Komplexität von Haplotypisierung mittels perfekten Phylogenien und kleinsten Haplotypmengen

Diplomarbeit
im Rahmen des Diplomstudiengangs Informatik

vorgelegt von
Michael Elberfeld

Betreuer: Prof. Dr. Till Tantau

Institut für Theoretische Informatik
Technisch-Naturwissenschaftliche Fakultät
Universität zu Lübeck

Lübeck, September 2007

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Lübeck, ...

Inhaltsverzeichnis

1 Einleitung

In dieser Arbeit wird die Komplexität von Haplotypisierungsverfahren, die auf perfekten Phylogenien und kleinsten Haplotypmengen basieren, untersucht.

Haplotypisierungsverfahren sind rechnergestützte Verfahren, die in der Humangenomforschung verwendet werden, um die Erbinformation eines Menschen möglichst genau und trotzdem mit geringem Aufwand zu ermitteln. Für die Erforschung des Zusammenhangs zwischen der Varianz im menschlichen Genom und dem Auftreten von Krankheiten sind sie dabei Teil eines ersten Schrittes, in dem die Erbinformation bei einer Gruppe von Menschen ermittelt wird. Das Ermitteln der Erbinformation geschieht hier im Labor und die Erbinformation liegt je nach verwendeter Methode in einer von zwei verschiedenen Formen vor. Diese beiden Formen der Kodierungen existieren, da das menschliche Genom auf Chromosomen verteilt ist, die zu Chromosomenpaaren gruppiert sind. Entweder man erhält pro Chromosom einen Haplotyp, der genau die Erbinformation in dem Chromosom beschreibt oder man erhält pro Chromosomenpaar einen Genotyp, der die Erbinformation in dem Chromosomenpaar auf eine kombinierte Weise beschreibt. Da die Haplotypen, im Gegensatz zu den Genotypen, die Erbinformation eines Menschen auf eindeutige Weise beschreiben, zieht man bei der Erforschung der menschlichen Genomvarianz die Haplotypen den Genotypen vor. Das Problem ist aber, dass man im großen Umfang nur Genotypen relativ günstig im Labor auslesen kann. Das Ermitteln von Haplotypen ist aufwendiger [32].

Die Idee ist nun, die Erbinformation zuerst in Form von Genotypen auszulesen und danach die zugrunde liegenden Haplotypen aus den Genotypen zu berechnen. Solche rechnergestützten Verfahren, die für Genotypen zugrunde liegende Haplotypen berechnen, nennt man Haplotypisierungsverfahren. Einige Verfahren verwenden hierbei statistische Methoden und versuchen zum Beispiel die Verteilung von Haplotypen in einer Population zu ermitteln, um daraufhin zugrunde liegende Haplotypen zu wählen [4]. Andere Ansätze sind rein kombinatorisch und suchen nach Haplotypen, die bestimmte Eigenschaften erfüllen.

In dieser Arbeit werden verschiedene kombinatorische Ansätze zur Haplotypisierung betrachtet. Zum einen die Haplotypisierung mittels kleinsten Haplotypmengen, bei der man möglichst wenige paarweise verschiedene Haplotypen ermittelt. Zum anderen die Haplotypisierung mittels perfekten Phylogenien, bei der man Haplotypen sucht, die sich als perfekte Phylogenie (Stammbaum mit zusätzlichen Eigenschaften) anordnen lassen. Jeder dieser Ansätze basiert auf evolutionstheoretischen Annahmen, durch die bis zu einem bestimmten Grad gesichert ist, dass die berechneten Haplotypen den real vorliegenden Haplotypen entsprechen. Durch Kombination der beiden Ansätze versucht man diesen Grad der Sicherheit, dass die berechneten Haplotypen den real Vorliegenden entsprechen, zu erhöhen. Der daraus resultierende Ansatz ist die Haplotypisierung mittels minimalen perfekten Phylogenien, bei der man möglichst wenige paarweise verschiedene Haplotypen sucht, die sich als perfekte Phylogenie anordnen lassen.

1.1 Ziele und Beiträge der Arbeit

Wenn man die Haplotypisierung mit dem Ansatz über kleinste Haplotypmengen, dem Ansatz über perfekte Phylogenien oder dem kombinierten Ansatz verwenden möchte, taucht die Frage auf, wie groß der Aufwand ist, um einen dieser Lösungsansätze zu berechnen. Lässt sich ein Ansatz effizient berechnen, effizient parallelisieren oder brauchen wir nicht zu hoffen, dass wir den Ansatz effizient berechnen können? Das Ziel der vorliegenden Arbeit ist es, diese Frage möglichst umfassend mit den Werkzeugen der Komplexitätstheorie zu beantworten. Das heißt, die Komplexität der Ansätze möglichst genau durch obere und untere Schranken einzuordnen. Hierzu werden bekannte Resultate aus der Literatur angegeben und es werden neue Resultate erarbeitet. Zu beachten ist, dass für die Ansätze zur Haplotypisierung nicht die entsprechenden Konstruktionprobleme (Haplotypen mit bestimmten Eigenschaften für Genotypen konstruieren), sondern die entsprechenden Entscheidungsprobleme (entscheiden, ob es möglich ist, Haplotypen mit bestimmten Eigenschaften für Genotypen zu konstruieren) untersucht werden.

Folgende neue Resultate werden in dieser Arbeit vorgestellt: Es wird gezeigt, dass die Haplotypisierung mittels perfekten Phylogenien in Mod_2L liegt und L-hart ist, wobei die Beweise hierzu mündlich überlieferten Beweisskizzen von Arfst Nickelsen und Till Tantau basiert. Für zwei Teilprobleme der Haplotypisierung mittels perfekten Phylogenien wird außerdem gezeigt, dass sie in FO liegen. Lässt man bei dem Ansatz mit perfekten Phylogenien nur Phylogenien zu, die die Form eines Pfades haben, so erhält man den Ansatz mit perfekten Pfadphylogenien. In dieser Arbeit wird gezeigt, dass die gerichtete Variante dieses Problems, bei der mindestens ein Haplotyp aus der Population bekannt ist, in FO liegt. Der Beweis hierzu baut auf Resultaten von Gramm et al. [17] auf. Auch bei dem Ansatz mit gerichteten perfekten Pfadphylogenien lässt sich nach möglichst wenigen paarweise verschiedenen Haplotypen suchen. Das Hauptresultat dieser Arbeit ist, dass dieses Problem, die Haplotypisierung mittels minimalen gerichteten perfekten Pfadphylogenien, in L liegt.

1.2 Aufbau der Arbeit

Neben diesem einleitenden Abschnitt ist diese Arbeit in drei Abschnitte aufgeteilt. Zuerst werden in Abschnitt 2 das Haplotypisierungsproblem und die verschiedenen Lösungsansätze zur Haplotypisierung genau beschrieben. Zusammen mit den Lösungsansätzen werden dort auch die evolutionstheoretischen Annahmen, auf welchen die Lösungsansätze basieren, beschrieben. Außerdem werden in Abschnitt 2 die verschiedenen Lösungsansätze zur Haplotypisierung als Entscheidungsprobleme definiert. Als nächstes folgt Abschnitt 3, in dem die Komplexität der verschiedenen Haplotypisierungsprobleme betrachtet wird. In diesem Abschnitt werden neben den Beiträgen dieser Arbeit auch bekannte Resultate aus der Literatur vorgestellt. Um die verschiedenen Resultate nach Ansätzen sortiert darzustellen und vergleichen zu können, ist Abschnitt 3 in drei Teile gegliedert, von denen der er-

ste Teil die Komplexität der Haplotypisierung mittels kleinsten Haplotypmengen, der zweite Teil die Komplexität der Haplotypisierung mittels perfekten Phylogenien und der dritte Teil die Komplexität der kombinierten Ansätze behandelt. Die Arbeit schließt mit Abschnitt 4 ab, in dem eine Übersicht über die komplexitätstheoretischen Resultate zu den Haplotypisierungproblemen gegeben wird. Außerdem wird an dieser Stelle diskutiert, für welche Probleme noch genauere Resultate wünschenswert sind.

2 Das Haplotypisierungsproblem

In diesem Abschnitt wird das Problem der Haplotypisierung von Populationen eingeführt. Die ersten beiden Teilabschnitte geben ohne formale Definitionen eine Einführung in das Haplotypisierungsproblem. Ziel dieser beiden Abschnitte ist es, das Problem und seine Lösungen im biologischen Kontext zu motivieren und zu beschreiben. Im dritten Abschnitt wird das Problem durch formale Definitionen eingeführt. Ziel dieses Abschnitts ist es, die Haplotypisierungsprobleme so als kombinatorische Probleme zu formulieren, dass eine Einordnung bezüglich ihrer Komplexität möglich wird. Die komplexitätstheoretische Einordnung der Haplotypisierungprobleme behandelt dann Abschnitt 3.

2.1 Biologische Problemstellung

Dieser Abschnitt gibt eine Einführung in die biologische Problemstellung der Haplotypisierung von Populationen. Zuerst wird das Teilgebiet der Biologie, aus dem das Problem stammt, beschrieben. Danach wird das Problem selbst beschrieben

Das menschliche Genom ist die bei der Fortpflanzung vererbbare Information eines Menschen. Das Genom liegt in Form von Desoxyribonukleinsäure (DNS) im Zellkern einer jeden Körperzelle vor und ist dabei auf mehrere DNS-Fäden aufgeteilt. Die biologische Struktur, die einen DNS-Faden enthält, wird als Chromosom bezeichnet. Ein Mensch besitzt 23 Chromosomenpaare, die jeweils aus einem Chromosom, das von mütterlicher Seite vererbt wurde, und einem Chromosom, das von väterlicher Seite vererbt wurde, bestehen.

Die DNS ist eine Kette von Nukleotiden, die jeweils eine Base enthalten. Eine Base besitzt eine von maximal vier Ausprägungen, welche Adenin (A), Guanin (G), Thymin (T) und Cytosin (C) sind und Allele genannt werden. Die DNS wird entsprechend der vorkommenden Allele in den Nukleotiden durch eine Folge von Allelzeichen beschrieben. Die Position eines Nukleotids in der DNS nennt man Basenposition.

Ein Haplotyp beschreibt die Erbinformation, die in einem Chromosom vorliegt. Zum Beispiel ist $ATCCC$ ein Haplotyp, der einen Chromosomenabschnitt beschreibt, der aus den Basen Adenin, Thymin, und dreimal Cytosin besteht. Ein Genotyp beschreibt die Erbinformation, die in einem Chromosomenpaar vorliegt, durch Kombination zweier Haplotypen. Für zwei Haplotypen AAT und AGC ist $A\{A,G\}\{T,C\}$ der zugehörige Genotyp. Der Genotyp ist an der ersten Position homozytisch, da die beiden Haplotypen an dieser Position das gleiche Allel enthalten. An den Positionen zwei und drei ist der Genotyp heterozytisch, da die beiden Haplotypen an diesen Positionen unterschiedliche Allele enthalten. Aus zwei Haplotypen lässt sich auf eindeutige Weise der zugehörige Genotyp ermitteln, aber aus einem Genotyp lassen sich die zugehörigen Haplotypen nicht immer eindeutig bestimmen. Zum Beispiel können zu dem Genotyp $A\{A,G\}\{T,C\}$ auch die Haplotypen AAC und AGT gehören. Falls nur der Genotyp eines Chromosomenpaares bekannt ist, ist also nicht klar, welches Allel an einer heterozytischen Position

zu welchem Chromosom gehört. Die tatsächlich zugrunde liegenden Haplotypen lassen sich für einen Genotyp mit zwei oder mehr heterozytischen Stellen nicht eindeutig bestimmen.

Das Genom zweier Menschen variiert annähernd an 0.1% der Basenpositionen [9]. Die Variation entsteht durch verschiedene Mutationsarten, von denen im Folgenden die Punktmutation und die Rekombination erläutert werden.

Bei einer Punktmutation wird das Allel an genau einer Basenposition verändert. Zum Beispiel geht der Haplotyp *AGC* durch eine Punktmutation an der zweiten Position aus dem Haplotyp *AAC* hervor. Als SNP (single nucleotide polymorphism – gesprochen: „snip“) bezeichnet man eine Basenposition, an der mindestens zwei Allele mit einer Frequenz von jeweils mehr als 1% in einer Population auftreten. Einen SNP kann man als die Position einer erfolgreichen Punktmutation ansehen, die sich in mehr als einem Prozent einer Population durchgesetzt hat. Das menschliche Genom enthält ungefähr 10 Millionen SNPs (auf etwa 300 Basenpositionen ein SNP). Die SNPs bilden 90% der menschlichen Genomvarianz [9] und daher wird ihnen eine große Bedeutung bei der Untersuchung genetisch bedingter Krankheiten beigemessen. An weniger als 0.1% der SNPs treten drei verschiedene Allele auf [27]. An den meisten SNPs treten nur zwei verschiedene Allele auf. Aus diesem Grund werden im Folgenden nur solche SNPs betrachtet, an denen genau zwei verschiedene Allele auftreten. Ein solcher SNP heißt biallelisch. Durch die Einschränkung auf biallelische SNPs lassen sich Haplotypen als binäre Zeichenketten und Genotypen als Zeichenketten über dem Alphabet $\{0, 1, 2\}$ kodieren, wobei der Eintrag 2 für eine heterozytische Stelle steht.

Das Internationale HapMap Projekt ermittelt mithilfe von SNPs eine Karte häufig auftretender Genomvarianzen [9]. Die Beschreibung der Varianz im menschlichen Genom wird in zwei Schritten erarbeitet. Im ersten Schritt werden alle Basenpositionen, die keine SNPs sind, als konstant angenommen und ausgespart. Im zweiten Schritt werden die SNPs auf wenige SNPs, welche die Genomvarianz noch ausreichend beschreiben und markierende SNPs genannt werden, reduziert. Durch die relativ geringe Anzahl von SNPs im Genom und die Reduktion auf markierende SNPs wird eine hohe Datenreduktion erreicht. Wenn die Haplotypenkarte vollständig vorliegt, benötigt man zum Bestimmen der Erbinformation eines Menschen nur noch die Allele der markierenden SNPs, da sich die Allele an den übrigen Basenpositionen aus den markierenden SNPs ergeben.

Bei der Rekombination werden zwei Chromosome aufgetrennt und in anderer Kombination neu verbunden. Das Crossing-over, ein Spezialfall von Rekombination, kann bei der Meiose, bei der eine Zelle mit einem diploiden Chromosomensatz in zwei Zellen mit einem haploiden Chromosomensatz geteilt wird, auftreten. Bei einem Crossing-over werden die Chromosomen eines Chromosomenpaares zwischen zwei Basenpositionen aufgetrennt und über Kreuz neu verbunden. Tritt zum Beispiel bei dem Haplotyppaar $AATTT$ und $CCAGG$ zwischen der zweiten und dritten Position ein Crossing-over auf, so resultiert dies in dem Haplotyppaar $AAAGG$ und $CCTTT$.

Die individuelle Ausprägung der Erbinformation eines Menschen steht im Zu-

sammenhang mit Herz- und Gefäßkrankheiten, Krebs, Fettsucht und psychischen Krankheiten [9]. Es wurde zum Beispiel die genetische Basis der Chorea Huntington und der Alzheimer-Krankheit entdeckt [33]. Das Wissen über genetisch bedingte Krankheiten verändern die Art der Diagnostik, Behandlung und Prävention von Krankheiten und lässt zudem eine genau Einteilung von Krankheiten bezüglich ihrer Ursache zu. Es lässt sich dadurch für eine Krankheit sagen, ob sie nur durch genetische Faktoren, nur durch Umwelteinflüsse oder durch eine Kombination genetischer Faktoren und Einflüsse aus der Umwelt hervorgerufen wird [8].

Um den Zusammenhang zwischen Genomvarianz und einer Krankheit herzustellen, werden unter anderem Assoziationstudien, bei denen man die Erbinformation von zwei Populationen vergleicht, durchgeführt. Dabei haben die Individuen der einen Population die Krankheit und die Individuen der anderen Population haben die Krankheit nicht. Für das Finden von Zusammenhängen zwischen Genomvarianz und Krankheiten haben Assoziationsstudien mit vorherigem umfangreichen Auslesen von Erbinformation ein großes Potential [33]. Haplotypen beschreiben die Erbinformation eines Menschen eindeutig und werden daher bei Assoziationsstudien den Genotypen vorgezogen. Das umfangreiche Auslesen von Erbinformation ist mit aktueller Technologie für Genotypen kostengünstig möglich, wohingegen sich das Auslesen von Haplotypen nur mit wesentlich höherem Aufwand realisieren lässt [32].

Um trotzdem Haplotypen zur Verfügung zu haben, leitet man Haplotypen aus Genotypen ab. Dies ist das Problem der Haplotypisierung von Populationen, bei dem für einen Bereich im Genom der Genotyp jedes Individuums einer Population gegeben ist und das Ziel ist, die tatsächlich zugrunde liegenden Haplotypen für jeden Genotyp zu ermitteln. Das heißt, es wird eine Menge von Haplotypen ermittelt, die für jeden Genotyp zugrunde liegende Haplotypen enthält. Da eine Ermittlung der Haplotypen im Labor zu aufwendig ist, möchte man dieses Problem mit dem Computer lösen. Dies ist aber ohne weitere Einschränkungen nicht möglich, da für jeden Genotyp mit k heterozytischen Positionen 2^{k-1} mögliche zugrunde liegende Haplotyppaare existieren und es ist nicht klar, welches Haplotyppaar einem Genotyp tatsächlich zugrunde liegt. Im nächsten Abschnitt werden Ansätze besprochen, die das Problem der Haplotypisierung so einschränken, dass man biologisch sinnvolle Lösungen erhält.

2.2 Lösungsansätze für das Haplotypisierungsproblem

Das Ermitteln von Haplotypen aus den Genotypen einer Population ist wegen der großen Anzahl möglicher zugrunde liegender Haplotyppaare ohne weitere Annahmen nicht möglich. Lösungsansätze für die Haplotypisierung stellen Bedingungen an die Haplotypen, die den Genotypen einer Population zugrunde liegen. Dadurch wird die Menge der möglichen Lösungen eingeschränkt und es ergeben sich Restriktionen des Haplotypisierungsproblems. In diesem Kapitel werden die Haplotypisierung mittels kleinsten Haplotypmengen, die Haplotypisierung mittels perfekten Phylogenien und ein Ansatz, der die beiden Verfahren kombiniert, vorge-

stellt. In der Literatur werden neben diesen Ansätzen noch weitere Lösungsansätze für die Haplotypisierung vorgeschlagen. Übersichten hierzu findet sich bei Gusfield [22], Halldórsson et al. [23] oder Bonizzoni et al. [4].

2.2.1 Haplotypisierung mittels kleinsten Haplotypmengen

Daly et al. [10] haben die Struktur des menschlichen Genoms untersucht und festgestellt, dass sich das Genom in Blöcke einteilen lässt, so dass für jeden Haplotypblock in einer Population nur wenige verschiedene Allelsequenzen vorkommen. Gabriel et al. [13] stellten zudem fest, dass die Blockgrenzen und die Allelsequenzen in den Blöcken über Populationsgrenzen hinweg weitestgehend erhalten bleiben. Wenn der Genombereich, für den die Genotypen gegeben sind, innerhalb eines Blocks liegt, reicht es daher nach Haplotypmengen zu suchen, die möglichst wenige verschiedene Haplotypen besitzen. Dies ist die Idee der Haplotypisierung mittels kleinsten Haplotypmengen, die von Gusfield [21] vorgeschlagen wurde und bei der man für eine Menge von Genotypen nach der kleinsten Menge von Haplotypen sucht, die den Genotypen zugrunde liegt. Die Haplotypisierung mittels kleinsten Haplotypmengen wird in Problem 2.3 als Entscheidungsproblem formuliert und die Komplexität wird in Abschnitt 3.2 untersucht.

2.2.2 Haplotypisierung mittels perfekten Phylogenien

Die Haplotypisierung mittels perfekten Phylogenien wurde von Gusfield [20] vorgeschlagen. Dieser Ansatz stützt sich auf zwei genetische Eigenschaften:

- Es existieren Blöcke im Genom, die keiner Rekombination unterliegen.
- Die Mutationfrequenz pro Basenposition ist sehr gering.

Im Folgenden werden diese Eigenschaften erläutert und die Haplotypisierung mittels perfekten Phylogenien vorgestellt.

Neben der Eigenschaft, dass für einen Haplotypblock nur wenige verschiedene Allelsequenzen vorkommen, stellten Daly et al. [10] fest, dass es in den Haplotypblöcken keine Anzeichen für Crossing-over Ereignisse während der Vererbungsgeschichte gibt. Die Varianz in den Haplotypblöcken wird nur durch Punktmutationen hervorgerufen und lässt sich durch die SNPs beschreiben.

Im menschlichen Genom treten nur selten Mutationen auf, die Anzahl der Basenpositionen im menschlichen Genom ist sehr groß (ungefähr 3 Milliarden Basenpositionen) und der Zeitraum, in dem die Haplotypen einer Population, die man bei der Haplotypisierung betrachtet, entstanden sind, ist relativ klein. Wegen diesen Eigenschaften wird angenommen, dass an einer Basenposition maximal eine Punktmutation in der Vererbungsgeschichte einer Population auftritt.

Falls bei der Haplotypisierung ein Genombereich betrachtet wird, der in einem Haplotypblock liegt, dann lässt sich fordern, dass die Menge der zugrunde liegenden Haplotypen die beiden oben beschriebenen genetischen Eigenschaften erfüllt.

Eine Menge von Haplotypen, die beide Eigenschaften erfüllt, lässt sich als perfekte Phylogenie anordnen. Eine perfekte Phylogenie stellt die Vererbungsgeschichte einer Population als Baum dar, dessen Knoten in der Weise mit Haplotypen markiert sind, dass die Haplotypen, die an einer Basenposition das gleiche Allel besitzen, eine Komponente im Baum bilden. In einer perfekten Phylogenie ist ein Knoten als Wurzel ausgezeichnet. Von dem Haplotyp, der die Wurzel markiert, stammen alle anderen Haplotypen ab. Dieser Haplotyp ist entweder festgelegt, dann spricht man von gerichteten perfekten Phylogenien oder kann beliebig gewählt sein, dann spricht man von ungerichteten perfekten Phylogenien. Wie schon beschrieben, werden im Folgenden nur biallelische SNPs betrachtet, weshalb für eine Basenposition nur zwei mögliche Allele existieren und das Modell perfekter Phylogenien auf binäre perfekte Phylogenien eingeschränkt werden kann. In einer binären perfekten Phylogenie existiert für jede Basenposition maximal ein Knotenpaar, das durch eine Kante verbunden ist und dessen Haplotypen sich an der Basenposition unterscheiden. Die Kante stellt die einmalige Mutation an der Basenposition in der Vererbungsgeschichte dar und wird mit der Basenposition markiert. In Definition 2.1 werden binäre perfekte Phylogenien formal eingeführt.

Bei der Haplotypisierung mittels perfekten Phylogenien sucht man für eine Menge von Genotypen nach einer Menge zugrunde liegender Haplotypen, die sich als perfekte Phylogenie anordnen lassen. Für die Haplotypisierung mittels perfekten Phylogenien wurden effiziente Algorithmen vorgestellt. Ein solcher Algorithmus bekommt eine Menge von Genotypen als Eingabe und leitet, unter der Bedingung, dass sich die Haplotypen als perfekte Phylogenie anordnen lassen, aus den Genotypen die zugrunde liegenden Haplotypen ab. Die Genotypen legen dabei nicht immer jeden Haplotyp und die gesamte Struktur einer perfekten Phylogenie fest, weshalb mehrere Möglichkeiten existieren, Haplotypen zu wählen, die dem Modell der perfekten Phylogenie entsprechen und den Genotypen zugrunde liegen. Falls mehrere Lösungen vorhanden sind, werden entweder Lösungen zur Auswahl gestellt oder es wird eine beliebige Lösung vom Algorithmus gewählt. Falls keine Haplotypen existieren, die den Genotypen zugrunde liegen und sich als perfekte Phylogenie anordnen lassen, gibt der Algorithmus dies aus.

Durch das Auftreten von Crossing-over-Ereignissen kann die Eigenschaft, dass sich Haplotypen als perfekte Phylogenie anordnen lassen, verloren gehen. Seien *AA* und *GG* zwei Haplotypen. Durch Crossing-over zwischen der ersten und zweiten Position entstehen die Haplotypen *AG* und *GA*. In perfekten Phylogenien bilden Haplotypen, die an einer Position das gleiche Allel besitzen, eine Komponente. Für die Haplotypen *AA*, *GG*, *AG* und *GA* bedeutet dies, dass jeweils *AA* und *AG*, *AA* und *GA*, *GG* und *AG* sowie *GG* und *GA* eine Komponente bilden. Jeder Graph, der diese Bedingungen erfüllt, enthält einen Kreis. Da eine perfekte Phylogenie die Vererbungsgeschichte als Baum darstellt, lassen sich die vier Haplotypen nicht als perfekte Phylogenie anordnen.

Eine weitergehende Einschränkung an die Menge zugrunde liegender Haplotypen stellt der Ansatz mit perfekten Pfadphylogenien dar, bei dem nach perfekten Phylogenien gesucht wird, die die Form eines Pfades haben. Gramm et al. [17] un-

tersuchten reale Eingaben für das Haplotypisierungproblem und stellten fest, dass 70% der untersuchten Genotypmengen, die dem Ansatz mit perfekten Phylogenien genügen, auch dem Ansatz mit perfekten Pfadphylogenien genügen.

Problem 2.4 formuliert die Haplotypisierung mittels perfekten Phylogenien als Entscheidungsproblem, woraus sich analog ein Entscheidungsproblem für die Haplotypisierung mittels perfekten Pfadphylogenien ergibt. Die Komplexität der beiden Entscheidungsprobleme behandelt Abschnitt 3.3.3.

2.2.3 Haplotypisierung mittels kombinierten Ansätzen

Für eine Menge von Genotypen kann es verschiedene Mengen von Haplotypen geben, die der Haplotypisierung mittels perfekten Phylogenien genügen. Ebenfalls kann es verschiedene Mengen von Haplotypen geben, die der Haplotypisierung mittels kleinsten Haplotypmengen genügen. Um möglichst die biologisch sinnvollste Menge von Haplotypen zu ermitteln, wird der Ansatz mit perfekten Phylogenien und der Ansatz mit kleinsten Haplotypmengen kombiniert. Es wird nach der kleinsten Haplotypmenge gesucht, die sich als perfekte Phylogenie anordnen lässt und den Genotypen zugrunde liegt. Das zugehörige Entscheidungsproblem wird in Problem 2.5 formuliert. Analog lässt sich nach kleinsten perfekten Pfadphylogenien suchen und das entsprechende Entscheidungsproblem formulieren.

2.3 Mathematische Modellierung der Problemstellungen

In diesem Kapitel werden die Haplotypisierungprobleme mathematisch gefasst. Dazu wird zuerst die Kodierung von Haplotypen und Genotypen festgelegt und es werden Begriffe aus dem vorangegangen Abschnitt formal eingeführt. Dann wird für jede Art von Haplotypisierung das entsprechende Entscheidungsproblem definiert. Dieser Abschnitt umfasst nicht alle mathematischen Begriffe, die in dieser Arbeit verwendet werden. Weitere Begriffe werden bei ihrer ersten Verwendung in den weiteren Abschnitten dieser Arbeit eingeführt.

