 150.578600
      SE003/00008 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          0.000000 102.543600 107.001000 107.001000 138.426200 138.426200
      SE008/00008 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          102.543600 0.000000 107.001000 107.001000 138.426200 138.426200
      SE003/00022 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          107.001000 107.001000 0.000000 102.543600 138.426200 138.426200
      SE008/00022 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          107.001000 107.001000 102.543600 0.000000 138.426200 138.426200
      SE006/00008 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          138.426200 138.426200 138.426200 138.426200 0.000000 58.117000
      SE006/00047 164.864800 164.864800 164.864800 164.864800 150.578600 150.578600 150.578600
          138.426200 138.426200 138.426200 138.426200 58.117000 0.000000
      ;
    ...
END;
```

```
BEGIN RELATIONS;
  ...
  relation
    type=RCB name=G8 triangle=both
      SE001/00008  0  0  0  0  0  0  0  0  0  0  0  0  0
      SE009/00008  0  0  0  0  0  0  0  0  0  0  0  0  0
      SE010/00008  0  0  0  0  0  0  0  0  0  0  0  0  0
      SE005/00008  0  0  0  0  0  0  0  0  0  0  0  0  0
      SE004/00008  0  0  0  0  0  0 -1  0  0  0  0  0  0
      SE007/00008  0  0  0  0  0  0 -1  0  0  0  0  0  0
      SE008/00034  0  0  0  0 -1 -1  0  1  1  1  1  1  1
      SE003/00008  0  0  0  0  0  0  1  0  0  1  1  0  0
      SE008/00008  0  0  0  0  0  0  1  0  0  1  1  0  0
      SE003/00022  0  0  0  0  0  0  1  1  1  0  0  0  0
      SE008/00022  0  0  0  0  0  0  1  1  1  0  0  0  0
      SE006/00008  0  0  0  0  0  0  1  0  0  0  0  0 -1
      SE006/00047  0  0  0  0  0  0  1  0  0  0  0 -1  0
      ;
  ...
END;
```

dists10.nex

```
BEGIN ALLDISTANCES;
  ...
  distances
    name=G8 triangle=upper
      SE001/00008 0.00 126.48 98.80 36.10 72.86  49.77 53.91  89.32  56.81 50.85 80.36  52.71 41.13
      SE003/00008 0.00 19.65  96.63 47.17 68.36  78.19 133.06 187.72 61.52 37.70 129.54 57.04
      SE003/00022 0.00 74.62  35.33 69.62 60.21  95.82 129.23 46.08  40.25 97.77 43.58
      SE004/00008 0.00 60.20  54.06 40.72 58.21  49.96 36.82  51.21  63.82 52.90
      SE005/00008 0.00 67.12  52.58 86.24 116.50 46.90 34.48  78.55  40.14
      SE006/00008 0.00 31.23  77.63 98.32 37.72  68.25 70.70  58.19
      SE006/00047 0.00 58.10  76.49 28.59 50.50  55.93 44.08
      SE007/00008 0.00 37.50  53.23 69.25 57.34  68.97
      SE008/00008 0.00 59.00 119.22 63.27 93.71
      SE008/00022 0.00 46.58  41.88 40.16
      SE008/00034 0.00 83.23  50.94
      SE009/00008 0.00 37.59
      SE010/00008 0.00
      ;
  ...
END;
```

B.2 ÜBERBLICK ÜBER HGTS EINIGER SPEZIES

Distanzen in `out10.nex` und Relationen zwischen Genen von `SE007` und `SE008`

Distanz	Baum	Typ
94.1424	G3	-1
94.1424	G3	-1
94.1424	G3	-1
94.1424	G3	-1
94.1424	G4	-1
94.1424	G4	-1
94.1424	G4	-1
94.1424	G5	-1
94.1424	G5	-1
94.1424	G5	-1
94.1424	G6	-1
94.1424	G7	-1
94.1424	G8	-1
94.1424	G9	-1
102.5436	G4	-1
102.5436	G4	-1
106.6418	G1	-1
123.3058	G3	-1
123.3058	G3	-1
123.3058	G3	-1
123.3058	G3	-1
139.412	G3	-1
148.8088	G4	-1
148.8088	G4	-1
150.5786	G1	0
150.5786	G1	0
150.5786	G10	0
150.5786	G10	0
150.5786	G2	0
150.5786	G2	0
150.5786	G3	0
150.5786	G3	0
150.5786	G3	0
150.5786	G4	0
150.5786	G4	0
150.5786	G4	0
150.5786	G4	0
150.5786	G4	0
150.5786	G4	0
150.5786	G5	0
150.5786	G5	0
150.5786	G5	0
150.5786	G6	0
150.5786	G7	0
150.5786	G8	0
150.5786	G8	0
150.5786	G9	0
150.5902	G4	1
150.5902	G4	1
150.5902	G5	1
164.8648	G1	1
164.8648	G1	1
164.8648	G1	1
164.8648	G2	1
164.8648	G2	1
164.8648	G3	1
164.8648	G3	1
164.8648	G3	1
164.8648	G3	1
164.8648	G3	1
164.8648	G3	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G4	1
164.8648	G5	1
164.8648	G5	1
164.8648	G5	1
164.8648	G5	1
164.8648	G5	1
164.8648	G5	1
164.8648	G5	1

Distanzen in `dists10.nex` und Relationen zwischen Genen von `SE007` und `SE008`

Distanz	Baum	Typ
20.81	G3	-1
26.02	G3	1
27.2	G3	-1
28.87	G4	1
30.49	G4	-1
33.62	G3	-1
34.74	G3	-1
35.46	G9	-1
37.5	G8	0
37.81	G1	0
38.97	G3	-1
39.46	G9	0
39.57	G3	-1
40.26	G4	0
40.95	G4	1
40.99	G4	0
41.42	G3	-1
41.91	G6	-1
42.65	G3	1
42.98	G1	1
43.6	G3	1
44.41	G3	-1
45.65	G10	0
45.97	G5	-1
46.01	G5	1
46.07	G4	0
46.54	G3	1
46.94	G4	1
49.42	G3	1
50.16	G7	0
50.21	G4	-1
50.64	G4	1
51.02	G4	-1
51.87	G3	0
52.14	G1	-1
52.93	G3	1
53.23	G8	0
53.25	G4	0
53.36	G1	1
53.55	G1	0
55.72	G3	1
57.34	G6	0
57.62	G5	-1
59.59	G7	-1
60.58	G4	1
60.7	G4	-1
61.06	G4	1
61.44	G4	1
62.34	G3	1
62.68	G4	-1
62.84	G5	1
62.86	G3	0
62.91	G5	1
62.93	G1	1
63.16	G3	-1
64.37	G5	-1
64.79	G5	1
64.9	G3	0
66.21	G5	0
66.62	G3	1
68.97	G4	-1
69.25	G8	-1
70.07	G4	1
71.45	G2	1
71.66	G4	0
75.64	G5	1
77.98	G4	1
82.4	G4	1
84.25	G5	1
85.04	G4	-1
89.24	G5	1
90.86	G4	1
91.48	G4	0
94.04	G4	1
94.11	G10	0
94.63	G2	1
97.74	G5	1
103.31	G5	1
106.46	G4	0
108.29	G4	1
112.32	G4	1
114.04	G4	1
119.15	G5	0
120.17	G2	0
129.2	G5	0
131.68	G2	0
132.73	G4	1
152.03	G4	1
157.24	G4	1

ERKLÄRUNG

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet.
Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.
Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, 14. September 2021

Stefan Grote