Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Algorithmen zur Rekonstruktion kophylogenetischer Ereignisse

Diplomarbeit

Leipzig, 17. Januar 2008

vorgelegt von

Nicolas Wieseke
geb. am: 22.04.1980

Studiengang Informatik

Betreuer: Dr. Daniel Merkle

Kurzbeschreibung

Das Problem der Rekonstruktion einer gemeinsamen evolutionären Entwicklung zwischen Wirts- und Parasitenspezies ist in der Forschung weit diskutiert. Dabei wird der Komplexität einer solchen Berechnung besondere Bedeutung beigemessen. In dieser Arbeit wird ein algorithmischer Ansatz vorgestellt, welcher auf Basis dynamischer Programmierung eine Rekonstruktion zweier phylogenetischer Stammbäume und einer gegebenen Abbildung von Parasiten auf zugehörige Wirte erzeugt. Grundlage dieser Berechnung ist ein ereignis-basiertes Modell der Koevolution, bei dem jedem Ereignis ein Kostenwert zugeordnet ist. Gesucht wird nach Rekonstruktionen, welche die Gesamtkosten der aufgetretenen Ereignisse minimieren. Es wird eine Vorgehensweise vorgestellt, mit welcher sich die Kosten der Ereignisse automatisch berechnen lassen. Dazu wurde ein Gütemaß entwickelt, um verschiedene Rekonstruktionen bezüglich der bei ihrer Berechnung verwendeten Ereigniskostenverteilung bewerten zu können. Im Gegensatz zu bisherigen Arbeiten unterstützt der vorgestellte Ansatz zudem die Verwendung von Stammbäumen mit mehrfach verzweigenden Knoten. Die algorithmischen Überlegungen wurden in einem Javaprogramm namens DynamicTreeMap umgesetzt.

Abstract

The problem of reconstructing the common evolutionary development between host- and parasite spezies has been strongly discussed in research. Hereby a special meaning has been attributed to the complexity of such a calculation. In this thesis an algorithmic approach based on dynamic programming will be introduced, that creates a reconstruction of two phylogenetic genealogical trees and a given mapping of parasites on appropriate hosts. The foundation of this calaculation is an event-driven model of coevolution where costs are assigned to each event. The algorithm searches for reconstructions, which minimize the total costs of all occurred events. A method will be introduced which calculates the event-costs automatically. Therefore a quality rate has been developed, to evaluate different reconstructions concerning the used event-costs. Unlike present approaches genealogical trees with multiple branching nodes can be considered. The described approach has been implemented in a java program named DynamicTreeMap.

Inhaltsverzeichnis

1 Einleitung

In der Natur beobachtet man häufig parasitäre Wechselwirkungen zwischen verschiedenen Organismen. Hierbei hält sich eine Spezies - der Parasit - zeitweise oder dauerhaft auf einer anderen Spezies - dem Wirt - auf, um sein Fortbestehen zu sichern. Diese Form der Wechselbeziehungen ist auch in der Evolutionsforschung von Interesse. Durch Beobachtung ist es zwar möglich, heute existierende Parasit-Wirt-Beziehungen zu untersuchen. Ob und inwieweit diese aber auch in früheren Stadien der Evolution existierten, kann so nicht geklärt werden. Um dennoch Aussagen über die gemeinsame Entwicklung von Wirts- und Parasitenarten treffen zu können, wird versucht eine beide Arten umfassende Historie zu rekonstruieren. Ausgangspunkt sind die zwei stammesgeschichtlichen Entwicklungen in Form von evolutionären Stammbäumen, sowie das Wissen über die existierenden Parasit-Wirt-Beziehungen beider Spezies.

In der Literatur wurden verschiedene Ansätze und Programme beschrieben, welche sich mit dem Problem der Koevolution beschäftigen. Ziel ist es dabei Koevolution für spezielle Ausgangsdaten grundsätzlich nachweisen zu können und für diese Daten eine gemeinsame evolutionäre Geschichte beider Spezies zu erzeugen. Ein Überblick über die wichtigsten Ansätze wird in [13] gegeben. In [4] wurde ein Verfahren namens „Brooks Parsimony Analysis“ (BPA) vorgestellt. Dieses wurde unter anderem in PAUP* 4.0 ([28]) integriert. Das Programme TreeFitter 1.0 ([23]) nutzt die von Ronquist in [22] beschriebene „Generalized Parsimony“. Page entwickelte in [16] den Ansatz der „Reconciliation Analysis“ und setzte diesen in TreeMap 1.0 um. Mit Hilfe der in [5] beschriebenen Datenstruktur der Jungles entwickelten Charleston und Page die zweite Version von TreeMap ([6]). Diese Datenstruktur wurde von Legat, Merkle und Middendorf in [12] und [14] um Zeitinformationen in den Ausgangsdaten erweitert und im Programm Tarzan ([11]) umgesetzt.

1.1 Motivation

Die drei letztgenannten Programme verwenden alle ein ereignis-basiertes Modell der Koevolution. Dabei wird eine Menge von Ereignissen definiert. Diese beschreiben, was im Laufe der Lebensspanne einer Wirtsart mit der auf ihr lebenden Parasitenart geschehen sein könnte. Jedem Ereignis wird ein bestimmter Kostenwert zugewiesen. Für eine er-

zeugte Rekonstruktion wird gefordert, dass sie die Gesamtkosten der in ihr auftretenden Ereignisse minimiert. Eine so berechnete Rekonstruktion ist daher maßgeblich von der Wahl dieser Kosten abhängig. Ändert man diese, so entstehen zum Teil deutlich unterschiedliche kostenminimale Rekonstruktionen. Da aber keine allgemeingültigen Kostenwerte für die verwendeten Ereignisse existieren, bleibt es dem Endanwender überlassen diese manuell zu justieren.

Ein weiterer Nachteil dieser Programme ist, dass diese im Allgemeinen keine Mehrfachverzweigungen in den Ausgangsdaten zulassen. Zwar wurde in Tarzan eine Möglichkeit integriert, mit welcher mehrere Abbildungen ein und des selben Parasiten auf unterschiedliche Wirte behandelt werden können. Dafür müssen jedoch alle binären Kombinationen gebildet und ausgewertet. Diese Vorgehensweise stellt sicher, dass unter den Kombinationen auch die kostengünstigste Variante gefunden wird, obwohl diese sehr rechenintensiv ist und die Laufzeit signifikant erhöht.

1.2 Herangehensweise

In der hier vorliegenden Arbeit sollen Lösungsmöglichkeiten für die oben genannten Kritikpunkte aufgezeigt werden, ohne jedoch das zugrunde gelegte Modell entscheidend ändern zu müssen. Dafür werden in Kapitel 2 zunächst die biologischen Grundlagen des ereignis-basierten Modells betrachtet.

Um eine reine Funktionserweiterung zu den anderen Herangehensweisen sicherstellen zu können, sollen für den Fall binärer Ausgangsdaten die gleichen Rekonstruktionen erzeugt werden. Dies gewährleistet eine Vergleichbarkeit zwischen den Ansätzen und den durch sie gefundenen Lösungen.

Allerdings soll zur Umsetzung nicht die von Charleston beschriebene Datenstruktur der Jungles verwendet werden, da diese bei großen Ausgangsdaten nicht mehr praktikabel ist. Vielmehr soll auf Basis von Teilrekonstruktionen eine Gesamtlösung erzeugt werden. Dazu eignet sich besonders das dynamische Programmierparadigma[1]. Eine Beschreibung des Rekonstruktionsproblems in Form dieses Paradigmas soll in Kapitel 3 gegeben werden.

Die genaue Umsetzung folgt in Kapitel 4. Hierbei wird auch eine exakte Beschreibung der Verfahrensweise bei Multifurkationen im Stammbaum des Wirtes bzw. des Parasiten gegeben. Die verwendeten Datenstrukturen werden vorgestellt und eine Komplexitätsanalyse durchgeführt.

In Kapitel 5 werden die zur automatischen Berechnung der Ereigniskosten verwendeten Methoden erläutert. Es wird ein Gütemaß entwickelt, welches den Zusammenhang

[1] vgl. [3]

zwischen einer gegebenen Kostenverteilung und den Anzahlen der aufgetretenen Ereignisse einer mit diesen Kosten berechneten günstigsten Rekonstruktion bewertet.

Im Rahmen dieser Arbeit wurde ein kommandozeilen-basiertes Javaprogramm namens DynamicTreeMap entwickelt, welches es dem Anwender erlaubt, für gegebene Ausgangsdaten eine Rekonstruktion der gemeinsamen Geschichte zu konstruieren. Sowohl die automatische Berechnung der Ereigniskosten als auch die Behandlung von mehrfach verzweigenden Stammbäumen wurden dabei umgesetzt. Die Verwendung von Zeitinformationen in den Knoten der Stammbäume wurde von Tarzen adaptiert. Als Ausgabe wird ein XML-Dokument erzeugt, welches die Daten der Rekonstruktion enthält. Diese können mit einer eigens dafür geschriebenen externen Anwendung angezeigt werden. Die Funktionsweise beider Programme wird in Kapitel 6 erläutert.

Abschließend wurden für einige sowohl konstruierte als auch reale Datensätze Berechnungen durchgeführt. Die Ergebnisse dieser Rechnungen sind in Kapitel 7 aufgeführt.

2 Theoretische Grundlagen des Rekonstruktionsproblems

Im Folgenden Abschnitt sollen die theoretischen Grundlagen für die Berechnung einer Rekonstruktion der gemeinsamen evolutionären Geschichte von Wirt- und Parasitenarten erläutert werden. Zuerst werden allerdings einige allgemeine Begriffe und Definitionen für die weitere Verwendung im Verlauf der Arbeit formalisiert. Es soll darauf eingegangen werden, auf welchen Ausgangsdaten die Berechnungen basieren und in welcher Form diese zur Verfügung stehen. Weiterhin wird das Ziel einer solchen Rekonstruktion näher beleuchtet, und es wird erläutert, welche Aussagen aus den Ergebnissen geschlossen werden können. Ein Absatz beschäftigt sich mit dem hier verwendeten Prinzip der koevolutionären Ereignisse. Diese sind nötig, um nicht nur Zuweisungen von Parasitenarten zu Wirtsarten zu postulieren, sondern ebenfalls Aussagen über genauere Vorgänge bei der gemeinsamen Evolution treffen zu können. Es folgen einige Einschränkungen biologischer Modelle und die Einführung einer Zeitfunktion. Dadurch kann die Anzahl möglicher Rekonstruktionen reduziert und durch zusätzliche Informationen qualitativ verbessert werden.

2.1 Allgemeine Definitionen

Für die Beschreibung von Stammbäumen kann das graphentheoretische Konstrukt der Bäume als Datenstruktur herangezogen werden. Diese wird ausführlich in [15] vorgestellt. In Erweiterung der allgemeinen Definition als Menge (V, E) von Knoten und Kanten, sollen Bäume des Weiteren wie folgt definiert werden.

Definition 2.1 (Baum bzw. Baumbereich)**.** *Sei $\mathbb{N}$ die Menge der natürlichen Zahlen ohne 0, dann ist ein Baum bzw. Baumbereich definiert als endliche nicht-leere Teilmenge $B \subseteq \mathbb{N}^*$ mit:*

$$u.v \in B \rightarrow u \in B, \forall u, v \in \mathbb{N}^*$$

Der Knoten eines Baumes soll somit als eine durch Punkte separierte Liste natürlicher Zahlen beschrieben werden.
Weiterhin gilt:

$$u.(i+1) \in B \rightarrow u.i \in B, \forall u \in \mathbb{N}^*, i \in \mathbb{N}$$

Es ist ϵ die Wurzel von B, $u.i$ Kind von u und u Vater von $u.i$.

Definition 2.2 (Blätter). *Mit*

$$L(B) = \{u : u \in B \wedge u.i \notin B, \forall i \in \mathbb{N}\}$$

wird die Menge der Blätter von B bezeichnet.

Definition 2.3 (Höhe). *Mit*

$$höhe(B) = sup(i : \exists u \in \mathbb{N}^i, u \in B)$$

wird die Länge des längsten Pfades im Baum B von der Wurzel zu einem seiner Blätter angegeben.

Definition 2.4 (Grad). *Mit*

$$grad(u) = sup(i : u.i \in B)$$

wird die Anzahl der Kindknoten von u bezeichnet.

Definition 2.5 (Vorgänger, Nachfolger). *Seien $u, u.v \in B$ mit $v \in \mathbb{N}^*/\epsilon$, dann gilt:*

u ist Vorgänger von $u.v$ und $u.v$ ist Nachfolger von u

Man schreibt $u <_B u.v$ bzw. $u.v >_B u$. Ist eine Gleichheit möglich, so wird $u \leq_B w$ bzw. $v \geq_B w$ verwendet.

Im Folgenden soll immer die Konvention getroffen werden, dass $i, j \in \mathbb{N}$ und $v, w \in \mathbb{N}^*$. Mit $u.i$ bzw. $u.j$ ist somit immer ein direktes Kind von u gemeint, $u.v$ bzw. $u.w$ bezeichnen hingegen einen weiter entfernteren Nachfolger.

2.2 Ausgangsdaten

Um Aussagen über die gemeinsame evolutionären Geschichte von Wirts- und Parasitenarten treffen zu können, müssen die jeweiligen Phylogenien beider Spezies in Form von evolutionären Stammbäumen bekannt sein. Hinzu kommen Informationen, welche Aufschluss darüber geben, welche Parasitenart auf welcher Wirtsart beobachtet werden konnte.

Der Stammbaum einer Art gibt den evolutionären Werdegang verwandter Spezies wieder. Ausgehend von einer einzelnen frühen Spezies zeigt er auf, in welcher Reihenfolge sich durch Speziationen Unterarten entwickelt haben. Diese Speziationen werden durch Knoten im Baum repräsentiert. Die Kanten zwischen den Knoten entsprechen der Lebensspanne einer Spezies. Damit ist der Zeitabschnitt von der Speziation der Ursprungsart aus der sie hervorgegangen ist, bis zu ihrer eigenen Speziation gemeint. Die Blätter repräsentieren die in der Regel noch existierenden Arten.[1] Ein Stammbaum beherbergt somit nur jene Arten, deren Nachfahren heute noch existieren, oder sich zumindest, zum Beispiel durch Fossilienfunde, nachweisen lassen. Es ist aber möglich, dass im Laufe der Evolution eine komplette Spezies ausstarb, ohne Spuren zu hinterlassen. In einem solchen Fall findet sich im Stammbaum natürlich weder ein Hinweis auf diese, noch auf die Speziation aus der sie hervor ging. Derartige Situationen müssen demzufolge bei der Erzeugung einer Rekonstruktion vernachlässigt werden.

Die beobachteten Parasit-Wirt-Beziehungen werden durch eine Abbildung $\varphi_{P,H}$ von $L(P)$ in $L(H)$ gekennzeichnet. Hierbei bedeutet $(p, h) \in \varphi_{P,H}$, dass der zum Knoten p gehörige Parasit auf dem zum Knoten h gehörigen Wirt ansässig war oder noch ist.[2] Oftmals wird diese Abbildung als eindeutig angenommen. Es wird also jedem Parasit nur ein Wirt zugeordnet. Es können jedoch mehrere Parasiten auf ein und demselben Wirt heimisch sein. Abbildung 2.1 zeigt ein Beispiel für die Stammbäume zweier Arten, sowie die zugehörige Abbildung $\varphi_{P,H}$.[3]

Wenn im Zusammenhang mit einem Knoten von einem bestimmten Parasit bzw. Wirt die Rede ist, so soll im Folgenden immer diejenige Spezies gemeint sein, deren Lebensspanne mit der durch den Knoten repräsentierten Speziation endete. Diese Konvention wird zum besseren Verständnis getroffen.

[1] vgl. [12] S.3
[2] vgl. [12] S.9 und [5] S.193
[3] Beispiel entnommen aus [19]

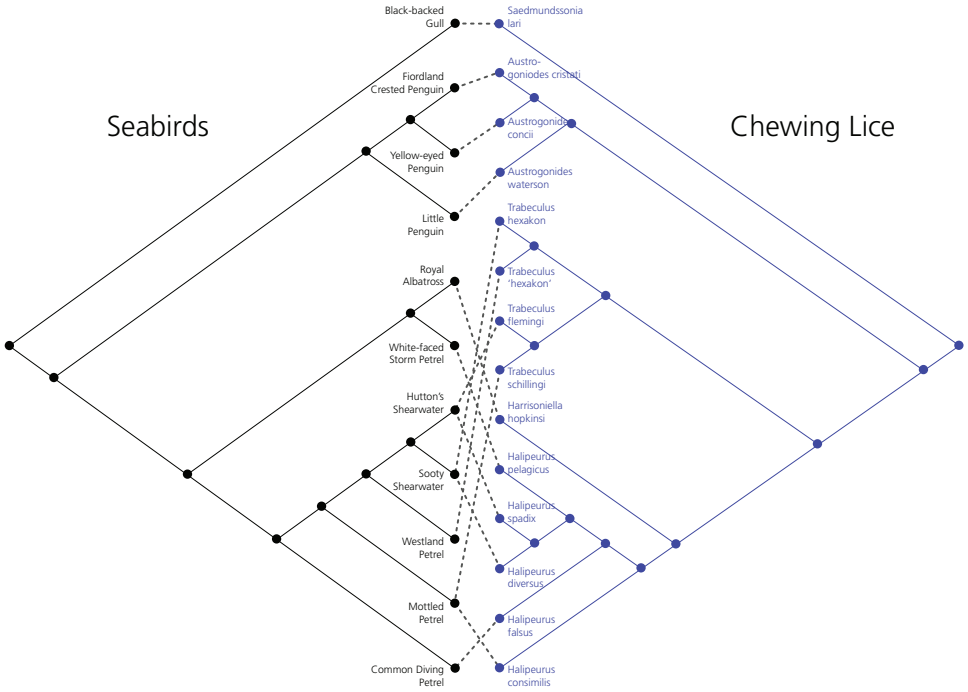


Abbildung 2.1: Stammbäume von Seabirds (Seevögel) und Chewing Lice (Kieferläuse) mit Abbildung $\varphi_{P,H}$. Dabei stehen die Kieferläuse in parasitärer Beziehung zu den Seevögeln.

2.3 Art und Ziel der Rekonstruktion

Ausgehend von der Annahme korrekter Ausgangsdaten[4] soll herausgefunden werden, welche im Laufe der Evolution existierende Parasitenart auf welcher Wirtsart heimisch war. Es soll geklärt werden, welche Umstände dazu führten, dass die heute beobachteten Parasit-Wirt-Beziehungen entstehen konnten und wie stabil diese Abhängigkeiten im Laufe der Evolution waren.

Eine Herangehensweise an diese Fragen liegt in der Rekonstruktion einer gemeinsamen evolutionären Geschichte. Dabei werden die Knoten des Parasitenbaumes auf die Knoten bzw. Kanten des Wirtsbaumes abgebildet. Die Zuweisung eines Parasitenknotens bedeutet dabei, dass der Parasit vor seiner Speziation zuletzt auf dem jeweiligen Wirt gelebt hat. Wurde der Parasitenknoten auf eine Kante des Wirtsbaumes abgebildet, so hat der Parasit seine Speziation zeitlich vor der Speziation des zugehörigen Wirtes durchgeführt. Bei einer Zuweisung direkt auf einen Wirtsknoten, fanden die Speziationen von Wirt und Parasit zeitgleich statt. Ausgehend von der Abbildung des

[4] vgl. [13] S.8

Wurzelknotens bis zu den Abbildungen der Blätter kann somit der koevolutionäre Verlauf dieser rekonstruierten Geschichte verfolgt werden.

Jedoch gibt es in den meisten Fällen sehr viele verschiedene Rekonstruktionen und es ist offensichtlich, dass es nicht ohne weiteres möglich ist, den tatsächlichen Verlauf der Evolution mit absoluter Sicherheit zu rekonstruieren. Ein entsprechendes Rekonstruktionsverfahren sollte demnach die biologisch wahrscheinlichste Abbildung finden. Diese wird wiederum maßgeblich von dem zugrunde gelegten evolutionären Modell bestimmt.

Abbildung 2.2 zeigt eine mögliche Rekonstruktion der Ausgangsdaten aus Abbildung 2.1.

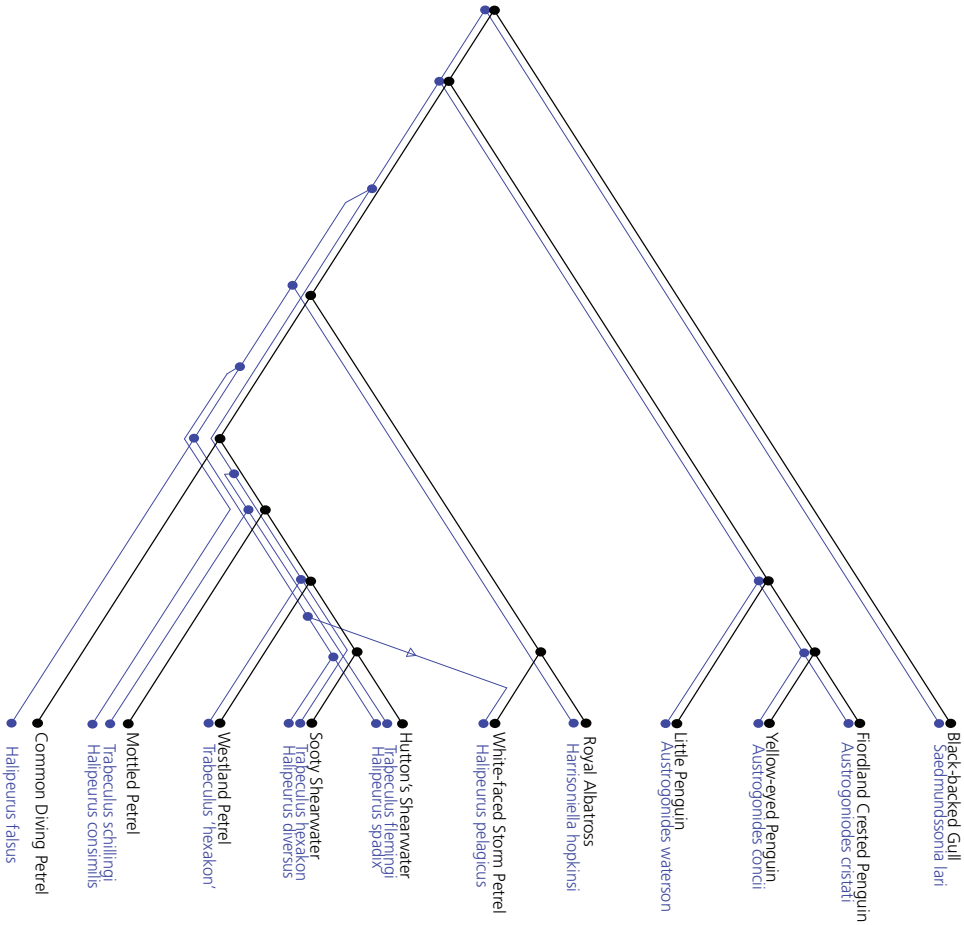


Abbildung 2.2: Rekonstruktion der gemeinsamen Evolution von Seabirds und Chewing Lice. Dabei ist der Stammbaum der Wirtsart schwarz, der der Parasitenart blau dargestellt.

2.4 Koevolutionäre Ereignisse

Zur Beschreibung von Koevolution zwischen Wirts- und Parasitenarten wurden in der Literatur verschiedene Ereignisse vorgeschlagen.[5] Die wichtigsten sollen im Folgenden beschrieben werden und als Grundlage des hier vorgestellten Ansatzes dienen.

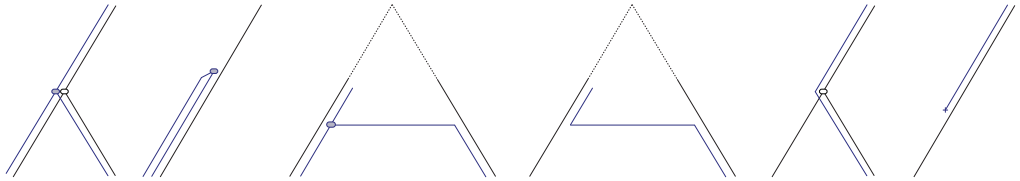


Abbildung 2.3: Die Abbildung zeigt die möglichen koevolutionären Ereignisse (a) Kospeziation, (b) Duplikation, (c) partieller und (d) kompletter Wirtswechsel, (e) Sorting sowie (f) Extinktion.

Kospeziation: Die Kospeziation beschreibt den Fall, dass ein Wirt im Laufe seiner Evolution eine Speziation durchführte und sich in zwei oder mehr Unterarten aufgliederte. Der auf diesem Wirt heimische Parasit passte sich nahezu zeitgleich an die sich ändernden Lebensbedingungen an und führte seinerseits ebenfalls eine Speziation durch. Jede der neu entstandenen Parasitenspezies siedelte sich daraufhin auf einer der neuen Wirtsspezies an.

Duplikation: Bei einer Duplikation führt nur der Parasit eine Speziation durch, während der Wirt unverändert bleibt. Dabei bleiben die neuen Unterspezies des Parasiten auf dem gleichen Wirt.

Host switch: Ein Host switch, oder auch Wirtswechsel, tritt auf, wenn eine Parasitenspezies seine Wirtsspezies verlässt und sich auf einer anderen, zeitgleich existierenden Art niederlässt. Der Zeitpunkt zu dem der Parasit seinen Wirt verlässt wird als *take-off site* bezeichnet. Analog dazu bezeichnet man den Zeitpunkt, an dem der Parasit nach einem Wirtswechsel einen neuen Wirt bevölkert mit *landing site*. Beide Punkte liegen nach Definition jeweils auf einer Kante im Wirtsstammbaum. Sie fallen somit nicht mit der Speziation eines Wirtes zusammen.

Es gibt zwei unterschiedliche Arten von Wirtswechseln. Man spricht von einem partiellen Wirtswechsel, wenn der Parasit sich durch Speziation aufteilt und eine der neu entstandenen Parasitenspezies einen Wirtswechsel durchführt, während die andere auf

[5] vgl. [24] S.27, [5] S.196, [9] S.312

dem Wirt verweilt. Im Gegensatz dazu handelt es sich um einen kompletten Wirtswechsel, wenn der Parasit ohne Speziation den Wirt verlässt, so dass auf diesem Wirt keine Nachfahren der Parasitenspezies mehr heimisch sind.

Sorting: Ein Sorting beschreibt die Speziation eines Wirtes ohne eine zeitgleich stattfindende Speziation des Parasiten. Der Parasit entscheidet sich in diesem Fall für eine der neu entstandenen Wirtsspezies.

Extinktion: Bei einer Extinktion handelt es sich um die Auslöschung einer Parasitenspezies, die zuvor auf einem Wirt heimisch war.

Bei kompletten Wirtswechseln (d) und Extinktionen (f) ist es im Unterschied zu den anderen Ereignissen nicht möglich die jeweilige Zuordnung von Parasit zu Wirt bis zu heute noch existierenden Spezies zurückverfolgen zu können. In der zugrunde gelegten Abbildung $\varphi_{P,H}$ der Ausgangsdaten findet sich kein Hinweis auf eine Assoziation des betreffenden Zweiges des Parasitenstammbaumes mit dem jeweiligen Zweig des Wirtsstammbaumes. Bei Extinktionen ist dies offensichtlich, da der Parasitenzweig vollständig ausstarb. Aber auch bei kompletten Wirtswechseln ist eine Verbindung zwischen Parasit und Ursprungswirt aus den Ausgangsdaten nicht ablesbar. Anders als im partiellen Fall finden sich keine Nachfolger des Parasiten auf Blättern des Teilbaumes unterhalb des Ursprungswirtes. Es existiert somit kein Anhaltspunkt dafür, auf welchem Wirt sich der Parasit direkt vor seinem Wirtswechsel aufgehalten haben könnte. Am ehesten könnte man den Wirtswechsel des Parasiten mit dem zugehörigen Wirt des Vater in Verbindung bringen. Eine solche Rekonstruktion entspricht einem partiellen Wirtswechsel.

Aus diesem Grund werden im Folgenden nur die vier koevolutionären Ereignisse Kospeziation, Sorting, Duplikation und partieller Wirtswechsel[6] verwendet.[7]

Ausgehend von der Fahrenholz-Regel[8] wird der Kospeziation eine besondere Bedeutung beigemessen. Bei dieser fallen die Speziationen von Wirt und Parasit nahezu zeitgleich aufeinander. Biologischer Hintergrund ist die Annahme, dass der Wirt seine eigene Speziation initiiert und der Parasit sich kurze Zeit später an die veränderten Umstände anpasst. Daraus lässt sich eine gewisse strukturelle Ähnlichkeit beider Stammbäume schlussfolgern. Im Extremfall sind beide Stammbäume isomorph. Da dies bei realen Daten nur sehr selten der Fall ist, muss es noch weitere Einflussfaktoren als die Speziation des Wirtes geben. Im Allgemeinen erscheint es dennoch sinnvoll Rekonstruktionen

[6] Partielle Wirtswechsel werden der Einfachheit halber im Weiteren als Wirtswechsel bezeichnet

[7] Die Beschränkung auf diese vier koevolutionären Ereignisse wurde auch von Charleston und Perkins in [8] vorgenommen.