Begriffsdefinitionen. Durch die Einschränkung auf biallelische SNPs existieren für jede Position in einem Haplotyp nur zwei mögliche Allele, die mit 0 und 1 kodiert werden. Seien beispielsweise AGC und TGG zwei Haplotypen. Nun sei an der ersten Position A mit 0 und T mit 1 kodiert. An der zweiten Position sei G mit 0 kodiert. An der dritten Position sei C mit 0 und G mit 1 kodiert. Der Haplotyp AGC wird dann durch 000 und der Haplotyp TGG durch 101 kodiert. Ein Haplotyp ist somit eine Zeichenkette aus $\{0,1\}$. Man beachte, dass ein Allel nicht an jeder Position mit der gleichen Zahl kodiert werden muss. Genotypen werden durch Einträge aus $\{0,1,2\}$ kodiert und sind somit Zeichenketten über $\{0,1,2\}$. Der Eintrag 0 oder 1 bedeutet, dass der Genotyp an dieser Position homozytisch ist und das Allel besitzt, welches durch 0 oder 1 kodiert wird. Der Eintrag 2 sagt aus, dass diese Position heterozytisch ist. Für die Haplotypen AGC und TGG ist

$\{A,T\}G\{C,G\}$ der Genotyp und für die kodierten Haplotypen 000 und 101 ist 202 der kodierte Genotyp.

Die Eigenschaft, dass zwei Haplotypen einen Genotyp erklären wird nun definiert. Zwei Haplotypen $h,h' \in \{0,1\}^m$ *erklären* einen Genotyp $g \in \{0,1,2\}^m$, falls für jedes $i \in \{1,\ldots,m\}$ gilt: Wenn $g[i] \in \{0,1\}$, dann $h[i] = h'[i] = g[i]$ und wenn $g[i] = 2$, dann $h[i] \neq h'[i]$.

Eine *Haplotypmatrix H* ist eine Matrix mit Einträgen aus $\{0,1\}$ in der jede Zeile einen Haplotyp kodiert . Eine *Genotypmatrix G* ist eine Matrix mit Einträgen aus $\{0,1,2\}$ in der jede Zeile einen Genotyp kodiert.

Sei G eine $n \times m$ Genotypmatrix und H eine $2n \times m$ Haplotypmatrix. Die Haplotypmatrix H *erklärt* die Genotypmatrix G, falls für jedes $k \in \{1,\ldots,n\}$ die Zeile k aus G durch die Zeilen $2k-1$ und $2k$ aus H erklärt wird.

In dieser Arbeit wird oft über Spalten von Haplotyp- und Genotypmatrizen gesprochen.Spalten werden dabei auf zwei verschiedene Arten bezeichnet. Entweder wird eine Spalte in einer Matrix durch den jeweiligen Index referenziert oder eine Spalte wird durch einen Spaltenvekor dargestellt, der den Inhalt der Spalte enthält. Spaltenvektoren werden in den Abschnitten 3.3.4 und 3.4.2 verwendet, in denen der Aufbau einer partiellen Ordnung über Spaltenvektoren betrachtet wird. In allen anderen Abschnitten werden Spalten durch Indizes referenziert. In diesen Abschnitten sprechen wir von einer Spalte s und meinen damit den Spaltenindex s.

Eine Spalte in einer Haplotyp- oder Genotypmatrix heißt 0-*Spalte*, falls sie in jeder Zeile den Eintrag 0 enthält. Analog ist eine 1-*Spalte* definiert. Sei g ein Genotyp aus einer Genotypmatrix G. Der Genotyp g *umfasst* die Spalte s, falls $g[s] \neq 0$ gilt. Analog umfasst ein Haplotyp h die Spalte s, falls $h[s] \neq 0$ gilt.

Nun wird das Modell der perfekten Phylogenie definiert. In der Literatur werden perfekte Phylogenien auf unterschiedliche Weise eingeführt. Die in dieser Arbeit verwendete Definition folgt der Definition von perfekten Phylogenien bei Gramm et al. [17]. Sie ist äquivalent zur Definition perfekter Phylogenien bei Gusfield [18]. Ein Vergleich verschiedener Definitionen findet sich bei Gramm, Nickelsen und Tantau [15, 16].

Definition 2.1. [Perfekte Phylogenie] Eine $n \times m$ Haplotypmatrix H *lässt eine perfekte Phylogenie zu*, falls ein Baum B_H mit ausgezeichneter Wurzel w existiert, so dass folgende Aussagen gelten:

1. Jeder Haplotyp aus H markiert genau einen Knoten in B_H.
2. Jede Spalte aus H markiert genau eine Kante in B_H.
3. Jede Kante in B_H ist markiert.
4. Für jedes Paar von Haplotypen h, h' aus H und jede Spalte $s \in \{1,\ldots,m\}$ gilt $h[s] \neq h'[s]$ genau dann, wenn s eine Kante auf dem Weg von h nach h' in B_H markiert.

Wenn die Wurzel einer perfekten Phylogenie B_H mit dem Haplotyp $0\ldots0$ markiert werden kann, dann ist B_H eine *gerichtete perfekte Phylogenie*. Jede perfekte

Phylogenie ist eine *ungerichtete perfekte Phylogenie*. Eine *perfekte Pfadpylogenie* ist eine perfekte Phylogenie mit maximal zwei disjunkten Zweigen, die an der Wurzel beginnen. Eine Haplotypmatrix H *lässt eine perfekte Pfadphylogenie* zu, falls sich die Haplotypen in H als perfekte Pfadphylogenie anordnen lassen. Für eine perfekte Pfadphylogenie B_H werden die beiden Zweige mit $B_{H,l}$ und $B_{H,r}$ bezeichnet, wobei ein Zweig leer sein kann. Abbildung 1 zeigt beispielhaft Haplotypmatrizen mit perfekten Phylogenien und perfekten Pfadphylogenien.

Entscheidungsprobleme. Bei der Haplotypisierung erstellt man für eine Genotypmatrix G eine Haplotypmatrix H, die G erklärt. Für einen Genotyp g mit k heterozytischen Stellen existieren 2^{k-1} verschiedene Paare von Haplotypen, die g erklären und somit existiert für eine Genotypmatrix eine große Anzahl möglicher Haplotypmatrizen. Lösungsansätze zur Haplotypisierung schränken diese Menge von Haplotypmatrizen dadurch ein, dass eine Haplotypmatrix nur dann in Betracht gezogen wird, wenn sie eine bestimmte Eigenschaft besitzt. Ein Algorithmus, der ein solches Problem löst, gibt eine Haplotypmatrix mit den Eigenschaften aus, falls diese existiert, oder bricht die Berechnung ab, falls eine Haplotypmatrix mit den gewünschten Eigenschaften nicht existiert.

Im Folgenden werden die Entscheidungsprobleme für verschiedene Arten von Haplotypisierung definiert. Bei Entscheidungsproblemen wird nur nach der Existenz einer Lösung gefragt. Für Haplotypisierungsprobleme heißt dies, man fragt nach der Existenz einer Haplotypmatrix, die die jeweiligen Eigenschaften erfüllt. Ein Algorithmus, der ein Entscheidungproblem löst, antwortet entweder mit „ja“ oder „nein“.

Folgendes Entscheidungsproblem fragt, ob eine Haplotypmatrix eine perfekte Phylogenie zulässt.

Problem 2.2. [Perfekte Phylogenie (PP)]
Eingabe: Eine $n \times m$ Haplotypmatrix H.
Frage: Lässt H eine perfekte Phylogenie zu?

Das Problem PP erhält als Eingabe eine Haplotypmatrix, also eine Matrix in der jeder Eintrag einen von zwei Zuständen einnimmt. Dies entspricht dem Problem binärer perfekte Phylogenien, für welches Gusfield einen Linearzeitalgorithmus vorgestellt hat [18]. Beim Problem allgemeiner perfekter Phylogenien ist die Anzahl der möglichen Zustände in den Einträgen dagegen nicht beschränkt. Dieses Problem ist NP-vollständig. Beschränkt man die Anzahl der Zustände mit einem beliebigen Wert, dann lässt sich aber immer ein Polynomialzeitalgorithmus finden. Ein Überblick über die verschiedenen Varianten von PP und deren Komplexität findet sich bei Gramm, Nickelsen und Tantau [15, 16].

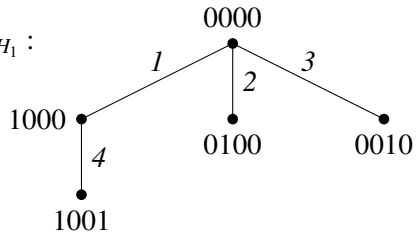
Die Komplexität von PP wird in Abschnitt 3.3.1 angesprochen.

Nun formulieren wir für die Haplotypisierung mittels kleinsten Haplotypmengen, für die Haplotypisierung mittels perfekten Phylogenien, und für den kombinierte Ansatz Entscheidungsprobleme.

Abbildung 1: Die Abbildung zeigt Haplotypmatrizen mit perfekten Phylogenien und perfekten Pfadphylogenien. Die Haplotypmatrix H_1 kodiert vier Haplotypen, die sich als perfekte Phylogenie B_{H_1} anordnen lassen. Für H_1 lässt sich aber keine perfekte Pfadphylogenie finden. Die Haplotypmatrix H_2 kodiert vier Haplotypen, die sich als perfekte Pfadphylogenie B_{H_2} anordnen lassen. Die Haplotypmatrix H_3 enthält ebenfalls vier Haplotypen, die sich aber nicht als perfekte Phylogenie anordnen lassen. In den perfekten Phylogenien ist die Wurzel jeweils nicht ausgezeichnet. Man kann einen beliebigen Knoten als Wurzel wählen. In B_{H_1} kommt der Haplotyp 0000 vor, aber H_1 enthält diesen Haplotyp nicht. Das heißt, nicht jeder Knoten einer perfekten Phylogenie muss mit einem Haplotyp aus der Haplotypmatrix markiert sein. Markierungen können auch hilfsweise angegeben werden. Da eine perfekte Phylogenie für H_1 den Haplotyp 0000 enthält, lässt H_1 eine gerichtete perfekte Phylogenie zu. Die Haplotypmatrix H_2 lässt aus folgendem Grund eine gerichtete perfekte Pfadphylogenie zu: Ersetzt man in B_{H_2} die Kante $1,2$ durch eine Kante 1, einen Knoten 0000, und eine Kante 2, so erhält man eine gerichtete perfekte Pfadphylogenie für H_2.

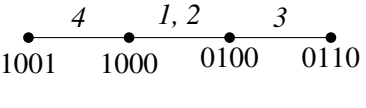
$$H_1 = \begin{array}{c} \begin{matrix} \textit{1} & \textit{2} & \textit{3} & \textit{4} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

B_{H_1} :



$$H_2 = \begin{array}{c} \begin{matrix} \textit{1} & \textit{2} & \textit{3} & \textit{4} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

B_{H_2} :



$$H_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Problem 2.3. [Minimale Haplotypisierung (MH)]
Eingabe: Eine $n \times m$ Genotypmatrix G und eine natürliche Zahl d.
Frage: Existiert eine $2n \times m$ Haplotypmatrix H, so dass:

1. H erklärt G und
2. H enthält d oder weniger paarweise verschiedene Haplotypen?

Das Finden kleinster Haplotypmengen ist ein Optimierungproblem, denn die Anzahl der paarweise verschiedenen Haplotypen wird unter der Bedingung, dass die Haplotypmatrix H die Genotypmatrix G erklärt, minimiert. Zur Formulierung des Entscheidungsproblems, wird die Eingabe um den Budgetwert d erweitert. Ergebnisse zur Komplexität von MH werden in Abschnitt 3.2 vorgestellt.

Problem 2.4. [Perfekte Phylogenie Haplotypisierung (PPH)]
Eingabe: Eine $n \times m$ Genotypmatrix G.
Frage: Existiert eine $2n \times m$ Haplotypmatrix H, so dass:

1. H erklärt G und
2. H lässt eine perfekte Phylogenie zu?

Analog zu PPH lässt sich das Entscheidungsproblem PPPH (Perfekte Pfadphylogenie Haplotypisierung) definieren. Beim Problem PPPH wird bei Eingabe einer Genotypmatrix nach der Existenz einer Haplotypmatrix gefragt, die die Genotypmatrix erklärt und eine perfekte Pfadphylogenie zulässt. In Abschnitt 3.3 wird die Komplexität von PPH und PPPH behandelt,

Falls eine Haplotypmatrix H eine Genotypmatrix G erklärt und sich als perfekte Phylogenie anordnen lässt, dann ist das Tupel (H, B_H) eine PP-*Lösung* für G. Eine PP-Lösung (H, B_H) ist eine PP-*Lösung der Größe d*, falls H genau d paarweise verschiedene Haplotypen enthält. Analog ist eine PPP-*Lösung* definiert. Falls für eine Genotypmatrix G eine PP-Lösung existiert, dann lässt G eine perfekte Phylogenie zu, ansonsten lässt G keine perfekte Phylogenie zu.

Problem 2.5. [Minimale Perfekte Phylogenie Haplotypisierung (MPPH)]
Eingabe: Eine $n \times m$ Genotypmatrix G und eine natürliche Zahl d.
Frage: Existiert eine $2n \times m$ Haplotypmatrix H, so dass:

1. H erklärt G,
2. H lässt eine perfekte Phylogenie zu und
3. H enthält d oder weniger paarweise verschiedene Haplotypen?

Analog zu MPPH lässt sich das Entscheidungsproblem MPPPH (Minimale Perfekte Pfadphylogenie Haplotypisierung) definieren, das den Ansatz mit perfekten Pfadphylogenien und kleinsten Haplotypmengen kombiniert. Die Komplexität der kombinierten Ansätze MPPH und MPPPH wird in Abschnitt 3.4 behandelt.

Eine Genotypmatrix heißt (k, l)-*beschränkt*, falls jede Zeile maximal k-mal den Eintrag 2 enthält und jede Spalte maximal l-mal den Eintrag 2 enthält. Es gilt

$k, l \in \mathbb{N} \cup \{\infty\}$, wobei ∞ beschreibt, dass für einen Parameter keine Beschränkung besteht. Zum Beispiel enthält die Menge aller $(2,\infty)$-beschränkten Genotypmatrizen genau die Genotypmatrizen, die in jeder Zeile maximal zweimal den Eintrag 2 besitzen. Die Anzahl der heterozytischen Einträge pro Spalte ist nicht beschränkt. Sharan, Halldórsson und Istrail [34] verwenden die (k,l)-Beschränkung, um Haplotypisierungsprobleme zu parametrisieren. Zum Beispiel ist das Problem PPH(k,l) analog zu PPH, aber es werden nur (k,l)-beschränkte Genotypmatrizen als Eingabe zugelassen. In dieser Arbeit wird die Komplexität von (k,l)-beschränkten Varianten für die Probleme MH, PPH und MPPH betrachtet.

Nun wird in Abschnitt 3 die Komplexität der verschiedenen Entscheidungsprobleme betrachtet. Eine Übersicht über die Komplexität der Entscheidungsprobleme findet sich in Abschnitt 4.

3 Komplexität von Haplotypisierungsproblemen

Im diesem Abschnitt wird die Komplexität von Haplotypisierungproblemen untersucht. Die Konzepte und Werkzeuge, die dabei aus der Komplexitätstheorie verwendet werden, sind in Abschnitt 3.1 beschrieben. In Abschnitt 3.2 wird auf den Ansatz mit kleinsten Haplotypmengen eingegangen. Abschnitt 3.3 behandelt die Komplexität der Haplotypisierung mittels perfekten Phylogenien und in Abschnitt 3.4 wird auf die kombinierten Ansätze zur Haplotypisierung eingegangen.

3.1 Komplexitätstheoretische Konzepte und Werkzeuge

Die Komplexitätstheorie versucht festzustellen, welchen Ressourcenaufwand man benötigt, um algorithmische Probleme zu lösen. Die Ressourcen sind hierbei meist Zeit- oder Platzbedarf bei Turingmaschinen, wobei die Zeit durch das Zählen von Zustandsübergängen und der Platz durch die Anzahl der besuchten Bandzellen gemessen wird. Probleminstanzen werden für Turingmaschinen als Zeichenketten kodiert. Der Ressourcenaufwand einer Maschine für eine Eingabe wird in Relation zur Länge der Eingabe gemessen. Dieses relative Maß erscheint sinnvoll, da sich zum Beispiel die in Abschnitt 2.3 definierten Entscheidungsprobleme einfacher für kleine und schwerer für große Matrizen lösen lassen. In der Komplexitätstheorie wird nun einerseits nach oberen Schranken für ein Problem gesucht, also Aussagen der Form: „Es gibt eine Maschine, die jede n lange Eingabe mit maximal $f(n)$ Ressourcenaufwand entscheidet.“ Andererseits werden aber auch untere Schranken für ein Problem gesucht, deren Kernaussage idealerweise Folgende ist: „Jede Maschine benötigt für unendlich viele Eingaben mindestens den Ressourcenaufwand $g(n)$.“ Je dichter obere und untere Schranken für ein Problem zusammenliegen, desto genauer ist dessen Komplexität bestimmt. In dieser Arbeit wird beispielsweise für PPH eine obere (Satz 3.12) und eine untere Schranke (Satz 3.10) vorgestellt.

Um Probleme bezüglich ihres Ressourcenaufwandes zu ordnen, werden sie in Komplexitätsklassen einsortiert und somit klassifiziert. In dieser Arbeit werden die bekannten Komplexitätsklassen L, P und NP verwendet. Es wird zum Beispiel gezeigt, dass MPPPH in L liegt (Satz 3.28). Um die Komplexität zweier Problemen zu vergleichen, verwendet man Reduktionen. In dieser Arbeit werden für die Probleme MH (Satz 3.1) und MPPH (Satz 3.21) NP-Vollständigkeitsbeweise aus der Literatur vorgestellt und dabei deterministische Polynomialzeitreduktionen verwendet. Eine genaue Definition der angesprochenen Konzepte findet sich bei Papadimitriou [31].

Anstatt den Zeit- oder Platzaufwand eines Problems bei Turingmaschinen zu betrachten, lassen sich Probleme auch bezüglich des Aufwands in andersartigen Modellen klassifizieren.

Schaltkreiskomplexität. Zum einen kann man die Tiefe und Größe von Schaltkreisen betrachten, die ein bestimmtes Problem lösen. Da ein fester Schaltkreis nur

für eine feste Eingabelänge verwendet werden kann, betrachtete man Familien von Schaltkreisen, die sich mit bestimmtem Aufwand konstruieren lassen. In dieser Arbeit werden aus der Schaltkreiskomplexität die Klassen AC^0 und NC^2 verwendet, die wie folgt definiert sind: AC^0 umfasst genau die Probleme, für die eine Schaltkreisfamilie konstanter Tiefe und polynomieller Größe existiert, deren Schaltkreise sich mit logarithmischem Platz konstruieren lassen. Dabei können Gatter mit beliebiger Anzahl von Eingängen verwendet werden. Die Klasse NC^2 umfasst genau die Probleme, für die eine Schaltkreisfamilie mit Tiefe $O(\log^2 n)$ und polynomieller Größe existiert, deren Schaltkreise sich ebenfalls mit logarithmischem Platz berechnen lassen, wobei die Anzahl der Eingänge bei den Gattern beschränkt ist.

Beschreibungskomplexität. Ein anderes Komplexitätsmaß ergibt sich, wenn betrachtet wird, mit welcher mathematischen Logik man ein Problem beschreiben kann. Eingaben liegen hierbei nicht als Zeichenketten vor, sondern werden als logische Strukturen kodiert. Die Syntax von Strukturen wird durch eine Signatur beschrieben. Die Signatur einer Genotypmatrix $\tau_G = (Z^1, S^1, G_0^2, G_1^2, G_2^2)$ besteht beispielsweise aus den einstelligen Relationssymbolen Z und S und den zweistelligen Relationssymbolen G_0, G_1 und G_2. Eine τ_G-Struktur (I, Z, S, G_0, G_1, G_2) kodiert eine Genotypmatrix folgendermaßen: Die Menge I umfasst alle Indizes, die Mengen Z und S umfassen die Indizes aus I, die Zeilen- bzw. Spaltenindizes sind, und G_0, G_1 und G_2 sind zweistellige Relationen über den Elementen von I, also Teilmengen von $I \times I$. Die zweistelligen Relationen stellen eine Genotypmatrix folgendermaßen dar: Falls $(z,s) \in I \times I$ in G_0 enthalten ist, dann besitzt die kodierte Genotypmatrix in Zeile z an Position s den Eintrag 0. Für die Relationen G_1 und G_2 ergibt sich die Bedeutung analog. Das heißt, falls (z,s) in G_1 liegt, dann steht in Zeile z an Position s der Eintrag 1 und falls (z,s) in G_2 liegt, dann steht in Zeile z an Position s der Eintrag 2. Neben der Signatur für Genotypmatrizen werden in dieser Arbeit auch Signaturen für Graphen (Beweis zu Lemma 3.9) und lineare Gleichungssysteme (Beweis zu Lemma 3.11) verwendet. Über einer Signatur lassen sich Formeln in Prädikatenlogik erster Stufe bilden (im Weiteren kurz *Formeln* genannt). Formeln über der Signatur τ_G heißen τ_G-Formeln und bestehen aus den Relationssymbolen $=$, G_0, G_1 und G_2, den Junktoren $\neg$, $\wedge$, $\vee$ und $\rightarrow$, den Quantoren $\exists$ und $\forall$ und Symbolen für Variablen. Wir werden in dieser Arbeit voraussetzen, dass die Elemente einer Struktur total geordnet sind. Für eine τ_G Struktur mit $|I| = n$ existiert also eine $1:1$-Beziehung zwischen den Elementen aus I und den Elementen aus $\{1, \ldots, n\}$. Wir werden weiter voraussetzen, dass die darauf aufbauende Relation $<$ und die Funktionen $+$ und $\cdot$ in Formeln verwendet werden können (vgl. Definition von Formeln bei Immerman [25]). In Formeln können außerdem *beschränkte Quantoren* verwendet werden, die wie folgt definiert sind (vgl. Seite 10 bei Immerman [25]): $(\exists x.\psi)\phi \equiv (\exists x)[\psi \wedge \phi]$ und $(\forall x.\psi)\phi \equiv (\forall x)[\psi \rightarrow \phi]$. Beschränkte Quantoren verwendet man dann, wenn man nur Elemente in Betracht ziehen möchte, die eine bestimmte Eigenschaft (im obigen Fall durch ψ beschrieben) besitzen.

Eine Formel wird für eine bestimmte Struktur entweder zu wahr oder zu falsch

ausgewertet. Beispielsweise ist die Formel $\phi_{\text{existert2erZeile}} \equiv (\exists z)(\forall s)[G_2(z,s)]$ genau dann für eine τ_G-Struktur wahr, wenn die dadurch kodierte Genotypmatrix eine Zeile enthält, die in jeder Spalte den Eintrag 2 besitzt. Formeln können auch Parameter enthalten. Zum Beispiel ist die Formel $\phi_{\text{ist2erZeile}}(z) \equiv (\forall s)[G_2(z,s)]$ genau dann für eine τ_G-Struktur wahr, wenn die dadurch kodierte Genotypmatrix an dem übergebenen Zeilenindex z in jeder Spalte den Eintrag 2 besitzt. Die Formel $\phi_{\text{existert2erZeile}}(z)$ lässt sich somit auch durch $(\exists z)[\phi_{\text{ist2erZeile}}(z)]$ definieren. Falls eine Formel ϕ für eine Struktur A zu wahr ausgewertet wird, dann schreiben wir $A \models \phi$ und sagen: *A ist ein Modell von* ϕ.

Ein Probleme *lässt sich in Prädikatenlogik erster Stufe beschreiben*, falls eine Formel existiert, für die genau die Strukturen Modelle sind, die Elemente der Problemsprache kodieren. Der Unterschied zwischen der Kodierung von Eingaben durch Strukturen und der Kodierung von Eingaben durch Zeichenketten fällt hierbei nicht ins Gewicht (vgl. [25, Seite 24f]). In Satz 3.18 wird gezeigt, dass sich das Problem ger-PPPH durch eine τ_G-Formel $\phi_{\text{ger-PPPH}}$ beschreiben lässt. Dies bedeutet, dass eine τ_G-Struktur (I,Z,S,G_0,G_1,G_2) genau dann ein Modell von $\phi_{\text{ger-PPPH}}$ ist, wenn die durch (I,Z,S,G_0,G_1,G_2) kodierte Genotypmatrix eine gerichtete perfekte Pfadphylogenie zulässt. Probleme lassen sich auch bezüglich ihrer Beschreibungskomplexität klassifiziert. So umfasst die Klasse FO genau die Probleme, die sich durch eine Formel in Prädikatenlogik erster Stufe beschreiben lassen.

Wie schon erwähnt werden Reduktionen verwendet, um die Komplexität zweier Probleme zu vergleichen. In dieser Arbeit werden unter anderem Reduktionen verwendet, die sich als Anfrage in Prädikatenlogik erster Stufe formulieren lassen. Für zwei Signaturen τ und σ bildet eine Anfrage in Prädikatenlogik erster Stufe jede τ-Struktur B auf eine σ-Struktur C ab. Die Elemente und die Relationen von C werden dabei durch τ-Formeln beschrieben, die für B ausgewertet werden. Beispielsweise lässt sich die Abbildung, die in einer Genotypmatrix jede 0 mit einer 1 und jede 1 mit einer 0 ersetzt, durch die Anfrage $A = \lambda_{zs}(\phi_I, \phi_Z, \phi_S, \phi_{G_0}, \phi_{G_1}, \phi_{G_2})$ mit den Formeln $\phi_I(i) \equiv \mathit{wahr}$, $\phi_Z(z) \equiv \mathit{wahr}$, $\phi_S(s) \equiv \mathit{wahr}$, $\phi_{G_0}(z,s) \equiv G_1(z,s)$, $\phi_{G_1}(z,s) \equiv G_0(z,s)$ und $\phi_{G_2}(z,s) \equiv G_2(z,s)$ formulieren. Die Anfrage A bildet nun nach folgendem Schema jede τ_G-Struktur (I,Z,S,G_0,G_1,G_2) auf eine τ_G-Struktur $(I',Z',S',G_0',G_1',G_2')$ ab: Ein Element $i \in I$ ist in I', wenn (I,Z,S,G_0,G_1,G_2) ein Modell von $\phi_I(i)$ ist; ein Tupel $(z,s) \in I \times I$ ist in G_0', wenn (I,Z,S,G_0,G_1,G_2) ein Modell von $\phi_{G_0}(z,s)$ ist; für die Relationen Z, S, G_1' und G_2' gilt dies analog. Es lässt sich erkennen, dass jedes Element aus I nach I' übernommen wird und dass die Relationen, welche 0-Einträge und 1-Einträge kodieren, vertauscht werden. Im Beweis zu Lemma 3.6 wird eine Anfrage vorgestellt, die jede τ_G-Struktur G auf eine τ_G-Struktur G' abbildet, so dass die durch G kodierte Genotypmatrix genau dann eine perfekte Phylogenie zulässt, wenn die durch G' kodierte Genotypmatrix eine gerichtete perfekte Phylogenie zulässt. Das Problem ung-PPH lässt sich also durch eine Anfrage in Prädikatenlogik erster Stufe auf ger-PPH reduzieren, wofür wir kurz ung-PPH $\leq_{\text{fo}}$ ger-PPH schreiben. Die vorgestellten Konzepte zur Beschreibungskomplexität werden umfassend bei Immerman [25] definiert.

Zählkomplexität. Die Komplexität von Problemen lässt sich auch betrachten, indem man die Anzahl der akzeptierenden Pfade von Turingmaschinen zählt. In dieser Arbeit wird die Zählklasse $\mathrm{Mod}_2\mathrm{L}$ verwendet, die wie folgt definiert ist: Ein Entscheidungsproblem E ist in $\mathrm{Mod}_2\mathrm{L}$, falls eine nichtdeterministische logarithmisch platzbeschränkte Turingmaschine M existiert, so dass eine Eingabe x genau dann in E liegt, wenn M auf Eingabe x eine ungerade Anzahl akzeptierender Berechnungspfade besitzt [6]. In dieser Arbeit wird gezeigt, dass PPH in $\mathrm{Mod}_2\mathrm{L}$ liegt (Satz 3.12).