[8] Die Fahrenholz-Regel besagt, dass die Entwicklung der Wirtsspezies die Entwicklung der Parasitenspezies maßgeblich mitbestimmt. [10]

mit vielen Kospeziationen zu bevorzugen. Diesen Ansatz verfolgen auch Charleston und Perkins[9].

2.5 Kostenmodell

Um verschiedene Rekonstruktionen hinsichtlich ihrer biologischen Relevanz direkt miteinander vergleichen zu können, werden für jedes Ereignis Kosten angenommen. Die Gesamtkosten einer Rekonstruktion ergeben sich aus der Summe der Kosten aller aufgetretenen Ereignisse. Unter allen möglichen Rekonstruktionen ist somit diejenige am plausibelsten, welche die geringsten Gesamtkosten aufweist.

Sowohl in [7] als auch in [14] wurden für die Kosten der verwendeten vier Ereignisse folgende Bedingungen festgelegt. Die Kosten einer Kospeziation müssen kleiner oder gleich 0 sein und die Kosten von Duplikation, Sorting und Wirtswechsel größer ([7]) bzw. größer oder gleich 0 ([14]). Diese Einschränkung wurde getroffen, da man der Kospeziation, wie oben erwähnt, eine besondere Rolle beimisst. Andererseits hat die Wahl einer solchen Kostenbedingung erheblichen Einfluss auf die Komplexität der verwendeten Datenstruktur.[10]

2.6 Einschränkungen des biologischen Modells

In der Forschung werden zusätzliche Einschränkungen eingeführt, welche entsprechende Modelle vereinfachen sollen. Einerseits sind diese biologisch motiviert, andererseits reduzieren sie die Komplexität des Problems. Im Folgenden sollen die von Ronquist[11] verwendeten Einschränkungen erläutert werden.

One-host-per-parasite-Annahme Diese Einschränkung reduziert die Anzahl der Wirte für einen Parasit. Sie besagt, dass jede Parasitenspezies im Laufe ihrer Evolution zu einem Zeitpunkt nur einer einzelnen Wirtsspezies zugeordnet ist. Aus algorithmischer Sicht bedeutet dies, dass für eine koevolutionäre Rekonstruktion der evolutionäre Stammbaum der Parasitenart auf den der Wirtsart eindeutig abgebildet werden kann.

[9] vgl. [7]

[10] Diese Datenstruktur, genannt *Jungles* (vgl. [5]), ist ein gerichteter Graph, in dem für jede mögliche Abbildung eines Parasiten p auf einen Wirt h in aller Regel zwei Knoten vorhanden sind. Ein Knoten $(p : h)$, steht dabei für eine Kospeziation von p auf h und ein weiterer Knoten $(p : h)^*$ repräsentiert die koevolutionären Ereignisse Duplikation und Wirtswechsel zwischen p und h. Eine Kante von einem Knoten $(p : h)$ zu einem Knoten $(p' : h')$ gibt dabei an, dass das Parasitenkind p' von p auf den Wirt h' abgebildet werden kann, wenn p seinerseits auf h abgebildet wurde. Die bei dieser Abbildung auftretenden Ereignisse werden mit den jeweiligen Kanten assoziiert.

[11] vgl. [24] S.26-27

Ohne diese Einschränkung müsste ein Knoten im Parasitenbaum gegebenenfalls mehrfach in den Wirtsbaum abgebildet werden. Dies hätte zur Folge, dass sich einerseits die Menge der Parasit-Wirt-Abbildungen drastisch erhöhen würde und andererseits, dass verschiedene Teile der Rekonstruktion stark abhängig voneinander wären, da gleiche Parasitenspezies an unterschiedlichen Stellen auftreten können.

Biologisch ist dieser Ansatz dadurch motiviert, dass eine Koevolution sehr viel wahrscheinlicher wird, wenn eine Parasitenspezies sehr stark an ihre Wirtsspezies angepasst ist und somit nur mit dieser interagiert. Jede Veränderung des Wirtes zwingt den Parasiten sich anzupassen. Nach einer Speziation des Wirtes steigt dadurch die Wahrscheinlichkeit für eine Speziation des Parasiten.

Sollte ein Parasit auf mehreren Wirten heimisch sein, so wird durch das Modell ein Hauptwirt angenommen.

Independant-parasites-Annahme Bei der Independant-parasites-Annahme wird davon ausgegangen, dass verschiedene Parasitenspezies, welche auf der gleichen Wirtsspezies heimisch sind, sich gegenseitig in ihrer Evolution nicht beeinflussen. Diese Einschränkung ermöglicht eine koevolutionäre Rekonstruktion auf Basis der zugrunde liegenden Wirtsphylogenie ohne die Anzahl und Art der Parasiten auf einem Wirt in Betracht ziehen zu müssen. Dadurch können disjunkte Teile des Parasitenstammbaumes unabhängig voneinander betrachtet und so gültige Teilrekonstruktionen erzeugt werden. Aufgrund dieser Annahme ist es möglich das Rekonstruktionsproblem in Form dynamischer Programmierung zu modellieren.

Biologisch kann diese Einschränkung jedoch nicht begründet werden. Da allerdings ohne die Annahme das Problem algorithmisch wesentlich komplexer wäre, wird sie dennoch beibehalten.

2.7 Definition von Zeitfunktionen für Wirts- und Parasitenspezies

In [14] wird ein Modell vorgestellt, welches die Integration von Zeitinformationen ermöglicht. Dieses soll ebenfalls für den in dieser Arbeit vorgestellten Ansatz verwendet werden. Dafür werden für die Speziationen in den Stammbäumen Zeitzonen in Form von ganzzahligen Werten definiert. Die genaue Länge dieser Zeitzonen bleibt dabei ungeklärt. Auch können die definierten Zeitzonen von unterschiedlicher Dauer sein. Es wird nur gefordert, dass sich die Zonen auf einer Zeitachse aufsteigend anordnen lassen. Jedem Knoten, und somit jeder Speziation, kann auf diese Weise eine Zeitperiode t zugewiesen werden, in der dieses Ereignis stattfand.

Um eine korrekte zeitliche Abfolge der Ereignisse zu gewährleisten, kann einem Knoten nur eine Zeitzone t zugewiesen werden, welche größer oder gleich der Zeitzone t' seines Vaterknoten ist.

Definition 2.6 (Gültige Zeitfunktion $T_1(B)$)**.** *Es ist $T_1(B)$ eine gültige Zeitfunktion, wenn gilt:*

$$(T_1(u) = t') \wedge (T_1(u.v) = t) \wedge (t' \leq t) : \forall u, u.v \in B$$

Da es oftmals nicht möglich ist, jedem Knoten eine bestimmte Zeitzone zuzuordnen, und da sich die Zeitzonen verschiedener Phylogenien auch nicht immer genau aufeinander abbilden lassen, wurde des Weiteren vorgeschlagen Intervalle von Zeitzonen einzuführen. Einem Knoten kann somit eine Reihe aufeinander folgender Zeitzonen zugewiesen werden, ohne festlegen zu müssen, in welcher dieser Zeitzonen die Speziation genau stattfand. Für diese Art der zeitlichen Einordnung wird ebenfalls gefordert, dass einem Knoten nur ein Zeitintervall zugewiesen werden kann, wenn es nicht vor dem festgelegten Zeitintervall des Vaterknotens liegt.

Definition 2.7 (Gültige Zeitfunktion $T_2(B)$)**.** *Es ist $T_2(B)$ eine gültige Zeitfunktion, wenn gilt:*

$$(T_2(u) = [s', t']) \wedge (T_2(u.v) = [s, t]) \wedge (s' \leq s) \wedge (t' \leq t) : \forall u, u.v \in B, v \neq \epsilon$$

Wie auch in [14] soll in unseren Ansatz für den Wirtsstammbaum H eine gültige Zeitfunktion $T_1(P)$ von einzelnen Zeitzonen und für den Parasitenstammbaum P eine gültige Zeitfunktion $T_2(H)$ von Zeitzonenintervallen verwendet werden.

Zur vereinfachten Schreibweise sollen im Folgenden für die verschiedenen Zeitfunktionen vergleichende Ordnungsrelationen eingeführt werden.

Definition 2.8 (Ordnungsrelationen für $T_1(H)$)**.** *Seien $T_1(h_1) = t_1$ und $T_1(h_2) = t_2$, dann ist*

$$T_1(h_1) <_{T1} T_1(h_2) \leftrightarrow t_1 < t_2$$

$$T_1(h_1) \leq_{T1} T_1(h_2) \leftrightarrow t_1 \leq t_2$$

$$T_1(h_1) =_{T1} T_1(h_2) \leftrightarrow t_1 = t_2$$

$$T_1(h_1) \geq_{T1} T_1(h_2) \leftrightarrow t_1 \geq t_2$$

$$T_1(h_1) >_{T1} T_1(h_2) \leftrightarrow t_1 > t_2$$

Definition 2.9 (Ordnungsrelationen für $T_2(P)$)**.** *Seien* $T_2(p_1) = [s_1, t_1]$ *und* $T_2(p_2) = [s_2, t_2]$*, dann ist*

$$T_2(p_1) <_{T2} T_2(p_2) \leftrightarrow t_1 < s_2$$

$$T_2(p_1) \leq_{T2} T_2(p_2) \leftrightarrow t_1 \leq t_2$$

$$T_2(p_1) =_{T2} T_2(p_2) \leftrightarrow t_1 \geq s_2 \wedge s_1 \leq t_2$$

$$T_2(p_1) \geq_{T2} T_2(p_2) \leftrightarrow s_1 \geq s_2$$

$$T_2(p_1) >_{T2} T_2(p_2) \leftrightarrow s_1 > t_2$$

Definition 2.10 (Ordnungsrelationen für $T_1(H)$ in Bezug auf $T_2(P)$)**.** *Seien* $T_1(h) = t_1$ *und* $T_2(p) = [s_2, t_2]$*, dann ist*

$$T_1(h) <_{T12} T_2(p) \leftrightarrow t_1 < s_2$$

$$T_1(h) \leq_{T12} T_2(p) \leftrightarrow t_1 \leq t_2$$

$$T_1(h) =_{T12} T_2(p) \leftrightarrow t_1 \geq s_2 \wedge t_1 \leq t_2$$

$$T_1(h) \geq_{T12} T_2(p) \leftrightarrow t_1 \geq s_2$$

$$T_1(h) >_{T12} T_2(p) \leftrightarrow t_1 > t_2$$

Analog lässt sich eine Ordnungsrelation für $T_2(P)$ in Bezug auf $T_1(H)$ definieren.

Es werden im Folgenden für Wirtsstammbäume H nur Zeitfunktionen $T_1(H)$ und für Parasitenstammbäume P nur Zeitfunktionen $T_2(P)$ verwendet. Wenn aus dem Kontext heraus klar ist, welche Zeitfunktion bzw. welche Ordnungsrelation gemeint ist, werden der Einfachheit halber nur die Schreibweisen $T(H)$ und $T(P)$ sowie $<_T$, $\leq_T$, $=_T$, $\geq_T$ und $>_T$ benutzt. Durch die obigen Definitionen werden nur Zeitzonen bzw. Zeitzonenintervalle spezifiziert. Sollte im Verlauf dieser Arbeit ein genauer Zeitpunkt innerhalb einer solchen Zeitzone für einen Knoten p oder h angegeben werden, so wird dafür $t(p)$ bzw. $t(h)$ geschrieben.

Die Einführung von Zeitinformationen hat zwei wesentliche Vorteile. Zum Einen können durch die Hinzunahme dieser zusätzlichen Informationen realistischere Rekonstruktionen erzeugt werden. Zum Anderen reduziert sich die Anzahl der gültigen Lösungen. Für eine gültige Rekonstruktion kann gefordert werden, dass die Abbildung eines Parasitenknotens p auf einen Wirtsknoten h nur dann erlaubt ist, wenn $T(p) =_T T(h)$ und somit die Zeitzone von h im Zeitzonenintervall von p liegt.

Für ein eventuell aufgetretenen Wirtswechsel, den ein Parasit p während seiner Speziation nach $p.i$ durchführte, bedeutet dies zusätzlich, dass sich take-off und landing site dieses Wirtswechsels in der gleichen Zeitzone befinden müssen. Allgemein wird vorausgesetzt, dass take-off und landing zeitgleich stattfinden. Da diese Punkte jedoch auf Kanten des Wirtsbaumes liegen, kann nicht immer eine eindeutige Zeitzone bestimmt werden. Zumindest müssen sich die Intervalle der zugehörigen Wirtsknoten überlappen. Angenommen die take-off site befindet sich auf einer Kante $(h_1, h_1.j)$ und die landing site auf der Kante $(h_2, h_2.k)$, dann bedeutet dies formal, dass $T(h_1.j) \geq_T T(h_2)$ und $T(h_1) \leq_T T(h_2.k)$ gelten muss.

Da wir, wie in Abschnitt 2.4 erwähnt, nur partielle Wirtswechsel betrachten wollen, wird in unserem Fall Parasit p immer dem Wirt $h_1.j$ zugewiesen sein. Jedoch muss nicht notwendigerweise auch $p.i$ immer auf $h_2.k$ abgebildet werden. Es ist durchaus denkbar, dass die Ansiedlung auf einem neuen Wirt deutlich vor der nächsten Speziation stattgefunden hat. Dies ist insbesondere dann zwingend notwendig, wenn die Lebensspanne des Wirtes, auf den $p.i$ abgebildet wird, nicht mit dem zur take-off site gehörenden Zeitintervall $[T(h_1), T(h_1.j)]$ überlappen würde. In einem solchen Fall muss der Parasit früher auf einem Wirt $h_2.k$ gelandet sein und von dort führte er bis zu seiner Speziation Sortings durch. Abbildung 2.4 stellt dieses Szenario grafisch dar.

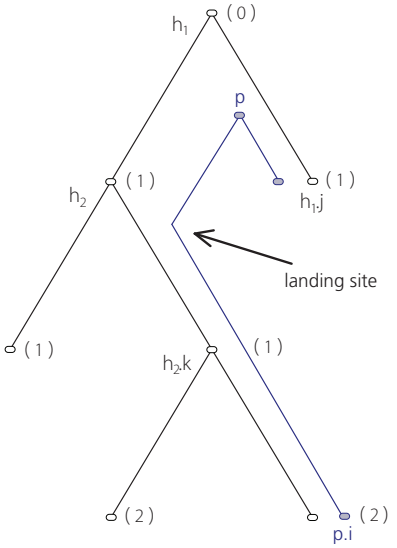


Abbildung 2.4: Der Parasit verlässt nach der Speziation den Wirt $h_1.j$ und siedelt sich auf $h_2.k$ an. Bis zu seiner Speziation führt er noch ein zusätzliches Sorting durch. Die Zeitzonen von take-off ($[T(h_1), T(h_1.j)] = [0, 1]$) und landing site ($[T(h_2), T(h_2.k)] = [1, 1]$) überlappen sich.

Anhand dieser Illustration wird deutlich, dass die Verwendung zusätzlicher Zeitinformationen nicht nur die Anzahl der Rekonstruktionen vermindert, sondern dass ebenfalls das Ereignis des Wirtswechsels differenzierter betrachtet werden kann. Einerseits ist nicht jeder Wirtswechsel im Baum erlaubt, und andererseits werden die Wirtswechsel durch die zusätzlich einzufügenden Sortings teurer, je weiter die Zeitzonen der Abbilder von p und $p.i$ im Wirtsbaum auseinander liegen.

3 Betrachtung des Rekonstruktionsproblems unter dem Gesichtspunkt dynamischer Programmierung

Im folgenden Kapitel soll ein dynamischer Programmieransatz vorgestellt werden, der eine kostenminimale Rekonstruktion der gemeinsamen Evolution von Wirts- und Parasitenspezies für die zwei gegebenen Phylogenien, die jeweiligen Ereigniskosten und die Abbildung $\varphi_{P,H}$ der Blätter erzeugt. Mit Hilfe dieses Ansatzes ist es möglich ein Allgemeineres, als das in Abschnitt 2.4 vorgestellte Kostenmodell zu verwenden.

Auch betrachteten bisherige Ansätze aus Komplexitätsgründen oft nur koevolutionäre Ereignisse, bei denen die beteiligten Arten ausschließlich binäre Speziationen durchführten. Dabei gehen bei einer Speziation von Parasiten- oder Wirtsarten immer genau zwei Unterarten aus der Ursprungsart hervor. Da aber die Ausgangsdaten nicht immer eine zweifelsfreie binäre Interpretation zulassen, soll auf Basis des dynamischen Ansatzes eine Möglichkeit vorgestellt werden, mit der auch Multifurkationen, also Mehrfachverzweigungen in den evolutionären Stammbäumen betrachtet werden können.

Ein grundsätzliches Problem beim Finden von Rekonstruktionen ist die Gültigkeit der Lösungen in Bezug auf chronologische Konsistenz. Da die gemeinsame evolutionäre Geschichte zweier Arten einem zeitlich wohl definierten Ablauf entspricht, muss eine gültige Rekonstruktion dieses Kriterium erfüllen. Zwar löst der in dieser Diplomarbeit entwickelte algorithmische Ansatz diese chronologischen Inkonsistenzen nicht explizit auf, doch werden diese in den folgenden Abschnitten klassifiziert und ausführlich erörtert.

3.1 Dynamischer Ansatz

Aus algorithmischer Sicht bleiben beim Auffinden einer koevolutionären Rekonstruktion Wirts- und Parasitenbaum strukturell unverändert. Die Knoten des Parasitenbaumes werden in den Wirtsbaum abgebildet. Grundsätzlich kann bei einer solchen Abbildung ein Parasitenknoten sowohl einem Knoten als auch einer Kante des Wirtsbaumes

zugewiesen werden. Man kann allerdings aus dem Kontext der koevolutionären Ereignisse heraus ablesen, um welche Art der Zuweisung des Parasiten es sich gehandelt haben muss. Fand eine Kospeziation statt, so müssen die Speziationen von Wirt und Parasit zeitgleich aufgetreten sein. Dies entspricht einer Knoten-zu-Knoten Abbildung. Bei einer Duplikation oder einem Wirtswechsel führte nur der Parasit eine Speziation durch. An dieser Stelle kann nach Definition nur eine Knoten-zu-Kanten Abbildung durchgeführt werden.

Aus technischen Gründen sollen formal jedoch nur Knoten-zu-Knoten Abbildungen betrachtet werden. Aus der Information des jeweilig aufgetretenen Ereignisses kann später rekonstruiert werden, ob der Knoten eventuell doch auf eine Kante abgebildet werden muss.

In Abgrenzung zu Charleston[1] werden demzufolge keine separaten Abbildungen von Knoten zu Kanten mehr unterschieden.

Der Algorithmus berechnet für jeden Teilbaum des Parasitenbaumes und für jede mögliche Abbildung der Wurzel dieses Teilbaumes auf einen Knoten des Wirtsbaumes die kostengünstigste Rekonstruktion. Für die einelementigen Teilbäume - also die Blattknoten des Parasitenbaumes - besteht die Rekonstruktion aus dem direkten Zuweisen der Blätter zu einem Wirtsknoten. Die Kosten einer solchen Rekonstruktion werden genau dann auf 0 gesetzt, wenn eine zugehörige Abbildung in $\varphi_{P,H}$ existiert. Anderenfalls werden die Kosten auf ∞ gesetzt, um diese Zuweisung zu verbieten.

Ausgehend von diesen Teilrekonstruktionen werden bottom-up die inneren Knoten des Parasitenbaumes ebenenweise von den Blättern zur Wurzel durchlaufen. Für jeden einzelnen Knoten wird eine Liste von kostengünstigsten Teilbaumrekonstruktionen erstellt - je eine pro möglicher Abbildung des Parasitenknotens auf einen der Wirtsknoten. Jede dieser Rekonstruktionen setzt sich aus den Teilrekonstruktionen der Kindknoten zusammen, deren Kombination wiederum kostenminimal ist. Dabei summieren sich die Kosten aus drei Werten auf:

1. aktuelle Kosten der am betrachteten Wirtsknoten auftretenden Kospeziation, Duplikation, Wirtswechsel sowie Sortings
2. Sortingkosten der Parasitenkinder vom aktuellen Wirt bis zu ihrer Abbildung in den Wirtsbaum.
3. Rekonstruktionskosten der Kindknoten

Diese Unterteilung spiegelt sich auch in zwei unterschiedlichen Typen von Sortings wieder. Als *direkte Sortings* sollen diejenigen Sortings bezeichnet werden, welche zum

[1] vgl. [5]

Zeitpunkt der Speziation des betrachteten Knotens h auftreten (Sortings aus 1.). Im Gegensatz dazu wird bei den Sortings Parasitenkinder $p.i$, welche auf dem Weg von h zu ihren Abbildern h_i auftreten von *zusätzlichen Sortings* gesprochen (Sortings aus 1.).

Am Ende der Berechnung existiert für jeden Parasitenknoten für den bei ihm beginnenden Teilbaum genau eine kostenminimale Rekonstruktion pro Abbildung auf einen Knoten im Wirtsbaum. Somit finden sich auch die günstigsten Rekonstruktionen für den gesamten Parasitenbaum in dessen Wurzel wieder. Unter diesen ist am Ende die kostenminimalste Lösung auszuwählen.

Wenn während der Berechnung zu jedem Kostenwert die dazugehörigen Abbildungen der Kindknoten des Parasitenbaumes in den Wirtsbaum gespeichert werden, so kann, wie später gezeigt wird, die günstigste Rekonstruktion mit Hilfe der ermittelten Kosten sehr einfach erzeugt werden. Aus diesem Grund soll sich die folgende formale Beschreibung nur auf die Berechnung der minimalen Kosten beziehen.

3.1.1 Formale Beschreibung des dynamischen Ansatzes

Es werden für jeden Parasitenknoten p und jeden Wirtsknoten h die minimalen Rekonstruktionskosten $C_{p,h}$ des Parasitenteilbaumes mit Wurzel p und der Abbildung von p auf h berechnet. Diese Kosten lassen sich in Form einer dynamischen Programmierung wie folgt formulieren.

$$C_{p,h}=\begin{cases} 0 & \text{für } p\in L(P),\ (p,h)\in\varphi_{P,H} \\ \infty & \text{für } p\in L(P),\ (p,h)\notin\varphi_{P,H} \\ \min\limits_{\substack{[h_1,\ldots,h_{grad(p)}]:\\ h_1,\ldots,h_{grad(p)}\in H}} \left(\left(\sum\limits_{i=1}^{grad(p)} (C_{p.i,h_i})\right)+\min(E(h,h_1,\ldots,h_{grad(p)}))\right) & \text{sonst} \end{cases} \tag{3.1}$$

Dabei bezeichnet $E(h, h_1, \ldots, h_{grad(p)})$ die Menge der möglichen Summen von Ereigniskosten, welche entstehen wenn nach der Speziation eines Parasiten p, der auf dem Wirt h heimisch war, sich dessen Kinder $p.i$ auf den jeweiligen Wirten h_i ansiedeln. Diese Kosten setzen sich einerseits aus den Kosten der koevolutionären Ereignisse zusammen, an denen die Speziation von p unmittelbar beteiligt war (Kosten aus 1.). Andererseits kommen noch zusätzliche Sortings für die Abbildungen der $p.i$ auf die h_i hinzu (Kosten aus 2.). Dabei wird jede mögliche Variation der h_i betrachtet. Für eine spezielle Auswahl der h_i werden diejenigen koevolutionären Ereignisse gesucht, welche eine Abbildung der Kindknoten von p auf die jeweiligen h_i erlauben. In Abbildung 3.1 sind beispielhaft zwei gültige und eine ungültige Rekonstruktion dargestellt.

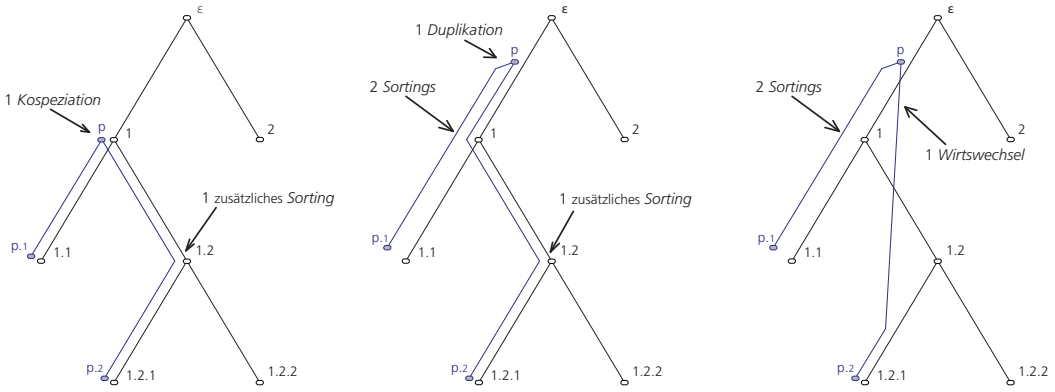


Abbildung 3.1: Zwei gültige (a), (b) und eine ungültige (c) Rekonstruktion ein und derselben Ausgangssituation. Der Parasitenknoten p wird auf den Wirt 1 abgebildet. Seine Kindknoten auf die jeweiligen Wirte 1.1 bzw. 1.2.1.

In den beiden gültigen Fällen ist ein zusätzliches Sorting nötig, damit der Parasit $p.2$ auf den Wirt 1.2.1 abgebildet werden kann. Das Beispiel zeigt ebenfalls auf, dass es verschiedene Konstellationen von Ereignissen geben kann, welche bei gleicher Wahl der h_i für die jeweiligen $p.i$ gültige Rekonstruktionen liefern. Um die günstigste unter diesen auszuwählen, müssen die sie erzeugenden Ereignismengen verglichen werden. Der Term $\min(E(h, h_1, \ldots, h_{grad(p)}))$ drückt dies aus.

3.1.2 Integration der Zeitinformationen

Um die zusätzlichen Zeitinformationen mit zu berücksichtigen muss ein weiterer Term $Z(p, h)$ eingeführt werden. Dieser ist nötig, da nur Abbildungen erlaubt sein sollen, bei denen die Zeitzonen der Wirte innerhalb der Zeitzonenintervalle der auf sie abgebildeten Parasiten liegen. Formal bedeutet dies:

$$Z(p,h) = \begin{cases} 0 & \text{für } T(p) \leq_T T(h) \wedge (h = \epsilon \vee (T(p) \geq_T T(h') \text{ mit } \exists i \in \mathbb{N} : h'.i = h)) \\ \infty & \text{sonst} \end{cases} \tag{3.2}$$

$Z(p, h)$ ist also 0, wenn es eine Zeitzone im Zeitzonenintervall $T(p)$ gibt, welche kleiner oder gleich der Zeitzone von h und größer oder gleich der Zeitzone von h', des Vorgängers von h, ist. Für die Abbildung eines Parasiten p auf einen Wirt h und die jeweiligen Abbildungen der Kinder von p auf die zugehörigen Wirte h_1 bis $h_{grad(p)}$ lässt sich dieser Term wie folgt erweitern.

$$Z(p, p.1, \ldots, p.grad(p), h, h_1, \ldots, h_{grad(p)}) = Z(p,h) + Z(p.1, h_1) + \ldots + Z(p.grad(p), h_{grad(p)}) \tag{3.3}$$

Der Übersichtlichkeit halber soll im Folgenden für diesen Term der Ausdruck Z verwendet werden. Z ist also gleich 0, wenn die Abbildungen der Parasitenknoten auf die jeweiligen Wirtsknoten unter Berücksichtigung der Zeitzonen gültig sind. Anderenfalls ist $Z = \infty$.

Somit ergibt sich für den dynamischen Ansatz folgende Funktion:

$$C_{p,h}=\begin{cases} Z(p,h) & \text{für } p\in L(P),\ (p,h)\in\varphi_{P,H} \\ \infty & \text{für } p\in L(P),\ (p,h)\notin\varphi_{P,H} \\ \min\limits_{\substack{[h_1,\ldots,h_{grad(p)}]:\\ h_1,\ldots,h_{grad(p)}\in H}} \left(\left(\sum\limits_{i=1}^{grad(p)} (C_{p.i,h_i})\right)+\min(E(h,h_1,\ldots,h_{grad(p)}))+Z\right) & \text{sonst} \end{cases} \tag{3.4}$$

3.2 Erweiterung des Kostenmodells

Das in Abschnitt 2.4 vorgestellte Kostenmodell hat die Einschränkung, dass die Kospeziationskosten kleiner oder gleich 0 sein müssen. Dies hat nicht nur biologische, sondern auch algorithmische Gründe.

Im Fall, dass ein Parasit $p.1$ auf einem Wirt $h.1$ und ein zweiter Parasit $p.2$ auf einem Wirt $h.2$ heimisch sind, wird von Charleston nachgewiesen, dass für eine kostenminimale Rekonstruktion der gemeinsame Vater p von $p.1$ und $p.2$ auf den gemeinsamen Vater h von $h.1$ und $h.2$ abgebildet sein muss.[2] Diese Situation entspricht der Kospeziation, welche durch die negativen Kosten immer günstiger ist als jede andere Möglichkeit. Somit können andere Ereignisse für eine kostenminimale Rekonstruktion ausgeschlossen werden. Dies führt bei Charleston zu einer enormen Reduktion der Komplexität der verwendeten Datenstruktur. Theoretisch könnte aber eine Duplikation oder auch ein Wirtswechsel stattgefunden haben. Im Falle der Duplikation wäre p auf die Kante vor h gesetzt worden. Bei einem Wirtswechsel müsste p auf eine der Kanten $(h, h.1)$ bzw. $(h, h.2)$ abgebildet werden.

Diese zusätzlichen Varianten, welche in Abbildung 3.2 (b)-(d) dargestellt sind, können bei Kospeziationskosten größer 0 und entsprechend gewählten Kosten der anderen Ereignisse günstiger sein als Variante (a). So treten bei binären Stammbäumen beispielsweise nie Kospeziationen auf, wenn die zugehörigen Kosten größer sind als die Summe der Kosten einer Duplikation und zweier Sortings. In einem solchen Fall würde Variante (b) immer bevorzugt.

[2] vgl. [5] S.14 Lemma 3

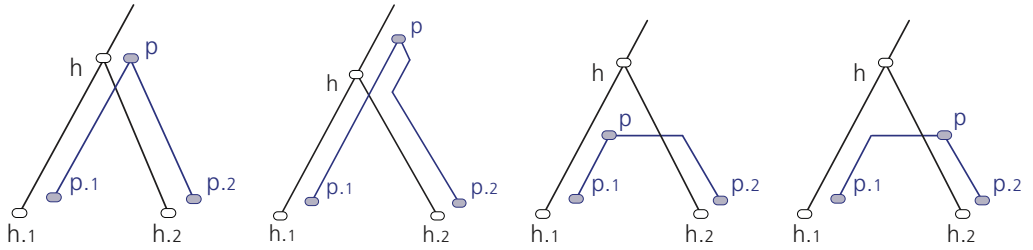


Abbildung 3.2: Für die festgelegten Abbildungen $p.1$ auf $h.1$ und $p.2$ auf $h.2$ wird im Falle der Kostenbeschränkung von Charleston nur Variante (a) betrachtet. Bei beliebigen Kostenvorgaben können auch andere Varianten günstiger sein. (b) zeigt eine mögliche Rekonstruktion mit einer Duplikation und zwei direkten Sortings. Bei (c) und (d) ist je ein Wirtswechsel nötig.

Auch könnte der Knoten p noch früher im Wirtsbaum, auf einen Vorgänger von h abgebildet werden. Dann wäre wie im Fall 3.2 (b) eine Duplikation nötig. Es würden allerdings noch weitere Sortings hinzukommen. Da aber auch negative Kosten für Sortings nicht ausgeschlossen werden sollen, ist auch dies eine nicht zu vernachlässigende Möglichkeit. Durch den beschriebenen dynamischen Ansatz werden all diese Kombinationen betrachtet und eine derartige Einschränkung des Kostenmodells ist nicht mehr nötig. Wie aber in Kapitel 5 näher erläutert wird, ist es unter Umständen sinnvoll alle Kosten als positiv anzunehmen.

3.3 Behandlung von Multifurkationen

3.3.1 Allgemeine Vorgehensweise

Wie bereits erwähnt, werden für viele vergleichbare Ansätze[3] gewurzelte Binärbäume als Ausgangsdaten vorausgesetzt. Da jedoch nicht immer eine exakte binäre Phylogenie der Wirts- und Parasitenarten gegeben ist, soll im Folgenden beschrieben werden, wie im Fall von mehrfach verzweigenden Stammbäumen verfahren werden kann. Ein allgemeiner Ansatz wäre, dass man versucht alle sich ergebenden binären Varianten einzeln zu betrachten. Das würde bedeuten, dass man für einen mehrfach verzweigenden Baum alle binären Varianten erzeugen müsste, um diese dann als Ausgangsbasis für je eine Berechnung zu verwenden. Allerdings würde sich dabei die Komplexität des Verfahrens drastisch erhöhen. Aus diesem Grund soll hier eine andere Herangehensweise vorgeschlagen werden.

[3] so auch die schon erwähnten algorithmischen Umsetzungen von [5] und [14]

Unter einer Multifurkation im Stammbaum vestehen wir im Folgenden die Entwicklung mehrerer neuer Spezies aus einer Ursprungsspezies. Wir gehen davon aus, dass die neu entstandenen Spezies so zeitnah auseinander hervor gingen, dass nicht mehr nachvollzogen werden kann, in welcher Reihenfolge sie dies taten. Aus Sicht der Wirtsspezies soll deshalb angenommen werden, dass diese Speziationen wirklich zeitgleich abliefen. Die Speziationen werden alle zu einem Knoten im Baum zusammengefasst. Es wird demnach nicht versucht den Wirtsbaum in eine binäre Form zu transformieren.

Für den Parasitenbaum sieht dies allerdings ein wenig anders aus. Bei der binären Speziation eines Parasiten wird eine gewählte Abbildung der beiden Kinder, abgesehen von Sortings, immer mit einer Kospeziation, einer Duplikation oder einem Wirtswechsel koevolutionär erklärt. Diese Ereignisse wurden binär definiert, was wegen der Vergleichbarkeit mit anderen Algorithmen, so beibehalten werden soll. Im Falle eines mehrfach verzweigenden Parasitenstammbaumes muss für diesen somit eine „quasi"-binäre Form gefunden werden, welche die Multifurkationen auflöst. Wird ein mehrfach verzweigender Parasitenknoten auf einen Wirtsknoten abgebildet, so kann nicht mehr nur eines der binären Ereignisse die gemeinsame Entwicklung erklären. Vielmehr ist eine Abfolge von koevolutionären Ereignissen erforderlich. Der hier lax verwendete Begriff der „quasi"-binären Form soll lediglich andeuten, dass auch mehrfach verzweigende Duplikationen sowie Wirtswechsel erlaubt sind. Diese scheinbar nicht-binären Ereignisse spiegeln aber lediglich eine ganze Reihe binärer Varianten wieder, welche von den Kosten her nicht voneinander zu unterscheiden sind. Eine Duplikation an der n Parasiten beteiligt sind bekommt somit die Kosten von $n-1$ einzelnen Duplikationen. Ebenso verhält es sich mit den Kosten für Wirtswechsel. Die Abbildungen 3.3 verdeutlichen diesen Sachverhalt anhand einer dreifachen Verzweigung des Parasitenknotens p. Dabei geben die Abbildungen 3.3 (a) und (b) die „quasi"-binäre Form wieder, wohingegen (a'), (a"), (b') und (b") die von (a) bzw. (b) repräsentierten binären Formen zeigen.

Eine Kospeziation hingegen kann nicht auf diese Weise dargestellt werden. Ein Merkmal von Kospeziationen ist, dass bei diesen die Speziationen von Wirt und Parasit zeitgleich ablaufen. Da bei einer mehrfach verzweigenden Speziation des Wirtes gefordert wurde, dass diese nur zu einem Zeitpunkt stattfindet, existiert auch nur dieser eine Zeitpunkt für mögliche Kospeziationen. Eine mehrfach verzweigende Speziation des Parasiten kann zwar als mehrere Kospeziationen interpretiert werden, diese müssen aber alle zum exakt gleichen Zeitpunkt stattfinden. Somit ist eine baumförmige Aufteilung, wie sie bei Duplikationen und Wirtswechseln verwendet wurde, nicht möglich. Statt dessen werden voneinander unabhängige Kospeziationen verwendet, wie in Abbildung 3.4 (b) dargestellt ist.

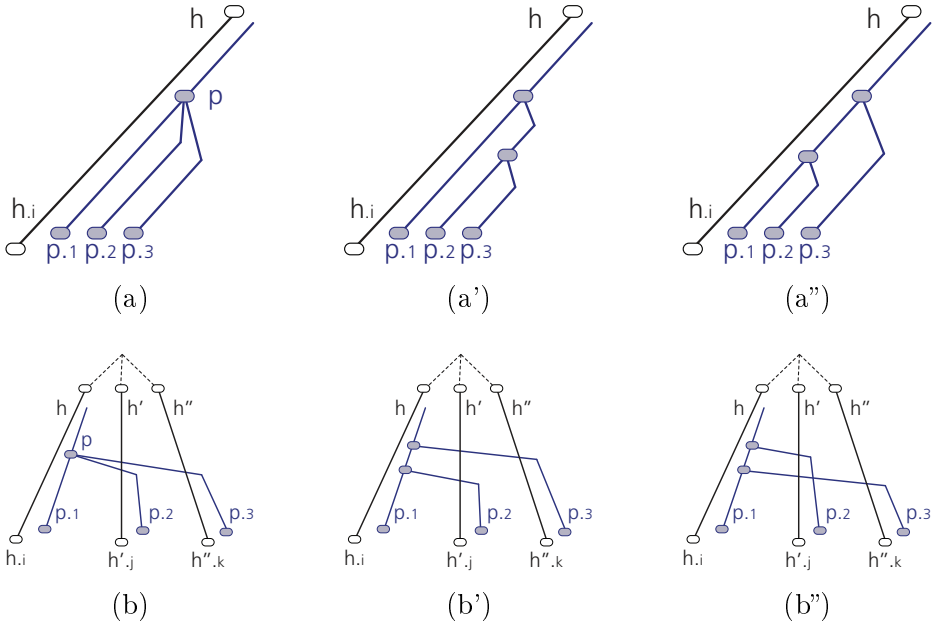


Abbildung 3.3: Die Abbildungen (a') und (a") bzw. (b') und (b") zeigen die beiden sich ergebenden binären Varianten der in Abbildung (a) bzw. (b) dargestellten mehrfach verzweigenden koevolutionären Ereignisse.

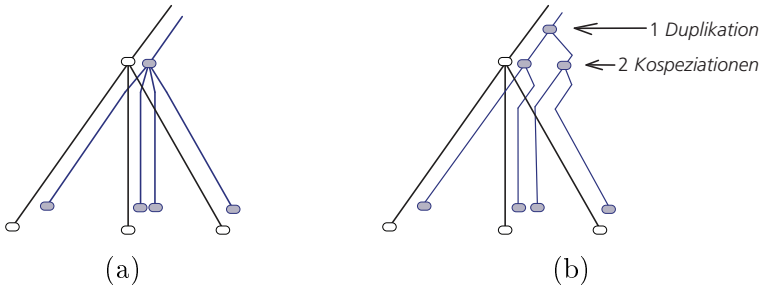
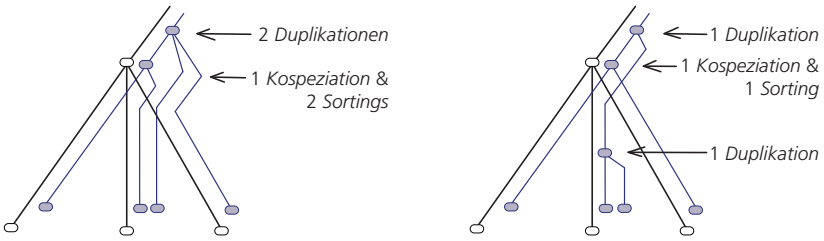


Abbildung 3.4: In (a) wird die Abbildung eines vierfach verzweigenden Parasiten auf einen dreifach verzweigenden Wirt schematisch dargestellt. Abbildung (b) zeigt die sich daraus ergebende binäre Variante mit zwei Kospeziationen.

Es wird deutlich, dass im Falle mehrerer Kospeziationen immer auch Duplikationen nötig sind. Der Parasit muss vor seinen Kospeziationen so oft durch Duplikationen verzweigen bis genügend unabhängige Knoten für diese Kospeziationen vorhanden sind.

Wie das Beispiel aus Abbildung 3.5 zeigt ist es ebenfalls möglich, dass zum Zeitpunkt der Speziation des Wirtes direkte Sortings auftreten können (a), und dass Duplikationen auch unterhalb, also zeitlich nach dieser Speziation, möglich sind (b).



(a) (b)

Abbildung 3.5: (a) zeigt eine Rekonstruktion der Ausgansdaten aus 3.4, welche direkte Sortings zum Zeitpunkt der Speziation des Wirtes verwendet. In (b) ist eine Rekonstrukton mit nach unten verlagerter Duplikation dargestellt.

Die eben genannten Beispiele zeigen auf, dass es im Falle von Multifurkationen im Parasitenbaum deutlich mehr mögliche Rekonstruktionen geben kann als dies bei Binärbäumen der Fall ist. Um diese Kombinationen zu strukturieren, soll im Folgenden bei der mehrfach verzweigenden Speziation eines Parasiten die Rekonstruktion immer als eine Abfolge koevolutionärer Ereignissen betrachtet werden. Eventuelle Wirtswechsel werden dabei immer oberhalb aller anderen Ereignisse dargestellt. Nach den Wirtswechseln folgt eine Gruppe von Duplikationen, welche zeitlich vor der Speziation des Wirtes stattfinden, gefolgt von Kospeziationen und direkten Sortings. Abschließend kommt noch eine weitere Gruppe von Duplikationen unterhalb der Speziation des Wirtes hinzu.

Jedoch muss dabei mindestens ein koevolutionäres Ereignis wärend der Lebensspanne der Wirtes auftreten. Dieses zeitlich vor oder während der Speziation stattfindende Ereignis wird gefordert, da nach Definition ein auf einen Wirtsknoten abgebildeter Parasit vor seiner Speziation zuletzt auf diesem Wirt heimisch war. Es ist folglich nicht erlaubt eine Rekostruktion aus ausschließlich unterhalb des Wirtsknotens stattfindenden Duplikationen zu erzeugen.

Mehrfach verzweigende Duplikationen und Wirtswechsel werden wie in Abbildung 3.3 (a) und (b) auch als solche dargestellt. Auf eine konkrete binäre Rekonstruktion soll verzichtet werden, da die Kosten der verschiedenen binären Varianten dieser Ereignisse wie oben erwähnt gleich sind und somit keinen Einfluss auf die Gesamtkosten der Abbildung haben. Auch die Konvention alle Wirstwechsel vor den Duplikationen zu modellieren ist eine Technik zur Reduktion der Anzahl möglicher Rekonstruktionen. Nach Definition von partiellen Wirtswechseln wird nur gefordert, dass diese vor der Speziation des Wirtes stattfinden - potentiell also auch innerhalb bzw. nach der ersten Gruppe von Duplikationen. Jedoch hat die genaue Position dieser Wirtswechsel keinen Einfluss auf die Kosten und kann deshalb vernachlässigt werden.

3.3.2 Multifurkationen durch nicht eindeutige Abbildungen $\varphi_{P,H}$

Wie in Kapitel 2.2 festgelegt wurde, muss die Abbildung $\varphi_{P,H}$ eindeutig sein. Das heisst, jedem Blattknoten aus P darf höchstens ein Blattknoten aus H zugewiesen sein. Um diese Beschränkung jedoch umgehen zu können werden Pseudoknoten $p.i$ eingeführt. Existieren mehrere $\varphi(p, h_i)$, so ist dabei jedes der neu erzeugten $p.i$ ein Kindknoten von p. Die Abbildung $\varphi_{P,H}$ wird dann so abgeändert, dass jedes der $p.i$ auf genau eines der h_i abgebildet wird. Da es sich bereits bei p um einen Blattknoten handelte - und nicht um eine Speziation - dürfen bei einer Rekonstruktion an diesem Knoten keine Kosten für Kospeziationen oder Duplikationen berechnet werden. Sortings und Wirtswechsel werden wie gehabt weiter in Rechnung gestellt.

3.4 Chronologische Inkompatibilitäten

Wie in Abschnitt 2.7 verdeutlicht, ist ein einzelner Wirtswechsel gültig, wenn sich die Zeitintervalle von take-off und landing site überlappen. Es können jedoch durch die Kombination mehrerer unabhängiger Wirtswechsel von verschiedenen Parasiten zeitliche Abhängigkeiten entstehen, welche gemeinsam zu Inkompatibilitäten in der Chronologie der Rekonstruktion führen.

Einfachstes Beispiel dafür sind zwei Wirtswechsel von Parasiten p_1 und p_2, wobei der Parasit p_1 von einem Wirt $h_1.i$ auf einen Wirt h_2 wechselt, wohingegen p_2 von $h_2.j$ auf h_1 springt. Dieses Szenario ist in Abbildung 3.6 grafisch dargestellt. Für die jeweiligen Wirte kann es Zeitzonen geben, so dass die Wirtswechsel einzeln betracht gültig sind. Es könnte z.B. allen vier Wirten ein und die selbe Zeitzone zugewiesen sein. Duch die Definition der Zeitzonen muss nur $T(h_1) \leq_T T(h_1.i)$ und $T(h_2) \leq_T T(h_2.j)$ gelten. Welcher genaue Zeitpunkt innerhalb dieser Zeitzone gemeint ist wird durch die Definition nicht näher spezifiziert. Durch den Wirtswechsel von p_1 wird aber gefordert, dass der Sprung von $h_1.i$ nach h_2 zeitlich nach h_1 stattfinden muss. Genauso muss aber auch für p_2 der Sprung von $h_2.j$ nach h_1 zeitlich nach h_2 stattfinden. Dadurch entstehen zirkuläre Abhängigkeiten zwischen den exakten Zeitpunkten innerhalb der jeweiligen Zeitzonen. Es müssten sowohl $t(h_1) < t(h2)$ als auch $t(h_2) < t(h_1)$ gelten, was zum Widerspruch führt.

Im folgenden Abschnitt sollen diese Inkompatibilitäten kategorisiert werden. Dazu werden - ähnlich des Ansatzes aus [12][4] - Abhängigkeiten betrachtet, welche durch Kombinationen mehrerer Wirtswechsel entstehen. Im Gegensatz zu Legat sollen hier

[4] In diesem wird aus dem Wirtsstammbaum und den durch Wirtswechsel erzeugten Abhängigkeiten zwischen den Wirtsknoten ein gerichteter Multigraph erzeugt und auf kreisfreiheit geprüft.

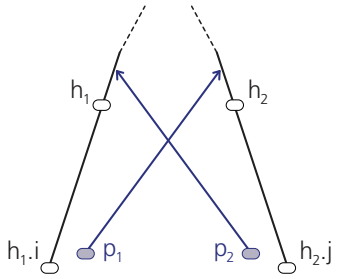


Abbildung 3.6: Zwei sich gegenseitig zeitlich ausschließende Wirtswechsel. Der Sprung von p_1 muss vor dem von p_2 stattfinden und umgekehrt.

jedoch Inkompatibilitäten nur gefunden und nicht aufgelöst werden.[5] Deshalb soll das Problem so umformuliert werden, dass nur Vergleiche zweier parasitärer Lebenslinien betrachtet werden müssen, um zu entscheiden ob eine Inkompatibilität vorliegt.

Zusätzlich sollen auch durch kaskadierende Wirtswechsel entstehende Abhängigkeiten untersucht werden. Diese treten auf, wenn während der Lebensline eines Parasiten mehrere nicht notwendigerweise nacheinander folgende Wirtswechsel stattfinden.

3.4.1 Einfache Inkompatibilitäten

Um aufzuzeigen, wann solche Inkompatibilitäten durch Wirtswechsel auftreten können, werden die Kombinationsmöglichkeiten zweier parasitärer Lebenslinien in Bezug auf durchgeführte Wirtswechsel zwischen zwei Wirten in den Abbildungen 3.7 und 3.8 grafisch dargestellt. Einerseits wird der Fall betrachtet, dass es sich bei der parasitären Lebenslinie um einen einzelnen Wirtswechsel handelt, andererseits können es aber auch mehrere nacheinander stattfindende Wirstwechsel sein. Bei diesen werden allerdings die Zwischenstationen auf denen sich der Parasit eventuell aufhielt gesondert betrachtet. Das Augenmerk liegt ausschließlich auf den Abhängigkeiten der Zeitpunkte von take-off site auf dem einen Wirt und landing site auf dem anderen - ganz gleich wieviele evolutionäre Ereignisse zwischendurch aufgetreten sein mögen. Bei einem einzelnen Wirtswechsel müssen take-off und landing site zum gleichen Zeitpunkt stattfinden. Bei mehreren Wirtswechseln wird gefordert, dass die take-off site zeitlich vor der landing site liegt.

Die Kombinationen lassen sich wie folgt gruppieren:

1. Die Wirtswechsel von p_1 finden zeitlich vor denen von p_2 statt (Abbildung 3.7 (a) bis (h)).

[5] Dazu wird in [12] ein Verfahren beschrieben durch welches sich durch Zurückziehen der landing site einige inkompatible Wirtswechsel auflösen lassen. Dabei erhöhen sich allerdings die Gesammtkosten der Rekonstruktion durch hinzugefügte Sortings.

2. Die Wirtswechsel von p_1 und p_2 finden zeitlich gleichläufig zueinander statt, d.h. die betrachtete Lebenslinie von p_1 landet auf einem Nachfolger der take-off site von p_2 und umgekehrt (Abbildung 3.7 (i) und (j) sowie 3.8 (m)).

3. Die Wirtswechsel von p_2 finden innerhalb der Zeitspanne der Lebenslinie von p_1 statt, d.h. im Intervall von der take-off site von p_1 bis zur landing site der betrachteten Lebenslinie von p_1 (Abbildung 3.7 (k) und (l), sowie 3.8 (n) und (o)).

4. Die Wirtswechsel von p_1 und p_2 finden zeitlich gegenläufig zueinander statt, d.h. die betrachtete Lebenslinie von p_1 landet auf einem Vorgänger der take-off site von p_2 und umgekehrt (Abbildung 3.8 (p) bis (r)).

Als chronologie-erhaltend werden die Kombinationen bezeichnet, für die eine gültige zeitliche Anordnung der Ereignisse existiert. Chronologie-verletzend sind jene, für die es keine solche Anordnung gibt.

In Abbildung 3.7 sind die chronologie-erhaltenden, in 3.8 die chronologie-verletzenden Kombinationen aufgeführt. Dabei wird ein einzelner Wirtswechsel durch eine durchgezogene Linie dargestellt. Mehrere Wirtswechsel im Verlauf einer Lebenslinie werden durch eine gestrichelte Linie repräsentiert.

Für die nachfolgenden Erläuterungen verwenden wir die Schreibweisen $t(tos(p))$ für den Zeitpunkt der take-off site beim Wirtswechsel von p und $t(ls(p.x))$ für den Zeitpunkt der landing site beim Sprung vor den Parasiten $p.x$. Das x deutet an, dass es sich bei $p.x$ sowohl um ein direktes Kind von p, als auch um einen weiter entfernten Nachfolger handeln kann.

Grundsätzlich wird angenommen, dass für einen einzelnen Wirtswechsel take-off und landing zeitgleich ablaufen. Es muss also $t(tos(p)) = t(ls(p.i))$ gelten. Für eine Abfolge von Wirtswechseln, die während der Lebenslinie von p nach $p.v$ stattfanden, gilt hingegen $t(tos(p)) > t(ls(p.v))$ gelten. Dies ist der Fall, da während der Entwicklung von p zu $p.v$ mehrere Speziationen stattfanden, wobei die Zeitpunkte $t(p) < t(p.i) < \ldots < t(p.v)$ stetig voran schritten.

Bei den Abbildungen 3.7 (a) bis (h) ist es offensichtlich, dass eine chronilogie-erhaltende Anordnung existiert, denn es gilt immer $t(ls(p_1.x)) < t(tos(p_2))$. Für diese Abbildungen können somit die Zeitpunkte der Speziationen in der Form $t(tos(p_1)) \leq t(ls(p_1.x)) < t(tos(p_2)) \leq t(ls(p_2.x))$ angeordnet werden.

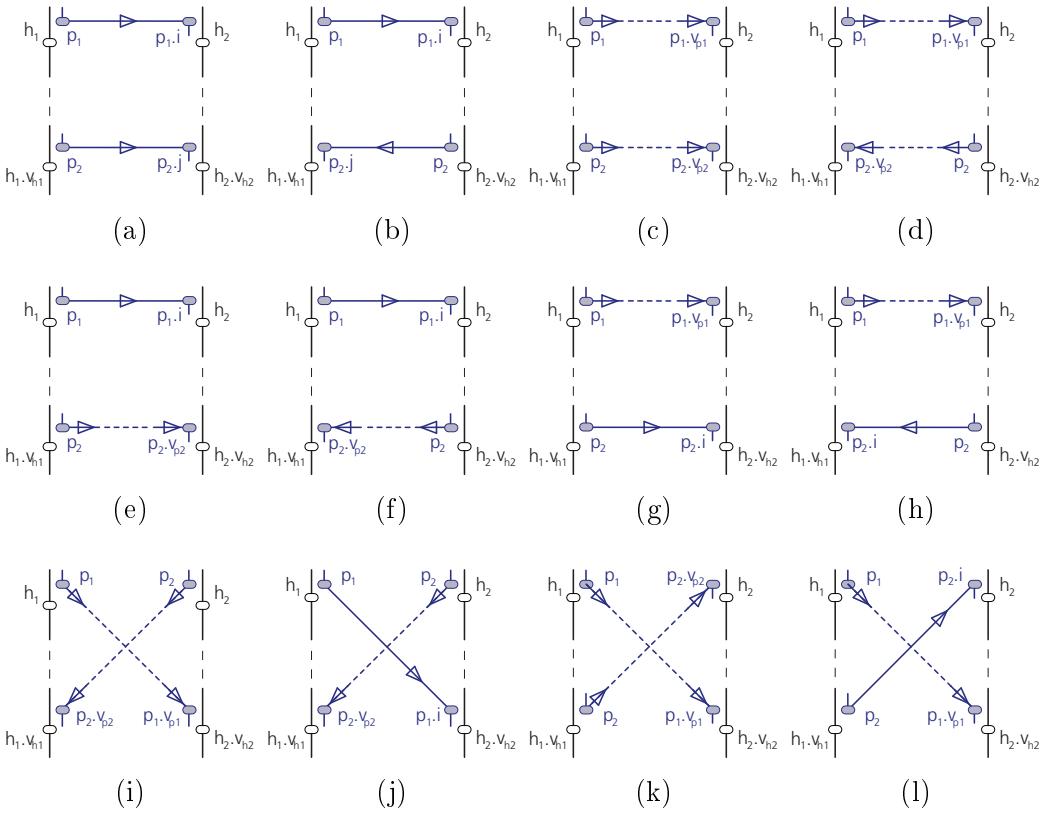


Abbildung 3.7: Chronologie-erhaltende Wirtswechselkombinationen

Für die Abbildungen (i) und (j) existiert eine Anordnung $t(tos(p_2) < t(tos(p_1)) \leq t(ls(p_1.x)) < t(ls(p_2.v_{p1}))$. Dies ist der Fall, da bei p_2 mehr als ein Wirtswechsel stattfindet und somit $t(tos(p_2)) > t(ls(p_2.v_{p2}))$ gilt.

Anders liegt der Fall in Abbildung 3.8 (m). Es muss dabei muss $t(tos(p_1)) < t(h_1) < t(ls(p_2.j)) < t(h_1.v_{h1})$ und $t(tos(p_2)) < t(h_2) < t(ls(p_1.i)) < t(h_2.v_{h2})$ gelten. Mit $t(tos(p_1)) = t(ls(p_1.i))$ und $t(tos(p_2)) = t(ls(p_2.j))$ ergibt sich $t(tos(p_1)) < t(tos(p_2))$ und $t(tos(p_2)) < t(tos(p_1))$, was zum Widerspruch führt.

In den Varianten 3.7 (k) und (l) existiert eine chonologie-erhaltende Anordnung mit $t(tos(p_1)) < t(tos(p_2) \leq t(ls(p_2.y)) < t(ls(p_1.v_{p1}))$. Die Abfolge von Wirtswechseln der Lebensline von p_1 umschließt somit die der von p_2 zeitlich. Dies ist möglich, da $t(tos(p_1)) < t(ls(p_1.i))$ gilt.