Zusammenhang zwischen Komplexitätsklassen Die Komplexitätstheorie vergleicht auch die verschiedenen Komplexitätsklassen und damit die Mächtigkeit der verschiedenen ressourcenbeschränkten Berechnungsmodelle. Zum Beispiel gilt $\mathrm{FO} \subseteq \mathrm{AC}^0$ [25], was heißt, dass für jedes Problem, das sich durch eine Formel in Prädikatenlogik erster Stufe beschreiben lässt, eine Schaltkreisfamilie konstanter Tiefe existiert. Durch die Beschreibung eines Problems erhält man somit auch dessen Schaltkreiskomplexität. Weiter gilt die Inklusion $\mathrm{AC}^0 \subseteq \mathrm{L}$, die eine Beziehung zwischen einer Klasse, die mit Schaltkreisen definiert wird und einer Klassen, die mit Turingmaschinen definiert wird, herstellt. Eine weitere Inklusion, die sich direkt aus der Definition von $\mathrm{Mod}_2\mathrm{L}$ ergibt, ist $\mathrm{L} \subseteq \mathrm{Mod}_2\mathrm{L}$. Außerdem gilt, wie von Buntrock et al. [6] bemerkt wird, $\mathrm{Mod}_2\mathrm{L} \subseteq \mathrm{NC}^2$. Insgesamt gelten für die in dieser Arbeit verwendeten Komplexitätsklassen folgende Inklusionen:

$$\mathrm{FO} \subseteq \mathrm{AC}^0 \subseteq \mathrm{NC}^1 \subseteq L \subseteq \mathrm{Mod}_2\mathrm{L} \subseteq \mathrm{NC}_2 \subseteq \mathrm{P} \subseteq \mathrm{NP}.$$

Approximationsalgorithmen. Bei einigen algorithmischen Problemen, den so genannten Optimierungsproblemen, sucht man nicht eine beliebige Lösung, sondern eine optimale Lösung. Zum Beispiel sucht man bei der Haplotypisierung mittels kleinsten Haplotypmengen nach Haplotypmatrizen, die möglichst wenige paarweise verschiedene Haplotypen enthalten. Falls sich ein Optimierungsproblem nicht exakt mit einem bestimmten Ressourcenaufwand lösen lässt, kann man nach Verfahren suchen, die das Problem nicht optimal, aber mit geringem Aufwand lösen. Falls für ein Problem ein Algorithmus existiert, der Polynomialzeit verwendet und immer eine Lösung ausgibt, die maximal ε mal größer als eine optimale Lösung ist, dann nennt man diesen Algorithmus einen ε-Approximationsalgorithmus. Die Komplexitätsklasse APX umfasst genau die Optimierungsprobleme, für die ein ε-Approximationsalgorithmus mit $\varepsilon > 0$ existiert. Auch für die Klasse APX lassen sich harte Probleme finden. Ein Problem A ist dabei APX-hart, falls sich zeigen lässt, dass man aus einem beliebig guten Approximationsalgorithmus für das Problem A für jedes Problem aus APX einen beliebig guten Approximationsalgorithmus konstruieren kann. Eine genaue Definition zu diesen Konzepten findet sich bei Ausiello et al. [1].

3.2 Haplotypisierung mittels kleinsten Haplotypmengen

Dieser Abschnitt behandelt die Komplexität der Haplotypisierung mittels kleinsten Haplotypmengen. Es werden aus der Literatur bekannte Ergebnisse zu MH und (k,l)-beschränkten Varianten von MH zusammengefasst und danach wird ein Beweis für die NP-Vollständigkeit von MH vorgestellt, der auf einem Beweis zur APX-Härte von MH von Lancia, Pinotti und Rizzi [28] basiert.

Das Problem MH wurde von Gusfield [21] eingeführt, der bemerkte, dass MH NP-vollständig ist und dass ein erster Beweis hierzu auf Earl Hubbell (nicht veröffentlicht) zurück geht. Es wurde außerdem eine Formulierung von MH als ganzzahliges lineares Programm vorgestellt, mit dem sich, unter Verwendung eines Lösers für das ganzzahlige lineare Optimierungsproblem, für eine große Menge von Eingaben das Problem MH effizient entscheiden lässt [21]. Brown und Harrower [5] geben einen Überblick über verschiedene Formulierungen von MH als ganzzahliges lineares Programm.

Verschiedene komplexitätstheoretische Resultate sind für MH und (k,l)-beschränkte Varianten von MH bekannt. Lancia, Pinotti und Rizzi [28] zeigten, dass $\mathrm{MH}(3,\infty)$ APX-hart ist. Durch den APX-Härte-Beweis aus [28] wird implizit gezeigt, dass $\mathrm{MH}(3,\infty)$ vollständig für die Klasse NP ist. Diese beiden Ergebnisse übertragen sich auf MH und jedes Problem $\mathrm{MH}(k,\infty)$ mit $k \geq 3$. Cilibrasi et al. [7] und Lancia und Rizzi [29] zeigten unabhängig voneinander, dass $\mathrm{MH}(2,\infty)$ in Polynomialzeit lösbar ist. Lancia und Rizzi [29] zeigten dies mit einer Reduktion auf das Finden einer kleinsten Knotenüberdeckung in bipartiten Graphen. Von Sharan, Halldórsson und Istrail [34] wurde gezeigt, dass schon $\mathrm{MH}(4,3)$ NP-vollständig und APX-hart ist. Diese Ergebnisse übertragen sich auf alle (k,l)-beschränkten Varianten, für die $k \geq 4$ oder $l \geq 3$ gilt. In [34] wurde außerdem gezeigt, dass sich $\mathrm{MH}(\infty,2)$ in Polynomialzeit lösen lässt, falls der Kompatibilitätsgraph der gegebenen Genotypmatrix vollständig ist. Der Kompatibilitätsgraph einer Genotypmatrix enthält die Genotypen als Knoten und eine Kante zwischen zwei Knoten, falls ein Haplotyp existiert, der von beiden Genotypen zur Erklärung verwendet werden kann. Zum Beispiel können die Genotypen 1202 und 1102 beide den Haplotyp 1100 zur Erklärung verwenden. Von van Iersel et al. [35] wurde gezeigt, dass schon $\mathrm{MH}(3,3)$ für die Klasse NP vollständig ist und APX-hart ist. Dieses Ergebnis überträgt sich ebenfalls auf jede (k,l)-beschränkte Variante, die die Eingabe weniger stark beschränkt. Außerdem wurde gezeigt, dass $\mathrm{MH}(\infty,1)$ in Polynomialzeit lösbar ist. Abbildung 2 gibt einen Überblick über die Komplexität (k,l)-beschränkter Varianten von MH.

Wang und Xu [37] präsentierten einen Branch-and-Bound Algorithmus für MH. Der Algorithmus berechnet bei Eingabe einer Genotypmatrix G eine Haplotypmatrix H, die G erklärt und mit einem Greedy-Ansatz möglichst optimal gewählt wird. Der Algorithmus probiert danach jede Haplotypmatrix, die G erklärt, durch und findet eine optimale Lösung. Ein potentieller Geschwindigkeitsgewinn ergibt sich dadurch, dass mögliche Lösungen aus Teillösungen aufgebaut werden und eine Teillösung schon dann verworfen wird, wenn sie sich nicht mehr zu einer

Abbildung 2: Die (k,l)-beschränkten Varianten von MH haben eine unterschiedliche Komplexität. Einige Varianten sind in P, andere sind NP-vollständig und APX-hart. Für einige Problemvarianten ist nur bekannt, dass sie in NP liegen und eine genauere Bestimmung (in P, oder NP-hart) ist bisher nicht bekannt. In der Abbildung sind die (k,l)-beschränkten Varianten von MH nach den Werten k und l in Zeilen und Spalten angeordnet und nach ihrer Komplexität hinterlegt.

$\mathrm{MH}(1,1)$	$\mathrm{MH}(1,2)$	$\mathrm{MH}(1,3)$	$\cdots$	$\mathrm{MH}(1,\infty)$
$\mathrm{MH}(2,1)$	$\mathrm{MH}(2,2)$	$\mathrm{MH}(2,3)$	$\cdots$	$\mathrm{MH}(2,\infty)$
$\mathrm{MH}(3,1)$	$\mathrm{MH}(3,2)$	$\mathrm{MH}(3,3)$	$\cdots$	$\mathrm{MH}(3,\infty)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathrm{MH}(\infty,1)$	$\mathrm{MH}(\infty,2)$	$\mathrm{MH}(\infty,3)$	$\cdots$	$\mathrm{MH}(\infty,\infty)$

$(= \mathrm{MH})$

In Polynomialzeit entscheidbar.

In NP und Weiteres unbekannt.

NP-vollständig und APX-hart.

optimalen Lösung erweitern lässt.

Huang, Chao und Chen [24] untersuchten eine Problemvariante von MH, dessen Eingabe aus einer Genotypmatrix mit n Genotypen und einer Menge von Haplotypen, deren Größe polynomiell in n ist, besteht. Bei dieser Problemvariante sucht man nach einer kleinsten Haplotypmatrix, die die Genotypmatrix erklärt und nur aus Haplotypen besteht, die in der gegebenen Haplotypmenge vorkommen. Dieses Problem ist NP-vollständig. Es wurde ein Approximationsalgorithmus vorgestellt, der in Polynomialzeit eine Lösung erstellt, die maximal $O(\log n)$-mal größer als eine optimale Lösung ist.

Im Folgenden wird gezeigt, dass MH NP-vollständig ist. Der hier vorgestellte Beweis basiert auf dem Beweis zur APX-Härte von MH von Lancia, Pinotti und Rizzi [28]. Wie bereits erwähnt, wurde die NP-Vollständigkeit von MH zuerst von Hubbel bewiesen.

Satz 3.1. MH *ist* NP-*vollständig.*

Beweis. Für den Beweis der NP-Vollständigkeit von MH wird gezeigt, dass MH in NP enthalten ist und dass MH NP-hart ist.

MH *liegt in* NP*:* Es wird nun eine nichtdeterministische Turingmaschine beschrieben, die MH in Polynomialzeit akzeptiert. Die Turingmaschine für MH rät bei Eingabe einer $n \times m$ Genotypmatrix G und einem Budgetwert d nichtdeterministisch eine $2n \times m$ Haplotypmatrix H. Danach wird deterministisch geprüft, ob jede Zeile k in G durch die Zeilen $2k$ und $2k+1$ in H erklärt wird und ob H nicht mehr als d paarweise verschiedene Haplotypen enthält. Das Abzählen der paarweise verschiedenen Haplotypen in H ist effizient möglich.

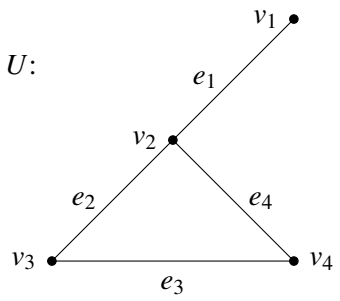
MH *ist* NP-*hart:* Es wird eine Reduktion von KNOTENÜBERDECKUNG auf MH angegeben. Eine Instanz von KNOTENÜBERDECKUNG besteht aus einem ungerichteten Graph $U = (V,E)$ und einer natürlichen Zahl d. Der ungerichtete Graph U besteht aus einer Menge von Knoten V und einer Menge ungerichteter Kanten $E \subseteq \{K \mid K \subseteq V, |K| = 2\}$. Das Problem ist nun, zu entscheiden, ob eine Menge $V' \subseteq V$ mit $|V'| \leq d$ existiert, so dass jede Kante aus E mit einem Knoten aus V' inzidiert. KNOTENÜBERDECKUNG ist NP-hart (siehe VERTEX COVER bei Garey und Johnson [14]).

Die Reduktion ist in drei Schritte aufgeteilt. Zuerst wird eine Abbildung von einem Graphen U auf eine Genotypmatrix G angegeben. Danach wird eine Aussage bewiesen, die einen Zusammenhang zwischen der Größe kleinster Knotenüberdeckungen für U und der minimalen Anzahl paarweise verschiedener Haplotypen in Haplotypmatrizen, die G erklären, hergestellt. Abschließend wird gezeigt, dass sich die Abbildung von U nach G in Polynomialzeit von einer deterministischen Turingmaschine berechnen lässt. In diesem Beweis werden die Spalten von Genotyp- und Haplotypmatrizen durch Indizes i und j referenziert, um untere anderem eine Beziehung zwischen einem Knoten v_i und einer Spalte i herzustellen.

Konstruktion: Sei $U = (V,E)$ mit $V = \{v_1, \ldots, v_n\}$ und $E = \{e_1, \ldots, e_m\}$ ein ungerichteter Graph. Der ungerichtete Graph U wird auf eine Genotypmatrix G mit $n+m$ Zeilen und $n+1$ Spalten abgebildet. Für jeden Knoten $v_i \in V$ enthält

G den Genotyp g_{v_i} mit $g_{v_i}[i] = 1$ und sonstigen Einträgen 0. Ein solcher Genotyp wird im Weiteren *Knotengenotyp* genannt. Für jede Kante $e_k = \{v_i, v_j\} \in E$ enthält G den Genotyp g_{e_k} mit $g_{e_k}[i] = g_{e_k}[j] = g_{e_k}[|V|+1] = 2$ und sonstigen Einträgen 0. Ein solcher Genotyp wird im Weiteren *Kantengenotyp* genannt. Abbildung 3 zeigt die Konstruktion an einem Beispiel.

Abbildung 3: Die Abbildung zeigt an einem Beispiel die Reduktion von KNOTENÜBERDECKUNG auf MH. Für einen Graphen U mit vier Knoten und vier Kanten wird eine Genotypmatrix G mit acht Zeilen und fünf Spalten konstruiert. Die oberen vier Genotypen sind Knotengenotypen und die unteren vier Genotypen sind Kantengenotypen. Die Knotenüberdeckung $V' = \{v_2, v_3\}$ für U enthält eine minimale Anzahl von Knoten. Zu V' korrespondiert eine Haplotypmatrix H, die neben den Knotengenotypen die Haplotypen 01001 und 00101 enthält und G mit einer minimalen Anzahl paarweise verschiedener Haplotypen erklärt.



G:

g_{v_1}	1 0 0 0 0
g_{v_2}	0 1 0 0 0
g_{v_3}	0 0 1 0 0
g_{v_4}	0 0 0 1 0
g_{e_1}	2 2 0 0 2
g_{e_2}	0 2 2 0 2
g_{e_3}	0 0 2 2 2
g_{e_4}	0 2 0 2 2

Korrektheit: Folgende Behauptung wird nun gezeigt: Es existiert genau dann eine Knotenüberdeckung $V' \subseteq V$ mit $|V'| \leq d$ für U, wenn eine Haplotypmatrix H existiert, die G erklärt und maximal $n+d$ paarweise verschiedene Haplotypen enthält.

Nur-wenn-Teil: Sei $V' \subseteq V$ mit $|V'| = d$ eine Knotenüberdeckung für U. Es wird nun eine Haplotypmatrix mit $n+d$ paarweise verschiedenen Haplotypen für G konstruiert. Aus der Konstruktion lässt sich erkennen, dass jeder Knotengenotyp nur 0 oder 1 als Eintrag enthält. Für jeden Knoten $v_i \in V$ wird daher ein Haplotyp $h_i = g_{v_i}$ erstellt. Dies sind n paarweise verschiedene Haplotypen. Für jeden Knoten $v_i \in V'$ wird ein Haplotyp h'_i mit $h'_i[i] = h'_i[n+1] = 1$ und sonstigen Einträgen 0 erstellt. Diese sind d paarweise verschiedene Haplotypen. Die h_i und h'_i sind zudem paarweise verschieden und es werden daher insgesamt $n+d$ paarweise verschiedene Haplotypen erstellt. Für jeden Knotengenotyp g_{v_i} existiert ein Haplotyp h_i, der diesen erklärt. Für jeden Kantengenotyp g_{e_k} mit $e_k = \{v_i, v_j\}$ gilt: Falls $v_j \in V'$, dann wird g_{e_k} durch h_i und h'_j erklärt und falls $v_i \in V'$, dann wird g_{e_k} durch h'_i und h_j erklärt. Folglich lässt sich mit den h_i und h'_i eine Haplotypmatrix erstellen, die G erklärt und $n+d$ paarweise verschiedene Haplotypen enthält.

Wenn-Teil: Es sei H eine Haplotypmatrix, die G erklärt und $n+d$ paarweise verschiedene Haplotypen enthält. Wir erstellen nun in zwei Schritten eine Knoten-

überdeckung $V' \subseteq V$ mit $|V'| \leq d$ für U. Im ersten Schritt werden die Haplotypen in H einzeln betrachtet und gegebenenfalls verändert. Im zweiten Schritt wird aus der veränderten Haplotypmatrix eine Knotenüberdeckung abgeleitet.

Jeder Knotengenotyp in G enthält nur Einträge aus $\{0,1\}$ und kommt daher als Haplotyp in H vor. Wir setzen wieder $h_i = g_{v_i}$ für jedes $i \in \{1,\ldots,n\}$. Nun betrachten wir die Haplotypen, die Kantengenotypen erklären. Es sei dazu h ein Haplotyp in H, der nicht gleich einem Haplotyp h_i ist.

Falls h nur für genau einen Kantengenotyp g_{e_k} zur Erklärung verwendet wird, dann werden die erklärenden Haplotypen von g_{e_k}, für den $e_k = \{v_i, v_j\}$ gelte, wie folgt ersetzt: Einer der Haplotypen von g_{e_k} wird durch h_i und der andere durch h'_j mit $h'_j[j] = h'_j[n+1] = 1$ und sonstigen Einträgen 0 ersetzt. Man erkennt nun, dass die Haplotypen h_i und h'_j den Kantengenotyp g_{e_k} erklären und dass die Anzahl der paarweise verschiedenen Haplotypen nicht durch den Austausch erhöht wird.

Falls h für mehr als einen Kantengenotyp zur Erklärung verwendet wird, besitzt h an höchstens einer Position aus $\{1,\ldots,n\}$ den Eintrag 1. Dies folgt aus der Eigenschaft, dass für zwei Genotypen maximal eine Spalte aus $\{1,\ldots,n\}$ existiert, in der beide Genotypen den Eintrag 2 enthalten. Wir nehmen nun an, dass h an jeder Position aus $\{1,\ldots,n\}$ den Eintrag 0 enthält und betrachten einen beliebigen Kantengenotyp g_{e_k}, für den $e_k = \{v_i, v_j\}$ gilt und der von h und einem zweiten Haplotyp h^* erklärt wird. Es gilt $h^*[i] = h^*[j] = 1$ und folglich kann der Haplotyp h^* von keinem weiteren Haplotyp zu Erklärung verwendet werden. Das heißt, der Haplotyp h^* wird nur für g_{e_k} zur Erklärung verwendet. Somit können die Haplotypen h und h^*, wie oben beschrieben, durch den Haplotyp h_i und den Haplotyp h'_j mit $h'_j[j] = h'_j[n+1] = 1$ und sonstigen Einträgen 0 ersetzt werden, ohne dabei die Anzahl der paarweise verschiedenen Haplotypen zu erhöhen.

Nach diesem ersten Schritt ist jeder Haplotyp in der veränderten Haplotypmatrix entweder gleich einem Knotengenotyp oder gleich einem Haplotyp h'_i, der in genau einer Spalte $i \in \{1,\ldots,n\}$ und in der Spalte $n+1$ den Eintrag 1 besitzt. Die veränderte Haplotypmatrix enthält maximal $n+d$ paarweise verschiedene Haplotypen. Dies sind n viele h_i und maximal d viele h'_i.

Im zweiten Schritt wird folgendermaßen aus den h' eine Knotenüberdeckung V' für U konstruiert: Falls der Haplotyp h'_i in der Haplotypmatrix vorkommt, wird der Knoten v_i zur Menge V' hinzugefügt. Jede Kante in U wird durch einen Knoten aus V' abgedeckt, da jeder Kantengenotyp in G einen Haplotyp h'_i zur Erklärung verwendet. Die Menge V' ist eine Knotenüberdeckung mit $|V'| \leq d$ für U.

Komplexität: Die Abbildung von U nach G lässt sich in Polynomialzeit von einer deterministischen Turingmaschine berechnen. Beim Durchlauf einer Adjazenzliste, die den Graph kodiert, werden Knoten-und Kantengenotypen ausgegeben und der Budgetwert d wird um n erhöht und ausgegeben.

Insgesamt folgt, dass MH NP-vollständig ist. □

Bei der Reduktion aus dem vorangegangen Beweis wird jeder Graph auf eine Genotypmatrix mit maximal drei heterozytischen Stellen pro Genotyp abgebildet. Es wurde also auch gezeigt, dass schon MH$(3,\infty)$ NP-vollständig ist.

3.3 Haplotypisierung mittels perfekten Phylogenien

In diesem Abschnitt wird die Komplexität von Haplotypisierungsproblemen, die auf dem Ansatz mit perfekten Phylogenien beruhen, untersucht. Ein Teilproblem von jedem Haplotypisierungsproblem, das auf perfekten Phylogenien beruht, ist die Frage, ob eine Haplotypmatrix eine perfekte Phylogenie zulässt. Die Komplexität dieses Problems wird in Abschnitt 3.3.1 betrachtet. Bei Ansätzen zur Haplotypisierung, die auf perfekten Phylogenien beruhen, werden gerichtete oder ungerichtete perfekte Phylogenien gesucht. Einen Zusammenhang zwischen der Komplexität ungerichteter und gerichteter Problemvarianten wird in Abschnitt 3.3.2 hergestellt. In Abschnitt 3.3.3 wird die Komplexität von PPH und (k,l)-beschränkten Varianten von PPH untersucht. Abschnitt 3.3.4 behandelt die Komplexität von PPPH.

3.3.1 Komplexität von PP

In diesem Anschnitt wird gezeigt, dass PP in FO liegt. Aus der Literatur (siehe Gramm, Nickelsen und Tantau [15, 16] für einen Überblick) ist bekannt, dass eine Haplotypmatrix genau dann eine perfekte Phylogenie zulässt, wenn kein Spaltenpaar die *verbotene Untermatrix*

$$V = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

enthält. Da dieser Zusammenhang in der vorliegenden Arbeit an vielen Stellen Verwendung findet, wird er in Lemma 3.4 bewiesen. Der Beweis zu Lemma 3.4 verwendet dabei die Aussagen der Lemmata 3.2 und 3.4.

Sei H eine Haplotypmatrix und s ein Spalte von H. Die Menge O_s enthält genau die Haplotypen aus H, die die Spalte s umfassen.

Lemma 3.2 (Gusfield [19])**.** *Eine Haplotypmatrix H lässt genau dann eine gerichtete perfekte Phylogenie zu, wenn $O_s \subseteq O_{s'}$, $O_{s'} \subseteq O_s$ oder $O_s \cap O_{s'} = \emptyset$ für jedes Spaltenpaar (s,s') von H gilt.*

Lemma 3.3. ung-PP $\leq_{\text{fo}}$ ger-PP.

Beweis. Der folgende Beweis ist in drei Schritte aufgeteilt. Zuerst wird eine Abbildung von einer Haplotypmatrix H auf eine Haplotypmatrix H' beschrieben. Danach wird gezeigt, dass H genau dann eine perfekte Phylogenie zulässt, wenn H' eine gerichtete perfekte Phylogenie zulässt. Abschließend wird gezeigt, dass sich die Abbildung als Anfrage in Prädikatenlogik erster Stufe formulieren lässt.

Konstruktion: Sei H eine Haplotypmatrix und h_1 die erste Zeile in H. Die Haplotypmatrix H' geht folgendermaßen aus H hervor: In jeder Spalte s von H, für die $h_1[s] = 1$ gilt, wird jede 1 durch eine 0 und jeder 0 durch eine 1 ersetzt. Durch diese Abbildung entspricht die erste Zeile in H' dem Haplotyp $0 \ldots 0$ ist.

Korrektheit: Folgende Aussage wird nun gezeigt: Es existiert genau dann eine PP-Lösung für H, wenn eine gerichtete PP-Lösung für H' existiert.

Nur-wenn-Teil: Es sei H eine Haplotypmatrix, deren Zeilen als perfekte Phylogenie B_H angeordnet werden können. Nun erstellen wir aus B_H folgendermaßen eine gerichtete perfekte Phylogenie $B_{H'}$ für H': Für jeden Knoten in $B_{H'}$ wird der markierende Haplotyp an den Stellen invertiert, an denen bei der Abbildung von H nach H' invertiert wird. Der Aufbau der perfekten Phylogenie und die Kantenmarkierungen ändern sich nicht. Es lässt sich außerdem erkennen, dass $B_{H'}$ jeden Haplotyp aus H' enthält und dass die Eigenschaft 4 von Definition 2.1 erhalten bleibt. Wegen der Abbildung von H nach H' ist mindestens ein Knoten in $B_{H'}$ mit dem Haplotyp $0 \dots 0$ markiert. Diesen Knoten zeichnen wir nun als Wurzel von $B_{H'}$ aus und erhalten eine gerichtete perfekte Phylogenie für H'.

Wenn-Teil: Es sei H' eine Haplotypmatrix, die aus H mit der beschriebenen Abbildung hervorgeht und $B_{H'}$ eine gerichtete perfekte Pfadphylogenie für H'. Analog zum ersten Beweisteil entsteht eine perfekte Phylogenie B_H für H aus $B_{H'}$, indem in den Spalten, in denen bei der Abbildung von H nach H' invertiert wird, zurück getauscht wird.

Komplexität: Für einen Beweis, dass sich die Abbildung als Anfrage in Prädikatenlogik erster Stufe beschreiben lässt, sei auf Lemma 3.6 verwiesen. Zieht man für die Anfrage aus dem Beweis zu Lemma 3.6 nur Haplotypmatrizen in Betracht, so erhält man eine Anfrage in Prädikatenlogik erster Stufe, die der beschriebenen Abbildung entspricht. □

Lemma 3.4. *Eine Haplotypmatrix H lässt genau dann eine perfekte Phylogenie zu, wenn kein Spaltenpaar in H die Untermatrix V enthält.*

Beweis. Nur-wenn-Teil: Die erste Beweisrichtung wird über Kontraposition gezeigt. Wir nehmen dafür an, dass H die Untermatrix V im Spaltenpaar (s, s') enthält und führen die Frage, ob H eine perfekte Phylogenie zulässt, auf den gerichteten Fall zurück. Sei nun H' die Haplotypmatrix, die sich aus H mit der Abbildung aus dem Beweis zu Lemma 3.3 ergibt. Es lässt sich erkennen, dass auch H' die Untermatrix V im Spaltenpaar (s, s') enthält, da bei Invertierung von Spalten die Untermatrix V bestehen bleibt. Betrachten wir nun die Mengen O_s und $O_{s'}$ zum Spaltenpaar (s, s'), dann ergibt sich, dass die Mengen nicht disjunkt sind und keine Menge die andere enthält. Mit Lemma 3.2 folgt, dass H' keine gerichtete perfekte Phylogenie zulässt und mit dem Beweis zu Lemma 3.3 folgt insgesamt, dass H keine perfekte Phylogenie zulässt.