Da bei 3.8 (n) und (o) $t(tos(p_1)) = t(ls(p_1.i))$ gilt, ist eine Anordnung wie in 3.7 (k) und (l) nicht möglich. Vielmehr muss gelten: $t(tos(p_1)) < t(h_1) < t(tos(p_2)) < t(h_1.v_{h1})$

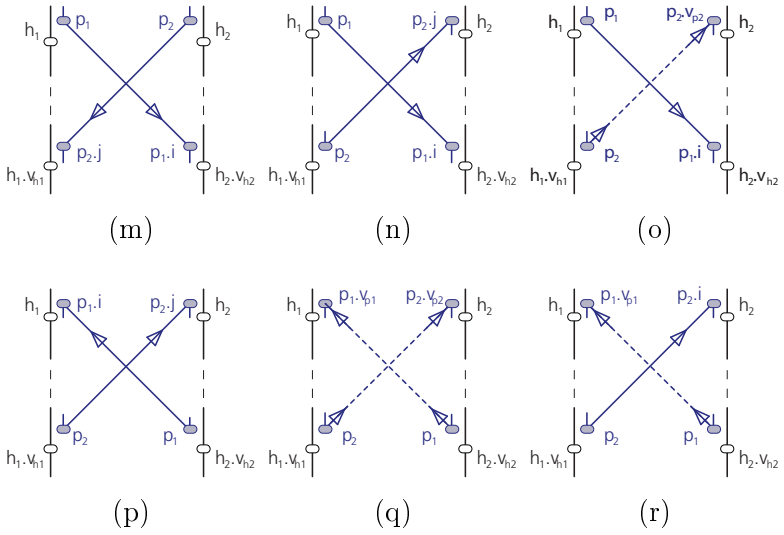


Abbildung 3.8: Chronologie-verletzende Wirtswechselkombinationen

und $t(ls(p_2.y)) < t(h_2) < t(ls(p_2.i)) < t(h_2.v_{h2})$, was mit der obigen Bedingung zum Widerspruch $t(tos(p_1)) < t(tos(p_2)) < t(tos(p_1))$ führt.

Die letzten drei Varianten 3.8 (p), (q) und (r) sind alle chronologie-verletzend, da bei diesen die widersprüchlichen Ungleichungen $t(ls(p_1.x)) < t(tos(p_2))$ und $t(ls(p_2.y)) < t(tos(p_1))$, sowie $t(tos(p_1)) \leq t(ls(p_1.x))$ und $t(tos(p_2)) \leq t(ls(p_2.y))$ erfüllt sein müssten.

3.4.2 Kaskadierende Inkompatibilitäten

Die oben genannten Varianten geben wie bereits erwähnt nur die Interaktionsmöglichkeiten zweier parasitärer Lebenslinien mit zwei Wirtslebenslinien wieder. Es können aber weitere Inkompatibilitäten auftreten, wenn mehr als zwei Parasit- und Wirtslebenslinien betrachtet werden. Diese Fälle lassen sich jedoch auf einfache Inkompatibilitäten reduzieren.

Kaskadierung zwischen zwei Wirtslebenslinien

Für den Fall, dass mehr als zwei Parasiten zwischen genau zwei Wirtslebenslinien wechseln, soll formal bewiesen werden, dass sich dabei auftretende kaskadierende Inkompatibilitäten auf den einfachen Fall reduzieren lassen.

Lemma 3.1. *Eine chronologische Inkompatibiliät zwischen mehr als zwei parasitären Lebenslinien, beim Sprung zwischen zwei Wirtslebenslinien tritt genau dann auf, wenn mindestens eine Inkompatibilität zwischen zwei der Parasitenlinien existiert.*

Beweis Lemma 3.1.

1. $\Longleftarrow$

Dieser Beweis ist trivial. Wenn eine Inkompatibilität zwischen zwei beliebigen Parasitenlebenslinien existiert, so ist nach Definition auch die komplette Rekonstruktion inkompatibel.

2. $\Longrightarrow$

Dieser Beweis wird indirekt geführt. Er zeigt, dass sich immer mindestens ein inkompatibles Paar finden lässt, wenn keine zeitliche Annordnung aller Zeitpunkte existiert.

Gegeben seien zwei Wirts- und n Parasitenlebenslinien, wobei jeder der Parasiten innerhalb der jeweiligen Zeitspanne durch einen oder mehrere Wirtsechsel von der einen Wirtslinie zur anderen springt. Für die zeitliche Anordnung der take-off und landing sites lassen sich folgende zwei Bedingungen formulieren, je eine pro Wirtslebenslinie: $t(a_1) < t(a_2) < \ldots < t(a_n)$ sowie $t(b_1) < t(b_2) < \ldots < t(b_n)$ mit $a_1, \ldots, a_n, b_1, \ldots, b_n \in tos(P) \bigcup ls(P)$ und $tos(p) \in \{a_1, \ldots, a_n\} \leftrightarrow ls(p.x) \in \{b_1, \ldots, b_n\}$. Dabei werden die Ereignisse $A = \{a_i\}$ der ersten und die Ereignisse $B = \{b_i\}$ der zweiten Wirtslebenslinie zugeordnet. Für jede parasitäre Lebenslinie gilt zusätzlich $t(tos(p)) = t(ls(p.i))$ für einen einzelnen Wirtswechsel bzw. $t(tos(p)) < t(ls(p.v))$ für eine Reihe von Wirtswechseln. Die ersten beiden Bedingungen folgen aus der gültigen Chronologie des Wirtsbaumes, die Letztere aus der des Parasitenbaumes. Da diese Kombination von Wirtswechseln als chronologie-verletzend vorausgesetzt wird, existiert keine zeitliche Anordnung der a_i und b_i, welche alle Bedingungen erfüllt. Es müssen also mindestens zwei Punkte auf einer der beiden Wirtsseiten existieren, für die jede der möglichen Anordnung zu einem Widerspruch führt. Ohne Beschränkung der Allgemeinheit seien dies a_x und a_y mit $a_x < a_y$. Seien weiterhin die dazugehörigen take-off bzw. landing sites $b_{x'}$ und $b_{y'}$. Dann lassen sich für diese beiden Punkte folgende mögliche Bedingungen aus der Chronologie der parasitären Lebenslinien ableiten:

1. $t(a_x) < t(b_{x'}) \wedge t(a_y) < t(b_{y'})$,	2. $t(a_x) < t(b_{x'}) \wedge t(a_y) = t(b_{y'})$,	3. $t(a_x) < t(b_{x'}) \wedge t(a_y) > t(b_{y'})$,
4. $t(a_x) = t(b_{x'}) \wedge t(a_y) < t(b_{y'})$,	5. $t(a_x) = t(b_{x'}) \wedge t(a_y) = t(b_{y'})$,	6. $t(a_x) = t(b_{x'}) \wedge t(a_y) > t(b_{y'})$,
7. $t(a_x) > t(b_{x'}) \wedge t(a_y) < t(b_{y'})$,	8. $t(a_x) > t(b_{x'}) \wedge t(a_y) = t(b_{y'})$,	9. $t(a_x) > t(b_{x'}) \wedge t(a_y) > t(b_{y'})$,

1. Fall: Sei $b_{x'} < b_{y'}$, dann existieren die möglichen Anordnungen

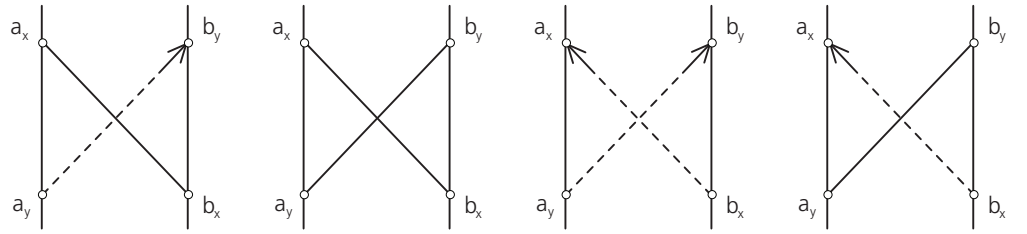
$t(a_x) < t(a_y) \leq t(b_{x'}) < t(b_{y'})$ (erfüllt 1),	$t(a_x) \leq t(b_{x'}) < t(a_y) \leq t(b_{y'})$ (erfüllt 1, 2, 4 und 5),
$t(a_x) < t(b_{x'}) \leq t(a_y) < t(b_{y'})$ (erfüllt 1),	$t(a_x) \leq t(b_{x'}) < t(b_{y'}) \leq t(a_y)$ (erfüllt 2, 3, 5 und 6),
$t(b_{x'}) < t(b_{y'}) \leq t(a_x) < t(a_y)$ (erfüllt 9),	$t(b_{x'}) \leq t(a_x) < t(b_{y'}) \leq t(a_y)$ (erfüllt 5, 6, 8 und 9),
$t(b_{x'}) < t(a_x) \leq t(b_{y'}) < t(a_y)$ (erfüllt 9),	$t(b_{x'}) \leq t(a_x) < t(a_y) \leq t(b_{y'})$ (erfüllt 4, 5, 7 und 8),

Da jede der neun Bedingungen von mindestens einer dieser Anordnungen erfüllt wird, muss die Annahme $b_{x'} < b_{y'}$ falsch gewesen sein.

2. Fall: Sei $b_{y'} < b_{x'}$, dann existieren die möglichen Anordnungen

1. $t(a_x) < t(a_y) \leq t(b_{y'}) < t(b_{x'})$ (erfüllt 1 und 2),
2. $t(a_x) \leq t(b_{y'}) < t(a_y) \leq t(b_{x'})$ (erfüllt 3),
3. $t(a_x) < t(b_{y'}) \leq t(a_y) < t(b_{x'})$ (erfüllt 2 und 3),
4. $t(a_x) \leq t(b_{y'}) < t(b_{x'}) \leq t(a_y)$ (erfüllt 3),
5. $t(b_{y'}) < t(b_{x'}) \leq t(a_x) < t(a_y)$ (erfüllt 6 und 9),
6. $t(b_{y'}) \leq t(a_x) < t(b_{x'}) \leq t(a_y)$ (erfüllt 3),
7. $t(b_{y'}) < t(a_x) \leq t(b_{x'}) < t(a_y)$ (erfüllt 3 und 6),
8. $t(b_{y'}) \leq t(a_x) < t(a_y) \leq t(b_{x'})$ (erfüllt 3),

In diesem Fall lassen sich für die Bedingungen 4, 5, 7 und 8 keine widerspruchsfreien Anordnungen der a_i und b_i finden. Daraus ergeben sich die in Abbildung 3.9 dargestellten vier Situationen.



(a) $t(a_x) \leq t(b_{y'}) < t(b_{x'}) \leq t(a_y)$ (b) $t(b_{y'}) < t(b_{x'}) \leq t(a_x) < t(a_y)$ (c) $t(b_{y'}) < t(a_x) \leq t(b_{x'}) < t(a_y)$ (d) $t(b_{y'}) \leq t(a_x) < t(a_y) \leq t(b_{x'})$

Abbildung 3.9: Die vier Situationen, für welche keine zeitliche Anordnung der Punkte existiert, die alle geforderten Bedingungen erfüllen.

In dieser Grafik sind die durchgezogenen Linien ungerichtet, da lediglich $t(a_x) = t(a_{x'})$ bzw. $t(b_y) = t(b_{y'})$ gefordert wird. Welche Seite take-off und welche landing ist, ist für die Chronologie unerheblich. Beide Varianten führen immer zu einer Inkompatibilität. Die in Abbildung 3.9 (a) und (d) aufgeführten Fälle sind equivalent zu denen der Abbildungen 3.8 (o) und (r). 3.9 (b) stellt die Fälle 3.8 (m), (n) und (p) dar und Abbildung 3.9 (c) spiegelt den Fall 3.8 (q) wieder. Jede diese vier Varianten entspricht somit einer Inkompatibilität zwischen den zwei betrachteten Parasitenlebenslinien. □

Kaskadierung zwischen mehr als zwei Wirtslebenslinien

In der vorangegangenen Betrachtung wurde gefordert, dass n parasitäre Lebenslinien während ihrer Evolution zwischen genau zwei Wirtslebenslinien Wechsel durchführen. Für diese ergaben sich die Abhängigkeiten $t(a_i) < t(a_{i+1})$ aus der Chronologie des Wirtsbaumes, und die $t(tos(p)) = t(ls(p.i))$ bzw. $t(tos(p)) \leq t(ls(p.v))$ aus der des Parasitenbaumes. Zieht man mehrere Wirtslebenslinien in Betracht, so können zusätzliche Bedingungen entstehen. Springt beispielsweise ein Parasit p_1 von einem Wirt h_1 zum Zeitpunkt $t(a)$ zu einem Wirt h_2 und landet dort zum Zeitpunkt $t(b_1)$. Des Weiteren springt ein zweiter Parasit p_2 von einem Wirt $h_2.v$ zum Zeitpunkt $t(b_2)$ zu einem Wirt

h_3 und landet zum Zeitpunkt $t(c)$, so muss wegen $t(b_1) < t(b_2)$ auch $t(a) < t(c)$ gelten. Abbildung 3.10 verdeutlicht dies.

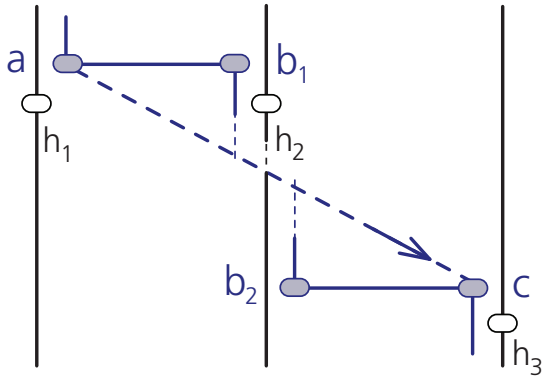


Abbildung 3.10: Das Schema zeigt zwei Wirtswechsel, welche durch ihre zeitliche Abfolge eine zusätzliche Bedingung für eine chronologisch gültige Rekonstruktion fordern.

Es wird deutlich, dass diese Bedingung auch durch eine Reihe von Wirtswechseln eines imaginären Parasiten p_3 erzeugt werden könnte, wenn dieser zum Zeitpunkt $t(a)$ von h_1 nach h_3 springen würde, und dort zum Zeitpunkt $t(c)$ landet.

Auf diese Weise können für alle Kombinationen von Sprüngen zweier Parasiten zwischen mehr als zwei Wirtslebenslinien Wirtswechsel imaginärer Parasiten eingeführt werden. In Abbildung 3.11 (a) bis (f) sind diese Kombinationen und die zugehörigen imaginären Sprünge dargestellt.

Für die Abbildungen 3.11 (a) bis (d) sind die Sprünge imaginärer Parasiten mit eingezeichnet. Bei (e) bis (i) existieren keine zusätzlichen Bedingungen, denn man kann chronologie-erhaltende Beispiele für jede der beiden zeitlichen Anordnungen $t(a) < t(c)$ und $t(c) < t(a)$ konstruieren. Somit müssen auch keine Abhängigkeiten durch Einfügen neuer Parasiten simuliert werden.

Wie im vorangegangenen Abschnitt gezeigt wurde, können durch die Einführung imaginärer Parasiten jene Bedingungen simuliert werden, welche durch die in Abbildung 3.11 dargestellten kaskadierenden Sprünge entstehen. Eine Prüfung auf chronologische Konsistenz kann somit für jede Kombination von zwei der Wirtslinien einzeln durchgeführt werden, ohne die restlichen Lebenslinien mitbetrachten zu müssen. Findet sich auf diese Weise keine Inkompatibilität, so ist nach Lemma 3.1 das komplette System valide.

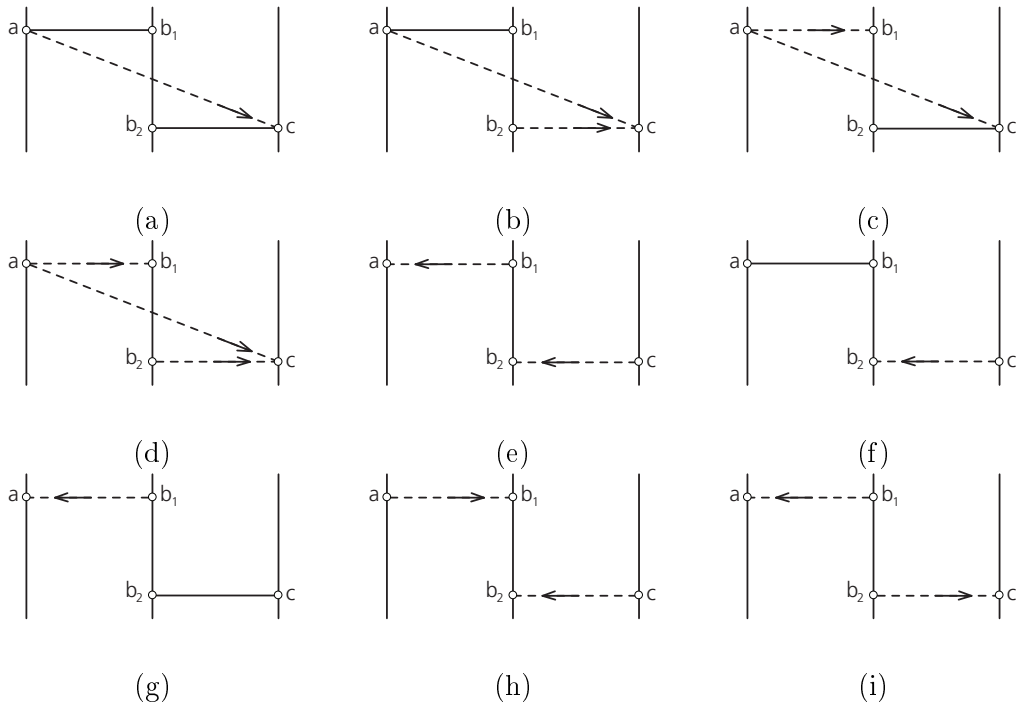


Abbildung 3.11: Die neun möglichen Kombinationen von Sprüngen zweier Parasiten zwischen drei Wirtslebenslinien.

4 Algorithmische Umsetzung

4.1 Berechnung der günstigsten Teilrekonstruktionen

Wie bereits erwähnt, reicht es für die Berechnung einer kostenminimalen Rekonstruktion aus die minimalen Ereigniskosten zu berechnen und sich für jeden Knoten des Parasitenbaumes sowohl die Abbildung seiner Kindknoten in den Wirtsbaum, als auch die aufgetretenen koevolutionären Ereignisse zu merken.

Zur Umsetzung des dynamischen Ansatzes wird - in der Reihenfolge von den Blättern zur Wurzel - jeder Knoten im Parasitenbaum auf jeden Knoten im Wirtsbaum abgebildet. In einem dieser Schritte werden wiederum alle Kombinationen von möglichen Abbildungen der Kindknoten des Parasiten betrachtet. Für jede Kombination werden die für eine Rekonstruktion notwendigen koevolutionären Ereignisse berechnet und unter allen Varianten wird die kostengünstigste ausgewählt und abgespeichert.

Der Algorithmus arbeitet somit in folgenden 4 verschachtelten Schleifen.

1. Alle Parasiten werden durchlaufen (n).
2. Für jeden Parasit werden alle Wirte durchlaufen (m).
3. Für jedes Parasit-Wirt-Paar werden alle Kombinationen von Parasit-Wirt-Paaren der Kindknoten durchlaufen ($m^{grad(p)}$).
4. Für jede dieser Kombinationen werden die Kosten der günstigsten Teilrekonstruktion berechnet.

Daraus ergibt sich folgender Pseudocode:

Algorithm 1 Berechnung der kostenminimalen Rekonstruktionskosten

1: Initialisieren des zweidimensionalen Arrays $C[n, m]$ mit ∞
2: **for all** Parasitenknoten p (bottom-up) **do**
3: **for all** Wirtsknoten h **do**
4: **if** p ist ein Blattknoten und es existiert eine Abbildung $\varphi(p, h)$ **then**
5: $C[p, h] = 0$;
6: **else if** p ist kein Blattknoten **then**
7: **for all** Kombinationen von Wirtsknoten h_i für jeden Parasitenkindknoten $p.i$ von p **do**
8: $C[p, h] = \min\left(C[p, h], \left(\sum_{i=1}^{grad(p)} C[p.i, h_i]\right) + \text{Kosten } E \text{ der Ereignisse am Knoten } h \text{ und zusätzlicher Sortings der } p.i \text{ bis zu den } h_i + Z\right)$;
9: **end for**
10: **end if**
11: **end for**
12: **end for**

4.2 Berechnung der günstigsten Ereigniskosten E einer Abbildung

Ausgangssituation ist immer die Annahme, dass ein Parasit p auf einen Wirt h und die Kindknoten $p.i$ auf die Wirte h_i abgebildet werden. Anhand der relativen Positionen der Wirte h_i zum Knoten h, wird geprüft ob Kospeziationen, Duplikationen oder Wirtswechsel stattgefunden haben können. Zuerst muss jedoch sichergestellt werden, dass die Zeitzonenintervalle der p und $p.i$ mit den Zeitzonen der Wirte übereinstimmen. Es werden $Z(p, h)$ sowie alle $Z(p.i, h_i)$ berechnet. Ist eines davon ∞, so sind auch die Kosten der gesamten Teilrekonstruktion ∞. Ebenso verhält es sich, wenn h_i Vorgänger von h, also $h_i <_B h$ ist. In einem solchen Fall wäre eine Rekonstruktion chronologie-verletzend. Für alle anderen Fälle werden die Kosten der jeweilig möglichen koevolutionären Ereignisse berechnet. Da nur partielle Wirtswechsel in Betracht gezogen werden, muss immer wenigstens eines der $h_i \geq_B h$ sein. Ist dem nicht so, werden ebenfalls Kosten von ∞ angenommen.

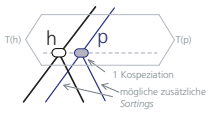
4.2.1 Verfahren bei binären Verzweigungen in den Stammbäumen

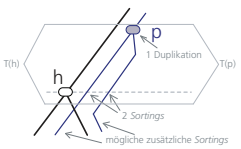
Binäre Verzweigungen sind ein Spezialfall der Multifurkationen. Um aufzuzeigen, dass sämtliche binären Rekonstruktionsmöglichkeiten betrachtet werden, sollen diese einführend erörtert werden. Im Speziellen wird damit gezeigt, dass Lösungen anderer in der

Forschung verwendeter Rekonstruktionsverfahren durch den hier vorgestellten Ansatz ebenfalls berücksichtigt werden. Als Referenz soll dabei das in [14] vorgestellte Programm *Tarzan* ([11]) dienen.

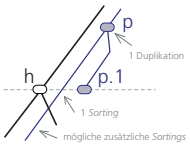
Bei binären Stammbäumen entstehen bei der Speziation eines Parasiten p genau zwei neue Unterarten $p.1$ und $p.2$. Wenn die Abbildungen der Parasitenknoten auf die jeweiligen Wirtsknoten h, h_1 und h_2 unter Berücksichtigung der Zeitzonen gültig sind, muss entschieden werden, welche koevolutionären Ereignisse in dieser Situation aufgetreten sein können. Dazu werden die Positionen der Knoten h_1 und h_2 in Bezug auf h und die Zeitzoneninformationen $T(p)$ und $T(h)$ verwendet. Folgende Ausgangssituationen erzeugen dabei die nachstehenden Kosten.

1. **Kospeziation oder Duplikation:** Die beiden Knoten h_1 und h_2 liegen in verschiedenen bei den Kindern von h beginnenden Teilbäumen und für das Zeitzonenintervall $T(p)$ gilt $T(p) =_T T(h)$. Dann können entweder eine Kospeziation oder eine Duplikation mit zwei direkten Sortings stattgefunden haben. Für beide Varianten werden die Kosten berechnet. Es wird die Kostengünstigere gewählt.

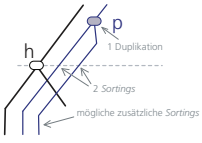




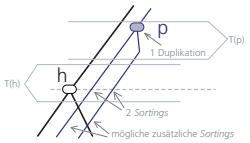
2. **Erzwungene Duplikation durch Abbildung mindestens eines Kindes auf h:** Die beiden Knoten h_1 und h_2 liegen im bei h beginnenden Teilbaum und mindestens einer der Knoten ist gleich dem Knoten h. Somit muss die Speziation von p vor der von h stattgefunden haben und es kann nur eine Duplikation aufgetreten sein. Es entstehen die Kosten für eine Duplikation und ein direktes Sorting.



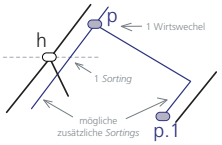
3. **Erzwungene Duplikation durch Abbildung beider Kinder in den gleichen bei h beginnenden Zweig:** Die beiden Knoten h_1 und h_2 liegen im gleichen, bei einem der Kinder von h beginnenden Teilbaum. Es kann daher nur eine Duplikation aufgetreten sein. Für diese muss die Speziation von p vor der von h stattgefunden haben. Es entstehen die Kosten für eine Duplikation und zwei direkte Sortings.



4. **Erzwungene Duplikation durch Zeitzonen:** Die beiden Knoten h_1 und h_2 liegen im bei h beginnenden Teilbaum und für das Zeitzonenintervall $T(p)$ gilt $T(p) <_T T(h)$. Die Speziation von p muss daher vor der von h stattgefunden haben. Somit kann nur eine Duplikation aufgetreten sein. Es entstehen die Kosten für eine Duplikation und zwei direkte Sortings.



5. **Wirtswechsel:** Genau einer der Knoten h_1 oder h_2 liegt im bei h beginnenden Teilbaum. Für den anderen Knoten gilt, dass er weder Vorgänger oder Nachfolger von h, noch gleich h ist. Es tritt dabei ein Wirtswechsel auf. Für diesen muss die Speziation von p vor der von h stattgefunden haben. Es entstehen die Kosten für einen Wirtswechsel und ein direktes Sorting.



   Für alle andere Fälle existiert keine gültige Rekonstruktion. Dafür werden Kosten von ∞ angenommen.

Trotz der gleichen koevolutionären Fallunterscheidungen können Unterschiede zu den durch *Tarzan* berechneten Rekonstruktionen auftreten. Gegebenenfalls entstehen bei *Tarzan* teurere Lösungen. Dies liegt an dem in [12] beschriebenen Verfahren, bei welchem schwach inkompatible Wirtswechsel[1] durch Zurückziehen der landing site aufgelöst werden. Wenn die kostenminimale Rekonstruktion keine gültige Chronologie aufweist, wird damit durch Einfügen zusätzlicher Sortings versucht, diese Inkompatibilitäten zu beseitigen. Es können jedoch nicht immer alle inkompatiblen Wirtswechsel aufgelöst werden. Insbesondere die in Kapitel 3.4 beschriebenen kaskadierenden Wirtswechsel werden nicht berücksichtigt.

Im Falle einer chronologisch inkompatiblen Gesamtlösung erzeugt der hier vorgestellte Algorithmus gegebenenfalls günstigere, jedoch inkompatible Rekonstruktionen.

4.2.2 Verfahren bei Multifurkationen in den Stammbäumen

Wird ein mehrfach verzweigender Parasit auf einen mehrfach verzweigenden Wirt abgebildet, so kann es deutlich mehr Varianten als im binären Fall geben. In Kapitel 3.3 wurde die Konvention getroffen, eine Rekonstruktion immer als eine vierstufige Abfolge koevolutionärer Ereignissen zu betrachten. Diese vier Stufen sind: 1. Wirtswechsel, 2.

[1] vgl. [12] S.23 Def. 1.17

frühe Duplikationen, 3. Kospeziationen und direkte Sortings sowie 4. späte Duplikationen.

Wirtswechsel

Für alle aufgetretenen Wirtswechsel wurde festgelegt, dass diese vor allen anderen Ereignissen stattfinden. Es entstehen somit immer die Kosten für diese Wirtswechsel plus die Kosten der dabei eventuell zusätzlich anfallenden Sortings.

Kospeziationen und Duplikationen

In einem zweiten Schritt wird geprüft, ob Kospeziationen aufgetreten sein können. Dafür müssen zwei Bedingungen erfüllt sein. Genau wie im binären Fall darf keines der Parasitenkinder auf den Wirtsknoten h abgebildet sein und für das Zeitzonenintervall $T(p)$ muss $T(p) =_T T(h)$ gelten.

Anderenfalls müsste wie in den 3 Fällen der erzwungenen Duplikation die Parasitenspeziation vor der des Wirtes stattgefunden haben. Die Speziationen werden dabei durch zeitlich vor dem Wirtsknoten h auftretende Duplikationen koevolutionär erklärt. Es entstehen die Kosten für diese Duplikationen plus die Kosten der am Wirtsknoten auftretenden direkten Sortings für jedes beteiligte Parasitenkind.

Für den Fall, dass Kospeziationen möglich sind, muss geprüft werden mit welcher Verteilung aus Kospeziationen und Duplikationen die Kosten dieser Abbildung minimal werden. Dazu werden alle möglichen Kombinationen dieser Verteilung betrachtet und für jede die günstigsten Ereigniskosten berechnet.