Wenn-Teil: Für diese Beweisrichtung nehmen wir an, dass eine Haplotypmatrix H nicht die Untermatrix V enthält und zeigen, dass dann H eine perfekte Phylogenie zulässt. Wie in der vorherigen Beweisrichtung wird die Frage, ob H eine perfekte Phylogenie zulässt, auf den gerichteten Fall zurückgeführt. Dazu entstehe die Haplotypmatrix H' aus H durch die Abbildung aus dem Beweis zu Lemma 3.3. Es lässt sich nun erkennen, dass jedes Spaltenpaar (s, s') in H' die Untermatrix $[0\ 0]$ enthält und kein Spaltenpaar in H' die Untermatrix V enthält, da V sonst schon in H enthalten wäre. Folglich enthält jedes Spaltenpaar (s, s') in H' höchstens drei verschiedene Zeilen und immer die Zeile $[0\ 0]$. Damit gilt $O_s \subseteq O_{s'}$, $O_{s'} \subseteq O_s$ oder

$O_s \cap O_{s'} = \emptyset$ für jedes Spaltenpaar (s, s'). Woraus mit Lemma 3.2 folgt, dass H' eine gerichtete perfekte Phylogenie zulässt und mit dem Beweis zu Lemma 3.3 weiter folgt, dass H eine perfekte Phylogenie zulässt. □

Satz 3.5. PP $\in$ FO.

Beweis. Zum Beweis, dass PP in FO liegt, wird eine Formel angegeben, die Haplotypmatrizen beschreibt, die in keinem Spaltenpaar die Untermatrix V enthalten. Die Signatur für Haplotypmatrizen $\tau_H = (Z^1, S^1, H_0^2, H_1^2)$ ist ähnlich der Signatur für Genoypmatrizen $\tau_G = (Z^1, S^1, G_0^2, G_1^2, G_2^2)$. Da eine Haplotypmatrix nur aus 0 und 1 besteht, wird die Relation, welche Einträge mit dem Wert 2 beschreibt, ausgelassen. Folgende Formel beschreibt nun PP:

$$\begin{aligned}\phi_{\text{PP}} \equiv (\forall s.S(s), s'.S(s'), z_1.Z(z_1), z_2.Z(z_2), z_3.Z(z_3), z_4.Z(z_4))[\\ \neg(H_0(z_1,s) \wedge H_0(z_1,s') \wedge \\ H_0(z_2,s) \wedge H_1(z_2,s') \wedge \\ H_1(z_3,s) \wedge H_0(z_3,s') \wedge \\ H_1(z_4,s) \wedge H_1(z_4,s'))].\end{aligned}$$

Eine Struktur (I, Z, S, H_0, H_1) ist genau dann ein Modell von ϕ_{PP}, wenn die durch (I, Z, S, H_0, H_1) kodierte Haplotypmatrix nicht die Untermatrix V enthält und daher mit Lemma 3.4 eine perfekte Phylogenie zulässt. □

Auch das Problem PPP lässt sich durch eine Anfrage in Prädikatenlogik erster Stufe beschreiben (Satz 3.20). Der Beweis hierzu ist technisch aufwändiger und baut unter anderem darauf auf, dass ger-PPPH in FO liegt.

3.3.2 Gerichtete und ungerichtete perfekte Phylogenien

Bei der Haplotypisierung mittels perfekten Phylogenien (PPH) kann die Wurzel einer perfekten Phylogenie mit einem beliebigen Haplotyp markiert sein. Dieses Problem ist der ungerichtete Fall der Haplotypisierung mittels perfekten Phylogenien und wird im Weiteren auch ung-PPH genannt. Bei der Haplotypisierung mittels gerichteten perfekten Phylogenien (ger-PPH) muss die Wurzel einer perfekten Phylogenie mit dem Haplotyp $0 \ldots 0$ markiert werden können. Dieser Abschnitt macht klar, dass sich die Komplexität der ungerichteten Problemvariante nur sehr wenig von der Komplexität der gerichteten Problemvariante unterscheidet. Es wird hierzu im Folgenden gezeigt, dass sich die Probleme ung-PPH und ger-PPH durch Anfragen in Prädikatenlogik erster Stufe aufeinander reduzieren lassen und dass dies ebenfalls für die Minimierungsvarianten der Probleme gilt. Für die Haplotypisierungprobleme, bei denen nach perfekten Pfadphylogenien gesucht wird, ist ein solcher Zusammenhang zwischen der gerichteten und der ungerichteten Variante nicht bekannt.

Im folgenden Lemma wird gezeigt, dass sich ung-PPH mit einer Anfrage in Prädikatenlogik erster Stufe auf ger-PPH reduzieren lässt. Die Abbildung von Genotypmatrizen nach Genotypmatrizen, die dabei verwendet wird, geht auf Eskin, Halperin und Karp [12] zurück. Neu an dem Beweis zum folgenden Lemma ist die Verwendung einer Anfrage in Prädikatenlogik erster Stufe, die die geringe Komplexität der Abbildung hervorhebt.

Lemma 3.6. ung-PPH $\leq_{\text{fo}}$ ger-PPH.

Beweis. Der Beweis zur Reduktion ist in drei Schritte aufgeteilt. Zuerst wird eine Abbildung von Genotypmatrizen nach Genotypmatrizen angegeben und danach gezeigt, dass die Abbildung eine Reduktion von ung-PPH auf ger-PPH darstellt. Abschließend wird eine Anfrage in Prädikatenlogik erster Stufe formulieren, die der Abbildung entspricht.

Konstruktion: Eine Genotypmatrix G wird folgendermaßen auf eine gleich große Genotypmatrix G' abgebildet: In jeder Spalte von G wird nach der ersten Zeile gesucht, die nicht den Eintrag 2 enthält. Falls an dieser Stelle eine 1 steht, wird in der Spalte jede 0 durch 1 und jede 1 durch 0 ersetzt.

Korrektheit: Folgende Behauptung wird nun gezeigt: Es existiert genau dann eine PP-Lösung der Größe d für G, wenn eine gerichtete PP-Lösung der Größe d für G' existiert.

Nur-wenn-Teil: Sei G eine Genotypmatrix und (H, B_H) eine PP-Lösung der Größe d. Es wird nun gezeigt, dass für G' eine gerichtete PP-Lösung der Größe d existiert. Hierzu entstehe die Haplotypmatrix H' aus der Haplotypmatrix H, indem H in den Spalten invertiert wird, in denen bei der Abbildung von G nach G' die Rollen von 0 und 1 vertauscht werden. Nun wird gezeigt, dass G' durch H' erklärt wird, H' eine perfekte Phylogenie zulässt und H' genau d paarweise verschiedene Haplotypen enthält.

Wir zeigen zuerst, dass G' durch H' erklärt wird. Hierzu sei g ein beliebiger Genotyp in G und h_1, h_2 die erklärenden Haplotypen zu g in H sowie g' der entsprechende Genotyp in G' und h'_1, h'_2 die entsprechenden Haplotypen in H'. Um nun zu zeigen, dass g' durch h'_1 und h'_2 erklärt wird sei s eine beliebige Spalte, in der bei der Transformation invertiert wird. Falls $g[s] = a$ mit $a \in \{0,1\}$ gilt, dann gilt nach Voraussetzung $g[s] = h_1[s] = h_2[s] = a$ und mit der Konstruktion von H ebenso $g'[s] = h'_1[s] = h'_2[s] = 1 - a$. Falls andererseits $g[s] = 2$, dann gilt nach Voraussetzung $h_1[s] \neq h_2[s]$ und mit der Konstruktion von H ebenso $h'_1[s] \neq h'_2[s]$. In jeder Spalte, in der nicht invertiert wird, sind die Einträge jeweils identisch und es folgt insgesamt, dass g' durch h'_1 und h'_2 erklärt wird. Da wir g' beliebig gewählt haben folgt, dass G' durch H' erklärt wird.

Als nächstes zeigen wir, dass H' eine gerichtete perfekte Phylogenie zulässt. Dazu sei (s, s') ein beliebiges Spaltenpaar und g_0 die erste Zeile in G, für welche $g_0[s] \neq 2$ oder $g_0[s'] \neq 2$ gilt. Schaut man sich nun die möglichen Einträge für $g'_0[s]$ und $g'_0[s']$ an, so lässt sich erkennen, dass einer der erklärenden Haplotypen für g'_0 in den Spalten s und s' den Eintrag 0 enthält. Analog zum Beweis von Lemma 3.4 lässt sich zeigen, dass das Spaltenpaar (s, s') in H' nicht die Untermatrix V enthält,

da das Spaltenpaar (s,s') in H nicht die Untermatrix V enthält. Erweitert man H um einen Haplotyp $0\ldots0$, so enthält weiterhin kein Spaltenpaar die Untermatrix V. Für die erweiterte Haplotypmatrix existiert daher eine perfekte Phylogenie, die den Haplotyp $0\ldots0$ enthält und eine gerichtete perfekte Phylogenie für H darstellt.

Abschließend zeigen wir nun, dass H genau d paarweise verschiedene Haplotypen enthält. Zu diesem Zweck seien h_1 und h_2 zwei Haplotypen in H und h'_1 und h'_2 die entsprechenden Haplotypen in H' sowie s eine beliebige Spalte. Wegen der Abbildung von G nach G' folgt, dass genau dann $h_1[s] = h_2[s]$ gilt, wenn $h'_1[s] = h'_2[s]$ gilt und somit sind zwei Haplotypen genau dann in H gleich, wenn sie in H' gleich sind. Dies bedeutet, H und H' enthalten die gleiche Anzahl paarweise verschiedener Haplotypen.

Wenn-Teil: Sei G' eine Genotypmatrix, die durch die Abbildung aus der Genotypmatrix G hervorgeht und $(H', B_{H'})$ eine gerichtete PP-Lösung der Größe d für G'. Wir zeigen nun, dass dann für G eine PP-Lösung (H, B_H) existiert. Hierzu entstehe die Haplotypmatrix H aus der Haplotypmatrix H', indem H' in den Spalten invertiert wird, in denen bei der Abbildung von G nach G' invertiert wird. Analog zur ersten Beweisrichtung wird G durch H erklärt und H enthält genau d paarweise verschiedene Haplotypen. Des Weiteren enthält H nicht die Untermatrix V, da H' nicht die Untermatrix V enthält und lässt somit eine perfekte Phylogenie zu.

Komplexität: Es wird nun gezeigt, dass sich die Abbildung von G nach G' durch eine Anfrage $A_{\text{ung}-\text{ger}}$ in Prädikatenlogik erster Stufe formulieren lässt. Die Anfrage $A_{\text{ung}-\text{ger}}$ bildet jede τ_G-Struktur (I, Z, S, G_0, G_1, G_2) auf eine τ_G-Struktur $(I', Z', S', G'_0, G'_1, G'_2)$ ab und wird durch die Formeln $\phi_{I'}(x)$, $\phi_{Z'}(z)$, $\phi_{S'}(s)$, $\phi_{G'_0}(z,s)$, $\phi_{G'_1}(z,s)$ und $\phi_{G'_2}(z,s)$ definiert. Folgende Formel wird dabei mehrfach als Teilformel verwendet und beschreibt eine Spalte, in der 0 und 1 bei der Abbildung vertauscht werden:

$$\phi_{\text{tausch}}(s) \equiv (\exists z'.Z(z'))[G_1(z',s) \wedge (\forall z.Z(z))[G_0(z,s) \rightarrow z' \leq z]].$$

Die Formeln für $A_{\text{ung}-\text{ger}}$ sind wie folgt definiert:

$$\begin{aligned}
\phi_{I'}(i) &\equiv \mathit{wahr},\\
\phi_{Z'}(z) &\equiv \mathit{wahr},\\
\phi_{S'}(s) &\equiv \mathit{wahr},\\
\phi_{G'_0}(z,s) &\equiv (G_1(z,s) \wedge \phi_{\text{tausch}}(s)) \vee (G_0(z,s) \wedge \neg\phi_{\text{tausch}}(s)),\\
\phi_{G'_1}(z,s) &\equiv (G_0(z,s) \wedge \phi_{\text{tausch}}(s)) \vee (G_1(z,s) \wedge \neg\phi_{\text{tausch}}(s)),\\
\phi_{G'_2}(z,s) &\equiv G_2(z,s).
\end{aligned}$$

Nun ist $A_{\text{ung}-\text{ger}} = \lambda_{z,s}(\phi_{I'}, \phi_{Z'}, \phi_{S'}, \phi_{G'_0}, \phi_{G'_1}, \phi_{G'_2})$ eine Anfrage in Prädikatenlogik erster Stufen, die der Abbildung von G nach G' entspricht. □

Die Reduktion im vorangegangenen Beweis verändert nicht die Größe von PP-Lösungen. Somit gilt auch ung-MPPH $\leq_{\text{fo}}$ ger-MPPH.

Lemma 3.7. ger-PPH $\leq_{\text{fo}}$ ung-PPH.

Beweis. Für eine Reduktion von ger-PPH auf ung-PPH bilden wir eine Genotypmatrix G wie folgt auf eine Genotypmatrix G' ab: Falls G die Zeile $0\ldots0$ enthält, dann ist G' gleich G und ansonsten entsteht G' durch das Hinzufügen der Zeile $0\ldots0$ zu G. Nun ist eine gerichtete PP-Lösung für G auch eine PP-Lösung für G' und aus einer PP-Lösung für G' entsteht eine gerichtete PP-Lösung für G, indem der Knoten, der mit dem Haplotyp $0\ldots0$ markiert ist, die Wurzel wird. Der Test, ob eine Genotypmatrix die Zeile $0\ldots0$ enthält und das Anfügen einer Zeile $0\ldots0$, das davon abhängt, lässt sich als Anfrage in Prädikatenlogik erster Stufe formulieren. □

Die im Beweis zu Lemma 3.7 beschriebene Reduktion lässt sich zu einer Reduktion von ger-MPPH auf ung-MPPH erweitern, indem eine Genotypmatrix G wie im Beweis zu Lemma 3.7 auf eine Genotypmatrix G' abgebildet wird und ein Budgetwert d folgendermaßen auf einen Budgetwert d' abgebildet wird: Falls die Zeile $0\ldots0$ in G enthalten ist, dann wird $d' = d$ gesetzt und ansonsten wird $d' = d+1$ gesetzt. Nun gilt, dass G genau dann eine gerichtete PP-Lösung der Größe d besitzt, wenn G' eine PP-Lösung der Größe d' besitzt. Die gesamte Reduktion lässt sich als Anfrage in Prädikatenlogik erster Stufe formulieren und folglich gilt ger-MPPH $\leq_{\text{fo}}$ ung-MPPH.

3.3.3 Komplexität von PPH

Dieser Abschnitt behandelt die Komplexität von PPH und (k,l)-beschränkten Varianten von PPH. Zuerst wird ein Überblick über bekannte Resultate zu PPH gegeben. Danach wird gezeigt, dass PPH L-hart ist und dass PPH in Mod_2L liegt. Die Beweisideen hierzu stammen von Arfst Nickelsen und Till Tantau, die diese aber noch nicht schriftlich festgehalten haben. In dieser Arbeit werden erste ausführliche Beweise zu diesen Resultaten gegeben. Im Anschluss wird gezeigt, dass PPH$(2,\infty)$ und PPH$(\infty,1)$ in FO liegen. Diese Resultate sind neu und helfen, die (k,l)-beschränkten Varianten von PPH bezüglich ihrer Komplexität zu ordnen.

Die Haplotypisierung mittels perfekten Phylogenien wurde von Gusfield [20] vorgeschlagen und ein erster Algorithmus, der PPH in Polynomialzeit löst, vorgestellt. Weitere Algorithmen, die PPH ebenfalls in Polynomialzeit lösen, aber einfacher zu implementieren sind, wurden daraufhin veröffentlicht [12][3]. Die Zeitkomplexität von PPH wurde weiter untersucht und es wurden Algorithmen vorgestellt, die PPH in Linearzeit entscheiden [11, 36, 30]. Von Gusfield [18] wurde gezeigt, dass ein Algorithmus, der PP löst, mindestens Linearzeit benötigt. Das Problem PP ist ein Teilproblem von PPH, da man Haplotypmatrizen als Genotypmatrizen ohne heterozytische Einträge ansehen kann. Folglich überträgt sich die untere Schranke zur Zeitkomplexität von PP auf PPH.

Im Weiteren werden induzierte Mengen für Spaltenpaare beschrieben. Induzierte Mengen wurden zuerst von Eskin, Halperin und Karp [12] eingeführt, um einen Zusammenhang zwischen einer Genotypmatrix und einer erklärenden Haplotypmatrix herzustellen. Es sei nun G eine Genotypmatrix und H eine Haplotypma-

trix, die G erklärt sowie (s,s') ein beliebiges Spaltenpaar. Es lässt sich erkennen, dass jede Untermatrix der Spalten s und s' in G, die nur Einträge aus $\{0,1\}$ enthält, eine Untermatrix der Spalten s und s' in H ist. Weiter gilt, dass H in den Spalten s und s' die Untermatrix $\left[\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right]$ enthält, falls G in den Spalten s und s' die Untermatrix $[2\ 0]$ enthält. Analog enthält H die Untermatrix $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ in den Spalten s und s', falls G die Untermatrix $[1\ 2]$ in den Spalten s und s' enthält. Aus einer Genotypmatrix können wir auf diese Weise Information über eine erklärende Haplotypmatrix gewinnen. Diese Eigenschaft wird für eine Genotypmatrix und ein Spaltenpaar (s,s') durch die *induzierte Menge* $I(s,s')$ formel gefasst, die wie folgt definiert ist: Falls (s,s') die Untermatrix $[a\ b]$ mit $a,b \in \{0,1\}$ enthält, dann gilt $(a,b) \in I(s,s')$. Falls (s,s') die Untermatrix $[a\ 2]$ mit $a \in \{0,1\}$ enthält, dann gilt $\{(a,0),(a,1)\} \subseteq I(s,s')$. Analog gilt $\{(0,a),(1,a)\} \subseteq I(s,s')$, falls (s,s') die Untermatrix $[2\ a]$ mit $a \in \{0,1\}$ enthält. Induzierte Mengen für Haplotypmatrizen sind auf gleiche Weise definiert. Es werden dabei nur die Fälle mit heterozytischen Einträgen ausgelassen. Abschließend können wir feststellen, dass die induzierte Menge für ein Spaltenpaar (s,s') in einer Genotypmatrix angibt, welche Einträge eine erklärende Haplotypmatrix mindestens in den Spalten (s,s') enthält. Es gilt also folgende Bemerkung:

Bemerkung 3.8. Sei G eine Genotypmatrix und H eine Haplotypmatrix, die G erklärt sowie (s,s') ein Spaltenpaar. Sei $I(s,s')$ die induzierte Menge von (s,s') in G und $J(s,s')$ die induzierte Menge von (s,s') in H. Dann gilt: $I(s,s') \subseteq J(s,s')$.

Für einen Genotyp, der mehr als einen heterozytischen Eintrag besitzt, existieren mehrere Paare erklärender Haplotypen. Falls eine Genotypmatrix einen solchen Genotyp enthält, existieren für sie mehrere erklärende Haplotypmatrizen. Für einen Genotyp g, der in einem Spaltenpaar (s,s') die Untermatrix $[2\ 2]$ enthält, enthalten die beiden erklärenden Haplotypen die Untermatrizen $\left[\begin{smallmatrix}0&0\\1&1\end{smallmatrix}\right]$ oder $\left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$ in (s,s'). Falls die erklärenden Haplotypen $\left[\begin{smallmatrix}0&0\\1&1\end{smallmatrix}\right]$ enthalten, sagt man: *g wird in (s,s') gleich aufgelöst*. Andernfalls sagt man: *g wird in (s,s') ungleich aufgelöst*. Sucht man für eine Genotypmatrix eine Haplotypmatrix, die diese nicht nur erklärt sondern auch eine perfekte Phylogenie zulässt, dann können in einem Spaltenpaar alle Genotypen entweder nur gleich oder nur ungleich aufgelöst werden, da sonst die verbotene Untermatrix V aus Lemma 3.4 entsteht. Aus ähnlichem Grund kann die Auflösung in einem Spaltenpaar auch schon durch die induzierte Menge vorgegeben sein. Falls $\{00,11\} \subseteq I(s,s')$ für eine Genotypmatrix G und ein Spaltenpaar (s,s') gilt, dann wird in einer Haplotypmatrix, die G erklärt und eine perfekte Phylogenie zulässt, jeder Genotyp in (s,s') gleich aufgelöst und man sagt: *(s,s') wird gleich aufgelöst*. Bei $\{01,10\} \subseteq I(s,s')$ wird jeder Genotyp in (s,s') ungleich aufgelöst und man sagt: *(s,s') wird ungleich aufgelöst*.

Nun wird gezeigt, dass sich UNGWEG, das Erreichbarkeitsproblem in ungerichteten Graphen, mit einer Anfrage in Prädikatenlogik erster Stufe auf $\overline{\text{PPH}}$ reduzieren lässt.

Lemma 3.9. UNGWEG $\leq_{\text{fo}} \overline{\text{PPH}(3,\infty)}$.

Beweis. Eine Instanz von UNGWEG setzt sich aus einem ungerichteten Graph $U = (V,E)$ und zwei ausgezeichneten Knoten $s,t \in V$ zusammen. Das Problem UNGWEG ist, zu entscheiden, ob in U ein Weg von s nach t existiert.

Der Beweis zur Reduktion ist in drei Schritte aufgeteilt. Zuerst wird eine Abbildung von einem Graph U mit zwei ausgezeichneten Knoten s und t auf eine Genotypmatrix G angegeben. Danach wird gezeigt, dass die Fragen, ob in U ein Weg von s nach t existiert und ob G keine perfekte Phylogenie zulässt, für U und G immer auf gleiche Weise beantwortet werden. Abschließend wird die Abbildung als Anfrage in Prädikatenlogik erster Stufe formuliert.

Konstruktion: Es sei eine Instanz von UNGWEG gegeben, die aus dem ungerichteten Graph $U = (V,E)$ und zwei ausgezeichneten Knoten $s,t \in V$ besteht. Der Graph U wird in zwei Schritten auf eine Genotypmatrix G abgebildet. Im ersten Schritt wird der Graph U folgendermaßen auf einen Graph $U' = (V',E')$ abgebildet: Jede Kante in U wird durch einen Pfad der Länge 2 ersetzt. Das heißt, $V \subseteq V'$ und für jede Kante $e = \{v,w\} \in E$ wird der Knoten u_e zu V' hinzugefügt und die Kanten $\{v,u_e\}$ und $\{u_e,w\}$ werden zu E' hinzugefügt. Im zweiten Schritt wird der Graph U' wie folgt auf eine Genotypmatrix G abgebildet: Es sei $V' = \{v_1,\ldots,v_n\}$ und $E' = \{e_1,\ldots,e_m\}$. Die Genotypmatrix G besteht aus $m+2$ Zeilen und $n+1$ Spalten. Die ersten m Zeilen entstehen dadurch, dass für jede Kante $e_k = \{v_i,v_j\} \in E'$ der Genotyp g_{e_k} mit $g_{e_k}[0] = g_{e_k}[i] = g_{e_k}[j] = 2$ und sonstigen Einträgen 0 erstellt wird. Sei nun $v_i = s$ und $v_j = t$. Weitere zwei Zeilen entstehen dadurch, dass die Genotypen g_1 mit $g_1[0] = g_1[i] = 1$ und sonstigen Einträgen 0 und g_2 mit $g_2[j] = 1$ und sonstigen Einträgen 0 erstellt werden. Es lässt sich erkennen, dass jeder Genotyp maximal drei heterozytische Einträge besitzt. Die auf diese Weise konstruierte Genotypmatrix ist somit eine Instanz von PPH$(3,\infty)$. Im vorangegangen Teil dieser Arbeit wurden für Zeilen und Spalten nur Indizes verwendet, die größer als 0 sind. Bei dieser Abbildung wird der Index 0 verwendet, was keine Erweiterung darstellt, da der Inhalt der Spalte 0 genauso in Spalte $n+1$ eingefügt werden kann.

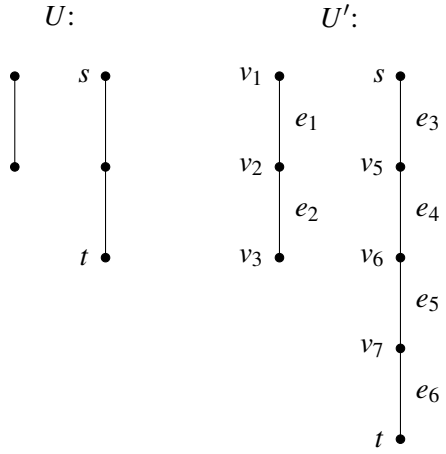
Korrektheit: Folgende Behauptung wird nun gezeigt: Es existiert genau dann ein Weg von s nach t in U, wenn keine PP-Lösung für G existiert.

Zwischen den Knoten von U' und m Spalten von G gibt es nach der Konstruktion eine $1:1$ Beziehung. Genauso gibt es zwischen n Zeilen von G, die aus Kanten konstruiert werden, und den Kanten von U' eine $1:1$ Beziehung. Im Folgenden bezeichnen wir die Spalte, die zu einem Knoten v gehört, ebenfalls mit v und einen Genotyp, der zu einer Kante e gehört, ebenfalls mit e.

Nur-wenn-Teil: Für diese Beweisrichtung nehmen wir an, dass in U ein Weg von s nach t existiert und zeigen, dass dann G keine perfekte Phylogenie zulässt. Seien $(w_1,\ldots,w_l)$ mit $w_1 = s$ und $w_l = t$ die Knoten und $(f_1,\ldots,f_{l-1})$ die Kanten auf einem Weg von s nach t. Es gilt also $\{w_i,w_{i+1}\} = f_i \in E'$ für jeden Index $i \in \{1,\ldots,l-1\}$. Wegen der Abbildung von U nach U' umfasst jeder Weg von s nach t in U' eine gerade Anzahl Kanten und l ist somit ungerade.

In G enthalten das Spaltenpaar (w_1,w_2) die Untermatrix $\left[\begin{smallmatrix}0&2\\1&0\end{smallmatrix}\right]$, jedes Spaltenpaar (w_i,w_{i+1}) mit $i \in \{2,\ldots,l-2\}$ die Untermatrix $\left[\begin{smallmatrix}2&0\\0&2\end{smallmatrix}\right]$ und das Spaltenpaar

Abbildung 4: Die Abbildung zeigt beispielhaft die Reduktion von UNGWEG auf $\overline{\text{PPH}}$. Im ersten Schritt wird aus dem Graph U der Graph U' konstruiert und im zweiten Schritt wird aus U' eine Genotypmatrix G abgeleitet. Die letzten acht Spalten von G korrespondieren zu den acht Knoten von U'. Die ersten sechs Zeilen von G korrespondieren zu den sechs Kanten von U'. Zu sehen ist, dass das Spaltenpaar $(0,s)$ gleich aufgelöst wird und dass das Spaltenpaar $(0,t)$ ungleich aufgelöst wird.



	0	v_1	v_2	v_3	s	v_5	v_6	v_7	t
g_{e_1}	2	2	2	0	0	0	0	0	0
g_{e_2}	2	0	2	2	0	0	0	0	0
g_{e_3}	2	0	0	0	2	2	0	0	0
g_{e_4}	2	0	0	0	0	2	2	0	0
g_{e_5}	2	0	0	0	0	0	2	2	0
g_{e_6}	2	0	0	0	0	0	0	2	2
g_1	1	0	0	0	1	0	0	0	0
g_2	0	0	0	0	0	0	0	0	1

(w_{l-1}, w_l) die Untermatrix $\left[\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Folglich wird für jedes $i \in \{1,\ldots,l-1\}$ das Spaltenpaar (w_i, w_{i+1}) in G ungleich aufgelöst.