Es sind jedoch nicht immer $\frac{n}{2}$ Kospeziationen möglich, denn für diese wird gefordert, dass zwei Parasitenkinder auf unterschiedliche, bei den Kindern von h beginnende Teilbäume aufgeteilt werden. Es müssen also für eine Kospeziation zwei Parasitenkinder $p.i$ und $p.j$ auf verschiedene Knoten $h.k$ und $h.l$ bzw. deren Nachfolger abgebildet sein. Die Anzahl der möglichen Kospeziationen ist von der jeweiligen Ausgangssituation abhängig. Für diese wird mit dem in Algorithmus 2 beschriebenen Verfahren eine maximal mögliche Anzahl berechnet.

Algorithm 2 Berechnung der maximal möglichen Anzahl von Kospeziationen

1: Speichere für jeden bei einem Kind von h beginnenden Teilbaum die Parasitenkinder von p, welche in diesen Teilbaum abgebildet werden;
2: Sortiere diese Liste absteigend nach der Anzahl der zugewiesenen Parasitenkinder;
3: $maxCospeciations = 0$
4: **while** Liste hat mehr als zwei Einträge **do**
5: Entferne je ein Parasitenkind aus dem ersten und zweiten Eintrag der Liste;
6: $maxCospeciations++$;
7: Sortiere die Liste neu;
8: Entferne alle leeren Einträge der Liste;
9: **end while**
10: **return** $maxCospeciations$;

Eine genauere Betrachtung der für die Implementierung dieses Algorithmus verwendeten Datenstruktur erfolgt in Kapitel 4.4.4.

Direkte und zusätzliche Sortings

Zu den Duplikations- und Kospeziationskosten kommen noch Sortingkosten hinzu. Diese entstehen, da jedes Parasitenkind $p.i$ noch zusätzliche Sortings bis zu seinem Abbild h_i durchführen muss. Ging der jeweilige Parasit $p.i$ aus einer frühen Duplikation hervor, so fällt auch ein direktes Sorting am Knoten h an. Bei einer Kospeziation oder einer späten Duplikation ist dies nicht der Fall. Es werden dann nur die zusätzlichen Sortings vom jeweiligen Kindknoten des Wirtes bis h_i erzeugt. Grundsätzlich kann man somit für negative Sortingkosten nach Teilrekonstruktionen mit ausschließlich vor der Speziation von h auftretenden Duplikationen suchen. Für positive Sortingkosten muss man dementsprechend möglichst viele der Duplikationen zeitlich nach unten verlagern.

In jedem Fall weiß man aber, dass für jedes der Parasitenkinder $p.i$ noch zusätzliche Sortings vom jeweiligen Kind des Wirtsknotens bis zum h_i hinzukommen. Es muss also nur noch die Anzahl der direkten Sortings bestimmt werden, welche am Knoten h auftreten.

Bei n Parasitenkindern, x Kospeziationen und negativen Sortingkosten sind dies $n - (2 * x)$ viele, denn wie oben erwähnt, gibt es in diesem Fall keine späten Duplikationen. Das bedeutet, dass alle Parasiten, die nicht an einer der x Kospeziation teilnahmen, aus frühen Duplikationen entstanden. Für diese fallen somit direkte Sortings am Knoten h an.

Bei positiven Sortingkosten hingegen werden möglichst späte Duplikationen verwendet. Wie die Abbildung 4.1 zeigt, sind jedoch auch da gegebenenfalls noch weitere Sortings am Knoten h nötig.

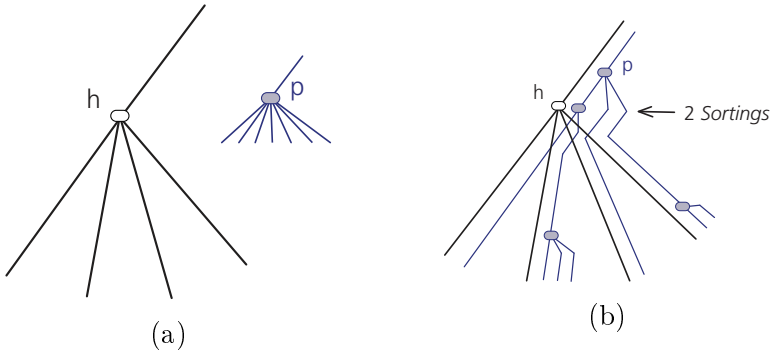


Abbildung 4.1: Ausgangsdaten (a) und Teilrekonstruktion (b) eines siebenfach verzweigenden Parasiten auf einen vierfach verzweigenden Wirt. Für die Rekonstruktion sind zwei Sortings am Knoten h nötig.

Diese direkten Sortings treten für jeden, bei den Kindern von h beginnenden Teilbaum auf, wenn in diesen eines der $p.i$ abgebildet wurde, welches nicht aus einer Kospeziation hervor ging. Um die Anzahl solcher Sortings zu minimieren, sollten Kospeziationen möglichst viele verschiedene dieser Teilbäume abdecken. Aus diesen Kospeziationen können dann, durch späte Duplikationen alle weiteren Parasitenkinder entstehen. Die minimale Anzahl solcher Sortings ist folglich abhängig von der Anzahl der Kospeziationen x und von der Anzahl der bei den Kindern von h beginnenden Teilbäume, in welche Kindknoten von p abgebildet wurden. Sei diese Anzahl y. Minimal nötig sind dann $\max\{0, y-(2*x)\}$, denn $2*x$ der Teilbäume können durch x Kospeziationen abgedeckt werden. Für alle anderen ist genau ein weiteres Sorting nötig.

4.2.3 Beispiel einer Rekonstruktion mit Multifurkationen

In den Abbildungen 4.2 und 4.3 sind die Ausgangsdaten und die sich ergebenden möglichen Rekonstruktionen eines konstruierten Beispiels dargestellt. Die sieben Rekonstruktionen geben die kostenunterschiedlichen Varianten der Abbildung wieder. Hierbei wird einerseits nach der Anzahl der Kospeziationen unterschieden und andererseits danach, ob späte Duplikationen vorhanden sind oder nicht.

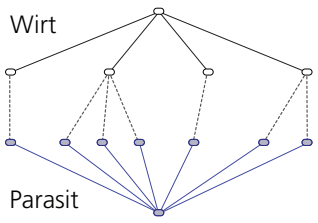
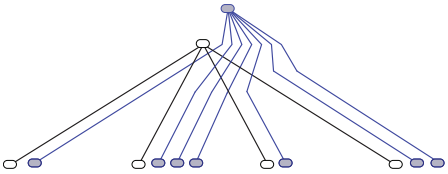
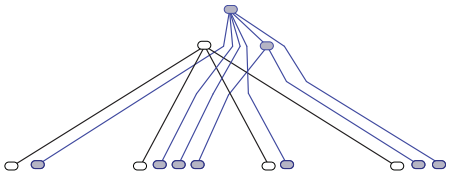


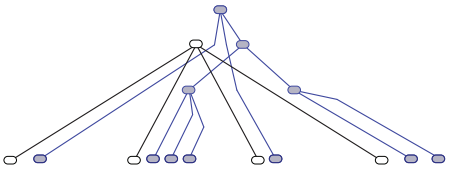
Abbildung 4.2: Ausgangsdaten für einen siebenfach verzweigenden Parasiten auf einem vierfach verzweigenden Wirt.



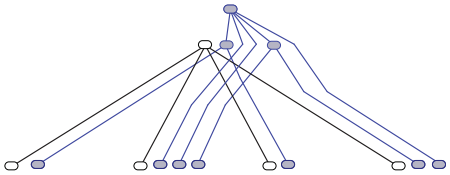
(a) 0 Kospeziationen (K), 7 Sortings (S), ohne späte Duplikationen (sD)



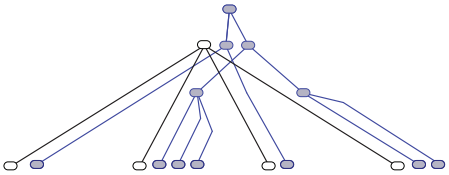
(b) 1 K, 5 S, ohne sD



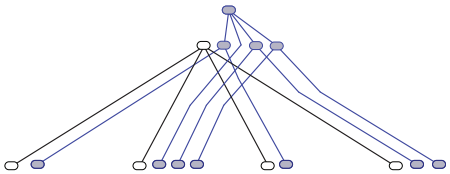
(c) 1 K, 2 S mit sD



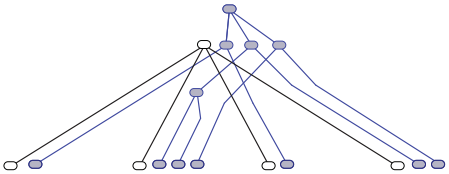
(d) 2 K, 2 S, ohne sD



(e) 2 K, 0 S mit sD



(f) 3 K, 1 S, ohne sD



(g) 3 K, 0 S, mit sD

Abbildung 4.3: Mögliche Rekonstruktionen der Ausgangsdaten aus Abbildung 4.2 mit unterschiedlicher Anzahl an Kospeziationen mit und ohne späten Duplikationen.

Das Beispiel zeigt auf, dass nicht alle möglichen Rekonstruktionen mit einer bestimmten Anzahl an Kospeziationen betrachtet werden, sondern nur zwei. Eine Version mit und eine ohne spätere Duplikationen. Um eine kostenminimale Lösung zu finden, ist es nicht nötig alle Varianten zu betrachten, denn wie oben erwähnt sind diese kosteninvariant. Da sich die mehrfach verzweigende Speziation aus den Ausgangsdaten nicht eindeutig in zeitlich geordnete binäre Speziationen aufteilen lässt, kann an dieser Stelle eine beliebige der kostengünstigsten Varianten angenommen werden.

4.3 Verfahren zur Reduktion betrachteter Parasit-Wirt-Paare

Um die günstigste Rekonstruktion des am Parasiten p beginnenden Teilbaumes mit der Abbildung von p auf h zu finden, wurden in der bisherigen Beschreibung des Algorithmuses alle Kombinationen von verschiedenen Wirten für jeden der Parasitenkindknoten $p.i$ betrachtet. Bei m Knoten im Wirtsbaum ergeben sich dabei $m^{grad(p)}$ mögliche Kombinationen. Diese Anzahl lässt sich jedoch noch stark einschränken. Dazu werden die Kosten in drei Gruppen aufgeteilt.

1. Kosten der Kospeziationen, Duplikationen und Wirtswechsel und direkten Sortings, welche zum Zeitpunkt der Speziation des Parasiten durch die Abbildung von p auf h entstanden.
2. Kosten der zusätzlichen Sortings der Parasitenkindknoten $p.i$, bis zu ihren jeweiligen Abbildern h_i.
3. Kosten der Teilrekonstruktionen $C[p.i, h_i]$ der Parasitenkinder $p.i$.

Wie im vorangegangenen Kapitel 4.2 deutlich wurde, sind die Kosten der Gruppe 1 abhängig von:

1. der Anzahl von Parasitenkindern, welche einen Wirtswechsel durchführen,
2. der Anzahl von Parasitenkindern, welche auf den Wirt h abgebildet werden, und
3. der Anzahlen von Parasitenkindern, welche auf die jeweiligen, bei den Kindern von h beginnenden Teilbäume der Wirtsphylogenie abgebildet werden.

Die genaue Wahl der h_i beeinflusst die Kosten aus Gruppe 1 nur in sofern, als dass sie die genannten Anzahlen verändert. Existieren für einen Parasit $p.i$ zwei mögliche Abbilder h_1 und h_2, für welche $p.i$ jeweils einen Wirtswechsel durchführen müsste, so bleiben diese Kosten unverändert. Genauso verhält es sich, wenn h_1 und h_2 im gleichen Teilbaum unterhalb eines Kindknotens von h liegen. Zwar wären die Kosten der Gruppen 3 und 4 eventuell verschieden, da es aber auf die restliche Berechnung keinen Einfluss hat, kann man an dieser Stelle das günstigere der beiden h_i wählen.

Dadurch muss man für die Parasit-Wirt-Paar-Kombinationen nicht alle m Wirte betrachten. Vielmehr genügt es, für jedes $p.i$ aus jedem der bei den Kindern von h beginnenden Teilbäumen einen der Wirte auszuwählen. Unter allen in Frage kommenden Wirten wird derjenige ausgewählt, für den die Summe der Kosten aus Gruppe 3 und 4 minimal ist. Hinzu kommen noch der kostengünstigste Wirtswechselkandidat und h selbst. Auf diese Weise müssen nicht mehr $m^{grad(p)}$, sondern nur noch $(grad(h)+2)^{grad(p)}$ Kombinationen betrachtet werden.

Der folgende Algorithmus 3 integriert die erläuterten Erweiterungen in den auf Seite 36 dargestellten Algorithmus 1

Algorithm 3 Berechnung die günstigsten Rekonstruktionskosten (erweitert)

1: Initialisieren des zweidimensionalen Arrays $C[n, m]$ mit ∞
2: **for all** Parasitenknoten p (bottom-up) **do**
3: **for all** Wirtsknoten h **do**
4: **if** p ist ein Blattknoten und es existiert eine Abbildung $\varphi(p, h)$ **then**
5: $C[p, h] = 0$;
6: **else if** p ist kein Blattknoten **then**
7: Initialisieren des zweidimensionalen Arrays $P[grad(p), grad(h) + 2]$ mit ∞
8: **for all** Parasitenkindknoten $p.i$ von p **do**
9: **for all** Wirtsknoten h' **do**
10: **if** $h == h'$ **then**
11: $P[p.i, 0] = C[p.i, h']$;
12: **else if** $h >_B h'$ **then**
13: berechne k, für das $h.k \leq_B h'$ ist;
14: $P[p.i, k] = \min(P[p.i, k], C[p.i, h']$ + zusätzliche Sortingkosten von h nach h');
15: **else if** $!(h \leq_B h')$ und $!(h \geq_B h')$ **then**
16: $P[p.i, grad(h) + 1] = \min(P[p.i, grad(h) + 1], C[p.i, h']$ + zusätzliche Sortingkosten für Wirtswechsel von h nach h');
17: **end if**
18: **end for**
19: **end for**
20: **for all** Kombinationen von k_i aus $P[p.i, k_i]$ für jeden Parasitenknoten $p.i$ **do**
21: **for all** Anzahlen möglicher Kospeziationen **do**
22: $C[p, h] = \min\left(C[p, h], \left(\sum_{i=1}^{grad(p)} P[p.i, k_i]\right) + \text{Kosten für Ereignisse am Knoten } h\right)$;
23: **end for**
24: **end for**
25: **end if**
26: **end for**
27: **end for**

4.4 Verwendete Datenstrukturen zur Reduktion des Berechnungsaufwandes

Im folgenden Kapitel werden die zur Datenhaltung verwendeten Speicherstrukturen erläutert. Da auch die Ausgangsdaten als baumförmige Phylogenien vorliegen, soll diese Struktur zumindest logisch beibehalten werden. Grundlage sind somit zwei Bäume, je einer für Wirts- und Parasitenart. Im Wirtsstammbaum werden nur die Ausgangsdaten der Wirtsphylogenie gespeichert. Der Parasitenstammbaum hingegen speichert in seinen Knoten zusätzlich die für die Rekonstruktion benötigten Kosten und die damit verbundenen Abbildungen der seiner Kinder in den Wirtsbaum.

4.4.1 Baumstruktur

Aus Effizienzgründen werden die Stammbäume nicht als lose Menge von Knoten und Zeigern gespeichert, sondern als Feld, welches in der Art einer Tiefensuche die Knotenmenge indiziert. Die Wurzel des Baumes befindet sich im Feld mit dem Index 0. Zusätzlich zu diesem Knotenfeld werden zwei weitere Ganzzahlfelder gespeichert. Eines davon enthält die der Größe nach geordneten Indizes der Blattknoten. Das zweite Feld speichert für jede Ebene des Baumes den am weitesten „links“ stehenden Knoten, also den Knoten der jeweiligen Ebene mit dem kleinsten Index.

4.4.2 Knotenstruktur

In den Knoten werden die Informationen aus den Ausgangsdaten gespeichert. Darunter sind sowohl der Namen der jeweiligen Spezies und die zugehörigen Zeitinformationen. Im Falle von Blattknoten, für welche eine Abbildung $\varphi(p, h)$ existiert, wird ebenfalls der Index des Abbildes dieses Knotens gespeichert.

Um eine Baumstruktur zu realisieren wird in jedem Knoten der Index des Vaterknotens gespeichert. Hinzu kommen die Anzahlen der direkten Kinder und die aller Nachfahren. Des Weiteren wird der Index des als nächsten „rechts“ gelegenen Knotens gleicher Ebene gespeichert und zusätzlich die Information, ob es sich dabei um einen Geschwisterknoten handelt.

Durch diese Implementierung lassen sich Vergleiche bezüglich der Position zweier Knoten im Baum sehr schnell realisieren. Es müssen dafür nur die zugehörigen Indices miteinander verglichen werden. Da ein Großteil der Berechnung des Algorithmuses aus solchen Vergleichen besteht wäre ein traversieren des Baumes entlang der Knoten nicht praktikabel.

Für den Parasitenstammbaum werden in den Knoten noch Informationen vorgehalten, welche für die Berechnung der Rekonstruktionen nötig sind. Für jede mögliche

Abbildung eines Knotens p auf einen Knoten h werden für p die Rekonstruktionskosten als Gleitkommawert gespeichert. Hinzu kommen die Indizes der h_i, auf welche die Kinder von p abgebildet werden, sowie die Anzahlen der für diese Teilrekonstruktion benötigten koevolutionären Ereignisse.

4.4.3 Statische Kostentabelle

Wie aus der formalen Beschreibung in Kapitel 3 ersichtlich wird, ist für die Ereigniskosten $E(h, h_1, \ldots, h_{grad(p)})$ aus Formel 3.4 die Wahl der zugehörigen Parasitenknoten p nur für die Bestimmung des $grad(p)$ von Interesse. Bei gleichem $grad(p_k)$ für verschiedene Knoten p_k entstehen die gleichen Ereigniskosten E. Aus diesem Grund bietet es sich an, einen Teil der für die Rekonstruktion benötigten Berechnungen einmalig in der Initialisierungsphase des Algorithmus durchzuführen und die berechneten Ergebnisse im weiteren Verlauf wieder zu verwenden. Zu diesem Zweck kann eine statische Kostentabelle $S(h, h_i)$ vorgeneriert werden. In dieser sind die Anzahlen der zusätzlichen Sortings enthalten, welche benötigt werden, um das Kind eines auf den Wirt h abgebildeten Parasiten dem Wirt h_i zuzuweisen.

Sollte h_i ein Nachfolger von h oder gleich h sein, so sind genau so viele zusätzliche Sortings nötig, wie es Knoten auf dem Pfad von h nach h_i gibt. Die direkt am Knoten h auftretenden Sortings werden nicht in der Kostentabelle gespeichert, sondern während der Rekonstruktion separat berechnet.

Sollte h_i ein Vorgänger von h sein, so sind die Kosten unendlich, da in diesem Fall eine Rekonstruktion chronologisch inkonsistent ist.

Wenn h_i weder Vorgänger noch Nachfolger von h ist, so muss ein Wirtswechsel stattgefunden haben. Je nachdem ob eine Zeitfunktion $T(H)$ für den Stammbaum des Wirtes verwendet wird, können dabei ebenfalls zusätzliche Sortingkosten auftreten. Für $T(h_i) <_T T_1(h)$ sind diese Kosten unendlich, da ein solches Ereignis die Chronologie verletzen würde. Wenn $T(h_i) =_T T(h)$ ist, dann sind keine weiteren Sortings nötig. Gilt $T(h_i) >_T T(h)$, so landet der Parasit nach dem Wirtswechsel nicht gleich auf h_i, sondern auf einem Wirt h_j, welcher Vorgänger von h_i ist und für den $T(h_j) =_T T(h)$ gelten muss. Somit kommen noch die Kosten der zusätzlichen Sortings für den Pfad von h_j zu h_i hinzu. Es ist auch denkbar, dass mehrere mögliche h_j mit $T(h_j) =_T T(h)$ existieren. In einem solchen Fall wird dasjenige h_j ausgewählt, welches die Sortingkosten minimiert.

Die folgende Abbildung und die dazugehörige Tabelle zeigen die statischen Sortingkosten anhand eines einfachen Beispiels.

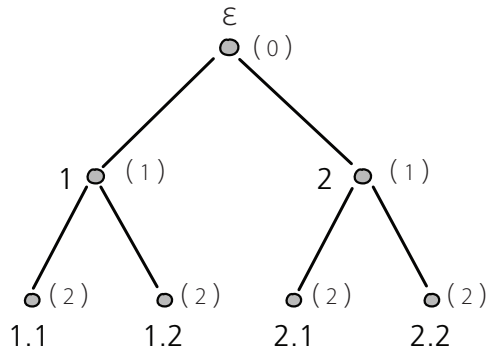


Abbildung 4.4: Die Abbildung zeigt einen einfachen Wirtsstammbaum mit zugehörigen Zeitinformationen.

	ϵ	1	1.1	1.2	2	2.1	2.2
ϵ	0	1	2	2	1	2	2
1	∞	0	1	1	0	1	1
1.1	∞	∞	0	0	∞	0	0
1.2	∞	∞	0	0	∞	0	0
2	∞	0	1	1	∞	1	1
2.1	∞	∞	0	0	∞	0	0
2.2	∞	∞	0	0	∞	0	0

Tabelle 4.1: Die Tabelle zeigt die Anzahl der statischen Sorting-Ereignisse zum in Abbildung 4.4 dargestellten Wirtsbaum. ∞ kennzeichnet dabei nicht gültige Abbildungen.

4.4.4 Datenstruktur zur Berechnung der maximalen Anzahl von Kospeziationen

Der im Abschnitt 4.2.2 vorgestellte Algorithmus zur Berechnung der maximalen Anzahl möglicher Kospeziationen soll im Folgenden näher betrachtet werden. Wie aus dem Algorithmus 2 ersichtlich wird, ist in jedem Schritt eine Sortierung der Liste nötig. Da jedoch pro Schritt immer nur die ersten zwei Elemente um eins reduziert werden, ist es möglich die Sortierung in konstanter Zeit durchzuführen. Dafür werden die Objekte in einer zweifach verketteten Liste gespeichert. Zusätzlich gibt es noch einen Zeiger auf das erste und auf das letzte Objekt. Die Liste wird in Blöcke eingeteilt, deren Elemente den gleichen Wert beinhalten. Für das erste Element eines Blockes wird jeweils noch ein Zeiger benötigt. Dieser zeigt auf das erste Element des nachfolgenden Blockes. Für die Sortierung nach einem Schritt müssen die ersten beiden Elemente gegebenenfalls bis an den Anfang des Folgeblockes verschoben werden. Ein Beispiel einer solchen Liste ist in Abbildung 4.5 dargestellt.

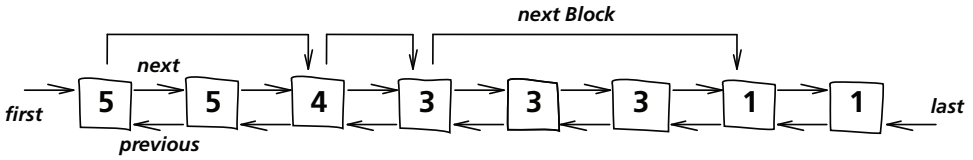


Abbildung 4.5: Beispiel der verwendeten Datenstruktur zur Berechnung der maximalen Anzahl möglicher Kospeziationen.

4.5 Ausgabe einer Gesamtlösung

Wie bereits in Kapitel 3.1 eingangs erwähnt, werden für jede Teilrekonstruktion die zu den günstigsten Kosten gehörenden Abbildungen der Kindknoten in den Parasitenknoten mitgespeichert. Am Ende des bottom-up-Prozesses existiert eine kostenminimale Gesamtrekonstruktion im Wurzelknoten des Parasitenbaumes. Um diese auszugeben muss für den Wurzelknoten nach dessen günstigster Abbildung in den Wirtsbaum gesucht werden. Ist diese gefunden, so werden die dazugehörigen Abbildungen der Kindknoten ausgelesen. Auf diese Weise wird der gesamte Baum in Form eines top-down-Verfahrens durchlaufen.

Da in jedem Knoten des Parasitenbaumes für jede Abbildung eines Parasiten auf einen Wirt nur eine Kombination von Abbildungen der Kindknoten gespeichert ist, wird auch nur eine der günstigsten Rekonstruktionen gefunden. Es können jedoch mehrere Rekonstruktionen mit minimalen Kosten existieren. Will man auch diese berücksichtigen, müssen alle zu kostenminimalen Rekonstruktionen führenden Abbildungen der Parasitenkindknoten gespeichert werden. Da diese bereits während der Berechnung betrachtet werden, bedarf es keiner Modifikation des vorgestellten Ansatzes. Aus Komplexitätsgründen wird allerdings in der Implementierung darauf verzichtet.

4.6 Komplexitätsanalyse

Für die Komplexitätsabschätzung werden folgende Konventionen getroffen:

1. Sei n die Anzahl der Knoten des Parasitenbaumes.
2. Sei m die Anzahl der Knoten des Wirtsbaumes.
3. Sei $grad(B) = \max_{b \in B}(grad(b))$ der maximale Grad eines Baumes B.

Zur Berechnung der statischen Kostentabelle muss die Anzahl der zusätzlichen Sortings von jedem Knoten des Wirtsbaumes zu jedem anderen Knoten berechnet werden. Im schlimmsten Fall gehört zu einer solchen Abbildung ein Wirtswechsel. Bei der Verwendung von Zeitinformationen, muss dann die landing site im Wirtsbaum bestimmt

werden. Da diese, wie in Kapitel 2.7 beschrieben, ein Vorgänger des Zielknotens ist, müssen maximal $höhe(H)$ Knoten traversiert werden. Es ergeben sich somit $m^2 * höhe(H)$ Operationen für die Berechnung der statischen Kostentabelle.

Für die Berechnung der Teilrekonstruktionen werden alle n Knoten des Parasitenbaumes durchlaufen (Zeile 2 des Algorithmus 3) und für jeden der m Knoten im Wirtsbaum werden die für eine Abbildung benötigten Ereigniskosten berechnet (Zeile 3). Dies ergibt n^m Berechnungen, von denen für jede die $P[p.i, h']$ berechnet werden müssen. Dazu werden für jeden der Kindknoten von p alle Wirtsknoten durchlaufen, also $grad(P) * m$ Operationen (Zeile 8 und 9). Des Weiteren werden für jede Kombination aus $P[p.i, h']$ (Zeile 20) die Kosten der verschiedenen Ereignisvarianten mit unterschiedlicher Anzahl von Kospeziationen berechnet (Zeile 21). Dies sind maximal $(grad(H)+2)^{grad(P)}$ Kombinationen und $grad(P)/2$ Ereignisvarianten. Für jede Variante müssen noch die zugehörigen $P[p.i, h']$ aufaddiert und die Kosten der jeweilig berechneten 4 Ereignisanzahlen zusammengezählt werden. Dies ergibt maximal $grad(P) + 4$ Operationen.

Zusammengefasst erhällt man eine Komplexität von:

$$O(m^2 * höhe(H) + n * m * (grad(P) * m + ((grad(H) + 2)^{(grad(P))} * grad(P)/2 * (grad(P) + 4)))) \tag{4.1}$$

mit $N = \max(n, m)$, $H = höhe(H)$ und $G = max(grad(P), grad(H))$ ergibt dies:

$$O(N^2 * H + N^2 * (G * N + (G^G * G^2) = \tag{4.2}$$

$$O(N^2 * H * (G * N + G^{G+2}) \tag{4.3}$$

5 Dynamisierung der Ereigniskosten

Für gegebene Wirts- und Parasitenbäume und eine gegebene Abbildung $\varphi_{P,H}$ der Blätter des Parasitenbaumes auf die Blätter des Wirtsbaumes ist die sich ergebende kostengünstigste Rekonstruktion maßgeblich von der Wahl der Ereigniskosten abhängig. Für verschiedene Ereigniskosten entstehen zum Teil sehr unterschiedliche Rekonstruktionen. Bislang mussten diese Kosten immer manuell gewählt werden. Da diese Wahl im Allgemeinen recht willkürlich ist, stellt sich die Frage nach einem Gütekriterium für eine Rekonstruktion bezüglich der bei ihrer Berechnung verwendeten Ereigniskostenverteilung.

Im folgenden Abschnitt wird ein mögliches Gütekriterium erläutert und aufgezeigt, wie mit dessen Hilfe eine möglichst plausible Verteilung der Ereigniskosten für gegebene Ausgangsdaten gefunden werden kann.

5.1 Verwendung von Ereigniswahrscheinlichkeiten anstelle von Ereigniskosten

Da man immer die kostengünstigste Rekonstruktion einer gemeinsamen Evolution von Parasit- und Wirtsarten sucht, können die jeweiligen Ereigniskosten auch als eine Art Wahrscheinlichkeit aufgefasst werden, mit der im Laufe der Evolution ein solches Ereignis auftrat.

> „If each event is associated with a cost that is inversely related to the likelihood of the event (...), then the most parsimonious reconstruction will also, in some sense, be the most likely explantion of the observed data.“[1]

Große Kosten bedeuten somit eine geringe Auftrittswahrscheinlichkeit, während niedrige Ereigniskosten für eine hohe Auftrittswahrscheinlichkeit stehen. Des Weiteren soll das Verhältnis zwischen den Kosten zweier Ereignisse angeben, wie groß die Wahrscheinlichkeit des einen Ereignisses im Verhältnis zur Wahrscheinlichkeit des anderen Ereignisses ist.

Eine direkte Umrechnung der Kosten in Wahrscheinlichkeiten ist somit nur möglich, wenn die Ereigniskosten größer 0 sind. Im Folgenden wird diese Einschränkung

[1] vgl. [24] S. 25

immer angenommen. Eine Kostenverteilung mit negativen Kospeziationskosten wie sie von Charleston vorgeschlagen wurde, ist demzufolge nicht mehr möglich.

Bei positiven Ereigniskosten c_1 bis c_n lassen sich die dazugehörigen Wahrscheinlichkeiten p_1 bis p_n wie folgt berechnen:

$$p_x = \frac{\frac{1}{c_x}}{\sum_{i=1}^{n} \frac{1}{c_i}} \tag{5.1}$$

Sie ergeben sich somit als die in der Summe auf 1 normierten Reziproken der Kosten.

5.2 Automatische Berechnung von Werten für die Ereigniskosten

Die zu einer Rekonstruktion gehörenden relativen Häufigkeiten r_1 bis r_n ergeben sich als die in der Summe auf 1 normierten Anzahlen der bei dieser Rekonstruktion aufgetretenen Ereignisse. Um diese Häufigkeiten zu bestimmen wird eine Kostenverteilung gewählt. Daraufhin wird die kostengünstigste Rekonstruktion berechnet und aus der Anzahl der dabei verwendeten Ereignisse ergeben sich die relativen Häufigkeiten.

Die Ereigniskosten sind dann am plausibelsten, wenn sich die aus der Kostenverteilung ergebenden Ereigniswahrscheinlichkeiten möglichst gering von den relativen Ereignishäufigkeiten der berechneten Rekonstruktion abweichen. Es wird somit die Kostenverteilung gesucht, bei der $\sum_{i=1}^{n} |p_i - r_i|$ minimal wird.

5.2.1 Gütekriterium

Um die Rekonstruktionen gleicher Ausgangsdaten mit verschiedenen Kostenwerten vergleichen zu können, ist ein Gütekriterium nötig. An dieser Stelle soll für das Gütekriterium im Folgenden die auf 1 normierte Summe der $|p_i - r_i|$ verwendet werden.

$$G = \sum_{i=1}^{n} \frac{|p_i - r_i|}{2} \tag{5.2}$$

Die Division durch 2 wird an dieser Stelle verwendet, da im ungünstigster Fall bei 4 verschiedenen Ereignissen $\sum_{i=1}^{n} |p_i - r_i| = 2$ ist. Dies tritt beispielsweise auf, wenn alle $p_i = 0,5$ sind und in der kostengünstigsten Rekonstruktion nur eine Ereignisart verwendet wird. Je kleiner der Gütewert ist, desto näher liegen die Wahrscheinlichkeiten und die relativen Häufigkeiten beieinander.

Auf diese Weise repräsentiert die berechnete Güte ein Maß für die Qualität einer Rekonstruktion. Ist der Gütewert sehr hoch (größer als 0.25), so bedeutet dies, dass die hinter den Kosten stehenden Wahrscheinlichkeiten stark von den gefundenen relativen

Häufigkeiten abweichen. In einem solchen Fall ist es sehr unwahrscheinlich, dass die rekonstruierte koevolutionäre Geschichte tatsächlich so stattfand.

5.2.2 Rekursive Annäherung an die optimalen Kostenwerte

Um mit Hilfe des vorgestellten Gütekriteriums die möglichst plausibelsten Kostenwerte zu ermitteln, wird der Algorithmus zur Berechnung der Rekonstruktion mehrfach mit unterschiedlichen Kostenwerten ausgeführt und die nach Formel 5.2 berechneten Gütewerte miteinander verglichen. Es wird diejenige Kostenverteilung gesucht, bei der die hinter den Ereigniskosten stehenden Wahrscheinlichkeiten mit den in einer Rekonstruktion gefundenen Anzahlen der Ereignisse möglichst gut übereinstimmen.

Da die Kostenwerte in Wahrscheinlichkeiten umgerechnet werden, deren Summe 1 ergibt, kann eine Kostenverteilung auch als 3-dimensionaler Vektor von Wahrscheinlichkeitswerten betrachtet werden, wobei die drei Werte den berechneten Wahrscheinlichkeiten von Sorting, Kospeziation und Duplikation entsprechen. Die Wahrscheinlichkeit des Auftreten eines Wirtswechsels berechnet sich dann als 1 minus diese Werte. Auf diese Weise kann eine Wahrscheinlichkeitsverteilung als Punkt im 3-dimensionalen Raum aufgefasst werden. Jeder dieser Punkte liegt innerhalb eines Tetraeders, welcher durch die Punkte $P_0 = (0, 0, 0)$, $P_1 = (1, 0, 0)$, $P_2 = (0, 1, 0)$ und $P_3 = (0, 0, 1)$ aufgespannt wird. Dabei entspricht der Punkt P_0 einer Wahrscheinlichkeitsverteilung mit 1 für Wirtswechsel und 0 für alle anderen Ereignisse. Analog dazu ergeben die Punkte P_1, P_2 und P_3 Wahrscheinlichkeitsverteilungen mit je 1 für Sortings, Kospeziationen und Duplikationen.

Um eine möglichst gute Kostenverteilung zu finden, wird rekursiv vorgegangen. In jedem Schritt wird, ausgehend von den vier gegebenen Eckpunkten eines Tetraeders, oder genauer von den vier Eckpunkten eines 3-dimensionalen Simplex, für diese Punkte die kostengünstigste Rekonstruktion berechnet. Zusätzlich wird auch die Rekonstruktion des Schwerpunktes diese Simplex berechnet. Danach wird der Simplex in acht neue Simplizes aufgeteilt. Dazu werden die Mittelpunkte der Kanten (P_0, P_1), (P_0, P_2), (P_0, P_3), (P_1, P_2), (P_1, P_3) und (P_2, P_3) berechnet. Seien diese Punkte P_{01}, P_{02}, P_{03}, P_{12}, P_{13} und P_{23}. Die acht neuen Simplizes werden dann durch folgende Punkte aufgespannt.

1. P_0, P_{01}, P_{02} und P_{03}
2. P_1, P_{01}, P_{12} und P_{13}
3. P_2, P_{02}, P_{12} und P_{23}
4. P_3, P_{03}, P_{13} und P_{23}
5. P_{01}, P_{23}, P_{02} und P_{03}
6. P_{01}, P_{23}, P_{12} und P_{13}
7. P_{01}, P_{23}, P_{02} und P_{12}
8. P_{01}, P_{23}, P_{03} und P_{13}

Für diese neuen Punkte werden rekursiv die kostengünstigsten Rekonstruktionen und deren acht neue Simplizes berechnet.

Da für die Ereigniskosten Werte größer als 0 gefordert wurden, können für den Beginn der Rekonstruktion nicht die Punkte $(0,0,0)$, $(1,0,0)$, $(0,1,0)$ und $(0,0,1)$ verwendet werden. Statt dessen sollten Werte in deren unmittelbaren Umgebung genutzt werden. So wurden hier standardmäßig die Punkte $(0.001, 0.001, 0.001)$, $(0.997, 0.001, 0.001)$, $(0.001, 0.997, 0.001)$ und $(0.001, 0.001, 0.997)$ verwendet. Diese Annäherung ist mathematisch nötig, hat aber biologisch kaum eine Relevanz. Zwar wird der Suchraum für eine gute Kostenverteilung eingeschränkt, aber eine solche Verteilung mit hoher Güte würde erfordern, dass nur Ereignisse ein und desselben Typs auftreten. Dass diese Fälle auch auftreten können, zeigt die folgende Überlegung.

Angenommen es ließe sich für gegebene Ausgangsdaten eine Rekonstruktion erzeugen, welche nur Wirtswechsel verwendet. Dies ist nahezu immer möglich, wenn keine Zeitinformationen für den Wirtsstammbaum vorhanden sind.[2] Für einen solchen Fall würde man im Punkt $(0,0,0)$ theoretisch eine optimale Kostenverteilung finden. Diese ergibt sich allerdings nur aus der Mächtigkeit des Wirtswechselereignisses und ist biologisch eher unwahrscheinlich.

Aus diesem Grund sollen Kostenverteilungen mit hoher Güte, deren kostengünstigste Rekonstruktion nur aus Wirtswechseln besteht, bei der automatischen Berechnung der Kosten nicht berücksichtigt werden. Statt dessen werden nur Kostenverteilungen in Betracht gezogen, bei denen eine kostenminimale Rekonstruktion mit mindestens einer Kospeziation gefunden wurde. Diese Einschränkung erscheint auch dann sinnvoll, wenn man die Aussage der bereits erwähnten Fahrenholz-Regel in Betracht zieht. Wenn man davon ausgeht, dass die Speziationen der Wirte Anpassungsprozesse bei den Parasiten verursachen, so will man Rekonstruktionen finden, welche auch Kospeziationen enthalten. Somit werden auch Lösungen ausgeschlossen, bei denen Duplikationen und Sortings recht billig, Kospeziationen jedoch im Verhältnis dazu sehr teuer sind. In solchen Fällen könnte eine Kospeziation durch Duplikationen mit anschließenden Sortings ersetzt werden und Kospeziationen würden gar nicht erst auftreten.

5.2.3 Abbruchkriterien

Da die oben beschriebene rekursive Annäherung nie enden würde, müssen für eine Implementierung Abbruchkriterien definiert werden. Diese sollen eine Rekursion möglichst frühzeitig beenden, wenn deutlich wird, dass in dem betrachteten Simplex höchstwahrscheinlich keine besseren Kostenverteilungen mehr vorzufinden sind.

[2] Es könnten alle Knoten des Parasitenstammbaumes auf Blattknoten des Wirtsbaumes abgebildet werden, bzw. auf die Kanten davor. Ausgehend von dem Wurzelknoten würde bei jeder Speziation eines Parasiten einer der neu entstandenen Parasiten zu einem neuen Blattknoten des Wirtsbaumes springen und der andere würde auf dem vorherigen Blattknoten verweilen.

Maximale Rekursionstiefe

Die einfachste Möglichkeit besteht darin, die Rekursionstiefe zu beschränken. Auf diese Weise wird sichergestellt, dass die Berechnung der besten Kostenwerte endet. Abhängig von der Komplexität der Ausgangsdaten sind Rekursionstiefen zwischen 3 und 6 sinnvoll. Wenn man berücksichtigt, dass in jedem Schritt mindestens acht neue Simplizes entstehen, für die jeweils sieben Rekonstruktionen berechnet werden, erhält man bei einer Rekursionstiefe von 7 bereits mehr als 10 Millionen Berechnungen.

Statische Eckpunkte

Wenn bei der Betrachtung eines Simplex für die Kostenverteilungen an den Eckpunkten jeweils die gleichen Rekonstruktionen erzeugt werden, so wird davon ausgegangen, dass sich diese Kostenwerte nur noch sehr wenig voneinander unterscheiden. Folglich wird angenommen, dass auch alle Kostenverteilungen von Punkten innerhalb des Simplex die gleichen Lösungen erzeugen würden. Die Rekursion wird an dieser Stelle fortgesetzt, jedoch ohne die Rekonstruktionen real zu berechnen. Statt dessen wird die Güte der Kostenverteilung mit den aus der Rekonstruktion der Eckpunkte berechneten relativen Ereignishäufigkeiten approximiert. Für die auf diesem Weg gefundene lokale Bestlösung wird die Rekonstruktion erneut durchgeführt, um sicher zu stellen, dass auch hier die gleiche Rekonstruktion erzeugt wird.

Schlechte lokale Lösungen

Mit dem folgenden Abbruchkriterium sollen Rekursionen beendet werden, bei denen nur noch Kostenverteilungen betrachtet würden, deren Güte einen bestimmten Schwellwert wahrscheinlich nicht unterschreiten würde. Dazu wird für jedes Ereignis der minimale und maximale Wert der von den Eckpunkten des Simplex verwendeten Wahrscheinlichkeiten gesucht. Des Weiteren wird der minimale und maximale Wert der bei den zugehörigen Rekonstruktionen gefundenen relativen Häufigkeiten für jedes Ereignis berechnet. Auf diese Weise werden pro Ereignis zwei Intervalle erzeugt, eines für die verwendeten Wahrscheinlichkeiten und eines für die gefundenen relativen Häufigkeiten. Überschneiden sich die beiden Intervalle eines Ereignisses so wird angenommen, dass innerhalb des Simplex ein Punkt existieren könnte, bei dem die Wahrscheinlichkeit und die relative Häufigkeit übereinstimmen. In einem solchen Fall wäre der Beitrag dieses Ereignisses zu Güte der Kostenverteilung 0. Wenn sich die beiden Intervalle jedoch nicht überschneiden so wird vorausgesetzt, dass der jeweilige Anteil vom Gütewert mindestens so groß ist, wie die Differenz zwischen den Intervallen dividiert durch 2. Die Summe der vier so berechneten Teilwerte wird als Schätzer für die maximal erreichbare

Güte genutzt. Übersteigt dieser Wert einen vorher gewählten Schwellwert, so kann die Rekursion an dieser Stelle abgebrochen werden.

Diese Vorgehensweise lässt sich wie folgt begründen. Es wird angenommen, dass innerhalb eines Simplex die bei der kostengünstigsten Rekonstruktion auftretenden relativen Häufigkeiten der Ereignisse stetig anwachsen oder abfallen. Das heißt, dass der minimale und maximale Wert der an den Eckpunkten des Simplex auftretenden relativen Häufigkeiten auch innerhalb des Simplex das Minimum bzw. Maximum darstellt. Für die Wahrscheinlichkeiten an den Eckpunkten gilt dies ohnehin. Unter dieser Annahme könnten somit innerhalb des Simplex nur Punkte gefunden werden, bei denen die relativen Häufigkeiten der Ereignisse zwischen diesen Werten liegen. Die an dieser Stelle geforderte Stetigkeitseigenschaft der relativen Häufigkeiten kann jedoch nicht immer gewährleistet werden. Vor allem bei sehr großen Simplizes, z.B. in der ersten und zweiten Rekursionsstufe ist die Eigenschaft oftmals nicht erfüllt. Da aber bei diesen Simplizes auch die zugehörigen Intervalle der Wahrscheinlichkeiten sehr groß sind, gibt es deutlich mehr Überschneidungen. Somit fallen diese bei der Berechnung der noch maximal erreichbaren Güte nicht ins Gewicht. Dennoch handelt es sich hierbei um eine Approximation und es kann nicht ausgeschlossen werden, dass eine Kostenverteilung innerhalb des Simplex eine bessere als die berechnete noch erreichbare Güte besitzt. Deshalb wird ein Schwellwert von nicht kleiner als 0.25 vorgeschlagen. Ab Werten größer 1 bleibt das Abbruchkriterium wirkungslos.

Lineare Interpolation

Wie das vorangegangene Abbruchkriterium verwendet auch dieses die Stetigkeitsannahme der relativen Häufigkeiten. Es soll eine lineare Interpolation für die relativen Häufigkeiten jedes einzelnen Ereignisses durchgeführt werden, abhängig von den Wahrscheinlichkeitsvektoren an den Eckpunkten des Simplex. Mit Hilfe dieser Interpolation kann eine Approximation für die relativen Häufigkeiten der Ereignisse an einem beliebigen Punkt innerhalb des Simplex berechnet werden. Dazu wird der Schwerpunkt des Simplex bestimmt. Für diesen Punkt wird die kostengünstigste Rekonstruktion berechnet. Weicht keiner der dabei aufgetretenen relativen Häufigkeiten der Ereignisse von den Schätzungen der linearen Interpolation um mehr als einen vorher festgelegten Schwellwert ab, so wird die lineare Interpolation als zuverlässig angenommen. In diesem Fall kann die Rekursion analog zum Abbruchkriterium mit statischen Eckpunkten ohne Berechnung der genauen Rekonstruktionen fortfahren. Es werden dabei lediglich die Ergebnisse der linearen Interpolation zur Bestimmung der Güte herangezogen. Nach dem Erreichen der maximalen Rekursionstiefe wird für die Kostenverteilung mit der am höchsten geschätzten Güte zur Kontrolle eine Rekonstruktion berechnet.

Für den Schwellwert wird ein Wert von 0.01 vorgeschlagen. Ab Werten kleiner 0 bleibt dieses Abbruchkriterium wirkungslos.

Für die lineare Interpolation der relativen Häufigkeiten eines jeden Ereignisses wird ein lineares Gleichungssystem mit vier Gleichungen aufgestellt, je eine pro Eckpunkt des Simplex. Seien dafür die für den Punkt P_i berechneten relativen Häufigkeiten des Ereignisses j gleich r_i^j. Das Gleichungssystem sieht dann wie folgt aus:

$$\begin{aligned} a_0^j * P_0.x + a_1^j * P_0.y + a_2^j * P_0.z + a_3^j = r_0^j \\ a_0^j * P_1.x + a_1^j * P_1.y + a_2^j * P_1.z + a_3^j = r_1^j \\ a_0^j * P_2.x + a_1^j * P_2.y + a_2^j * P_2.z + a_3^j = r_2^j \\ a_0^j * P_3.x + a_1^j * P_3.y + a_2^j * P_3.z + a_3^j = r_3^j \end{aligned} \tag{5.3}$$

Um das Gleichungssystem zu lösen wird eine QR-Faktorisierung der Matrix

$$\begin{pmatrix} P_0.x & P_0.y & P_0.z & r_0^j \\ P_1.x & P_1.y & P_1.z & r_1^j \\ P_2.x & P_2.y & P_2.z & r_2^j \\ P_3.x & P_3.y & P_3.z & r_3^j \end{pmatrix} \tag{5.4}$$

durchgeführt. Dadurch entsteht die orthogonale Matrix Q und die obere Dreiecksmatrix R. Es kann der Hilfsvektor

$$\vec{z} = Q^T * \vec{r^j} \tag{5.5}$$

berechnet werden. Durch Rückwärtseinsetzen in das Gleichungssystem

$$R\vec{a^j} = \vec{r^j} \tag{5.6}$$

erhält man die gesuchten Koeffizienten a_0^j bis a_3^j.

5.2.4 Aussagewert der gefundenen Bestlösung für eine Kostenverteilung

Resultat des zuvor beschriebenen rekursiven Verfahrens ist eine gefundene Bestlösung. Je näher die Güte dieser Lösung bei 0 liegt, desto besser passen die dabei verwendeten Kosten mit den aufgetretenen Anzahlen der Ereignisse zusammen. Sollten die so berechneten Kosten den in der Natur auftretenden Wahrscheinlichkeiten entsprechen, so ist auch die gefundene Rekonstruktion plausibel. Wenn eine Bestlösung mit einer sehr schlechten Güte (von 0,25 oder mehr) gefunden wurde, bedeutet dies, dass möglicherweise die Rekursionstiefe nicht ausreichend war und eine bessere Lösung noch nicht gefunden werden konnte. Es kann aber auch sein, dass für die Ausgangsdaten keine Kostenverteilung existiert, bei welcher die Ereigniswahrscheinlichkeiten mit den gefundenen relativen Häufigkeiten gut übereinstimmen. In einem solchen Fall können hieraus Rückschlüsse auf die Qualität dieser Ausgangsdaten gezogen werden.

6 Implementierung des Algorithmus und grafische Ausgabe

Im folgenden Kapitel sollen die Applikationen vorgestellt werden, welche auf Basis des algorithmischen Ansatzes aus Kapitel 4 erstellt wurden. Dies ist zum einen ein Programm namens DynamicTreeMap, welches für gegebene Ausgangsdaten eine kostenminimale Rekonstruktion berechnet. Weiterhin wurde eine zweite Anwendung erstellt, welche die Ergebnisse grafisch darstellt.

6.1 Implementierung des Algorithmus

DynamicTreeMap ist ein kommandozeilenbasiertes Java-Programm. Für die plattformunabhängige Umsetzung des Algorithmus wurde Java als Programmiersprache gewählt. Da das Augenmerk auf der Implementierung des Algorithmus lag, wurde auf eine grafische Oberfläche verzichtet. Statt dessen wird das Programm mit dem Befehl *java -jar DynamicTreeMap.jar* gestartet und durch Eingabe von Parametern konfiguriert. Diese Parameter sollen im folgenden vorgestellt werden.

6.1.1 Ausgangsdaten

Parameter: -n [Dateiname] (obligatorisch)
Mit diesem Parameter wird der Name der Datei angegeben, welche die im Nexusformat[1] gespeicherten Ausgangsdaten enthält. In dieser Datei werden in einem TAXA-Block die verwendeten TAXLABELS angegeben. Im TREE-Block müssen die beiden Bäume „host“ und „parasite“ definiert sein. Zusätzlich gibt es noch einen PRIVATE-Block welcher die Informationen der Abbildung $\varphi_{P,H}$ in der der Form

PHI-FUNCTION
(Parasit1,Wirt1,Type)
(Parasit2,Wirt2,Type)
...
(ParasitN,WirtN,Type)
;

[1] vgl. [27]

beinhaltet. Type ist ein numerischer Wert, welcher allerdings bei der Berechnung ignoriert wird. Im PRIVATE-Block befinden sich ebenfalls die Zeitinformationen sowie sie Standardkosten der koevolutionären Ereignisse. Diese werden syntaktisch wie folgt beschrieben.

```
RANKS
    Parasit1    Start    Ende,
    Wirt1       Start    Ende,
    Parasit2    Start    Ende,
    Wirt2       Start    Ende,
    ...
    ParasitN    Start    Ende,
    WirtN       Start    Ende
    ;
```

sowie

```
COST-TABLE
    Kospeziation, Duplikation, Auslöschung, Sorting, Wirtswechsel, 0
    ;
```

Obwohl für die Zeitinformationen von Wirtsknoten Start- und Endwert angegeben werden kann, wird dennoch nur der Startwert verwendet.

6.1.2 Ereigniskosten

Neben der Angabe in der Nexusdatei existiert für die Eingabe der Ereigniskosten eine weitere Möglichkeit in Form von Parametern. Dazu wird der jeweilige Parameter mit dem dazugehörigen Wert angegeben. Dieser Wert ist eine Gleitkommazahl mit Punkt als Trennzeichen für Nachkommastellen.

Kospeziationskosten: Parameter: -co [Wert] (Standard: -2.0)

Sortingkosten: Parameter: -so [Wert] (Standard: 1.0)

Duplikationskosten: Parameter: -du [Wert] (Standard: 2.0)

Wirtswechselkosten: Parameter: -hs [Wert] (Standard: 2.0)

Optionen für die Art der Kosteneingabe

Parameter: -wX (X$\in \{0,1\}$) (Standard: 0)
Es kann ebenfalls bestimmt werden, ob die Kosten als direkte Kostenwerte (0) oder als Wahrscheinlichkeiten (1) angegeben werden. Für den zweiten Fall müssen die vier Werte in der Summe 1 ergeben.

6.1.3 Implementierte Algorithmusvarianten

Um den Algorithmus auch für andere Problemklassen nutzen zu können wurden etliche Optionen hinzugefügt, welche maßgeblichen Einfluss auf die zu berechnende kostenminimale Rekonstruktion haben. Diese Optionen bestimmen teilweise welche Ereignisse in bestimmten Situationen auftreten dürfen. Andererseits beeinflussen sie die Kosten indem sie die Behandlung von Wirtswechseln genauer differenzieren.

Optionen für erlaubte Wirtswechsel

Parameter: -hX (X$\in \{0,1,2,3\}$) (Standard: 2)
Mit diesem Parameter kann bestimmt werden, welche Wirtswechsel für die Berechnung der Rekonstruktion verwendet werden sollen. Dabei werden Abbildungen der Parasitenkindknoten auf Vorgängerknoten im Wirtsbaum ebenfalls als Wirtswechsel betrachtet. Bei zusätzlicher Verwendung von Zeitinformationen sind Wirtswechsel unabhängig von dem hier gesetzten Parameterwert nur dann gültig, wenn die Abbildungen auch von den Zeitzonen her erlaubt sind. Bei 0 sind beliebige Sprünge gültig. Bei 1 werden Wirtswechsel verboten, bei welchen die Parasitenkindknoten auf Vorgänger im Wirtsbaum oder auf Knoten höherer Ebene abgebildet werden. Der Wert 2 schließt nur Abbildungen auf Vorgänger im Wirtsbaum aus und bei 3 werden Wirtswechsel sämtlichst verboten.

Optionen für Sortingkosten bei Wirtswechseln

Parameter: -sX (X$\in \{0,1,2\}$) (Standard: 1)
Wie bereits erläutert, können bei Wirtswechseln immer noch zusätzliche Sortings entstehen. Ob und welche das sind, kann mit diesem Parameter reguliert werden. Bei 0 werden keine zusätzlichen Sortings berechnet. Es wird somit von direkten Wirtswechseln ausgegangen. Mit 1 wird für einen Wirtswechsel angenommen, dass dieser nur in der gleichen Ebene bzw. bei Verwendung von Zeitinformationen in der gleichen Zeitzone stattfinden kann. Deshalb werden hierbei Sortings von der landing site zum Wirtsknoten des Parasitenkindes hinzugerechnet. Sollte die take-off site unterhalb des Wirtsknotens des Parasitenvaters liegen, fallen dafür ebenfalls Sortingkosten an. Bei 2 werden zusätzlich zu den Kosten des Wirtswechsels noch Sortingkosten für den kompletten Weg über den gemeinsamen Vaterknoten im Wirtsbaum berechnet.

Optionen für die take-off site bei einem Wirtswechsel

Parameter: -tX (X$\in \{0,1\}$) (Standard: 0)
In dem zugrunde gelegten evolutionären Modell wurde ursprünglich gefordert, dass die take-off site eines Wirtswechsels immer direkt vor dem Wirtsknoten statt findet, auf den der Parasit abgebildet wird (0). Manchmal kann es jedoch sinnvoll sein die take-off

site auf einen Nachfolger dieses Wirtsknotens zu verlagern. Dadurch können Lösungen entstehen, welche weniger Sortingkosten für diesen Wirtswechsel verursachen. Diese Variante wird durch 1 spezifiziert. Sie erlaubt somit komplette Wirtswechsel ohne dass eine Speziation zum Zeitpunkt dieses Wirtswechsels beim Parasiten auftritt.

Optionen für das Berücksichtigen vollständiger Wirtswechsel

Parameter: -fX (X$\in \{0,1\}$) (Standard: 0)
Wie der vorangegangene Parameter bestimmt auch dieser die Vorgehensweise bei kompletten Wirtswechseln. Standardmäßig wurde vom evolutionären Modell gefordert, dass nicht alle Parasitenkinder Wirtswechsel durchführen dürfen. Mindestens eines der Kinder muss auf den gleichen oder einen Nachfolger des Wirtsknotens abgebildet werden, auf dem der Parasitenvater lebte (0). Soll diese Einschränkung nicht gelten, so kann die Option 1 verwendet werden.

Optionen für das Berücksichtigen von Zeitzonen

Parameter: -zX (X$\in \{0,1\}$) (Standard: 0)
Mit diesem Parameter wird bestimmt, ob die in der Nexusdatei beschriebenen Zeitinformationen benutzt werden sollen (1) oder nicht (0).

Optionen für das Berücksichtigen eingefügter Speziationen bei Abbildungen eines Parasitenblattes auf mehrere Wirtsblätter

Parameter: -lX (X$\in \{0,1\}$) (Standard: 0)
Wurde in der Abbildung $\varphi_{P,H}$ ein Parasit auf mehrere Wirtsblattknoten abgebildet, so werden für diese, wie in Kapitel 3.3.2 beschrieben, Pseudoknoten eingefügt. Da der eigentliche Parasitenblattknoten allerdings keiner Speziation entspricht, kann hier festgelegt werden, ob an diesem auftretende Kospeziationen und Duplikationen Kosten erzeugen (1) oder nicht (0).