In jeder Haplotypmatrix, die G erklärt und eine perfekte Phylogenie zulässt, wird jedes Spaltenpaar entweder gleich oder ungleich aufgelöst. Es wird nun gezeigt, dass keine solche Haplotypmatrix für G existiert, indem die Genotypen zusammen mit möglichen erklärenden Haplotyppaaren Schritt für Schritt betrachtet werden. Wir beginnen mit dem Genotyp f_1, der genau in den Spalten 0, w_1 und w_2 den Eintrag 2 enthält. Das Spaltenpaar $(0, w_1)$ wird gleich aufgelöst, da es die Untermatrix $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ enthält und das Spaltenpaar (w_1, w_2) wird, wie schon beschrieben, ungleich aufgelöst. Seien nun h und h' die auf die Spalten 0, w_1 und w_2 eingeschränkten erklärenden Haplotypen zu f_1. Dann gilt $h = 001$ und $h' = 110$ oder $h = 110$ und $h' = 001$. Hieraus folgt, dass in jeder PP-Lösung für G das Spaltenpaar $(0, w_2)$ ungleich aufgelöst wird. Als nächstes betrachten wir den Genotyp f_2, der genau in den Spalten 0, w_2 und w_3 den Eintrag 2 enthält. Das Spaltenpaar $(0, w_2)$ wird, wie gerade gezeigt, ungleich aufgelöst und das Spaltenpaar (w_2, w_3) wird ebenfalls ungleich aufgelöst. Für die erklärenden Haplotypen h und h' von f_2, eingeschränkt auf die Spalten 0, w_2 und w_3, ergibt sich $h = 010$ und $h' = 101$ oder $h = 101$ und $h' = 010$. Folglich wird das Spaltenpaar $(0, w_3)$ in jeder PP-Lösung für G ungleich aufgelöst. Diese Argumentation lässt sich Schritt für Schritt für jeden Genotyp, dessen zugehörige Kante auf dem Weg von s nach t liegt, fortsetzen. Für jedes Spaltenpaar $(0, w_i)$ wird dabei eine Auflösung hergeleitet, die das Spaltenpaar $(0, w_i)$ in jeder PP-Lösung besitzt. Es lässt sich nun feststellen, dass das Spal-

tenpaar $(0, w_i)$ gleich aufgelöst wird, falls i ungerade ist und ungleich aufgelöst wird, falls i gerade ist. Es gilt also, dass in jeder PP-Lösung das Spaltenpaar $(0, w_l)$ gleich aufgelöst wird, da l ungerade ist. Da aber $(0, w_l)$ die Untermatrix $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ enthält, muss $(0, w_l)$ ungleich aufgelöst werden. Es existiert keine PP-Lösung für G, die beide Bedingungen erfüllt und somit lässt G keine perfekte Phylogenie zu.

Wenn-Teil: Diese Beweisrichtung wird über Kontraposition gezeigt. Wir nehmen dazu an, dass in U kein Weg von s nach t existiert und zeigen, dass dann G eine perfekte Phylogenie zulässt. Für jeden der Genotypen in G wird im Folgenden ein erklärendes Haplotypenpaar erstellt und es wird gezeigt, dass sich die so konstruierten Haplotypen als perfekte Phylogenie anordnen lassen. Zuerst stellen wir fest, dass in U' ebenfalls kein Weg von s nach t existiert und U' daher aus mindestens zwei Komponenten besteht. Wir betrachten nun Schritt für Schritt die verschiedenen Komponenten in U' und leiten für Genotypen, die zu Kanten in der jeweiligen Komponente korrespondieren, erklärende Haplotypen ab.

Sei V_s die Komponente in U', die s enthält und $f = \{v, w\}$ eine Kante in V_s. Nach Konstruktion von U' kommt genau ein Endknoten von f in U vor und der andere Endknoten wird erst bei der Abbildung hinzugefügt. Sei v der Knoten, der in U vorkommt und w der Knoten, der nicht in U vorkommt. Für den Genotyp f setzen wir die erklärenden Haplotypen h und h', eingeschränkt auf die Spalten 0, v und w, auf $h = 001$ und $h' = 110$. In den übrigen Spalten ist der Genotyp f homozytisch und die Einträge in den erklärenden Haplotypen sind folglich vorgegeben. Es lässt sich erkennen, dass die Spaltenpaare (v, w) und $(0, w)$ ungleich aufgelöst werden und dass das Spaltenpaar $(0, v)$ gleich aufgelöst wird. Weiterhin gilt, dass jeder Weg von s nach v in U' eine gerade Länge besitzt und dass jeder Weg von s nach w in U' eine ungerade Länge besitzt.

Nun betrachten wir die Komponenten V_t, die t enthält und eine beliebige Kante $f = \{v, w\}$ in V_t. Wie im vorherigen Fall sei v ein Knoten, der auch in U vorkommt und von t nur auf Wegen gerader Länge erreichbar ist und w ein Knoten, der nicht in U vorkommt und von t nur auf Wegen ungerader Länge erreichbar ist. Für den Genotyp f setzen wir die erklärenden Haplotypen h und h', eingeschränkt auf die Spalten 0, v und w, auf $h = 010$ und $h' = 010$. Dabei werden die Spaltenpaare (v, w) und $(0, v)$ ungleich aufgelöst und das Spaltenpaar $(0, w)$ wird gleich aufgelöst. An den übrigen Stellen ist der Genotyp f homozytisch und die Einträge sind somit vorgegeben.

Sei K eine Komponente in U', die nicht s und nicht t enthält und sei v^* ein beliebiger fest gewählter Knoten in K. Die erklärenden Haplotypen für die Genotypen der Kanten von K werden analog zur Komponente V_s erstellt, wobei der Knoten v^* die Rolle von s übernimmt.

Sei nun H eine Haplotypmatrix, die neben den oben erstellten Haplotypen nur die homozytischen Genotypen g_1 und g_2, die aus der Konstruktion bekannt sind, enthält. Wir zeigen, dass H eine perfekte Phylogenie zulässt. Dazu sei v ein Knoten in V_s. Falls ein Weg von s nach v gerader Länge existiert, dann wird jeder Genotyp, der v umfasst, im Spaltenpaar $(0, v)$ gleich aufgelöst. Es gilt damit, dass $(0, v)$ in H nicht die Untermatrix $\left[\begin{smallmatrix} 0 & 1 \end{smallmatrix}\right]$ enthält. Falls andererseits ein Weg von s nach v un-

gerader Länge existiert, dann wird $(0,v)$ ungleich aufgelöst und H enthält in $(0,v)$ nicht die Untermatrix $\left[\begin{smallmatrix}1 & 1\end{smallmatrix}\right]$. Für jeden Knoten in V_t und jeden Knoten einer beliebigen Komponente lässt sich diese Argumentation analog führen. Für zwei beliebige Knoten v und w aus U' wird das Spaltenpaar (v,w) ungleich aufgelöst, falls eine Kante $\{v,w\}$ in U' existiert und benötigt keine Auflösung, falls die Kante $\{v,w\}$ nicht in U' vorkommt. In beiden Fällen enthält (v,w) in H nicht die Untermatrix $\left[\begin{smallmatrix}1 & 1\end{smallmatrix}\right]$. Insgesamt gilt, dass kein Spaltenpaar in H die Untermatrix V aus Lemma 3.4 enthält und H damit eine PP-Lösung für G darstellt.

Komplexität: Im Folgenden wird eine Anfrage in Prädikatenlogik erster Stufe angegeben, die der Abbildung vom Graph U auf die Genotypmatrix G entspricht. Die Anfrage wird, wie die Abbildung, in zwei Schritten beschrieben.

Zuerst wird nun die Kodierung von Graphen als Struktur erläutert. Die Signatur eines Graphen mit zwei ausgezeichneten Knoten ist $\tau_U = (E^2, s, t)$. Eine τ_U-Struktur $U = (V, E, s, t)$ besteht aus der Knotenmenge V, der zweistelligen Kantenrelation $E \subseteq V \times V$ und zwei Knoten $s,t \in V$. Da die Relation E einer Menge von gerichteten Kanten entspricht, wird im Weiteren gefordert, dass die Formel $(\forall v,w)[E(v,w) \leftrightarrow E(w,v)]$ für jede τ_U-Struktur gilt. Knoten, die durch eine Kante verbunden sind, sind somit immer in beide Richtungen verbunden, was einem ungerichteten Graphen entspricht. Wie in Abschnitt 3.1 beschrieben, wird eine Genotypmatrix durch eine τ_G-Struktur (I, Z, S, G_0, G_1, G_2) kodiert.

Die Elemente einer Struktur sind total geordnet. So lässt sich für eine Struktur die Konstante max festlegen, die den Wert des maximalen Elements in der Struktur angibt. Wir nehmen im Folgenden an, dass max ein Teil jeder Struktur ist, was der Definition von Strukturen bei Immerman [25, Seite 13f] entspricht.

Nun wird die Anfrage $A_{U-U'}$, die der Abbildung von U nach U' entspricht, durch die Formeln $\phi_{V'}$ und $\phi_{E'}$ definiert. Die Menge der Knoten V' von U' wird wie folgt beschrieben:

$$\phi_{V'}(v) \equiv v \leq \max + \max^2.$$

Für jeden Knoten aus U und jede mögliche ungerichtete Kante existiert ein Knoten in U'. Die Anzahl der Knoten in U' ist quadratisch in der Anzahl der Knoten in U.

Zwei Knoten v und w sind in U' adjazent, falls w aus einer Kante entsteht, die in U mit v inzident ist oder v aus einer Kante entsteht, die in U mit w inzident ist. Folgende Formel beschreibt die Kantenrelation E' von U', wobei der Index eines Knotens, der für eine ungerichtete Kante steht, aus den Indizes der entsprechenden inzidenten Knoten berechnet wird:

$$\begin{aligned}\phi_{E'}(v,w) \equiv (\exists u)[&(E(v,u) \wedge \\ &(v < u \rightarrow w = \max + (v-1)\cdot\max + u - 1) \wedge \\ &(v > u \rightarrow w = \max + (u-1)\cdot\max + v - 1)) \vee \\ &(E(u,w) \wedge \\ &(u < w \rightarrow v = \max + (u-1)\cdot\max + w - 1) \wedge \\ &(u > w \rightarrow v = \max + (w-1)\cdot\max + u - 1))].\end{aligned}$$

Somit ist $A_{U-U'} = \lambda_{zs}(\phi_{V'}, \phi_{E'})$ eine Anfrage in Prädikatenlogik erster Stufen, die der Abbildung von U nach U' entspricht.

Nun wird die Anfrage $A_{U'-G}$, die der Abbildung von U' nach G entspricht, formuliert. Die Anfrage $A_{U'-G}$ bildet auf Genotypmatrizen ab und ist durch die folgenden sechs Formeln definiert:

$$\begin{aligned}
\phi_I(i) &\equiv i \leq \max^2 + 2, \\
\phi_Z(z) &\equiv (\exists v, w)[v < w \wedge E(v,w) \wedge z = (v-1) \cdot \max + (w-1)] \vee \\
&\quad z = \max^2 + 1 \vee z = \max^2 + 2 \\
\phi_S(c) &\equiv c \leq \max + 1 \\
\phi_{G_2}(z,c) &\equiv (\exists v, w)[v < w \wedge E(v,w) \wedge z = (v-1) \cdot \max + (w-1) \wedge \\
&\quad (c = v \vee c = w \vee c = \max + 1)], \\
\phi_{G_1}(z,c) &\equiv (z = \max^2 + 1 \wedge (c = s \vee c = \max + 1)) \vee \\
&\quad (z = \max^2 + 2 \wedge c = t), \\
\phi_{G_0}(z,c) &\equiv (\exists v, w)[v < w \wedge E(v,w) \wedge z = (v-1) \cdot \max + (w-1) \wedge \\
&\quad c \neq v \wedge c \neq w \wedge c \leq \max] \vee \\
&\quad (z = \max^2 + 1 \wedge c \neq s \wedge c \leq \max) \vee \\
&\quad (z = \max^2 + 2 \wedge c \neq t \wedge c \leq \max + 1)]
\end{aligned}$$

Eine Genotypmatrix, die aus der Anfrage entsteht, kann leere Zeilen enthalten. Genau heißt dies, falls zwei Knoten v und w mit $v < w$ in U nicht adjazent sind oder $v > w$ gilt, dann ist die Zeile $z = (v-1) \cdot \max + (w-1)$ leer. Keine Spalte enthält in diesen Zeilen den Eintrag 0, 1 oder 2. Dies wird aber durch die Relation Z, die genau die Indizes enhält, die Zeilenindizes sind, abgefangen. Die Relation Z umfasst nur die Zeilenindizes, die keine leeren Zeilen indizieren. Die Anfrage $A_{U'-G} = \lambda_{zs}(\phi_I, \phi_Z, \phi_S, \phi_{G_0}, \phi_{G_1}, \phi_{G_2})$ in Prädikatenlogik erster Stufen entspricht der Abbildung von U' nach G und die Komposition der Anfragen $A_{U-U'}$ und $A_{U'-G}$ entspricht der Abbildung von U nach G. Die gesamte Abbildung lässt sich somit als Anfrage in Prädikatenlogik erster Stufe formulieren, da Anfragen in Prädikatenlogik erster Stufe unter Komposition abgeschlossen sind. □

Die Klasse L enthält genau die Entscheidungsprobleme, die von einer deterministischen Turingmaschine auf logarithmischem Platz gelöst werden können. Ein Problem ist L-hart und damit mindestens so schwer wie jedes Problem in L, falls jedes Problem aus L auf dieses reduziert werden kann. Damit nicht schon durch die Reduktion selbst das zu reduzierende Problem gelöst wird, beschränkt man die Komplexität von Reduktionen. Im Folgenden wird gezeigt, dass $\mathrm{PPH}(3,\infty)$ bezüglich Reduktionen, die sich als Anfrage in Prädikatenlogik erster Stufe formulieren lassen, L-hart ist.

Satz 3.10. $\mathrm{PPH}(3,\infty)$ *ist* L*-hart.*

Beweis. Das Problem UNGWEG, welches häufig auch UPATH oder USTCON genannt wird, ist bezüglich $\leq_{\text{fo}}$-Reduktion hart für die Klasse L [25]. Mit Lemma 3.9 gilt UNGWEG $\leq_{\text{fo}} \overline{\text{PPH}(3,\infty)}$. Anfragen in Prädikatenlogik erster Stufe sind unter Komposition abgeschlossen und somit folgt, dass $\overline{\text{PPH}(3,\infty)}$ bezüglich $\leq_{\text{fo}}$-Reduktion L-hart ist. Dies gilt ebenfalls für $\text{PPH}(3,\infty)$, da L unter Komplement abgeschlossen ist. □

Das Problem $\text{PPH}(3,\infty)$ ist eine Teilmenge von PPH. Mit der Reduktion in Lemma 3.9 gilt somit auch, dass PPH L-hart ist.

Lemma 3.11. PPH *lässt sich mit einer Anfrage in Prädikatenlogik erster Stufe auf das Lösen linearer Gleichungssysteme über* $\mathbb{Z}/2\mathbb{Z}$ *reduzieren.*

Beweis. Die Reduktion wird, wie im Beweis zu Lemma 3.9, in drei Schritten beschrieben.

Konstruktion: Es wird im Folgenden beschrieben, wie eine Genotypmatrix G auf ein lineares Gleichungssystem abgebildet wird: Falls ein Spaltenpaar (s,s') existiert, für das $|I(s,s')| = 4$ in G gilt, dann lässt G keine perfekte Phylogenie zu, da jede erklärende Haplotypmatrix für G die Untermatrix V aus Lemma 3.4 in den Spalten s und s' enthält. In diesem Fall wird das Gleichungssystem $0 = 1$ erstellt, welches keine Lösung besitzt. Falls andererseits für jedes Spaltenpaar (s,s') die Ungleichung $|I(s,s')| \leq 3$ erfüllt ist, wird ein lineares Gleichungssystem aus G abgeleitet, indem für jeden Genotyp, der an mindestens drei Stellen heterozytisch ist, wie folgt Gleichungen erstellt werden: Sei g ein Genotyp mit mindestens drei heterozytischen Einträgen und s_r die kleinste Spalte für die $g[s_r] = 2$ gilt. Die Spalte s_r nennen wir die *Referenzspalte von* g. Für jedes Spaltenpaar (s_1,s_2) mit $s_r < s_1 < s_2$ und $g[s_1] = g[s_2] = 2$ wird die Gleichung $A(s_1,s_2) = A(s_r,s_1) + A(s_r,s_2)$ über $\mathbb{Z}/2\mathbb{Z}$ erstellt. In dem so konstruierten Gleichungssystem werden einige Variablen folgendermaßen konstant gesetzt: Sei (s_1,s_2) ein beliebiges Spaltenpaar. Falls $\{(0,0),(1,1)\} \subseteq I(s_1,s_2)$ gilt, dann wird $A(s_1,s_2) = 0$ gesetzt und falls $\{(0,1),(1,0)\} \subseteq I(s_1,s_2)$, dann $A(s_1,s_2) = 1$. Die Variable $A(s_1,s_2)$ gibt auf diese Weise an, ob (s_1,s_2) gleich $(A(s_1,s_2) = 0)$ oder ungleich $(A(s_1,s_2) = 1)$ aufgelöst wird. Abbildung 5 zeigt die Reduktion an einem Beispiel.

Korrektheit: Folgende Behauptung wird nun gezeigt: G lässt genau dann eine perfekte Phylogenie zu, wenn das lineare Gleichungssystem eine Lösung besitzt.

Nur-wenn-Teil: Im Folgenden nehmen wir an, dass G eine PP-Lösung (H,B_H) besitzt und leiten daraus eine Lösung des linearen Gleichungssystems ab. Hierfür sei g ein Genotyp, der an mindestens drei Stellen heterozytisch ist und s_r die Referenzspalte von g. Weiter seien s_1 und s_2 zwei Spalten, so dass $s_r < s_1 < s_2$ und $g[s_1] = g[s_2] = 2$ gilt sowie h und h' die erklärenden Haplotypen zu g, eingeschränkt auf die Spalten s_r, s_1 und s_2. Nun werden die Variablen der zugehörigen Gleichung $A(s_1,s_2) = A(s_r,s_1) + A(s_r,s_2)$ folgendermaßen belegt:
1. Falls $h = 000$ und $h' = 111$, dann $A(s_r,s_1) = A(s_1,s_2) = A(s_r,s_2) = 0$.
2. Falls $h = 001$ und $h' = 110$, dann $A(s_r,s_1) = 0$ und $A(s_1,s_2) = A(s_r,s_2) = 1$.
3. Falls $h = 010$ und $h' = 101$, dann $A(s_r,s_1) = A(s,s_2) = 1$ und $A(s_r,s_2) = 0$.

Abbildung 5: Die Abbildung zeigt beispielhaft die Reduktion von PPH auf das Lösen linearer Gleichungssysteme über $\mathbb{Z}/2\mathbb{Z}$. Die induzierte Menge jedes Spaltenpaares in G enthält weniger als vier Elemente und somit werden für jeden Genotyp, der mehr als drei heterozytische Einträge besitzt, Gleichungen erstellt. Die Genotypen g_3, g_4 und g_6 enthalten jeweils drei oder mehr heterozytische Stellen. Für diese Genotypen werden Gleichungen aufgestellt. In dem Beispiel gehört zu g_3 die Gleichung L_1 mit Referenzspalte s_1, zu g_4 gehört die Gleichung L_2 mit Referenzspalte s_2 und zu g_6 gehören die Gleichungen L_3, L_4 und L_5 mit Referenzspalte s_1. Da $\{(0,1),(1,0)\} \subseteq I(s_1,s_2)$, wird $A(s_1,s_2) = 1$ gesetzt. Außerdem werden $A(s_2,s_3) = 0$, $A(s_1,s_3) = 1$ und $A(s_1,s_4) = 1$ gesetzt. An dem Beispiel kann man sehen, dass mehrere Gleichungen die gleiche Variable umfassen können (zum Beispiel enthalten die Gleichungen L_1 und L_3 die Variable $A(s_1,s_2)$) und dass Gleichungen identisch sein können (zum Beispiel die Gleichungen L_1 und L_2). Die abgebildet Genotypmatrix G lässt eine perfekte Phylogenie zu und das Gleichungssystem ist lösbar.

G :

	s_1	s_2	s_3	s_4	s_5
g_1	1	0	0	0	0
g_2	0	1	1	0	0
g_3	2	2	2	0	0
g_4	0	2	2	2	0
g_5	0	1	2	0	2
g_6	2	2	2	2	0

LGLS :

$$
\begin{array}{ll}
L_1 & A(s_2,s_3) = A(s_1,s_2) + A(s_1,s_3) \\
L_2 & A(s_3,s_4) = A(s_2,s_3) + A(s_2,s_4) \\
L_3 & A(s_2,s_3) = A(s_1,s_2) + A(s_1,s_3) \\
L_4 & A(s_3,s_4) = A(s_1,s_3) + A(s_1,s_4) \\
L_5 & A(s_2,s_4) = A(s_1,s_2) + A(s_1,s_4)
\end{array}
$$

4. Falls $h = 011$ und $h' = 100$, dann $A(s_r, s_1) = A(s_r, s_2) = 1$ und $A(s_1, s_2) = 0$. Weitere vier Fälle ergeben sich, wenn wir h und h' in den einzelnen Fällen vertauschen.

Da (H, B_H) eine PP-Lösung für G ist, wird jedes Spaltenpaar entweder gleich oder ungleich aufgelöst und somit wird bei dem obigen Vorgehen keine Variable mit zwei verschiedenen Werten belegt. Aus dem gleichen Grund ist die Belegung der Variablen konsistent mit den Variablen, die in der Konstruktion konstant gesetzt werden. Die Variablen werden in den Fällen 1 bis 4 immer so belegt, dass die Gleichung erfüllt ist und jede Variable im Gleichungssystem wird auf diese Weise belegt. Zusammen gilt, dass durch die Belegung der Variablen eine Lösung des linearen Gleichungssystems gegeben ist.

Wenn-Teil: Für diese Beweisrichtung nehmen wir an, dass das lineare Gleichungssystem eine Lösung besitzt. Die Lösung entspricht einer Belegung der Variablen mit 0 und 1. Aus dieser Belegung leiten wir eine PP-Lösung (H, B_H) für G ab. Da das lineare Gleichungssystem eine Lösung besitzt, gilt $|I(s_1, s_2)| \leq 3$ für jedes Spaltenpaar (s_1, s_2) in G. Zuerst wird die Menge der Variablen erweitert, indem wir für jedes Spaltenpaar (s_1, s_2), dessen zugehörige Variable nicht im Gleichungssystem vorkommt, die Variable $A(s_1, s_2)$ einführen. Die neu eingeführten Variablen werden wie folgt mit Werten belegt: Falls $\{(0,1),(1,0)\} \subseteq I(s_1, s_2)$ in G gilt, dann wird $A(s_1, s_2) = 1$ gesetzt und ansonsten wird $A(s_1, s_2) = 0$ gesetzt. Für jedes Spaltenpaar existiert nun eine zugehörige Variable, die belegt ist.

Aus der Variablenbelegung werden für jeden Genotyp zwei erklärende Haplotypen abgeleitet. Hierzu sei g ein beliebiger Genotyp aus G. Die erklärenden Haplotypen h und h' für g werden wie folgt konstruiert: Für jede homozytische Stelle s in g wird $h[s] = h'[s] = g[s]$ gesetzt. Falls g nur homozytische Stellen besitzt, dann erklären die Haplotypen h und h' den Genotyp g. Falls g genau eine heterozytische Stelle s besitzt, wird $h[s] = 0$ und $h'[s] = 1$ gesetzt und die Haplotypen h und h' erklären g. Falls g genau zwei heterozytische Stellen s_1 und s_2 besitzt, werden diese entsprechend der Belegung von $A(s_1, s_2)$ gesetzt: Falls $A(s_1, s_2) = 0$, dann werden $h[s_1] = h[s_2] = 0$ und $h'[s_1] = h'[s_2] = 1$ gesetzt und falls $A(s_1, s_2) = 1$, dann werden $h[s_1] = h'[s_2] = 0$ und $h'[s_1] = 1 = h[s_2] = 1$ gesetzt. Es lässt sich nun erkennen, dass das Spaltenpaar (s_1, s_2) entsprechend der Belegung von $A(s_1, s_2)$ aufgelöst wird.

Falls g mindestens drei heterozytische Stellen besitzt, sei s_r die Referenzspalte von g und wir setzen $h[s_r] = 0$ und $h'[s_r] = 1$. Für jede andere heterozytische Stelle s von g werden h und h' entsprechend der Belegung von $A(s_r, s)$ gesetzt: Falls $A(s_r, s) = 0$, dann werden $h[s] = 0$ und $h'[s] = 1$ gesetzt und falls $A(s_r, s) = 1$, dann werden $h[s] = 1$ und $h'[s] = 0$ gesetzt. Es wird nun gezeigt, dass jedes Spaltenpaar (s_1, s_2), in dem g heterozytisch ist, entsprechend der Belegung von $A(s_1, s_2)$ aufgelöst wird. Für ein Spaltenpaar, welches die Referenzspalte enthält, gilt die Behauptung nach Konstruktion von h und h'. Für jedes andere Spaltenpaar wird die Behauptung nun mit einer Fallunterscheidung bezüglich der vier Möglichkeiten $A(s_r, s_1)$ und $A(s_r, s_1)$ zu belegen gezeigt. Falls $A(s_r, s_1) = A(s_r, s_2) = 0$, dann gilt $A(s_1, s_2) = A(s_r, s_1) + A(s_r, s_2) = 0$ und (s_1, s_2) wird in g gleich aufgelöst. Der Fall für $A(s_r, s_1) = A(s_r, s_2) = 1$ ist analog. Falls $A(s_r, s_1) = 0$ und $A(s_r, s_2) = 1$,

dann gilt $A(s_1,s_2) = A(s_r,s_1) + A(s_r,s_2) = 1$ und (s_1,s_2) wird in g ungleich aufgelöst. Der Fall für $A(s_r,s_1) = 1$ und $A(s_r,s_2) = 0$ ist analog. Insgesamt folgt, dass jedes Spaltenpaar entsprechend der Variablenbelegung aufgelöst wird.

Für jeden Genotyp liegen nun erklärende Haplotypen vor. Eine Haplotypmatrix H, die aus diesen Haplotypen zusammen gestellt wird, erklärt G und wir zeigen nun, dass H auch eine perfekte Phylogenie zulässt. Dazu sei (s_1,s_2) ein beliebiges Spaltenpaar, $I(s_1,s_2)$ bezeichne die induzierte Menge in G und $J(s_1,s_2)$ bezeichne die induzierte Menge in H. Falls in G kein Genotyp existiert, der in s_1 und s_2 heterozytisch ist, dann gilt $I(s_1,s_2) = J(s_1,s_2)$ und mit $|I(s_1,s_2)| \leq 3$ folgt $|J(s_1,s_2)| \leq 3$. Falls andererseits ein Genotyp existiert, der in s_1 und s_2 heterozytisch ist, unterscheiden wir zwischen Spaltenpaaren, deren Variablen bei der Konstruktion konstant gesetzt werden, und Spaltenpaaren, deren Variablen bei der Konstruktion nicht konstant gesetzt werden. Falls die Variable $A(s_1,s_2)$ in der Konstruktion konstant gesetzt wird, dann gilt $I(s_1,s_2) = J(s_1,s_2)$ und mit $|I(s_1,s_2)| \leq 3$ folgt $|J(s_1,s_2)| \leq 3$. Falls die Variable $A(s_1,s_2)$ nach der Konstruktion noch keinen Wert besitzt, also ihr Wert erst durch das Lösen des linearen Gleichungssystems festgelegt wird, dann gilt $\{(0,0),(1,1)\} \nsubseteq I(s_1,s_2)$ und $\{(0,1),(1,0)\} \nsubseteq I(s_1,s_2)$. Die Haplotypmatrix H löst die Genotypen in einem Spaltenpaar auf gleiche Weise auf. Dies entspricht entweder einer Erweiterung der induzierten Menge um $\{(0,0),(1,1)\}$ oder um $\{(0,1),(1,0)\}$. Somit gilt $|J(s,s')| \leq 3$. Für ein Spaltenpaar (s_1,s_2), dessen Variable im Beweis neu eingeführt wird, lässt sich auf ähnliche Weise argumentieren, dass $|J(s_1,s_2)| \leq 3$ gilt. Insgesamt folgt, dass H in keinem Spaltenpaar die Untermatrix V aus Lemma 3.4 enthält und somit eine perfekte Phylogenie zulässt.