Option für das Erzwingen einer Wurzel-zu-Wurzel-Abbildung

Parameter: -rX (X$\in \{0,1\}$) (Standard: 0)
Oftmals wird gefordert, dass der Wurzelknoten des Parasitenstammbaumes auf die Wurzel des Wirtsstammbaumes abgebildet wird. Diese erzwungene Abbildung kann mit 1 forciert werden. Bei 0 wird keine solche Bedingung an die Abbildung des Wurzelknotens gestellt.

6.1.4 Test auf chronologische Konsistenz

Parameter: -cX (X$\in \{0,1\}$) (Standard: 0)
Wie in Kapitel 3.4 erläutert, kann es zu chronologischen Inkompatibilitäten kommen, welche in einer gefundenen Rekonstruktion enthalten sein können. Will man die gefundene Lösung auf diese Inkompatibilitäten hin testen, so kann für diesen Parameter 1 angegeben werden. Soll kein Test stattfinden, so verwendet man 0.

6.1.5 Automatische Berechnung der Ereigniskosten

Parameter: -a [Wert1] [Wert2] [Wert3] (Standard: 3 0.01 0.25)
Der in Kapitel 5 vorgestellte Automatismus zur Berechnung der Ereigniskosten wird mit Angabe des Parameters -a verwendet. Die drei zugehörigen Werte bestimmen die Rekursionstiefe (Wert1), den Schwellwert für die lineare Interpolation (Wert2) und den Schwellwert für die minimale Güte (Wert3).

6.1.6 Textausgabe

Parameter: -o [Dateiname] (Standard: output.xml)
Es wird immer eine Textausgabe in der Kommandozeile erzeugt. In dieser stehen die Ergebnisdaten der gefundenen kostenminimalen Rekonstruktion. Diese Ausgaben haben die Form:

$$p\text{: } h^0\text{:}c^0\text{-}h^0_1,\ldots,h^0_n;\ h^1\text{:}c^1\text{-}h^1_1,\ldots,h^1_n;\ \ldots h^m\text{:}c^m\text{-}h0_1,\ldots,h^m_n;$$

Dabei ist p der Index eines Parasitenknotens, welcher bei seiner Abbildung auf den Knoten mit Index h^i die Kosten c^i verursacht. Ein Kindknoten $p.j$ von p wird bei dieser Rekonstruktion auf den Wirtsknoten mit Index h^i_j abgebildet. Ein Beispiel einer solchen Ausgabe findet sich in Kapitel 7.1.

Bei der Verwendung des Parameters -a werden noch Informationen zu jeder Teilberechnung angezeigt. Diese enthalten den Index des Wirtsknotens, auf den der Parasitenwurzelknoten bei der kostengünstigsten Rekonstruktion abgebildet wurde. Des Weiteren werden die berechnete Güte und die Gesamtkosten der Rekonstruktion angegeben, sowie die Ereigniskosten mit den dazugehörigen Wahrscheinlichkeiten und den aufgetretenen relativen Häufigkeiten. Beim Test auf chronologische Konsistenz wird noch das Ergebnis dieses Test angezeigt.

Zusätzlich kann aber mit dem Parameter -o eine Datei angegeben werden, in welche die berechnete kostenminimale Rekonstruktion im XML-Format gespeichert werden soll. Diese XML-Datei kann von der im Anschluss beschriebenen Applikation zur grafischen Ausgabe verwendet werden, um die berechnete Rekonstruktion anzuzeigen. Wird der Parameter nicht angegeben, so wird standardmäßig die Datei „output.xml“ erzeugt. Eine schon vorhandene Datei wird ohne Nachfrage überschrieben.

6.2 Grafische Ausgabe

Für die grafische Ausgabe der Ergebnisse wurde eine Adobe Flash Anwendung entwickelt. Mit dieser ist es möglich eine berechnete kostenminimale Rekonstruktion zu betrachten. Flash wurde als Framework verwendet, da es nach Installation eines frei verfügbaren Plugins mit allen gängigen Webbrowsern angezeigt werden kann.

Wie im Screenshot aus Abbildung 6.1 zu sehen ist, zeigt das Hauptfenster den auf den Wirtsstammbaum abgebildeten Parasitenstammbaum. Rechts davon werden zur Rekonstruktion gehörende Informationen eingeblendet. Dazu zählen der Zeitstempel der Berechnung und der Name der Nexusdatei in der die Ausgangsdaten gespeichert sind sowie die erzeugte Ausgabedatei. Hinzu kommen die Parameter, mit denen die Rekonstruktion berechnet wurde, die benutzten Ereigniskosten, die Anzahlen der aufgetretenen Ereignisse und die berechneten Gesamtkosten. Sollte auf chronologische Konsistenz geprüft worden sein, so wird auch das Ergebnis dieser Prüfung angegeben. Bei automatischer Berechnung der Kosten wird ebenfalls der berechnete Gütewert für diese Kostenverteilung dargestellt.

Wenn man mit dem Mauszeiger im Hauptfenster über einen der abgebildeten Knoten fährt, werden im Informationsfenster zu diesem Knoten gehörende Daten eingeblendet. Darunter sind der Knotenname und die gegebenenfalls vorhandenen Zeitinformationen. Bei Parasitenknoten werden zusätzlich die Kosten der bei diesem Knoten beginnenden Teilrekonstruktion sowie die dazugehörigen Anzahlen der verwendeten Ereignisse angezeigt. Auch sind die nur am dazugehörigen Wirtsknoten aufgetretenen Ereignisse ablesbar.

Zur besseren Rückverfolgbarkeit des Pfades von einem gerade selektierten Parasitenknoten zur Wurzel des Stammbaumes wird dieser farbig dargestellt. Zusätzlich ist es für jeden Knoten möglich durch Doppelklick den darunterliegenden Teilbaum auszublenden. Dies gilt sowohl für Wirts- als auch für Parasitenknoten. Wurde ein Teilbaum ausgeblendet, so erscheint in dem dazugehörigen Wurzelknoten ein Pluszeichen.

Im unteren Bereich des Fensters ist eine Menüleiste integriert. In dieser befindet sich eine Drop-Down-Box mit der man zwischen verschiedenen Rekonstruktionen hin und her wechseln kann. Die zu diesen Rekonstruktionen gehörenden Ausgabedateien müssen jedoch in einer separaten XML-Datei angegeben werden. Des Weiteren befinden sich in dieser Leiste Schaltflächen, welche das Zoomen im Hauptfenster sowie das An- und Abschalten des Informationsfensters am rechten Seitenrand und der Beschriftungen an den Knoten ermöglichen.

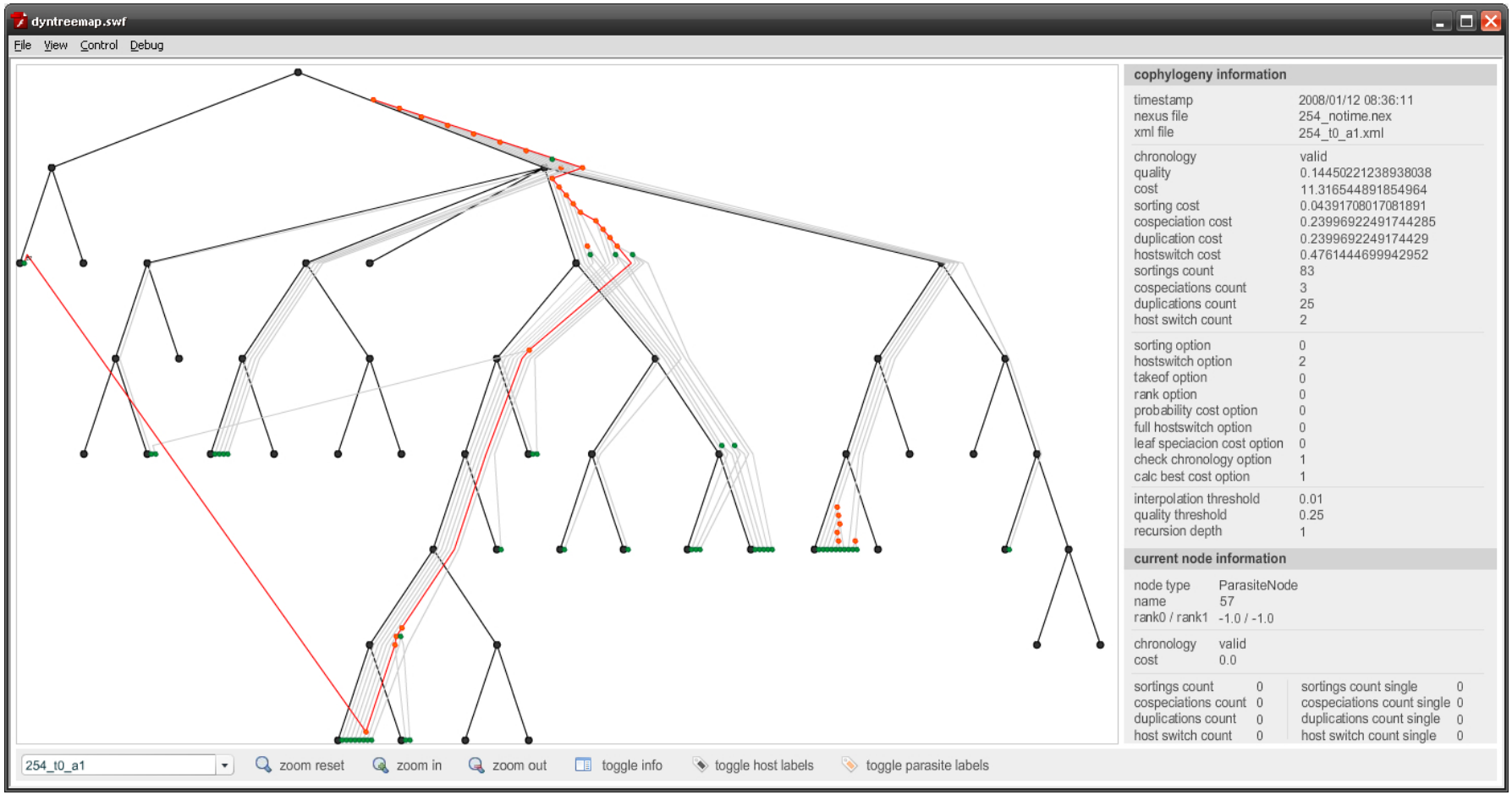


Abbildung 6.1: Screenshot der grafischen Ausgabe einer berechneten koevolutionären Rekonstruktion.

7 Beispielrechnungen

Im folgenden Kapitel sollen für einige Ausgangsdaten Beispielrechnungen und gefundene Rekonstruktionen vorgestellt werden. Die Rechnungen wurden auf einer Maschine mit Intel Pentuim 4 3000 MHz Prescott Prozessor durchgeführt. Alle Beispiele wurden mit den in [2] vorgeschlagenen Standardkosten von $co = -2$ für Kospeziationen, $so = 1$ für Sortings, $du = 2$ für Duplikationen und $hs = 2$ für Wirtswechsel durchgeführt. Diese Kosten entsprechen bei TreeMap ([6]) den Kosten von $c = -1$ für Kospeziationen und $s = 1$ für Sortings, $d = 1$ für Duplikationen und $h = 1$ für Wirtswechsel. Der Zusammenhang zwischen diesen unterschiedlichen Ereigniskosten wurde in [14] erläutert.[1]

Des Weiteren wurden die Kostenverteilungen mit dem in Kapitel 5 vorgestellten Verfahren automatisch berechnet. Dafür wurde eine Rekursionstiefe von 3 verwendet, sowie Schwellwerte für die Interpolation von 0.01 und für die minimale Güte von 0.25.

Alle Berechnungen mit den Parametern -h2, -s0, -t0, -f0, -z0, -l0, -c1 und -r0 durchgeführt. Diese Parameter entsprechen des von Tarzan verwendeten evolutionären Modelles.

7.1 Konstruiertes Beispiel von Charleston

Als erstes soll das von Charleston in [5] vorgestellte Standardbeispiel betrachtet werden. Im Wesentlichen geht es hierbei darum, die Gleichheit zu den von Tarzan gefundenen Rekonstruktionen aufzuzeigen.

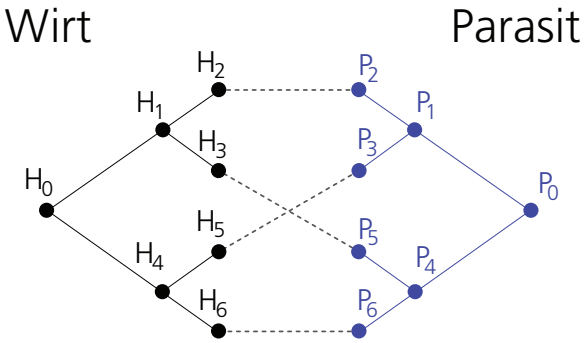


Abbildung 7.1: Konstruierte Phylogenien und Abbildung $\varphi_{P,H}$ des Standardbeispiels von Charleston.

[1] Da das verwendete Modell auf den in [14] beschriebenen Ereignissdefinitionen beruht gilt auch hier $co = 0.5 * c$, $so = s$, $du = 0.5 * d$, $hs = d + h$.

Die Berechnung mit Standardkosten erzeugt eine kostenminimale Rekonstruktion mit Gesamtkosten von 2. Dabei treten vier Sortings, zwei Kospeziationen und eine Duplikation auf.

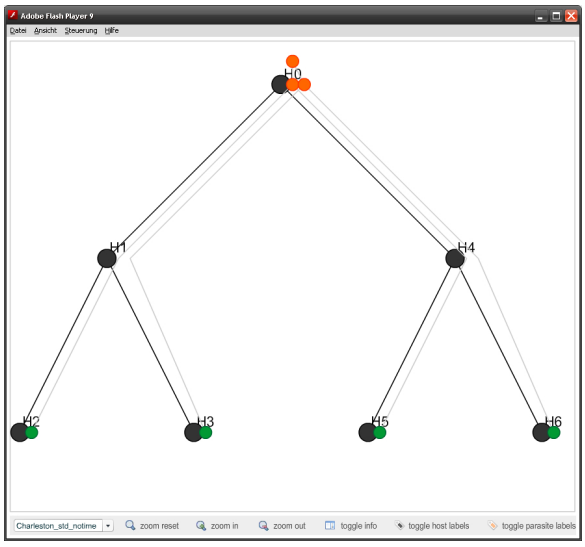
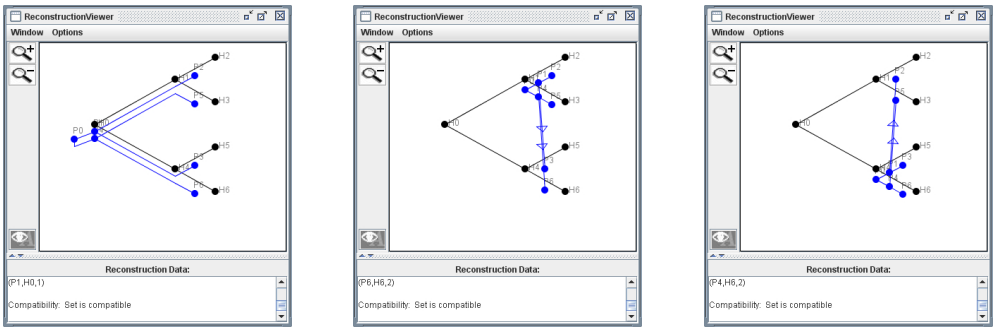


Abbildung 7.2: Kostenminimale Rekonstruktion des Standardbeispiels von Charleston (sowohl für Standard- als auch für automatische Kosten).

Tarzan berechnete folgende drei kostenminimale Rekonstruktionen.



(a) (b) (c)

Abbildung 7.3: Von Tarzan gefundene kostenminimale Rekonstruktion des Standardbeispiels von Charleston

Es ist ersichtlich, dass die Abbildungen 7.2 und 7.3 (a) die selbe Rekonstruktion beschreiben. Zwar wird von DynamicTreeMap immer nur eine Rekonstruktion ausgegeben, jedoch können die anderen beiden aus der folgenden Textausgabe der Komandozeile ausgelesen werden.

Ausgabe:

```
1: Nexus-Datei wird eingelesen.
2: Host-Baum wird geparst.
3: Parasite-Baum wird geparst.
4: Mapping der Blätter wird durchgeführt.
5: Zeitzonen werden zugewiesen.
6: Sorting- und Hostswitchkosten werden berechnet.
7: true 0 Kosten: 2.0 S(1.0): 4 C(-2.0): 2 D(2.0): 1 H(2.0): 0 Wahrscheinlichkeiten:
   S(0.0%): 57.14% C(0.0%): 28.57% D(0.0%): 14.29% H(0.0%): 0.0%
8: 0: 0:2.0-0,0; 1:2.0-2,3; 2:6.0-2,3; 3:6.0-2,3; 4:2.0-5,6; 5:6.0-5,3; 6:6.0-2,6;
9: 1: 0:0.0-2,5; 1:3.0-2,5; 2:2.0-2,5; 4:3.0-2,5; 5:2.0-2,5;
10: 2: 2:0.0;
11: 3: 5:0.0;
12: 4: 0:0.0-3,6; 1:3.0-3,6; 3:2.0-3,6; 4:3.0-3,6; 6:2.0-3,6;
13: 5: 3:0.0;
14: 6: 6:0.0;
15: Finish 1 calculations in 0 hours 0 minutes 0 seconds and 141 milliseconds
```

In Zeile 8 werden für den Wurzelknoten des Parasitenbaumes dessen mögliche Abbildungen in den Wirtsbaum aufgelistet. Dabei verweist „0:2.0-0,0;“ auf die ausgegebene Rekonstruktion. Die Wurzel wird mit Kosten 2.0 auf den Wirtsknoten mit Index 0 gesetzt. Die beiden Kindknoten werden ebenfalls auf den Wirtsknoten 0 abgebildet. Es existieren jedoch auch die Einträge „1:2.0-2,3;“ und „4:2.0-5,6;“. Diese entsprechen den Rekonstruktionen mit ebenfalls Kosten von 2.0, wobei die Wurzel diesmal auf die Wirtsknoten 1 bzw. 4 gesetzt wird. Die Kindknoten werden jeweils auf 2 und 3 bzw. 5 und 6 abgebildet. Diese Rekonstruktionen mit jeweils einer Kospeziation und zwei Wirtswechseln entsprechen denen aus den Abbildungen 7.3 (b) und (c).

Bei der Berechnung mit automatischen Kosten wurde die gleiche Rekonstruktion erzeugt, welche auch mit Standardkosten errechnet wurde. Wie Tabelle 7.1 zeigt, wurden dabei Ereigniskosten gefunden, bei denen die Verhältnisse zwischen den Kosten und den aufgetretenen Häufigkeiten fast exakt umgekehrt proportional zueinander sind. Daraus resultiert der sehr gute Wert von 0.00115 für die Güte diese Kostenverteilung. Es wurden 4186 Rekonstruktionen in 12 Sekunden und 468 Millisekunden berechnet.

	Ereigniskosten	Ereignisanzahl	Güte	Berechnungen	Laufzeit
Standard-kosten	co=-2.0 so=1.0 du=2.0 hs=2.0	co: 2 so: 4 du: 1 hs: 0	-	1	141ms
automatische Kosten Rekursions-tiefe 3	co=0.00347 so=0.00173 du=0.00691 hs=0.98790	co: 2 so: 4 du: 1 hs: 0	0.00115	4186	12sec 468ms

Tabelle 7.1:

7.2 Seabirds und Chewing Lice

Bei dem folgenden Beispiel handelt es sich um die in [19] verwendeten phylogenetischen Daten von Seevögeln und Kieferläusen. Diese wurden bereits in Abbildung 2.1 auf Seite 7 vorgestellt. Eine ausführliche Betrachtung bezüglich der koevolutionären Entwicklung beider Spezies wurde in [20] durchgefürht. Dabei wurden sowohl „Brooks Parsimony Analysis“ ([4]) als auch Page's „Reconciliation Analysis“([16]) verwendet.

Mit den Standardkosten wurde eine kostenminimale Rekonstruktion mit neun Kospeziationen, zwei Sortings und vier Wirtswechseln gefunden. Unter den 5503 Berechnungen mit automatischen Kosten erzeugte der beste Wert für die Kostenverteilung eine Rekonstruktion mit neun Kospeziationen, elf Sortings, drei Duplikationen und einem Wirtswechsel. Bis auf die take-off site des einen Wirtswechsels stimmt die von DynammicTreeMap berechnete Lösung mit der von Paterson vorgeschlagenen Rekonstruktion überein. Beide Varianten verursachen dabei die gleichen Gesammtkosten. Da der hier vorgestellte Algorithmus jedoch nur eine der kostenminimalen Rekonstruktionen betrachtet, wurde Patterson's Lösung zwar von DynamicTreeMap berechnet, jedoch nicht mit ausgegeben.

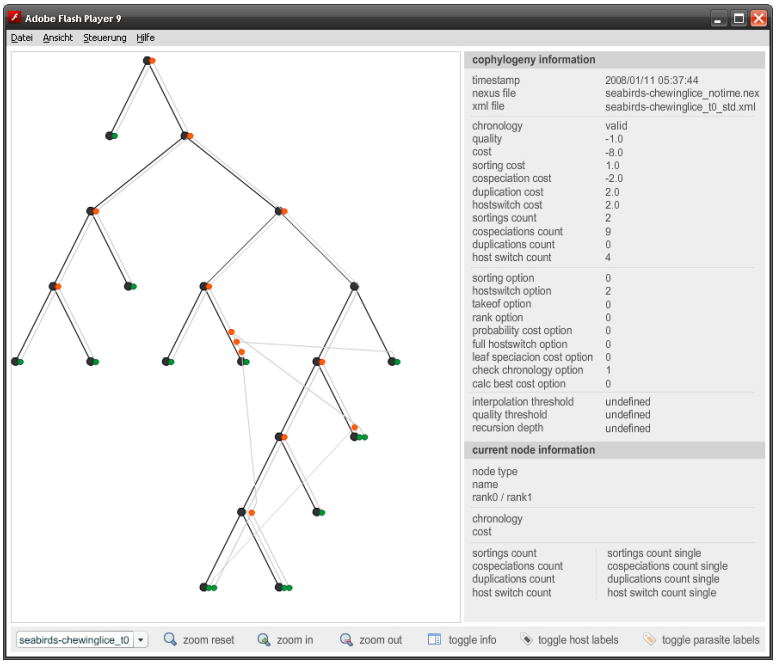


Abbildung 7.4: Rekonstruktion der koevolutionären Geschichte von Seevögeln und Kieferläusen mit Standardkosten.

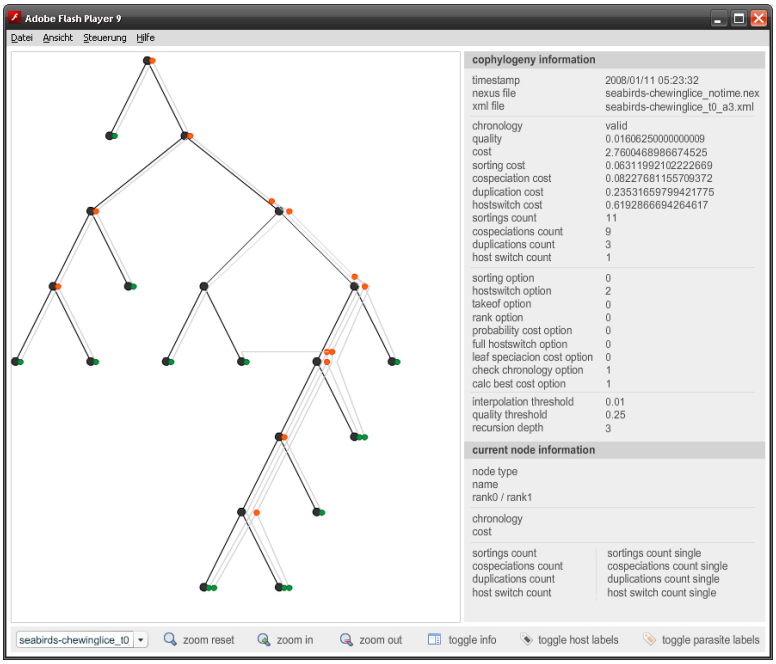


Abbildung 7.5: Rekonstruktion der koevolutionären Geschichte von Seevögeln und Kieferläusen mit automatisch berechneten Kosten und Rekursionstiefe 3.

	Ereigniskosten	Ereignisanzahl	Güte	Berechnungen	Laufzeit
Standard-kosten	co=-2.0 so=1.0 du=2.0 hs=2.0	co: 9 so: 2 du: 0 hs: 4	-	1	312ms
automatische Kosten Rekursions-tiefe 3	co=0.08228 so=0.06312 du=0.23532 hs=0.61929	co: 9 so: 11 du: 3 hs: 1	0.01606	5503	25sec 344ms

Tabelle 7.2:

7.3 Apis und Varroa

In [2] wurde der koevolutionäre Zusammenhang zwischen verschiedenen Bienenspezies und Varroamilben untersucht. Auf Basis der dort verwendeten Daten soll die Relevanz des Parameters der Rekursionstiefe für die automatische Berechnung der Ereigniskosten aufgezeigt werden. Dazu wurden die Rekonstruktionen mit Rekursionstiefen von 1 bis 7 berechnet. Wie aus Tabelle 7.3 ersichtlich wird, fand der Algorithmus in den ersten 5 Rekursionsstufen eine Rekonstruktion mit neun Kospeziationen, 69 Sortings und 16 Duplikationen. Dabei wurden lediglich leicht varierende Werte für die Verteilung der Ereigniskosten gefunden. Die berechnete kostenminimale Rekonstruktion blieb allerdings bis zu einer Rekursionstiefe von 5 gleich. Ab einer Tiefe von 6 wurde eine sich von den vorhergehenden Rekonstruktion unterscheidende Lösung gefunden. Diese verwendet 10 Kospeziationen, 49 Sortings, 11 Duplikationen und 4 Wirtswechsel.

Die Abbildungen 7.6 und 7.7 zeigen die berechneten Rekonstruktionen für Standardkosten und automatisch berechnete Kosten mit Rekursionstiefe 7. Es wird deutlich, dass bei der mit automatischen Kosten berechneten Lösung viel weniger Wirtswechsel auftreten. Grund dafür sind die im Verhälltnis zu den anderen Ereignissen höheren Wirtswechselkosten.

	Ereigniskosten	Ereignisanzahl	Güte	Berechnungen	Laufzeit
Standard-kosten	co=-2.0 so=1.0 du=2.0 hs=2.0	co: 13 so: 11 du: 1 hs: 13	-	1	516ms
automatische Kosten mit Rekursions-tiefe 1	co=0.23997 so=0.04392 du=0.23997 hs=0.47614	co: 9 so: 69 du: 16 hs: 0	0.09301	140	4sec 594ms
automatische Kosten mit Rekursions-tiefe 2	co=0.00783 so=0.00131 du=0.00783 hs=0.98302	co: 9 so: 69 du: 16 hs: 0	0.04471	933	27sec 391ms
automatische Kosten mit Rekursions-tiefe 3	co=0.00784 so=0.00136 du=0.00661 hs=0.98419	co: 9 so: 69 du: 16 hs: 0	0.03076	3994	1min 33sec 969ms
automatische Kosten mit Rekursions-tiefe 4	co=0.06922 so=0.01028 du=0.05394 hs=0.86656	co: 9 so: 69 du: 16 hs: 0	0.02915	19809	7min 30sec 266ms
automatische Kosten mit Rekursions-tiefe 5	co=0.00895 so=0.00134 du=0.00630 hs=0.98341	co: 9 so: 69 du: 16 hs: 0	0.01519	91713	34min 26sec 391ms
automatische Kosten mit Rekursions-tiefe 6	co=0.24728 so=0.04917 du=0.23345 hs=0.47009	co: 10 so: 49 du: 11 hs: 4	0.01503	412246	2h 52min 23sec 438ms
automatische Kosten mit Rekursions-tiefe 7	co=0.24117 so=0.04910 du=0.23438 hs=0.47536	co: 10 so: 49 du: 11 hs: 4	0.01406	1964573	11h 49min 22sec 188ms

Tabelle 7.3:

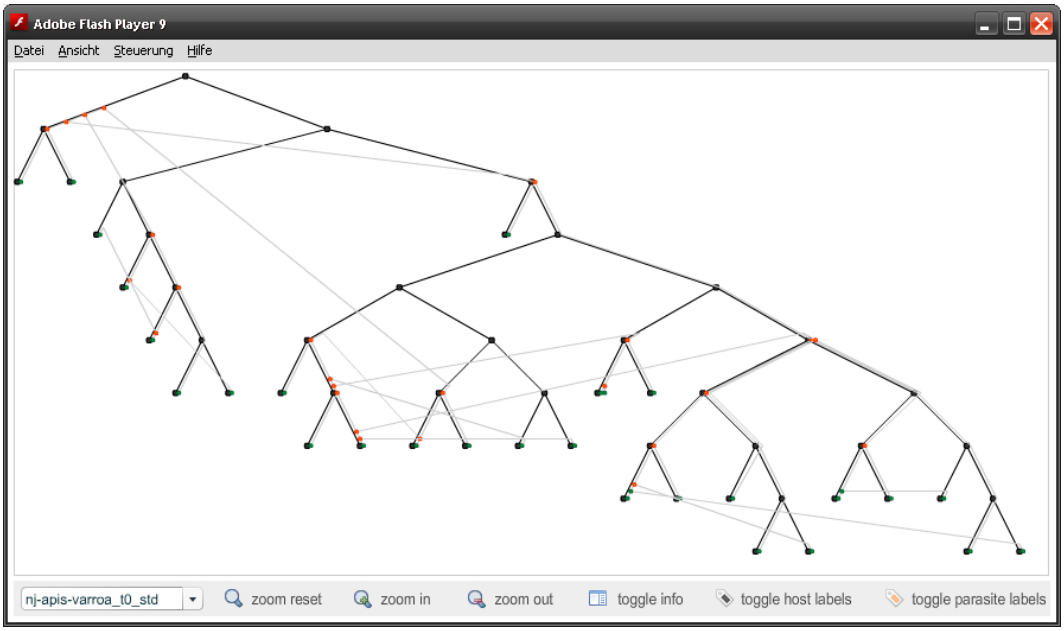


Abbildung 7.6: Rekonstruktion der koevolutionären Geschichte von Apis und Varroa mit Standardkosten.