Komplexität: Nun wird die oben beschriebene Abbildung von einer Genotypmatrix auf ein lineare Gleichungssystem als Anfrage in Prädikatenlogik erster Stufe formuliert. Als Erstes wird beschrieben, wie sich ein lineares Gleichungssystem als Struktur kodiert lässt. Die Gleichungen, die durch die Abbildung erstellt werden, lassen sich so äquivalent umformen, dass eine Seite der Gleichung immer konstant 0 ist. Zum Beispiel können wir die Gleichung $A(2,3) = A(1,2) + A(1,3)$ zur Gleichung $0 = A(1,2) + A(1,3) + A(2,3)$ umformt. Auf diese Weise entsteht ein homogenes Gleichungssystem. Im Folgenden ist $\tau_L = (L^2, K_0^1, K_1^1, K_?^1)$ die Signatur für lineare homogene Gleichungssysteme über $\mathbb{Z}/2\mathbb{Z}$. Eine τ_L-Struktur $(I, L, K_0, K_1, K_?)$ besteht aus der Menge der verwendeten Indizes I, einer zweistelligen Relation $L \subseteq I \times I$ und drei einstelligen Relationen $K_0, K_1, K_? \subseteq I$. Die Relationen haben folgende Bedeutung: Es gilt $L(g,v) \equiv \mathit{wahr}$, falls die Gleichung mit Index g die Variable mit Index v enthält und $K_0(v) \equiv \mathit{wahr}$, falls die Variable v den festen Wert 0 besitzt. Weiter ist ein Index v in K_1, falls die zugehörige Variable konstant 1 ist und ein Index v ist in $K_?$, falls der Wert der zugehörigen Variable nicht bekannt ist. Zum Beispiel kann das Gleichungsystem, welches aus den Gleichungen $0 = x + y + 1$, $0 = x + 1$ und $0 = y + 0$ besteht, durch eine τ_L-Struktur $(I, L, K_0, K_1, K_?)$ mit $I = \{1,2,3\}$, $K_0 = \{3\}$, $K_1 = \{4\}$, $K_? = \{1,2\}$ und $L = \{(1,1),(1,2),(1,4),(2,1),(2,4),(3,2),(3,3)\}$ kodiert werden.

Eine Gleichung $0 = A(s_1,s_2) + A(s_r,s_1) + A(s_r,s_2)$ wird durch die beschriebene

Konstruktion genau dann erstellt, wenn ein Genotyp existiert, dessen Referenzspalte s_r ist und der in zwei weiteren Spalten s_1 und s_2, für die insgesamt $s_r < s_1 < s_2$ gilt, heterozytisch ist. Folgende Formel beschreibt, dass eine Gleichung mit Referenzspalte s_r und zwei weiteren Spalten s_1 und s_2 im Gleichungssystem vorkommt:

$$\begin{aligned}\phi_{gl}(s_r,s_1,s_2) \equiv\ & s_r < s_1 \wedge s_1 < s_2 \wedge \\ & (\exists z.Z(z))[G_2(z,s_r) \wedge G_2(z,s_1) \wedge G_2(z,s_2) \wedge \\ & (\forall s.S(s))[s < s_r \rightarrow \neg G_2(z,s)]].\end{aligned}$$

Für ein Spaltenpaar (s,s') beschreiben folgende Formeln die Elemente, die in der induzierten Menge $I(s,s')$ vorkommen:

$$\begin{aligned}\phi_{(0,0)}(s,s') \equiv\ & (\exists z.Z(z))[(G_0(z,s) \wedge G_0(z,s')) \vee \\ & (G_0(z,s) \wedge G_2(z,s')) \vee \\ & (G_2(z,s) \wedge G_0(z,s'))], \\ \phi_{(0,1)}(s,s') \equiv\ & (\exists z.Z(z))[(G_0(z,s) \wedge G_1(z,s')) \vee \\ & (G_0(z,s) \wedge G_2(z,s')) \vee \\ & (G_2(z,s) \wedge G_1(z,s'))], \\ \phi_{(1,0)}(s,s') \equiv\ & (\exists z.Z(z))[(G_1(z,s) \wedge G_0(z,s')) \vee \\ & (G_1(z,s) \wedge G_2(z,s')) \vee \\ & (G_2(z,s) \wedge G_0(z,s'))], \\ \phi_{(1,1)}(s,s') \equiv\ & (\exists z.Z(z))[(G_1(z,s) \wedge G_1(z,s')) \vee \\ & (G_1(z,s) \wedge G_2(z,s')) \vee \\ & (G_2(z,s) \wedge G_1(z,s'))].\end{aligned}$$

Zum Beispiel ist die Formel $\phi_{(0,0)}(s,s')$ für ein Spaltenpaar (s,s') genau dann wahr, wenn $(0,0) \in I(s,s')$ gilt. Hierauf aufbauend beschreibt folgende Formel ein Spaltenpaar, für das $|I(s,s')| \leq 3$ gilt:

$$\begin{aligned}\phi_{\text{IND}\leq 3}(s,s') \equiv\ & \neg(\phi_{(0,0)}(s,s') \wedge \phi_{(0,1)}(s,s') \wedge \\ & \phi_{(1,0)}(s,s') \wedge \phi_{(1,1)}(s,s')).\end{aligned}$$

Die Anfrage A_{G-L} wird nun durch folgende Formeln definiert:

$$
\begin{aligned}
\phi_I(i) &\equiv i \leq \max^3,\\
\phi_L(g,v) &\equiv (((\forall s.S(s), s'.S(\mathit{anderes}'))[\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow\\
&\quad (\exists s_r.S(s_r), s.S(s), s'.S(s'))[\phi_{gl}(s_r,s,s') \wedge\\
&\quad g = (s_r-1)\cdot\max^2 + (s-1)\cdot\max + s' - 1 \wedge\\
&\quad (v = (s-1)\cdot\max + (s'-1) \vee\\
&\quad v = (s_r-1)\cdot\max + (s-1) \vee\\
&\quad v = (s_r-1)\cdot\max + (s'-1))] \wedge\\
&\quad (((\exists s.S(s), s'.S(s'))[\neg\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow (g = 1 \wedge (v = 1 \vee v = 2))),\\
\phi_{K_0}(v) &\equiv (((\forall s.S(s), s'.S(s'))[\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow\\
&\quad (\exists s.S(s), s'.S(s'))[s < s' \wedge\\
&\quad v = (s-1)\cdot\max + (s'-1) \wedge \phi_{(0,0)}(s,s') \wedge \phi_{(1,1)}(s,s')]) \wedge\\
&\quad (((\exists s.S(s), s'.S(s'))[\neg\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow v = 0),\\
\phi_{K_1}(v) &\equiv (((\forall s.S(s), s'.S(s'))[\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow\\
&\quad (\exists s.S(s), s'.S(s'))[s < s' \wedge\\
&\quad v = (s-1)\cdot\max + (s'-1) \wedge \phi_{(0,1)}(s,s') \wedge \phi_{(1,0)}(s,s')]) \wedge\\
&\quad (((\exists s.S(s), s'.S(s'))[\neg\phi_{\mathrm{IND}\leq 3}(s,s')]) \rightarrow v = 1),\\
\phi_{K_?}(v) &\equiv \neg\phi_{K_0}(v) \wedge \neg\phi_{K_1}(v).
\end{aligned}
$$

Die Formel $\phi_I(i)$ beschreibt die Element der Indexmenge des Gleichungssystems, die Formel $\phi_L(g,v)$ beschreibt, ob eine Variable v in einer Gleichung g vorkommt und die Formeln $\phi_{K_0}(v)$, $\phi_{K_1}(v)$ und $\phi_{K_?}(v)$ beschreiben, ob eine Variable mit einem Wert konstant gesetzt ist oder der Wert noch offen ist. Auf die Gleichung $0 = A(s_1,s_2) + A(s_r,s_1) + A(s_r,s_2)$ wird durch $(s_r-1)\cdot\max^2 + (s_1-1)\cdot\max + (s_2-1)$ referenziert und eine Variable $A(s,s')$ wird durch $(s-1)\cdot\max + (s'-1)$ referenziert. Die Konstante max ist gleich dem maximalen Element in einer Struktur. Für τ_G-Strukturen ist dies der Wert vom maximalen Element in I. Wie im Beweis zu Lemma 3.9 nehmen wird an, dass die Konstante max in jeder Struktur vorkommt. Nun entspricht die Anfrage $A_{G-L} = \lambda_{gv}(\phi_I, \phi_L, \phi_{K_0}, \phi_{K_1}, \phi_{K_?})$ der Abbildung von G auf das lineare Gleichungssystem. Zu beachten ist, dass das lineare Gleichungssystem, welches durch die Anfrage erstellt wird, Lücken in der Indizierung der Gleichungen und der Indizierung der Variablen enthalten kann. □

Satz 3.12. PPH $\in$ $\mathrm{Mod}_2\mathrm{L}$.

Beweis. Die Klasse $\mathrm{Mod}_2\mathrm{L}$ ist unter NC^1-Reduktion abgeschlossen [6]. Hieraus folgt, dass $\mathrm{Mod}_2\mathrm{L}$ auch unter Reduktionen abgeschlossen ist, die sich als Anfrage in Prädikatenlogik erster Stufe formulieren lassen. Lemma 3.11 besagt, dass PPH mit einer Anfrage in Prädikatenlogik erster Stufe auf das Lösen linearer Gleichungssysteme über $\mathbb{Z}/2\mathbb{Z}$ reduzierbar ist. Da das Problem, zu entscheiden, ob ein

lineares Gleichungssystem über $\mathbb{Z}/2\mathbb{Z}$ lösbar ist, in Mod_2L liegt ist folglich auch PPH in Mod_2L enthalten. □

Das Problem PP ist eine Teilmenge von PPH, da man Haplotypmatrizen als Genotymatrizen ansehen kann, die keinen heterozytischen Eintrag enthalten. Das Problem PP liegt in FO (Satz 3.5) und ist somit besonders einfach. Die Haplotypisierung mittels perfekten Phylogenien wird schwerer, wenn man mehrere heterozytische Einträge zulässt. So ist schon das Problem PPH$(3,\infty)$ L-hart. Um einen Zusammenhang zwischen der Anzahl heterozytischer Einträge in der Eingabe und der Komplexität der entsprechenden $\{k,l\}$-beschränkten Variante von PPH herzustellen, wird im Folgenden gezeigt, dass PPH$(2,\infty)$ und PPH$(\infty,1)$ in FO liegen.

Lemma 3.13. *Sei G eine Genotypmatrix, in der jeder Genotyp an maximal zwei Stellen heterozytisch ist. Dann lässt G genau dann eine perfekte Phylogenie zu, wenn für jedes Spaltenpaar (s,s') die Ungleichung $|I(s,s')| \leq 3$ erfüllt ist.*

Beweis. Die Beweisrichtung, die als erstes gezeigt wird, gilt für beliebige Genotypmatrizen. Im zweiten Beweisteil schränken wir die Menge der Genotypmatrizen auf solche ein, die maximal zwei heterozytische Einträge in einem Genotyp enthalten.

Nur-wenn-Teil: Sei G eine beliebige Genotypmatrix, (H,B_H) eine PP-Lösung für G und (s,s') ein beliebiges Spaltenpaar. Weiter sei $I(s,s')$ die induzierte Menge von (s,s') in G und $J(s,s')$ die induzierte Menge von (s,s') in H. Mit Lemma 3.4 enthält H nicht die Untermatrix V und somit gilt $|J(s,s')| \leq 3$. Mit $I(s,s') \subseteq J(s,s')$ gilt folglich $|I(s,s')| \leq |J(s,s')| \leq 3$, was zu zeigen war.

Wenn-Teil: Für diese Beweisrichtung betrachten wir eine Genotypmatrix G, die maximal zwei heterozytische Einträge pro Genotyp enthält und nehmen an, dass $|I(s,s')| \leq 3$ für jedes Spaltenpaar (s,s') gilt. Nun wird folgendermaßen eine PP-Lösung für G konstruiert: Sei g ein beliebiger Genotyp in G. An jeder homozytischen Stelle von g werden die erklärenden Haplotypen h und h' auf den entsprechenden Wert von g gesetzt. Falls g nur homozytische Einträge besitzt, erklären h und h' den Genotyp g. Falls g genau eine homozytische Stelle besitzt, wird diese Stelle in h und h' verschieden gesetzt. Auf diese Weise entstehen erklärende Haplotypen h und h' für g. Falls g in genau zwei Spalten s und s' heterozytisch ist, werden diese Stellen entsprechend der induzierten Menge $I(s,s')$ gesetzt. Dies bedeutet, falls $\{(0,0),(1,1)\} \subseteq I(s,s')$, wird g in (s,s') gleich aufgelöst und falls $\{(0,1),(1,0)\} \subseteq I(s,s')$, wird g in (s,s') ungleich aufgelöst. Falls keiner der beiden Fälle zutrifft, wird g gleich aufgelöst. Die Haplotypmatrix H, die auf diese Weise konstruiert wird, lässt aus folgendem Grund eine perfekte Phylogenie zu: Ein beliebiges Spaltenpaar (s,s') ist entweder in einem Genotyp heterozytisch, dann werden alle Genotypen in dem Spaltenpaar nach dem selben Schema aufgelöst und die induzierte Menge wird nicht bis auf 4 Elemente erweitert oder (s,s') ist in keinem Genotyp heterozytisch und die induzierten Mengen in H und G sind gleich. Insgesamt folgt, dass kein Spaltenpaar in H die verbotene Untermatrix V enthält und somit G eine perfekte Phylogenie zulässt. □

Satz 3.14. $\text{PPH}(2,\infty) \in \text{FO}$.

Beweis. Folgende Formel in Prädikatenlogik erster Stufe entspricht der rechten Seite der Aussage von Lemma 3.13 und beschreibt somit $\text{PPH}(2,\infty)$:

$$\phi_{\text{PPH}(2,\infty)} \equiv (\forall s.S(s), s'.S(s'))[\phi_{\text{IND}\leq 3}(s,s')].$$

Die Formel $\phi_{\text{IND}\leq 3}(s,s')$ wird im Beweis zu Lemma 3.11 eingeführt und beschreibt ein Spaltenpaar, dessen induzierte Menge weniger als vier Elemente enthält. □

Sei G eine $n \times m$ Genotypmatrix und g ein Genotyp in G. Die Genotypmatrix G_g ist eine Untermatrix von G und besteht genau aus den Spalten, in denen g den Eintrag 2 enthält. Falls beispielsweise ein Genotyp g in 9 Spalten heterozytisch ist, dann ist G_g eine $n \times 9$ Genotypmatrix.

Lemma 3.15. *Sei G eine Genotypmatrix, in der jede Spalte maximal einen heterozytischen Eintrag besitzt. Dann lässt G genau dann eine perfekte Phylogenie zu, wenn für jedes Spaltenpaar (s,s') die Ungleichung $|I(s,s')| \leq 3$ erfüllt ist und für jeden Genotyp g die Genotypmatrix G_g eine perfekte Pfadphylogenie zulässt.*

Beweis. Ähnlich dem Beweis zu Lemma 3.13 wird die erste Beweisrichtung für beliebige Genotypmatrizen und die zweite Beweisrichtung für Genotypmatrizen, die in jeder Spalte maximal einen heterozytischen Eintrag enthalten, gezeigt.

Nur-wenn-Teil: Sei G eine beliebige Genotypmatrix und (H, B_H) eine PP-Lösung für G. Weiter seien (s,s') ein beliebiges Spaltenpaar, $I(s,s')$ die induzierte Menge von (s,s') in G und $J(s,s')$ die induzierte Menge von (s,s') in H. Es gilt $I(s,s') \subseteq J(s,s')$ und mit $|J(s,s')| \leq 3$ folgt der erste Aussagenteil: $|I(s,s')| \leq 3$.

Nun zeigen wir, dass für jeden Genotyp g die Untermatrix G_g eine perfekte Pfadphylogenie zulässt. Sei dazu g ein beliebiger Genotyp in G sowie h und h' die erklärenden Haplotypen zu g in H. Nach Definition 2.1 liegen auf dem Pfad von h nach h' in B_H genau die Spalten, in denen g heterozytisch ist. Diesen Pfad nennen wir den *2er-Pfad von g*. Nun erstellen wir eine PP-Lösung (H_g, B_{H_g}) für G_g, indem wir (H, B_H) wie folgt auf den 2er-Pfad von g reduzieren: Die Haplotypmatrix H_g entsteht aus H, indem jede Spalte, die nicht auf dem 2er-Pfad von g liegt, gelöscht wird. Die perfekte Phylogenie B_{H_g} entsteht aus B_H, indem für jede Spalte s, die nicht auf dem 2er-Pfad von g liegt, die Kante s aus B_H gelöscht wird, die Endknoten von s verschmolzen werden und jede Knotenmarkierung um die Stelle s reduziert wird. Nun gilt, dass G_g durch H_g erklärt wird, dass B_{H_g} jeden Haplotypen aus H_g enthält und dass Punkt 4 von Definition 2.1 weiterhin für jedes Paar von Haplotypen und jede Spalte erfüllt ist. Der Baum B_{H_g} hat nach Konstruktion die Form eines Pfades und damit gilt, dass (H_g, B_{H_g}) eine PPP-Lösung für G darstellt. Insgesamt folgt, dass die Aussage auf der rechten Seite für jede Genotypmatrix, die eine perfekte Phylogenie zulässt, gilt.

Wenn-Teil: Sei G eine $n \times m$ Genotypmatrix, in der jede Spalte maximal einen heterozytischen Eintrag besitzt und gelte die Aussage auf der rechten Seite. Seien nun $g_1, \ldots, g_n$ die Genotypen in G und $H_{g_1}, \ldots, H_{g_n}$ die PPP-Lösungen zu den

Untermatrizen $G_{g_1}, \ldots, G_{g_n}$. Eine $2n \times m$ Haplotypmatrix H, die G erklärt und eine perfekte Phylogenie zulässt, wird nun folgendermaßen spaltenweise zusammengesetzt: Sei s eine beliebige Spalte aus G. Falls ein Genotyp g existiert, für den $g[s] = 2$ gilt, wird die Spalte s in G gleich der entsprechenden Spalte in H_g gesetzt. Falls andererseits die Spalte s keinen heterozytischen Eintrag besitzt, wird die Spalte s in H gleich der Spalte s in G gesetzt. Hierbei wird jeder Eintrag der Spalte s aus G verdoppelt, da H aus $2n$ Zeilen und G aus n Zeilen besteht. Die so konstruierte Haplotypmatrix erklärt G. Um den Beweis zu vervollständigen wird nun gezeigt, dass H auch eine perfekte Phylogenie zulässt. Dazu sei (s, s') ein beliebiges Spaltenpaar in G. Falls (s, s') für einen Genotyp g in H_g liegt, dann enthält (s, s') in H_g und damit auch in H nicht die verbotene Untermatrix V. Falls (s, s') für keinen Genotyp g in H_g liegt, dann enthält G keinen Genotyp, der in s und s' heterozytisch ist. In diesem Fall gilt $I(s, s') = J(s, s')$, wobei $I(s, s')$ die induzierte Menge von (s, s') in G bezeichnet und $J(s, s')$ die induzierte Menge von (s, s') in H bezeichnet. Nach Voraussetzung gilt $|I(s, s')| \leq 3$ und somit folgt $|J(s, s')| = |I(s, s')| \leq 3$. Insgesamt folgt, dass kein Spaltenpaar in H die Untermatrix V enthält und G somit eine perfekte Phylogenie zulässt. □

Satz 3.16. $\mathrm{PPH}(\infty, 1) \in \mathrm{FO}$.

Beweis. Die Formeln, die im Weiteren vorgestellt werden, machen unter anderem Aussagen über perfekte Pfadphylogenie und verwenden Begriffe und Konzepte, die erst in Abschnitt 3.3.4 eingeführt werden. Für das Verständnis dieses Beweises ist es daher sinnvoll erst Abschnitt 3.3.4 zu lesen und dann an diese Stelle zurück zu kehren.

Zum Beweis wird nun eine Formel $\phi_{\mathrm{PPH}(\infty,1)}$ in Prädikatenlogik erster Stufe angegeben, die $\mathrm{PPH}(\infty, 1)$ beschreibt. Die Formel entspricht der Aussage von Lemma 3.15 und besteht aus zwei Teilen. Im ersten Teil wird beschrieben, dass die induzierte Menge von jedem Spaltenpaar in einer Genotypmatrix nicht mehr als drei Elemente enthält. Im zweiten Teil wird beschrieben, dass für jeden Genotyp g die Genotypmatrix G_g eine perfekte Pfadphylogenie zulässt. Falls eine Genotypmatrix G aus mehr als einer Zeile besteht, dann enthält jede nichtleere Genotypmatrix G_g eine Zeile, die nur homozytische Einträge besitzt. Dies ist der Fall, da nach Voraussetzung jede Spalte höchstens einen heterozytischen Eintrag enthält und der Genotyp g in G_g gleich dem Genotyp $2 \ldots 2$ ist. Um zu entscheiden, ob G_g eine perfekte Pfadphylogenie zulässt, wird G_g durch eine Anfrage, die ähnlich zu der aus Lemma 3.19 ist, auf eine Genotypmatrix G'_g abgebildet. Die Anfrage entspricht dabei einer Reduktion auf die gerichtete Problemvariante, die mit einer Formel, ähnlich zu der in Satz 3.18, beschrieben werden kann.

Es wird nun schrittweise die Formel $\phi_{\mathrm{PPPHauf2er}}(g)$ hergeleitet, die eine Genotypmatrix beschreibt, in der die Untermatrix G_g eine perfekte Pfadphylogenie zulässt.

Durch folgende Formel wird eine Zeile, die in G_g nur homozytische Einträge

besitzt, eindeutig bestimmt:

$$\phi_{\text{kleinste}}(g,z) \equiv z \neq g \wedge (\forall z'.Z(z'))[z' \neq g \rightarrow z' \geq z].$$

Die nun folgende Formel ist ähnlich zur Formel $\phi_{\text{tausch}}(s)$ aus Lemma 3.19 und beschreibt eine Spalte, in der bei der Reduktion auf den gerichteten Fall die Rollen von 0 und 1 vertauscht werden:

$$\phi_{\text{tausch}}(g,s) \equiv (\exists z.Z(z))[\phi_{\text{kleinste}}(g,z) \wedge G_1(z,s)].$$

Nun folgen zwei Formeln, die die Reduktion auf den gerichteten Fall formulieren:

$$\phi_{G_0}(g,z,s) \equiv (G_1(z,s) \wedge \phi_{\text{tausch}}(g,s)) \vee (G_0(z,s) \wedge \neg\phi_{\text{tausch}}(g,s)),$$
$$\phi_{G_1}(g,z,s) \equiv (G_0(z,s) \wedge \phi_{\text{tausch}}(g,s)) \vee (G_1(z,s) \wedge \neg\phi_{\text{tausch}}(g,s)).$$

Die Formel $\phi_{G_0}(g,z,s)$ wird im Weiteren anstatt der Relation $G_0(z,s)$ verwendet. Sie gibt an, ob in Zeile z an Stelle s in G'_g der Eintrag 0 steht. Die Bedeutung der Formel $\phi_{G_1}(g,z,s)$ ist analog. Durch die Reduktion auf den gerichteten Fall werden Einträge mit dem Wert 2 nicht verändert. Die Relation $G_2(z,s)$ wird daher nicht durch eine Formel ersetzt, sondern weiterhin verwendet.

Durch die vorangegangenen Formeln wird eine Abbildung von G_g nach G'_g beschrieben, die einer Reduktion von ung-PPPH auf ger-PPPH entspricht. Nun wird die Formel hergeleitet, die entscheidet, ob G'_g eine gerichtete perfekte Pfadphylogenie zulässt und G_g eine perfekte Pfadphylogenie zulässt.

Folgende Formel ist ähnlich zu Formel $\phi_{\succeq}(s,t)$ aus Satz 3.18 und beschreibt die partielle Ordnung $\succeq$ auf Spalten:

$$\phi_{\succeq}(g,s,t) \equiv (\forall z.Z(z))[\phi_{G_0}(g,z,t) \vee \phi_{G_1}(g,z,s) \vee (G_2(z,s) \wedge G_2(z,t))].$$

Sei S'_g die Menge der Spalten in der Genotypmatrix G'_g. Folgende Formel ist ähnlich zu Formel $\phi_{\text{ZweiKetten}}$ aus Satz 3.18 und beschreibt, dass sich S'_g mit zwei Ketten überdecken lässt:

$$\begin{aligned}\phi_{\text{ZweiKetten}}(g) \equiv (\forall s_1.S(s_1) \wedge G_2(g,s_1), s_2.S(s_2) \wedge G_2(g,s_2), s_3.S(s_3) \wedge G_2(g,s_3))[\\ \phi_{\succeq}(s_1,s_2) \vee \phi_{\succeq}(s_2,s_1) \vee\\ \phi_{\succeq}(s_2,s_3) \vee \phi_{\succeq}(s_3,s_2) \vee\\ \phi_{\succeq}(s_3,s_1) \vee \phi_{\succeq}(s_1,s_3)].\end{aligned}$$

Die folgenden Formeln beschreiben eine Spalte s^*, für die $\text{hma}_1(S'_g) = \{s^*\}$

gilt bzw. zwei Spalten s_1 und s_2, für die $\text{hma}_2(S'_g) = \{s_1, s_2\}$ gilt:

$$\begin{aligned}
\phi_{\text{hma}_1}(g, s^*) &\equiv (\forall s.S(s) \wedge G_2(g,s))[\phi_{\succeq}(s^*, s)], \\
\phi_{\text{hma}_2}(g, s_1, s_2) &\equiv \neg(\phi_{\succeq}(s_1, s_2) \vee \phi_{\succeq}(s_2, s_1)) \wedge \\
&\quad (\forall s.S(s) \wedge G_2(g,s))[\phi_{\succeq}(s_1, s) \vee \phi_{\succeq}(s, s_1) \vee \\
&\quad \phi_{\succeq}(s_2, s) \vee \phi_{\succeq}(s, s_2)] \wedge \\
&\quad (\forall s.S(s) \wedge G_2(g,s))[\\
&\quad ((\phi_{\succeq}(s, s_1) \wedge s \neq s_1) \vee (\phi_{\succeq}(s, s_2) \wedge s \neq s_2)) \rightarrow \\
&\quad (\forall t.S(t) \wedge G_2(g,t))[\phi_{\succeq}(s,t) \vee \phi_{\succeq}(t,s)]].
\end{aligned}$$

Folgende Formel ist ähnlich zur Formel $\phi_{\text{sep}}(s, s')$ aus Satz 3.18 und beschreibt zwei Spalten, die separierbar sind:

$$\begin{aligned}
\phi_{\text{sep}}(g, s, s') \equiv (\forall z.Z(z))[&(\phi_{G_1}(g, z, s) \rightarrow \phi_{G_0}(g, z, s')) \wedge \\
&(\phi_{G_1}(g, z, s') \rightarrow \phi_{G_0}(g, z, s))].
\end{aligned}$$

Nun werden analog zum Beweis von Satz 3.18 drei Formeln angegeben, so dass eine der Formeln wahr ist, falls G'_g eine gerichtete perfekte Pfadphylogenie zulässt und keine Formel wahr ist, falls G'_g keine gerichtete perfekte Phylogenie zulässt. Die drei Formeln entsprechen dabei drei Möglichkeiten zum Aufbau von S'_g. Das heißt, entweder eine der Mengen $\text{hma}_1(S'_g)$ und $\text{hma}_2(S'_g)$ ist leer oder beide sind nicht leer. Die Formeln sind folgendermaßen definiert:

$$\begin{aligned}
\phi_{\text{Fall}_1}(g) &\equiv (\exists s^*.S(s^*) \wedge G_2(g, s^*))[\phi_{\text{hma}_1}(g, s^*)] \wedge \\
&\quad (\forall s_1.S(s_1) \wedge G_2(g,s), s_2.S(s_2) \wedge G_2(g,s))[\neg\phi_{\text{hma}_2}(g, s_1, s_2)], \\
\phi_{\text{Fall}_2}(g) &\equiv (\forall s^*.S(s^*) \wedge G_2(g, s^*))[\neg\phi_{\text{hma}_1}(g, s^*)] \wedge \\
&\quad (\exists s_1.S(s_1) \wedge G_2(g, s_1), s_2.S(s_2) \wedge G_2(g, s_2))[\\
&\quad \phi_{\text{hma}_2}(g, s_1, s_2) \wedge \phi_{\text{sep}}(g, s_1, s_2)], \\
\phi_{\text{Fall}_3}(g) &\equiv (\exists s^*.S(s^*) \wedge G_2(g, s^*), s_1.S(s_1) \wedge G_2(g, s_1), s_2.S(s_2) \wedge G_2(g, s_2))[\\
&\quad \phi_{\text{hma}_1}(g, s^*) \wedge \phi_{\text{hma}_2}(g, s_1, s_2) \wedge \\
&\quad (\phi_{\text{sep}}(g, s_1, s^*) \vee \phi_{\text{sep}}(g, s_2, s^*) \vee (\exists s.S(s) \wedge G_2(g,s))[\\
&\quad \phi_{\succeq}(g, s, s_1) \wedge \phi_{\succeq}(g, s, s_2) \wedge s \neq s^* \wedge \phi_{\text{sep}}(g, s, s^*)])].
\end{aligned}$$

Folgende Formel beschreibt nun eine Genotypmatrix, in der die Untermatrix G_g eine perfekte Phylogenie zulässt:

$$\phi_{\text{PPPHauf2er}}(g) \equiv \phi_{\text{ZweiKetten}}(g) \wedge (\phi_{\text{Fall}_1}(g) \vee \phi_{\text{Fall}_2}(g) \vee \phi_{\text{Fall}_3}(g)).$$

Wie im Beweis zu Satz 3.18 muss auch diese Formel noch so erweitert werden, dass von doppelten Spalten nur genau eine betrachtet wird.

Eine Genotypmatrix mit maximal einem heterozytischen Eintrag pro Spalte lässt eine perfekte Phylogenie zu, falls jedes Spaltenpaar maximal drei Elemente induziert und außerdem die Genotypmatrix nur eine Zeile enthält oder für jeden

Genotyp g die Untermatrix G_g eine perfekte Pfadphylogenie zulässt. Folgende Formel beschreibt diese Eigenschaft und somit PPH$(\infty,1)$:

$$\begin{aligned}\phi_{\text{PPH}(\infty,1)} \equiv\ & (\forall s.S(s), s'.S(s'))[\phi_{\text{IND}\leq 3}(s,s')]\wedge \\ & ((\exists z.Z(z))[\forall z'.Z(z')[z=z']]\vee \\ & (\forall z.Z(z))[(\exists s.S(s))[G_2(z,s)]\rightarrow \phi_{\text{PPPHauf2er}}(z)]).\end{aligned}$$

Die Formel $\phi_{\text{IND}\leq 3}(s,s')$ wird im Beweis zu Lemma 3.11 eingeführt und beschreibt ein Spaltenpaar (s,s'), dessen induzierte Menge maximal 3 Elemente enthält. □

3.3.4 Komplexität von PPPH

In diesem Abschnitt wird die Komplexität der Haplotypisierung mittels perfekten Pfadphylogenien betrachtet. Gramm et al. [17] führten diesen Ansatz zur Haplotypisierung ein und zeigten, dass ger-PPPH in Linearzeit lösbar ist. In diesem Abschnitt wird gezeigt, dass ger-PPPH in FO liegt. Dieses Ergebnis ist bereits bekannt, aber einen ersten Beweis hierzu gibt diese Arbeit.

Im Folgenden werden die Spalten einer Genotypmatrix nicht durch einen Index benannt, sondern die Spalten werden durch Spaltenvektoren dargestellt, die den Inhalt einer Spalte enthalten. Falls beispielsweise G eine $n\times m$ Genotypmatrix ist und wir über eine Spalte s aus G sprechen, dann bezeichnet $s[i]$ für $i\in\{1,\ldots,n\}$ den Eintrag von s in Zeile i. Nun werden zwei partielle Ordnungen auf den Spalten von Genotypmatrizen eingeführt und einige Begriffe aus der Ordnungstheorie erläutert. Im Kontext perfekter Pfadphylogenien wurden diese Konzepte zuerst von Gramm et al. [17] eingeführt und verwendet.

Seien G eine $n\times m$ Genotypmatrix, (H,B_H) eine gerichtete PP-Lösung für G und S die Menge der Spaltenvektoren von G. Zwei partielle Ordnungen auf S werden folgendermaßen definiert:

- Die Nachfolgerelation $\rightarrow$, die sich aus den Kantenmarkierungen von B_H ergibt. Für $s,s'\in S$ gilt $s\rightarrow s'$, falls in B_H die Spalte s auf dem Weg von der Wurzel zu s' liegt.
- Die partielle Ordnung $\succeq$. Sei $1\succ 2\succ 0$. Die Ordnung wird folgendermaßen auf die Spalten einer Genotypmatrix erweitert: Für $s,s'\in S$ gilt $s\succeq s'$, falls $s[i]\succeq s'[i]$ für jedes $i\in\{1,\ldots,n\}$ gilt.

Nun wird folgende Aussage gezeigt: Für zwei Spalten $s,s'\in S$ gilt $s\succeq s'$, falls $s\rightarrow s'$ gilt. Sei G eine $n\times m$ Genotypmatrix und (H,B_H) eine PP-Lösung für G. Weiter sei i eine beliebige Zeile in G, die den Genotyp g enthält sowie seien s und s' zwei Spalten, für die $s\rightarrow s'$ gilt. Seien h und h' die erklärenden Haplotypen zu g in (H,B_H). Falls nun $s[i]=1$, dann gilt $s[i]\succeq s'[i]$. Falls $s[i]=2$, dann liegt genau einer der Haplotypen h und h' in (H,B_H) unterhalb von s. Folglich liegt maximal einer der Haplotypen h und h' unterhalb von s' und es gilt $s'[i]=2$ oder $s'[i]=0$ und damit $s[i]\succeq s'[i]$. Falls $s[i]=0$, liegt keiner der Haplotypen h und h' unterhalb von

s und damit auch nicht unterhalb von $s'[i]$. In diesem Fall gilt $s'[i] = 0$ und somit $s[i] \succeq s'[i]$. Insgesamt folgt, dass $s \succeq s'$ gilt. Die Vergleichbarkeit zweier Spalten ist somit eine notwendige Bedingung dafür, dass in einer gerichteten perfekten Phylogenie ein Weg von der Wurzel zu einem Blatt existiert, auf dem die beiden Spalten vorkommen und kein Knoten mehrfach besucht wird.

Die durch $\succeq$ partiell geordnete Spaltenmenge S wird mit $(S, \succeq)$ bezeichnet. Eine Teilmenge $S' \subseteq S$ heißt *Kette*, falls die Element in S' paarweise vergleichbar sind. Das heißt, für zwei beliebige Spalten $s, s' \in S'$ gilt $s \succeq s'$ oder $s' \succeq s$. Eine Teilmenge $S' \subseteq S$ heißt *Antikette*, falls die Elemente in S' paarweise nicht vergleichbar sind. Das heißt, für zwei beliebige Spalten $s, s' \in S'$ gilt $s \not\succeq s'$ und $s' \not\succeq s$. Eine Antikette S' ist *maximal*, wenn für jedes $s \in S \setminus S'$ die Menge $\{s\} \cup S'$ keine Antikette bildet. Eine maximale Antikette mit Größe k ist die *höchste maximale Antikette der Größe k*, wenn kein Element einer anderen Antikette der Größe genau k echt größer als ein Element aus S' ist. In $(S, \succeq)$ gibt es für jedes k nicht mehr als eine höchste maximale Antikette der Größe k, die wir mit $\text{hma}_k(S)$ bezeichnen. Falls keine höchste maximale Antikette der Größe k existiert, sei $\text{hma}_k(S) = \emptyset$.

Ein Spaltenpaar (s, s') heißt *separierbar*, falls folgende zwei Implikationen gelten: wenn $s[i] = 1$, dann $s'[i] = 0$ und wenn $s[i] = 1$, dann $s'[i] = 0$. Zum Beispiel sind die Spalten $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ und $\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$ separierbar und die Spalten $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ und $\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$ nicht separierbar.

Gramm et al. [17] stellten einen Algorithmus vor, der ger-PPPH in Linearzeit löst. Der Algorithmus prüft, ob eine Genotypmatrix die Bedingungen 1 und 2 aus dem folgenden Lemma erfüllt.

Lemma 3.17 (Gramm et al. [17])**.** *Sei G eine Genotypmatrix und S die Menge der Spalten von S. Die Genotypmatrix G lässt genau dann eine gerichtete perfekte Pfadphylogenie zu, wenn*

1. *sich $(S, \succeq)$ mit maximal zwei Ketten überdecken lässt, so dass*
2. *die maximalen Elemente der beiden Ketten separierbar sind.*

Satz 3.18. ger-PPPH $\in$ FO.

Beweis. Im Folgenden wird eine Formel $\phi_{\text{ger-PPPH}}$ schrittweise hergeleitet, so dass genau dann $(I, Z, S, G_0, G_1, G_2) \models \phi_{\text{ger-PPPH}}$ gilt, wenn die durch (I, Z, S, G_0, G_1, G_2) kodierte Genotypmatrix eine gerichtete perfekte Pfadphylogenie zulässt.

Nachfolgende Formel beschreibt die partielle Ordnung $s \succeq t$ auf den Spalten einer Genotypmatrix:

$$\phi_{\succeq}(s,t) \equiv (\forall z.Z(z))[G_0(z,t) \vee G_1(z,s) \vee (G_2(z,s) \wedge G_2(z,t))].$$

Folgende Formel beschreibt, dass sich die partiell geordnete Menge von Spalten $(S, \succeq)$ einer Genotypmatrix mit maximal zwei Ketten überdecken lässt:

$$\begin{aligned} \phi_{\text{ZweiKetten}} \equiv (\forall s_1.S(s_1), s_2.S(s_2), s_3.S(s_3))[&\phi_{\succeq}(s_1,s_2) \vee \phi_{\succeq}(s_2,s_1) \vee \\ &\phi_{\succeq}(s_2,s_3) \vee \phi_{\succeq}(s_3,s_2) \vee \\ &\phi_{\succeq}(s_3,s_1) \vee \phi_{\succeq}(s_1,s_3)]. \end{aligned}$$

Formel $\phi_{\text{ZweiKetten}}$ beschreibt, dass $(S, \succeq)$ keine Antikette der Größe 3 enthält. Die Weite von $(S, \succeq)$ ist damit maximal 2. Die Korrektheit von $\phi_{\text{ZweiKetten}}$ folgt mit dem Satz von Dilworth, der besagt, dass die Weite einer partiell geordneten Menge gleich der minimalen Anzahl von Ketten ist, mit der man die Menge überdecken kann. Die Formel $\phi_{\text{ZweiKetten}}$ entspricht somit genau der Bedingung 1 von Lemma 3.17. Wenn eine Struktur (I, Z, S, G_0, G_1, G_2) ein Modell von $\phi_{\text{ZweiKetten}}$ ist, dann besitzt die zugehörige partielle Ordnung $(S, \succeq)$ eine Weite von maximal 2 und für $k \geq 3$ ist die Menge $\text{hma}_k(S)$ leer.

Im Folgenden wird unterschieden, ob für eine partielle Ordnung $(S, \succeq)$ genau eine der Menge $\text{hma}_1(S)$ oder $\text{hma}_2(S)$ leer ist oder beide Mengen nicht leer sind. Auf diese Weise ergeben sich drei Fälle zum Aufbau von $(S, \succeq)$. Für jeden dieser Fälle wird eine Formel eingeführt, die beschreibt, ob sich unter den möglichen Partitionierungen von $(S, \succeq)$ in maximal zwei Ketten eine Partitionierung befindet, die Bedingung 2 von Lemma 3.17 erfüllt.

Die folgenden zwei Formeln beschreiben, dass eine Spalte in $\text{hma}_1(S)$ enthalten ist bzw. dass zwei Spalten in $\text{hma}_2(S)$ enthalten sind:

$$\begin{aligned}
\phi_{\text{hma}_1}(s^*) &\equiv (\forall s.S(s))[\phi_{\succeq}(s^*, s)], \\
\phi_{\text{hma}_2}(s_1, s_2) &\equiv \neg(\phi_{\succeq}(s_1, s_2) \vee \phi_{\succeq}(s_2, s_1)) \wedge \\
&\quad (\forall s.S(s))[\phi_{\succeq}(s_1, s) \vee \phi_{\succeq}(s, s_1) \vee \phi_{\succeq}(s_2, s) \vee \phi_{\succeq}(s, s_2)] \wedge \\
&\quad (\forall s.S(s))[((\phi_{\succeq}(s, s_1) \wedge s \neq s_1) \vee (\phi_{\succeq}(s, s_2) \wedge s \neq s_2)) \rightarrow \\
&\qquad (\forall t.S(t))[\phi_{\succeq}(s, t) \vee \phi_{\succeq}(t, s)]].
\end{aligned}$$

Die Formel $\phi_{\text{hma}_1}(s^*)$ besagt, dass das Element in $\text{hma}_1(S)$ größer als jedes andere Element ist. Die Formel $\phi_{\text{hma}_2}(s_1, s_2)$ beschreibt $\text{hma}_2(S)$ wie folgt: Die Elemente von $\text{hma}_2(S)$ sind nicht vergleichbar und $\text{hma}_2(S)$ ist eine maximale Antikette der Größe 2. Das heißt, kein Element kann zu $\text{hma}_2(S)$ hinzugefügt werden, so dass die resultierende Menge eine Antikette ist. Zudem ist $\text{hma}_2(S)$ die höchste maximale Antikette der Größe 2. Das heißt, es existiert keine Antikette der Größe 2 mit einem Element, welches echt größer als ein Element aus $\text{hma}_2(S)$ ist.

Folgende Formel beschreibt, dass ein Spaltenpaar separierbar ist:

$$\phi_{\text{sep}}(s, s') \equiv (\forall z.Z(z))[(G_1(z, s) \rightarrow G_0(z, s')) \wedge (G_1(z, s') \rightarrow G_0(z, s))].$$

Falls $\text{hma}_1(S) = \{s^*\}$ und $\text{hma}_2(S) = \emptyset$, dann ist $(S, \succeq)$ eine Kette. Diese Eigenschaft lässt sich per Induktion über die Größe von S zeigen. Man steigt dabei Schritt für Schritt die Ordnung $(S, \succeq)$ hinab und erkennt, dass keine Verzweigung zu zwei nicht vergleichbaren Elementen existiert. Für die Überdeckung (K_1, K_2) mit $K_1 = S$ und $K_2 = \emptyset$ ist somit Bedingung 2 von Lemma 3.17 erfüllt. Diesen Fall beschreibt folgende Formel:

$$\begin{aligned}
\phi_{\text{Fall}_1} \equiv\ & (\exists s^*.S(s^*))[\phi_{\text{hma}_1}(s^*)] \wedge \\
& (\forall s_1.S(s_1), s_2.S(s_2))[\neg\phi_{\text{hma}_2}(s_1, s_2)].
\end{aligned}$$

Falls $\text{hma}_1(S) = \emptyset$ und $\text{hma}_2(S) = \{s_1, s_2\}$, dann liegen s_1 und s_2 in verschiedenen Ketten jeder Partitionierung von $(S, \succeq)$ in Ketten. Sei (K_1, K_2) eine beliebige Partitionierung von $(S, \succeq)$ in zwei Ketten mit s_1 in K_1 und s_2 in K_2. Nach Voraussetzung gilt, dass s_1 das maximale Element von K_1 ist und dass s_2 das maximale Element von K_2 ist. Daraus folgt, dass eine Überdeckung mit Bedingung 2 von Lemma 3.17 genau dann existiert, wenn s_1 und s_2 separierbar sind. Dieser Fall wird durch folgende Formel beschrieben:

$$\begin{aligned}\phi_{\text{Fall}_2} \equiv\ & (\forall s^*.S(s^*))[\neg\phi_{\text{hma}_1}(s^*)]\wedge \\ & (\exists s_1.S(s_1), s_2.S(s_2))[\phi_{\text{hma}_2}(s_1, s_2) \wedge \phi_{\text{sep}}(s_1, s_2)].\end{aligned}$$

Falls $\text{hma}_1(S) = \{s^*\}$ und $\text{hma}_2(S) = \{s_1, s_2\}$, dann liegen s_1 und s_2 in verschiedenen Ketten jeder Überdeckung von $(S, \succeq)$ in Ketten und s^* ist maximales Element einer Kette. Sei (K_1, K_2) eine beliebige Partitionierung von $(S, \succeq)$ in zwei Ketten mit s_1 in K_1 und s_2 in K_2. Die Spalte s^* liegt entweder in K_1 oder in K_2. Wir nehmen an, dass s^* in K_1 liegt. Dann ist s^* das maximale Element von K_1. Die Spalten, die Größer als s_1 und s_2 und kleiner als s^* sind, können unter Einhaltung der Ordnung $\succeq$ beliebig auf K_1 und K_2 verteilt werden. Das oberste Element in K_2 ist somit s_2 oder eines dieser Elemente. Nun existiert eine Überdeckung mit Bedingung 2 von Lemma 3.17, falls eine Spalte existiert, die größer oder gleich s_2 ist und mit s^* separierbar ist. Diese Spalte wird bei der Aufteilung der Elemente nach K_1 und K_2 als maximales Element von K_2 gesetzt. Die Argumentation lässt sich analog für Partitionierungen führen, in denen s^* und s_2 in einer gemeinsamen Kette liegen. Nachfolgende Formel beschreibt diesen Fall:

$$\begin{aligned}\phi_{\text{Fall}_3} \equiv\ & (\exists s^*.S(s^*), s_1.S(s_1), s_2.S(s_2))[\phi_{\text{hma}_1}(s^*) \wedge \phi_{\text{hma}_2}(s_1, s_2)\wedge \\ & (\phi_{\text{sep}}(s_1, s^*) \vee \phi_{\text{sep}}(s_2, s^*)\vee \\ & (\exists s.S(s))[\phi_{\succeq}(s, s_1) \wedge \phi_{\succeq}(s, s_2) \wedge s \neq s^* \wedge \phi_{\text{sep}}(s, s^*)])].\end{aligned}$$

Die nachfolgende Formel besagt, dass sich $(S, \succeq)$ mit zwei Ketten überdecken lässt und einer der drei Fälle zutrifft:

$$\phi_{\text{ger-PPPH}} \equiv \phi_{\text{ZweiKetten}} \wedge (\phi_{\text{Fall}_1} \vee \phi_{\text{Fall}_2} \vee \phi_{\text{Fall}_3}).$$

Zu beachten ist noch, dass eine Genotypmatrix, für die die Formel $\phi_{\text{ger-PPPH}}$ ausgewertet wird, mehrere Spalten mit gleichem Inhalt enthalten kann. Für eine solche Genotypmatrix kann es zum Beispiel mehrere Spalten geben, die gleich dem Element aus der Menge $hma_1(S)$ sind. Um dies zu umgehen, kann man mit einer Anfrage in Prädikatenlogik erster Stufe so Spalten in einer Genotypmatrix löschen, dass von gleichen Spalten nur genau eine erhalten bleibt. Durch Kombination dieser Anfrage und der oben beschriebenen Formel ergibt sich eine vollständige Beschreibung von ger-MPPPH. □

Bei der Haplotypisierung mittels perfekten Phylogenien lässt sich die ungerichtete Problemvariante ung-PPH mit einer Anfrage in Prädikatenlogik erster Stufe auf die gerichtete Problemvariante ger-PPH reduzieren (siehe Lemma 3.6). Es

ist aber nicht klar, ob eine solche Reduktion auch von ung-PPPH auf ger-PPPH möglich ist. Ebenfalls ist nicht klar, ob die Reduktion aus Lemma 3.6 schon eine Reduktion von ung-PPPH auf ger-PPPH darstellt. Das Folgende Lemma zeigt, dass sich ung-PPPH mit einer Anfrage in Prädikatenlogik erster Stufe auf ger-PPPH reduzieren lässt, falls man nur Genotypmatrizen zulässt, die einen Genotyp enthalten, der an maximal einer Stelle heterozytisch ist.

Lemma 3.19. *Sei G eine Genotypmatrix, die einen Genotyp enthält, der an maximal einer Stelle heterozytisch ist. Es existiert eine Abbildung A von Genotypmatrizen auf Genotypmatrizen, die sich als Anfrage in Prädikatenlogik erster Stufe beschreiben lässt und für die gilt: $G \in$ ung-PPPH gdw. $A(G) \in$ ger-PPPH.*

Beweis. Im Folgenden betrachten wir eine Genotypmatrix G, die einen Genotyp enthält, der an maximal einer Stelle heterozytisch ist. Es wird nun zuerst die Abbildung A angegeben. Dann wird die obige Aussage zur Existenz ungerichteter und gerichteter perfekter Pfadphylogenien bewiesen und abschließend wird A als Anfrage in Prädikatenlogik erster Stufe formuliert.

Die Abbildung A bildet eine Genotypmatrix G wie folgt auf eine Genotypmatrix $A(G)$ ab: Sei g_0 ein Genotyp in G, der nur an maximal einer Stelle heterozytisch ist. Die Genotypmatrix $A(G)$ entsteht aus G, indem in jeder Spalte s, für die $g_0[s] = 1$ gilt, die Rollen von 0 und 1 vertauscht werden. Es lässt sich erkennen, dass diese Abbildung ähnlich der Reduktion im Beweis zu Lemma 3.3 ist.

Ebenfalls analog zu Lemma 3.3 wird nun gezeigt, dass G genau dann eine perfekte Pfadphylogenie zulässt, wenn $G' = A(G)$ eine gerichtete perfekte Pfadphylogenie zulässt. Hierzu sei einerseits (H, B_H) eine PPP-Lösung für G. Eine gerichtete PPP-Lösung $(H', B_{H'})$ für G' entsteht dadurch, dass in jeder Spalte, in der bei der Abbildung getauscht wird, auch in H und den Knotenmarkierungen von $B_{H'}$ die Rollen von 0 und 1 vertauscht werden. Der Genotyp g'_0 in G', der aus dem Genotyp g_0 in G entsteht, verwendet den Haplotyp $0 \ldots 0$ zur Erklärung und folglich kommt der Haplotyp $0 \ldots 0$ in $B_{H'}$ vor. Dies ist gleichbedeutend mit der Eigenschaft, dass $B_{H'}$ gerichtet ist. Auf der anderen Seite sei $(H', B_{H'})$ eine gerichtete PPP-Lösung für G'. Analog zum ersten Beweisschritt entsteht eine PPP-Lösung für G aus $(H', B_{H'})$, indem in den Spalten, in denen bei der Abbildung getauscht wird, zurück getauscht wird.

Die obige Abbildung wird durch eine Anfrage beschrieben, die bis auf eine kleine Änderung der Anfrage $A_{\text{ung-ger}}$ aus dem Beweis zu Lemma 3.6 gleicht. Es ändert sich nur die Definition der Formel $\phi_{\text{tausch}}(s)$, die angibt, ob in einer Spalte s die Rollen von 0 und 1 vertauscht werden. Die übrigen Formeln $\phi_{I'}$, $\phi_{Z'}$, $\phi_{S'}$, $\phi_{G'_0}$, $\phi_{G'_1}$ und $\phi_{G'_2}$ sind wie im Beweis zu Lemma 3.6 definiert. Aus diesem Grund wird an dieser Stelle nur die Formel $\phi_{\text{tausch}}(s)$ neu formuliert und für den restlichen Teil der Anfrage auf den Beweis zu Lemma 3.6 verwiesen.

Folgende Formel beschreibt eine Zeile, die maximal einen heterozytischen Eintrag enthält:

$$\phi_{\text{maxEine2}}(z) \equiv (\exists s.S(s))[(\forall s'.S(s'))[G_2(z,s') \rightarrow s = s']].$$

Die nachfolgende Formel beschreibt die kleinste Zeile, die einen Genotyp enthält, der an maximal einer Stelle heterozytisch ist:

$$\phi_{\text{kleinsterIndex}}(z) \equiv \phi_{\text{maxEine2}}(z) \wedge (\forall z'.Z(z'))[\phi_{\text{maxEine2}}(z') \rightarrow z' \geq z].$$

Nun folgt die neue Formulierung der Formel $\phi_{\text{tausch}}(s)$, die angibt ob in einer Spalte getauscht wird:

$$\phi_{\text{tausch}}(s) \equiv (\exists z.Z(z))[\phi_{\text{kleinsterIndex}}(z) \wedge G_1(z,s)].$$

□

Satz 3.20. PPP $\in$ FO.

Beweis. Lemma 3.19 sagt aus, dass sich für Genotypmatrizen, die einen Genotyp enthalten, der maximal einen heterozytischen Eintrag besitzt, das Problem PPPH auf ger-PPPH mit einer Anfrage in Prädikatenlogik erster Stufe reduzieren lässt. Satz 3.18 sagt aus, dass sich das Problem ger-PPPH durch eine Formel $\phi_{\text{ger-PPPH}}$ in Prädikatenlogik erster Stufe beschreiben lässt. Das Problem PPP umfasst nur Haplotypmatrizen und wie schon erwähnt, kann man Haplotypmatrizen als Genotypmatrizen ohne heterozytische Einträge auffassen. Wenn man die Eingabe auf Genotypmatrizen ohne heterozytische Einträge einschränkt (das heißt, die Relation G_2 ist in den betrachteten Strukturen leer), dann stellt die Reduktion aus Lemma 3.19 auch eine Reduktion von PPP auf ger-PPP dar und die Formel $\phi_{\text{ger-PPPH}}$ beschreibt auch ger-PPP. Die Anfrage aus Lemma 3.19 und die Formel $\phi_{\text{ger-PPPH}}$ lassen sich nun folgendermaßen zu einer Formel für PPP kombinieren: In der Formel $\phi_{\text{ger-PPPH}}$ wird jede Relation durch die entsprechende Formel aus der Anfrage von Lemma 3.19 ersetzt. Zum Beispiel wird die Relation $G_0(z,s)$ durch die Formel $\phi_{G'_0}(z,s)$ ersetzt. Auf diese Weise werden die Relationen der Genotypmatrix, auf die durch die Anfrage abgebildet wird, verwendet. Für beliebige Anfragen und passende Formeln (solche, die über der Bildsignatur der Anfrage formuliert werden) wird dieses Verfahren von Immerman [25, Seite 46f] beschrieben und als duale Abbildung von Anfragen bezeichnet. □

3.4 Haplotypisierung mittels kombinierten Ansätzen

In diesem Abschnitt wird auf die Komplexität von kombinierten Ansätzen zur Haplotypisierung eingegangen. Der Abschnitt 3.4.1 behandelt die Komplexität der Haplotypisierung mittels minimalen perfekten Phylogenien. Dieses Problem ist NP-vollständig und es wird ein Beweis hierzu aus der Literatur vorgestellt. Der Abschnitt 3.4.2 behandelt die Komplexität der Haplotypisierung mittels minimalen perfekten Pfadphylogenien. Die genaue Komplexität dieses Problems war bisher unbekannt. In dieser Arbeit wird gezeigt, dass MPPPH in der Klasse L liegt.