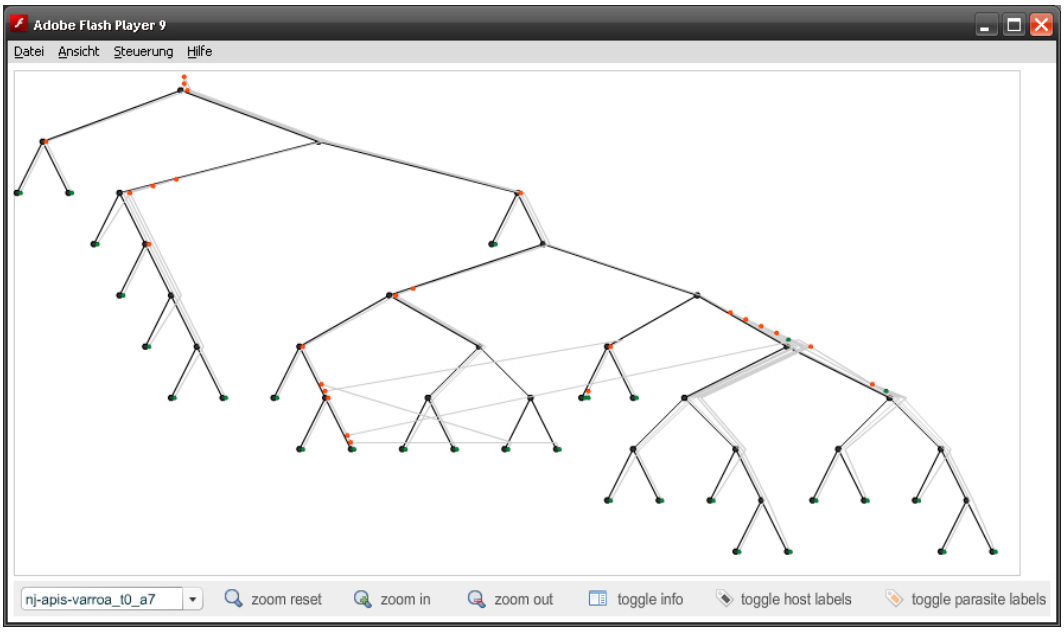


Abbildung 7.7: Rekonstruktion der koevolutionären Geschichte von Apis und Varroa mit automatisch berechneten Kosten und Rekursionstiefe 7.

Diese Rekonstruktion kann als Ausgangspunkt für weitere Betrachtungen verwendet werden. In Abbildung 7.8 ist exemplarisch versucht worden, den Bienenstammbaum geografisch darzustellen. Dabei wurden die Orte der inneren Knoten geschätzt. Es macht den Anschein, dass in diesem speziellen Fall die gefundenen Wirtswechsel alle in einem regional sehr begrenztem Gebiet um Borneo herum auftraten. Bei der mit Standardkosten berechneten Lösung war dies nicht der Fall.[2]

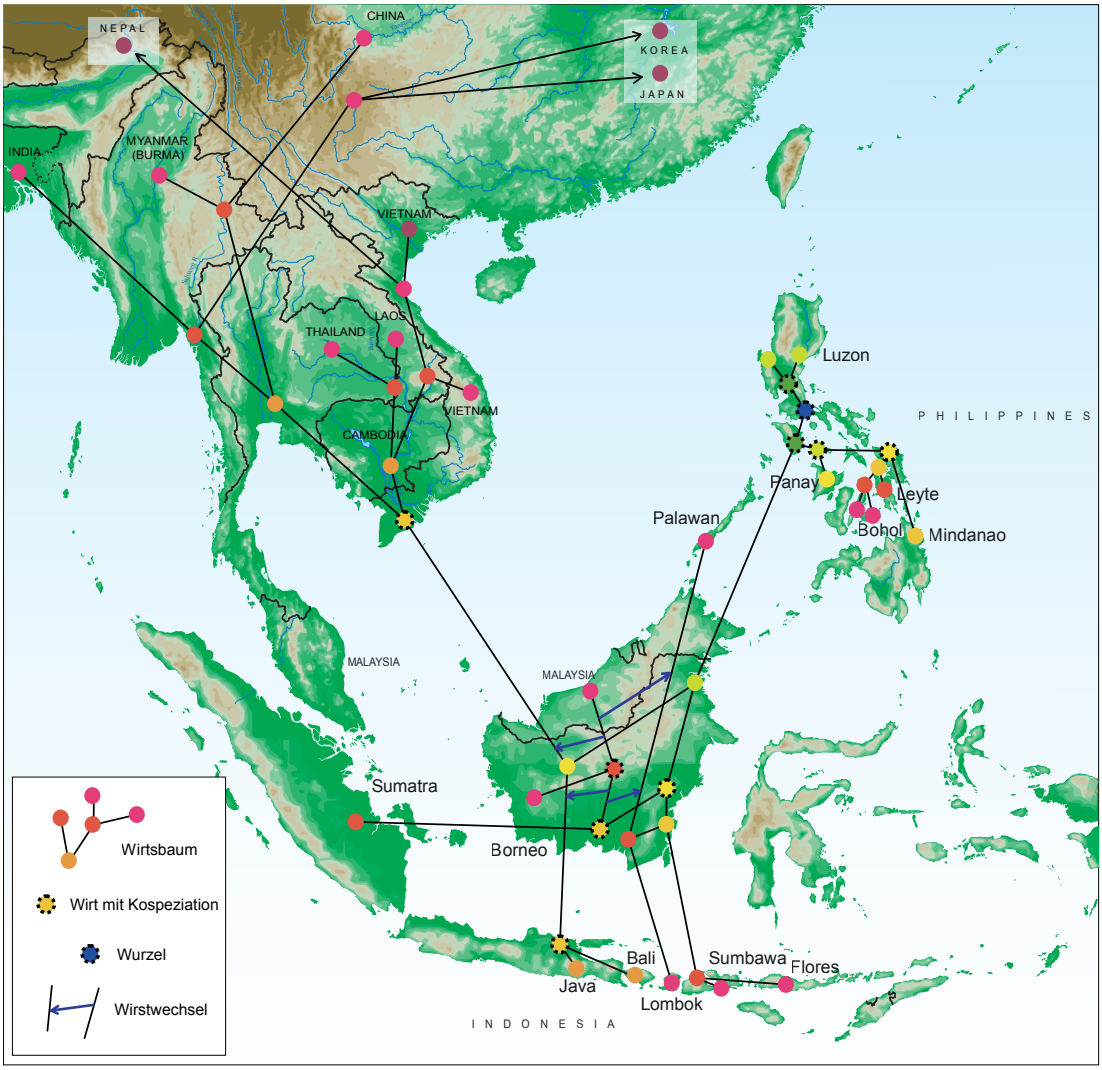


Abbildung 7.8: Mögliche geografische Abbildung der Wirtsphylogenie mit eingezeichneten Kospeziationen und Wirtswechsels der Parasiten.

[2]Die Einschätzung, inwieweit diese Aussage mit realen Gegebenheiten übereinstimmt, bleibt Biologen überlassen.

7.4 Legumes und Psyllids

Abschließend sollen noch Ausgangsdaten betrachtet werden, welche Multifurkationen in den Stammbäumen enthalten. Dazu wurden die Phylogenien von Blattflöhen und den sie beherbergenden Pflanzen betrachtet. Die Daten stammen aus [21] und wurden ebenfalls in [14] in einer binären Form untersucht.

Mit DynamicTreeMap wurden für die binäre und die Multifurkationen enthaltende Variante Berechnungen durchgeführt. Es wurden dabei keine Zeitinformationen verwendet. Es wurden die Rekonstruktionen sowohl mit Standardkosten, als auch mit automatisch berechneten Kosten und einer Rekursionstiefe von 1 bis 3 berechnet. Die gefundenen Rekonstruktionen mit Rekursionstiefe 3 sind in den Abbildungen 7.9 und 7.10 dargestellt. Beide Fälle ergaben Rekonstruktionen mit außergewöhnlich vielen Abbildungen unterschielicher Parasiten auf den Ursprungswirt. Dadurch sind deutlich mehr Sortings nötig. Der Wahrheitsgehalt einer solchen Rekonstruktion ist daher zweifelhaft. Zumindest zeigt das Beispiel die Laufzeitunterschiede zwischen den Berechnungen der binären und mehrfach verzweigenden Ausgangsdaten auf.

Wie der Tabelle 7.4 zu entnehmen ist, benötigten die Berechnungen der mehrfach verzweigenden Varianten nur geringfügig mehr Rechenzeit. Auch ist die Anzahl der Berechnungen pro Rekursionstiefe etwas höher als im binären Fall. Bei zunehmender Anzahl und Komplexität der Multifurkationen in den Stammbäumen können allerdings auch deutlich höhere Laufzeiten pro Berechnung entstehen.

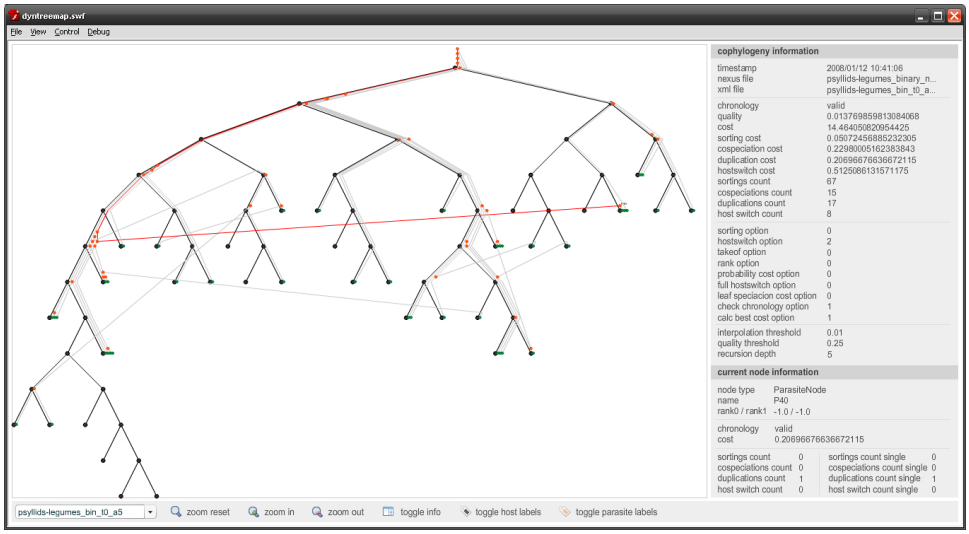


Abbildung 7.9: Rekonstruktion der koevolutionären Geschichte von Blattflöhen und ihren Wirtspflanzen, berechnet mit binären Phylogenien und Rekursionstiefe 3.

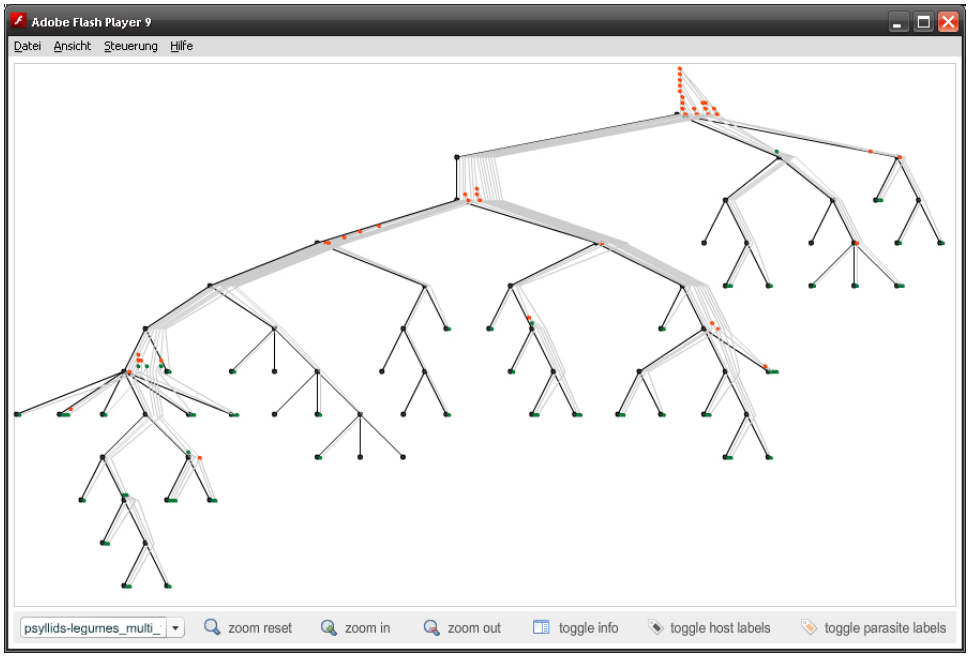


Abbildung 7.10: Rekonstruktion der koevolutionären Geschichte von Blattflöhen und ihren Wirtspflanzen, berechnet mit mehrfach verzweigenden Phylogenien und Rekursionstiefe 3.

	Ereigniskosten	Ereignisanzahl	Güte	Berechnungen	Laufzeit
Standard- kosten ohne Multifurk.	co=-2.0 so=1.0 du=2.0 hs=2.0	co: 16 so: 24 du: 4 hs: 20	-	1	687ms
Standard- kosten mit Multifurk.	co=-2.0 so=1.0 du=2.0 hs=2.0	co: 17 so: 13 du: 2 hs: 31	-	1	671ms
auto. Kosten Rekursions- tiefe 1 ohne Multifurk.	co=0.23999 so=0.04392 du=0.23999 hs=0.47614	co: 16 so: 61 du: 15 hs: 09	0.08179	138	9sec 31ms
auto. Kosten Rekursions- tiefe 1 mit Multifurk.	co=0.23997 so=0.04392 du=0.23997 hs=0.47614	co: 11 so: 70 du: 20 hs: 12	0.09444	137	11sec 0ms
auto. Kosten Rekursions- tiefe 2 ohne Multifurk.	co=0.00783 so=0.00131 du=0.00783 hs=0.98302	co: 15 so: 144 du: 25 hs: 0	0.04498	894	53sec 250ms
auto. Kosten Rekursions- tiefe 2 mit Multifurk.	co=0.00783 so=0.00131 du=0.00783 hs=0.98302	co: 14 so: 160 du: 29 hs: 0	0.05753	904	1min 8sec 109ms
auto. Kosten Rekursions- tiefe 3 mit Multifurk.	co=0.25948 so=0.05223 du=0.17344 hs=0.51485	co: 14 so: 75 du: 19 hs: 7	0.02867	3680	3min 8sec 859ms
auto. Kosten Rekursions- tiefe 3 mit Multifurk.	co=0.12793 so=0.15192 du=0.12793 hs=0.72895	co: 14 so: 160 du: 29 hs: 0	0.04848	3898	4min 7sec 312ms

Tabelle 7.4:

8 Zusammenfassung

In dieser Diplomarbeit wurde ein Algorithmus entwickelt, welcher auf Basis von dynamischer Programmierung für Wirts- und Parasitenstammbäume, für eine Abbildung $\varphi_{P,H}$ zwischen den Blättern und für gegebene Ereigniskosten eine kostenminimale Rekonstruktion der koevolutionären Geschichte erzeugt. Dafür ist eine Erweiterung des in der Forschung diskutierten Modells vorgenommen wurden, mit deren Hilfe es möglich ist, Ausgangsdaten mit Multifurkationen in den Stammbäumen zu verwenden. Diese Erweiterung wurde so konzipiert, dass der entwickelte Algorithmus im Falle binärer Stammbäume die gleichen Rekonstruktionen wie auch andere algorithmische Ansätze erzeugt.[1] Zur Prüfung der Rekonstruktionen bezüglich ihrer chronologischen Konsistenz, wurde ein Verfahren aufgezeigt, mit welchem sich diese Inkompatibilitäten durch Vergleiche jeweils zweier parasitärer Lebenslinien erkennen lassen.

Es wurde eine Vorgehensweise vorgestellt, mit der es möglich ist, ohne Informationen über Ereigniskosten zu, diese automatisch zu bestimmen. Dazu konnte ein Gütemaß aufgestellt werden, welches die Kostenverteilungen anhand der in der kostenminimalen Rekonstruktion aufgetretenen Anzahlen der Ereignisse bewertet. Auf diese Weise wurde die Kostenverteilung der Ereignisse direkt mit der Wahrscheinlichkeit des Auftretens dieser Ereignisse verknüpft. Mit der so berechneten Güte konnte somit ein Maß für die Qualität der gefundenen Rekonstruktion zur Verfügung gestellt werden.

Die beschriebenen algorithmischen Verfahren wurden in einem Javaprogramm namens DynamicTreeMap umgesetzt. Um diese Anwendung für ein breiteres Spektrum an Problemklassen nutzen zu können, wurden zahlreiche optionale Varianten implementiert, mit welchen sich die Verfahrensweise im Speziellen bei Wirtswechselereignissen genau spezifizieren lässt.

Zur Visualisierung der Ergebnisse wurde eine Flashanwendung entwickelt, welche die relevanten Informationen einer berechneten Rekonstruktion übersichtlich darstellt. Durch das Aus- und Einblenden von Teilen der Rekonstruktion, eine Zoomfunktionalität und dem Hervorheben der Pfade von fokussierten Knoten zur Wurzel lassen sich auch komplexe Rekonstruktionen einfach visuell erfassen.

[1] Primär sollten in diesem Fall die gleichen Berechnungen betrachtet werden, welche auch von dem in [14] vorgestellten Programm „Tarzan“ erzeugt wurden.

Als zukünftige Erweiterung ist geplant nicht nur eine, sondern alle kostenminimalen Rekonstruktionen auszugeben. Dies erfordert keine großen algorithmischen Anpassungen, jedoch wird sowohl der Speicherbedarf sowie der Berechnungsaufwand bei der Prüfung auf chronologische Inkompatibilität deutlich erhöht. In der momentanen Variante kann es gegebenenfalls vorkommen, dass eine kostenminimale Rekonstruktion erzeugt wird, welche chronologisch inkompatibel ist. Dies muss aber nicht bedeuten, dass nicht auch eine chronologisch gültige Rekonstruktion mit den gleichen Kosten existieren kann.

Des Weiteren sollen zusätzliche Optionen integriert werden, mit denen Lösungen auf ein bestimmtes Ereignis hin optimiert werden können. Somit könnte man beispielsweise nach Rekonstruktionen suchen, welche die maximale Anzahl möglicher Kospeziationen bei gleichzeitiger Beibehaltung der gewählten Kostenwerte enthalten.

Ebenfalls ist angedacht die Anwendung serverseitig über einen Browser zur Verfügung zu stellen. Über ein PHP-Skript könnten die Ausgangsdaten und die gewählten Optionen an einen zentralen Rechner übermittelt werden, welcher die Berechnung übernimmt und im Ergebnis einen Verweis auf die mit den Ergebnisdaten vorkonfigurierte Flashanwendung zurück gibt.

Für die einfachere Bedienbarkeit ist noch geplant die Parameter direkt in der grafischen Ausgabe über Schieberegler anpassen zu können. Somit kann jede Veränderung in den Rekonstruktionen sofort betrachtet werden. Für Anwender wäre dies ein geeignetes Mittel, um für spezielle Ausgangsdaten den Einfluß der Ereigniskosten auf die zugehörige Rekonstruktion sofort beurteilen zu können.

Eine sinnvolle Ergänzung der Funktionalität könnte weiterhin durch die Integration von statistischen Tests erzielt werden. Diese Tests, welche schon in den Programmen TreeMap und Tarzan zur Verfügung stehen, können helfen die Signifikanz einer gefundenen Lösung bezüglich der darin enthaltenen Anzahl an Kospeziationen im Vergleich zu zufällig gewählten Ausgangsdaten zu bestimmen. Dazu werden nach bestimmten Kriterien randomisierte Ausgangsdaten erzeugt. Diese können im Vergleich zu den gegebenen Daten einerseits in der Abbildung $\varphi_{P,H}$ der Blätter variieren.[2] Andererseits können auch zufällige Stammbäume für die Wirts- und/oder Parasitenphylogenie erzeugt werden.[3] Signifikante Unterschiede zwischen den Original- und Zufallsdaten können somit Aufschluss auf die Plausibilität einer gefundenen Rekonstruktion und auf den Wahrheitsgehalt der Ausgangsdaten geben.

[2] Diese Vorgehensweise wurde in [26] näher betrachtet.
[3] Verfahren dazu finden sich in [1].

In Kombination mit der hier vorgestellten Methode zur automatischen Berechnung der Ereigniskosten wäre dies ein geeignetes Mittel um speziesabhängige Unterschiede zwischen gefundenen Kostenverteilungen zu untersuchen.

Durch die schnelle Berechnung einzelner Rekonstruktionen wäre es weiterhin denkbar, nicht nur die Ereigniskosten wie beschrieben anzupassen, sondern ebenfalls Veränderungen an den Ausgangsdaten vorzunehmen. Wenn auf diese Weise deutlich bessere Lösungen entstehen, gäbe dies einen Hinweis auf eventuelle Fehler in den Ausgangsdaten. Auch könnte man einzelne kleinere Parasitenteilbäume testweise aus den betrachteten Stammbäumen entfernen und die Unterschiede zwischen den gefundenen Rekonstruktionen betrachten. Sind diese sehr gering, könnte auf eine recht stabile und damit wahrscheinlichere Rekonstruktion geschlossen werden.

Um einen Überblick über die gefundenen Lösungen und deren Güte mit verschiedenen Ereigniskosten zu bekommen, wäre es zudem nützlich eine Visualisierung des 3-dimensionalen Raumes der Ereigniskostenverteilung zu erstellen. Einerseits könnte dort durch Farbverläufe die Güte der gefundenen Rekonstruktionen dargestellt werden. Andererseits wäre es ebenfalls sinnvoll Bereiche zu markieren, in denen die gleichen Rekonstruktionen erzeugt werden. Im ersten Fall kann somit ein Überblick über die Stabilität der Kostenverteilung, im zweiten Fall über die der Rekonstruktion gegeben werden. Der Wahrheitsgehalt einer gefundenen Bestlösung für die Kostenverteilung wäre somit besser einschätzbar. Je größer der um diese Lösung herumliegende Bereich gleicher Rekonstruktionen ist und je weniger andere Kostenverteilungen hoher Güte existieren, desto plausibler ist das Ergebnis.

Abschließend wird deutlich, dass mit den vorgestellten Erweiterungen zu bestehenden Verfahren, wie z.B. der automatischen Kostenberechnung und der Betrachtung von Multifurkationen, mit Hilfe von einer schnellen Implementierung ein dynamischer Ansatz entwickelt wurde, welcher ein geeigntes Instrument für koevolutionäre Forschung zu Verfügung stellt.

Literaturverzeichnis

[1] Aldous, D.: *Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today*. Statistical Science, 16(1):23–34, 2001.

[2] Anderson, D., S. Junick, D. Merkle und M. Middendorf: *Apis cerana/Varroa and Apis dorsata/Tropilaelaps Cophylogenies*. Artikel noch in Bearbeitung, 2006.

[3] Bellman, R.: *Dynamic programming*. Princeton Univ. Pr., 3. Aufl., 1962.

[4] Brooks, D.: *Parsimony Analysis in Historical Biogeography and Coevolution: Methodological and Theoretical Update*. Systematic Zoology, 39(1):14–30, 1990.

[5] Charleston, M.: *Jungles: a new solution to the host/parasite phylogeny reconciliation problem*. Mathematical Biosiences, 149:191–223, 1998.

[6] Charleston, M. und R. Page: *A Macintosh program for the analysis of how dependent phylogenies are related, by cophylogeny mapping*, 2002. http://www.cs.usyd.edu.au/ mcharles/software/treemap/treemap.html.

[7] Charleston, M. und S. Perkins: *Lizards, Malaria, and Jungles in the Caribbean*. In: Page, R. (Hrsg.): *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, S. 65–92. The University of Chicago Press, 2003.

[8] Charleston, M. und S. Perkins: *Traversing the tangle: Algorithms and applications for cophylogenetic studies*. Journal of Biomedical Informatics, 39:62–71, 2006.

[9] Clayton, D., S. Al-Tamimi und K. Johnson: *The ecological basis of coevolutionary history*. In: Page, R. (Hrsg.): *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, S. 287–309. The University of Chicago Press, 2003.

[10] Fahrenholz, H.: *Ectoparasiten und Abstammungslehre*. Zoologischer Anzeiger, 41:371–374, 1913.

[11] Junick, S., D. Merkle und M. Middendorf: *Tarzan. Phylogeniesoftware zur Ermittlung von Cophylogenien*, 2005. http://pacosy.informatik.uni-leipzig.de/pv/Software/Tarzan/PV-Tarzan.html.

[12] Legat, R.: *Datenstrukturen zur Analyse der Phylogenie von Parasit-Wirt-Beziehungen*. Diplomarbeit, Institut für Informatik, Universität Hannover, 2001.

[13] Light, J.: *Host-parasite cophylogeny and rates of evolution in two rodent-louse assemblages*. Diplomarbeit, Department of Biological Sciences, University of Michigan, 2005.

[14] Merkle, D. und M. Middendorf: *Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information*. Theory in Biosciences, 123:277–299, 2005.

[15] Ottmann, T. und P. Widmayer: *Algorithmen und Datenstrukturen*. Spektrum Akademischer Verlag GmbH, 3. Aufl., 1996.

[16] Page, R.: *Parallel phylogenies: reconstructing the history of host-parasite assemblages*. Cladistics, 10:155–173, 1994.

[17] Page, R.: *Tangled Trees: Phylogeny, Cospeciation and Coevolution*. The University of Chicago Press, 2003.

[18] Page, R. und M. Charleston: *Trees within trees: phylogeny and historical associations*. Trends in Ecology and Evolution, 13(9):356–359, 1998.

[19] Paterson, A., R. Palma und R. Gray: *Drowning on arrival, missing the boat, and x-events: How likely are sorting events?*. In: Page, R. (Hrsg.): *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, S. 287–309. The University of Chicago Press, 2003.

[20] Paterson, A., G. Wallis, L. Wallis und R. Gray: *Seabird and Louse Coevolution: Complex Histories Revealed by 12S rRNA Sequences and Reconciliation Analyses*. Systematic Biology, 49:383–399, 2000.

[21] Percy, D.: *Diversification of legume-feeding psyllids (Psylloidea, Hemiptera) and their host plants (Leguminosae, Genisteae)*. Diplomarbeit, University of Glasgow, 2001.

[22] Ronquist, F.: *Reconstructing the history of host-parasite associations using generalised parsimony*. Cladistics, 11:73–89, 1995.

[23] Ronquist, F.: *TreeFitter 1.0*, 2001. http://www.ebc.uu.se/systzoo/research/treefitter/treefitter.html.

[24] Ronquist, F.: *Parsimony analysis of coevolving species associations*. In: Page, R. (Hrsg.): *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, S. 22–64. The University of Chicago Press, 2003.

[25] Schaback, R. und H. Wendland: *Nummerische Mathematik*. Springer, 5. Aufl., 2005.

[26] Siddall, M.: *Computer-Intensive Randomization in Systematics*. Cladistics, 17:35–52, 2001.

[27] Swofford, D.: *NEXUS: an extensible file format for systematic information*. Systematic Biology, 46(4):590–621, 1997.

[28] Swofford, D.: *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*, 2003.
http://paup.csit.fsu.edu/nfiles.html.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Leipzig, 17. Januar 2008