3.4.1 Komplexität von MPPH

In diesem Abschnitt werden bekannte Ergebnisse zur Komplexität von MPPH und (k,l)-beschränkten Varianten von MPPH vorgestellt. Weiter wird ein Beweis zur NP-Vollständigkeit von MPPH aus der Literatur vorgestellt.

Für das Problem MPPH zeigten zuerst Bafna et al. [2] die NP-Vollständigkeit. Van Iersel et al. [35] stellten Resultate zu (k,l)-beschränkten Variante von MPPH vor. Sie zeigten, dass MPPH$(2,\infty)$ und MPPH$(\infty,1)$ in Polynomialzeit lösbar sind und dass MPPH$(3,3)$ NP-vollständig und APX-hart ist. Außerdem wurde gezeigt, dass MPPH$(\infty,2)$ in Polynomialzeit lösbar ist, falls der Kompatibilitätgraph der Genotypmatrix vollständig ist. Die bekannte Komplexität (k,l)-beschränkter Varianten von MPPH ist damit analog zur Komplexität (k,l)-beschränkter Varianten von MH. Ein Überblick über die Komplexität der Varianten von MPPH lässt sich analog zu dem Überblick über die Varianten von MH in Abbildung 2 auf Seite 21 darstellen. Auch das Problem MPPH wurde als ganzzahliges lineares Programm formuliert. Eine erste Formulierung findet sich bei Brown und Harrower [5].

Im Folgenden wird gezeigt, dass MPPH NP-vollständig ist. Der hier vorgestellte Beweis basiert auf dem ersten Beweis zur NP-Vollständigkeit von MPPH [2].

Satz 3.21. MPPH *ist* NP-*vollständig.*

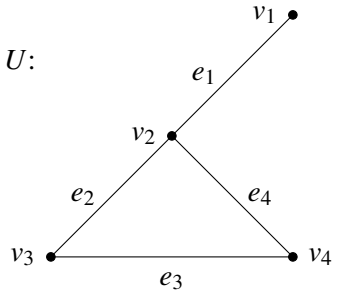
Beweis. Zum Beweis der NP-Vollständigkeit von MPPH wird gezeigt, dass MPPH in NP liegt und dass MPPH NP-hart ist.

MPPH *liegt in* NP*:* Eine nichtdeterministische Turingmaschine, die MPPH in Polynomialzeit akzeptiert, arbeitet analog zu der Turingmaschine im Beweis zu Satz 3.1. Die Turingmaschine prüft zusätzlich, ob die geratene Haplotypmatrix eine perfekte Phylogenie zulässt. Dieses Problem lässt sich als Formel in Prädikatenlogik erster Stufe beschreiben (siehe Satz 3.5) und ist somit deterministisch in Polynomialzeit lösbar.

MPPH *ist* NP-*hart:* Für den Beweis der NP-Härte wird eine Reduktion von KNOTENÜBERDECKUNG angegeben. KNOTENÜBERDECKUNG ist im Beweis zu Satz 3.1 beschrieben. Der Beweis zur Reduktion teilt sich, wie im Beweis zu Satz 3.1, in drei Schritte auf und die Spalten von Genotyp- und Haplotypmatrizen werden wieder durch Indizes i und j referenziert.

Konstruktion: Es sei ein ungerichteter Graph $U = (V,E)$ mit Knotenmenge $V = \{v_1,\ldots,v_n\}$ und Kantenmenge $E = \{e_1,\ldots,e_m\}$ gegeben. Der Graph U wird wie folgt auf eine Genotypmatrix G mit $n+m+1$ Zeilen und $2n+m$ Spalten abgebildet: Für jeden Knoten $v_i \in V$ enthält G den Genotyp g_{v_i} mit $g_{v_i}[i] = g_{v_i}[n+i] = 1$ und sonstigen Einträgen 0. Ein solcher Genotyp wird im Weiteren *Knotengenotyp* genannt. Für jede Kante $e_k = \{v_i, v_j\} \in E$ enthält G den Genotyp g_{e_k} mit $g_{e_k}[i] = g_{e_k}[j] = g_{e_k}[2n+k] = 2$ und sonstigen Einträgen 0. Ein solcher Genotyp wird im Weiteren *Kantengenotyp* genannt. Abschließend wird ein Genotyp eingefügt, der an jeder Position den Eintrag 0 besitzt. Dieser Genotyp wird im Weiteren mit g_0 bezeichnet. Die Konstruktion wird in Abbildung 6 an einem Beispiel gezeigt.

Abbildung 6: Die Abbildung zeigt an einem Beispiel die Reduktion von KNOTENÜBERDECKUNG auf MPPH. Für den Graph U mit vier Knoten und vier Kanten wird ein Genotypmatrix G mit neun Zeilen und zwölf Spalten erstellt. Die oberen vier Zeilen enthalten Knotengenotypen. Danach folgen vier Kantengenotypen und ein weiterer Genotyp, der an jeder Position den Eintrag 0 enthält. Die Menge $V' = \{v_2, v_3\}$ ist eine Knotenüberdeckung minimaler Größe für U. Dazu korrespondiert eine erklärende Haplotypmatrix für G, die die Haplotypen aus H_V, $H_E = \{100000001000, 010000000100, 000100000010, 000100000001\}$, $H_{V'} = \{010000000000, 001000000000\}$ und den Haplotyp $0 \ldots 0$ enthält, wobei H_V die Menge der Knotengenotypen in G bezeichnet.



G:

g_{v_1}	1 0 0 0 1 0 0 0 0 0 0 0
g_{v_2}	0 1 0 0 0 1 0 0 0 0 0 0
g_{v_3}	0 0 1 0 0 0 1 0 0 0 0 0
g_{v_4}	0 0 0 1 0 0 0 1 0 0 0 0
g_{e_1}	2 2 0 0 0 0 0 0 2 0 0 0
g_{e_2}	0 2 2 0 0 0 0 0 0 2 0 0
g_{e_3}	0 0 2 2 0 0 0 0 0 0 2 0
g_{e_4}	0 2 0 2 0 0 0 0 0 0 0 2
g_0	0 0 0 0 0 0 0 0 0 0 0 0

Korrektheit: Es wird nun gezeigt, dass genau dann eine Knotenüberdeckung $V' \subseteq V$ mit $|V'| \leq d$ für U existiert, wenn eine PP-Lösung mit maximal $n+m+d+1$ paarweise verschiedenen Haplotypen für G existiert.

Nur-wenn-Teil: Für die erste Beweisrichtung sei $V' \subseteq V$ mit $|V'| = d$ eine Knotenüberdeckung für U. Wir zeigen, dass dann eine PP-Lösung (H, B_H) der Größe $n+m+d+1$ für G existiert. Dazu wird im Folgenden eine Haplotypmatrix H, die G erklärt, schrittweise aufgebaut. Zuerst lässt sich erkennen, dass jeder Knotengenotyp nur 0 oder 1 als Eintrag enthält und damit als Haplotyp in H vorkommt. Sei H_V die Menge der Haplotypen, die aus den Knotengenotypen entstehen. Nach Konstruktion der Knotengenotypen besitzt jeder Haplotyp in H_V genau zweimal den Eintrag 1 und H_V enthält n Haplotypen. Für jeden Kantengenotyp wird folgendermaßen ein Haplotyp erstellt: Es sei g_{e_k} ein Kantengenotyp in G und $e_k = \{v_i, v_j\}$ die entsprechende Kante in U. Weiter sei ein Endknoten von e_k, der in V' liegt, fest gewählt. Wir wählen $v_i \in V'$ fest. Für g_{e_k} wird dann der Haplotyp h mit $h[j] = h[n+k] = 1$ und sonstigen Einträgen 0 erstellt. Sei H_E die Menge der Haplotypen, die auf diese Weise aus den Kantengenotypen entstehen. Jeder Haplotyp in H_E besitzt genau zweimal den Eintrag 1 und H_E enthält m Haplotypen. Abschließend wird für jeden Knoten $v_i \in V'$ ein Haplotyp h mit $h[i] = 1$ und sonstigen Einträgen 0 erstellt. Sei $H_{V'}$ Menge der Haplotypen, die aus der Knotenüberdeckung entstehen. Nach Konstruktion besitzt jeder Haplotyp in $H_{V'}$ genau einmal den Eintrag 1 und $H_{V'}$ enthält d Haplotypen. Es lässt sich erkennen, dass keiner der erstellten Haplotypen in mehr als einer der drei Mengen vorkommt. So-

mit enthält $H_V \cup H_E \cup H_{V'}$ genau $n+m+d$ Haplotypen.

Es gilt nun, dass jeder Knotengenotyp durch den entsprechenden Haplotyp aus H_V erklärt wird und dass jeder Kantengenotyp g_{e_k} durch den entsprechenden Haplotyp aus H_E und einen Haplotyp aus $H_{V'}$ erklärt wird. Der Haplotyp, den g_{e_k} aus $H_{V'}$ verwendet, entspricht dabei einem Knoten, mit dem e_k abgedeckt wird. Es lässt sich nun eine Haplotypmatrix H mit $n+m+d+1$ paarweise verschiedenen Haplotypen konstruieren, die G erklärt und neben den Haplotypen aus H_V, H_E und $H_{V'}$ nur den Haplotyp $0\ldots0$ enthält.

Es verbleibt zu zeigen, dass H einer perfekte Phylogenie zulässt. Seien hierzu i und j mit $i < j$ zwei Spalten, die in H die Untermatrix $\left[\begin{smallmatrix}1 & 1\end{smallmatrix}\right]$ enthalten. Ein Haplotyp, der zweimal den Eintrag 1 enthält, kommt nach Konstruktion der Haplotypen aus H_V oder H_E. Falls wir einerseits annehmen, dass ein Haplotyp aus H_V an Position i und j den Eintrag 1 enthält, dann gilt, dass kein anderer Haplotyp in H an Position j den Eintrag 1 enthält. Falls wir andererseits annehmen, dass ein Haplotyp aus H_E an Position i und j den Eintrag 1 enthält, dann gilt $j = 2n+k$ und ebenfalls, dass kein anderer Haplotyp in H an Position j den Eintrag 1 enthält. Insgesamt folgt, dass kein Spaltenpaar in H die Untermatrix $\left[\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right]$ enthält und somit auch nicht die Untermatrix V aus Lemma 3.4 enthält. Die Haplotypmatrix H lässt folglich eine perfekte Phylogenie zu.

Wenn-Teil: Für diese Beweisrichtung nehmen wir an, dass eine PP-Lösung (H, B_H) der Größe $n+m+d+1$ für G existiert und konstruieren eine Knotenüberdeckung $V' \subseteq V$ mit $|V'| \leq d$ für U.

Wie schon in der ersten Beweisrichtung angemerkt lässt sich erkennen, dass die Knotengenotypen in G nur den Eintrag 0 oder 1 enthalten und daher als Haplotypen in H vorkommen. Sei H_V die Menge dieser n vielen Haplotypen. Nun sei g_{e_k} ein beliebiger Kantengenotyp aus G und gelte $e_k = \{v_i, v_j\}$. Seien weiter h und h' die erklärenden Haplotypen zu g_{e_k} aus H. Da der Kantengenotyp g_{e_k} in Spalte $n+m+k$ den Eintrag 2 besitzt, enthält entweder h oder h' an Position $n+m+k$ den Eintrag 1. Wie nehmen im Weiteren an, dass $h[n+m+k] = 1$ gilt. Nur der Kantengenotyp g_{e_k} besitzt an Position $n+m+k$ den Eintrag 2 und folglich wird h nur von g_{e_k} zur Erklärung verwendet. Sei H_E die Menge dieser m Haplotypen für die Kantengenotypen von G. Für den Haplotyp h gilt weiter, dass er nicht an Position i und j den Eintrag 1 enthält, da ansonsten H in den Spalten i und j die Untermatrix V aus Lemma 3.4 enthält. Gleiches gilt für h'. Folglich besitzt h' an genau einer Position aus $\{1,\ldots,n\}$ den Eintrag 1. Sei $H_{V'}$ die Menge dieser Haplotypen.

Nun wird eine Knotenüberdeckung V' folgendermaßen aus $H_{V'}$ erstellt: Falls $H_{V'}$ einen Haplotyp enthält, der an Position i den Eintrag 1 besitzt, dann ist v_i Teil der Knotenüberdeckung V'. Die Haplotypen aus H_V, H_E und der Haplotyp $0\ldots0$ sind paarweise verschieden und umfassen genau $n+m+1$ Haplotypen. Folglich umfasst $H_{V'}$ maximal d Haplotypen und es gilt $|V'| \leq d$. Außerdem ist jede Kante in U mit einem Knoten aus V' inzident, da jeder Kantengenotyp einen Haplotyp aus $H_{V'}$ zur Erklärung verwendet. Somit ist V' eine Knotenüberdeckung mit maximaler Größe d für U'.

Komplexität: Eine Turingmaschine, die die Abbildung von U nach G berechnet,

arbeitet analog zur Turingmaschine aus dem Beweis zu Satz 3.1. Es ändern sich nur die Berechnungsvorschrift zum Erstellen der Knoten- und Kantengenotypen.

Insgesamt folgt, dass MPPH NP-vollständig ist. □

Bei der Reduktion aus dem vorangegangenen Beweis wird jeder Graph auf eine Genotypmatrix mit maximal drei heterozytischen Stellen pro Genotyp abgebildet. Es wurde also auch gezeigt, dass MPPH$(3,\infty)$ NP-vollständig ist. Wie oben bemerkt zeigten van Iersel et al. [35], dass schon MPPH$(3,3)$ NP-vollständig ist. Für diesen Beweis wurde die obige Reduktion verwendet, aber es wurde von einer Variante von KNOTENÜBERDECKUNG reduziert. Bei dieser Variante des Problems der Knotenüberdeckung sucht man für einen Graphen mit maximalem Grad drei nach der kleinsten Knotenüberdeckung. Wenn jeder Knoten in U mit maximal drei Kanten inzident ist, dann enthält jede Spalte in G maximal in drei Genotypen den Eintrag 2.

3.4.2 Komplexität von MPPPH

In diesem Abschnitt wird die Komplexität der Haplotypisierung mittels minimalen perfekten Pfadphylogenien betrachtet. Es wird gezeigt, dass ger-MPPPH in L liegt. Die Argumentation dazu teilt sich in mehrere Schritte auf. Zu Beginn wird eine Vorverarbeitung für Genotypmatrizen beschrieben, durch die bestimmte Spalten in Genotypmatrizen gelöscht werden. Die Größe kleinster perfekter Pfadphylogenien ändert sich durch diese Vorverarbeitung nicht. Danach wird der prinzipielle Aufbau und die Größe von Genotypmatrizen untersucht, die aus der Vorverarbeitung hervorgehen. Abschließend wird gezeigt, dass sich ger-MPPPH durch eine logarithmisch platzbeschränkte deterministische Turingmaschine entscheiden lässt.

Vorverarbeitung von Genotypmatrizen. Im weiteren Teil dieser Arbeit wird angenommen, dass die Genotypmatrizen keine 0-Spalte enthalten. Sei G eine Genotypmatrix mit einer 0-Spalte und (H, B_H) eine gerichtete PPP-Lösung für G. Kein Knoten, der in B_H unterhalb einer 0-Spalte liegt, wird von einem Haplotyp aus H markiert. Sei G eine Genotypmatrix mit einer 0-Spalte und G' entstehe aus G durch löschen der 0-Spalten. Mit der beschriebenen Eigenschaft können PPP-Lösungen für G und G' ineinander überführt werden ohne die Anzahl der paarweise verschiedenen Haplotypen in der Lösung zu verändern. Es folgt, dass 0-Spalten ausgespart werden können ohne die Größe kleinster gerichteter PPP-Lösungen zu verändern.

Falls eine Genotypmatrix zwei identische Spalten s und s' enthält, dann können diese in einer PPP-Lösung durch identische oder verschiedene Spalten erklärt werden. Eine PPP-Lösung mit verschiedenen Spalten s und s' kann dabei nur existieren, falls ein Genotyp in s und s' heterozytisch ist, der dann ungleich aufgelöst wird. Durch den Beweis zum nachfolgenden Lemma wird klar, dass man sich bei der Suche nach kleinsten PPP-Lösung auf solche Lösungen beschränken kann, in

denen identische Spalten in G immer durch identischen Spalten in H erklärt werden.

Lemma 3.22. *Sei G eine Genotypmatrix und Genotypmatrix G' entstehe aus G, indem von zwei gleichen Spalten eine gelöscht wird. Dann existiert eine* PPP-*Lösung der Größe maximal d für G genau dann, wenn eine* PPP-*Lösung der Größe maximal d für G' existiert.*

Beweis. Falls alle Spalten in G paarweise verschieden sind, dann ist G gleich G' und die Aussage ist richtig. Falls G zwei gleiche Spalten enthält, seien s und s' zwei solche Spalten und G' entstehe aus G durch Löschen der Spalte s'.

Nur-wenn-Teil: Da G' eine Untermatrix von G darstellt, ist diese Beweisrichtung einfach zu zeigen. Aus einer PPP-Lösung (H, B_H) für G lässt sich durch Löschen der Spalte s' in H und B_H eine PPP-Lösung für G' erstellen. Die Anzahl der paarweise verschiedenen Haplotypen kann sich dabei nicht erhöhen.

Wenn-Teil: Wir nehmen nun an, dass für die Genotypmatrix G' eine PPP-Lösung $(H', B_{H'})$ der Größe d existiert. Eine PPP-Lösung (H, B_H) für G entsteht nun folgendermaßen: Die Haplotypmatrix H entsteht aus H', indem die Spalte s nach Spalte s' kopiert wird und der Baum B_H entsteht aus $B_{H'}$, indem die Kante, welche mit s markiert ist, zusätzlich mit s' markiert wird und jede Knotenmarkierung um Spalte s' erweitert wird. Es gilt nun, dass G durch H erklärt wird und dass B_H eine perfekte Pfadphylogenie für H ist, da sich Eigenschaft 4 aus Definition 2.1 von Spalte s auf Spalte s' überträgt. Außerdem erhöht sich die Anzahl der paarweise verschiedenen Haplotypen nicht, da zwei Haplotypen, die in Spalte s' verschieden sind, auch in Spalte s verschieden sind. □

Der Beweis zu Lemma 3.22 lässt sich analog für perfekte Phylogenien führen.

Mit Lemma 3.22 folgt, dass sich die Größe einer minimalen gerichteten PPP-Lösung nicht ändert, wenn von gleichen Spalten alle bis auf eine gelöscht werden. Im Folgenden wird daher angenommen, dass die Spalten einer Genotypmatrix G paarweise verschieden sind. Insgesamt ergibt sich eine Vorverarbeitung, die jede 0-Spalte und von gleichen Spalten alle bis auf eine löscht. Eine Genotypmatrix, die durch diese Vorverarbeitung entsteht, wird im Weiteren eine *vorverarbeitete* Genotypmatrix genannt.

Aufbau und Größe gerichteter perfekter Pfadphylogenien. Für Genotypmatrizen, die aus der vorgestellten Vorverarbeitung entstehen, wird nun der prinzipielle Aufbau und die Größe gerichteter PPP-Lösung untersucht. Lemma 3.23 trifft eine Aussage über den Aufbau gerichteter PPP-Lösungen und Lemma 3.24 trifft darauf aufbauend eine Aussage über die Größe gerichteter PPP-Lösungen.

Lemma 3.23. *Sei G eine vorverarbeitete $n \times m$ Genotypmatrix und (H, B_H) eine gerichtete* PPP-*Lösung für G. Dann gilt:*

1. *Jede Kante von B_H ist mit genau einer Spalte markiert und*

2. *B_H besteht aus m Kanten, deren Markierungen paarweise verschieden sind.*

Beweis. Wir nehmen nun an, dass in B_H eine Kante existiert, die mit zwei Spalten markiert ist und führen dies zu einem Widerspruch. Seien hierzu s und s' zwei Spalten, die eine gemeinsame Kante in B_H markieren. Da B_H eine gerichtete perfekte Phylogenie ist, gilt mit Punkt 4 von Definition 2.1, dass ein Haplotyp in B_H genau dann in einer Spalte den Eintrag 1 besitzt, wenn diese Spalte auf dem Weg von der Wurzel zu dem Haplotyp vorkommt. Da s und s' die selbe Kante markieren, gilt somit für jeden Haplotyp h die Gleichung $h[s] = h[s']$. Die Spalten s und s' sind daher in H identisch. Sei nun g ein Genotyp in G sowie seien h und h' die erklärenden Haplotypen zu g aus H. Mit obiger Eigenschaft gilt $h[s] = h[s']$ und $h'[s] = h'[s']$ und folglich $g[s] = g[s']$. Somit sind die Spalten s und s' in G identisch. Dies steht aber im Wiederspruch zu der Annahme, dass G vorverarbeitet ist. Es folgt, dass B_H keine Kante besitzt, die mit mehr als einer Spalte markiert ist. Nach Punkt 3 von Definition 2.1 ist zudem jede Kante in B_H markiert. Insgesamt folgt, dass jede Kante in B_H mit genau einer Spalte markiert ist.

Mit Punkt 2 von Definition 2.1 kommt jede Spalte aus G als Kantenmarkierung in jeder PPP Lösung für G vor. Dies sind m paarweise verschiedene Spalten. Mit dem oben bewiesenen Punkt 1 existiert nun eine $1:1$-Zuordnung zwischen Spalten und Kanten. Es folgt, dass B_H aus m Kanten besteht, deren Markierungen paarweise verschieden sind. □

Genotypen, die den Eintrag 1 enthalten, werden im Weiteren 1-*Genotypen* genannt. Beispielsweise ist 0122 ein 1-Genotyp. Es lässt sich direkt erkennen, dass kein 1-Genotyp den Haplotyp $0\ldots0$ zur Erklärung verwendet. Genotypen, die nur Einträge aus $\{0,2\}$ enthalten, werden im Weiteren $\{0,2\}$-*Genotypen* genannt. Zum Beispiel ist 0220 ein $\{0,2\}$-Genotyp.

Lemma 3.24. *Sei G eine vorverarbeitete $n \times m$ Genotypmatrix und (H, B_H) eine gerichtete* PPP-*Lösung der Größe d für G. Dann gilt $m \leq d \leq m+1$ und folgende Aussagen sind äquivalent:*

1. *H enthält genau m paarweise verschiedene Haplotypen.*
2. *Die Wurzel von B_H ist nicht mit einem Haplotyp aus H markiert.*
3. *Der Haplotyp $0\ldots0$ kommt nicht in H vor.*
4. *Jeder $\{0,2\}$-Genotyp in G umfasst Spalten aus beiden Zweigen von B_H.*

Beweis. Es sei G eine vorverarbeitete $n \times m$ Genotypmatrix und (H, B_H) eine gerichtete PPP-Lösung der Größe d für G. Wir zeigen nun eine obere und eine untere Schranke für die Größe von (H, B_H) und beginnen mit der oberen Schranke. Mit Lemma 2 gibt es in B_H genau m viele Kanten und damit genau $m+1$ viele Knoten. Somit enthält B_H maximal $m+1$ paarweise verschiedene Haplotypen, da Haplotypen nur verschieden sein können, wenn sie verschiedene Knoten markieren. Da jeder Haplotyp aus H auch in B_H vorkommt, ist die Menge der paarweise verschiedenen Haplotypen in H durch $m+1$ nach oben beschränkt. Nun beschränken wir die Anzahl der paarweise verschiedene Haplotypen in (H, B_H) nach unten. Hierzu

sei w die Wurzel von B_H und v ein beliebiger Knoten in B_H, der nicht die Wurzel ist. Wir zeigen nun, dass v mit einem Haplotyp aus H markiert ist und unterschieden dazu zwei Fälle zur Lage von v in B_H:

Für den ersten Fall sei der Knoten v mit genau einer Kante, die mit der Spalte s markiert ist, inzident. Wir nehmen nun an, dass v mit keinem Haplotyp aus H markiert ist. Dann gilt $h[s] = 0$ für jeden Haplotyp h aus H und somit ist s eine 0-Spalte in H und daher eine 0-Spalte in G. Dies steht im Widerspruch dazu, dass G eine vorverarbeitete Genotypmatrix ist und insbesondere keine 0-Spalte enthält. Es folgt, dass in einer gerichtete PPP-Lösung für G jedes Blatt, das nicht die Wurzel ist, mit einem Haplotyp aus H markiert ist. Abbildung 7a zeigt diesen Fall.

Für den zweiten Fall sei der Knoten v mit genau zwei Kanten inzident. Eine der Kanten sei mit Spalte s und die andere mit Spalte s' markiert. Wir nehmen nun an, dass v mit keinem Haplotyp aus H markiert ist. Sei h ein beliebiger Haplotyp in H. Die Spalte s liegt in B_H genau dann auf dem Weg von der Wurzel zu h, wenn dies für s' gilt und mit Punkt 4 von Definition 2.1 folgt somit $h[s] = h[s']$. Die Spalten s und s' sind daher in H und in G identisch. Dies steht im Widerspruch dazu, dass G vorverarbeitet ist. Es folgt, dass in einer gerichteten PPP-Lösung für G jeder innere Knoten, der nicht die Wurzel ist, mit einem Haplotyp aus H markiert ist. Abbildung 7b zeigt diesen Fall.

Auf m Knoten einer perfekten Pfadphylogenie treffen der erste oder der zweite Fall zu. Jeder dieser Knoten ist mit einem Haplotyp aus H markiert und somit umfasst H mindestens m paarweise verschiedene Haplotypen. Hieraus folgt insgesamt $m \leq d \leq m+1$.

Nun wird die Äquivalenz der Aussagen 1 bis 4 gezeigt: Hierzu sei G wieder eine vorverarbeitete $n \times m$ Genotypmatrix und (H, B_H) eine PPP-Lösung der Größe d für G.

(1 ⇒ 2) Mit der obigen Argumentation folgt, dass jeder der m Knoten, der in B_H nicht Wurzelknoten ist, mit einem Haplotyp aus H markiert ist. Falls nun $d = m$ gilt, so verteilen sich die m Haplotypen auf die Knoten, die nicht die Wurzel sind und die Wurzel ist nicht markiert.

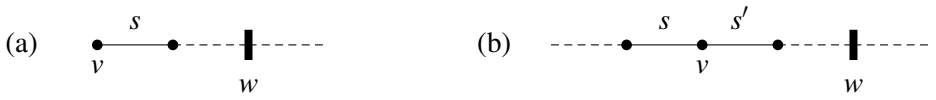
(2 ⇒ 3) Da B_H eine gerichtet ist, kann der Haplotyp $0 \ldots 0$ nur die Wurzel markieren. Falls die Wurzel in B_H nicht mit einem Haplotyp aus H markiert ist, kommt in H somit nicht der Haplotyp $0 \ldots 0$ vor.

(3 ⇒ 4) Für diesen Beweisteil nehmen wir an, dass in G ein $\{0,2\}$-Genotyp g existiert, der nur Spalten umfasst, die im Zweig $B_{H,l}$ von B_H liegen. Seien h und h' die Haplotypen in H, die g erklären. Es gilt dann, dass h und h' in $B_{H,l}$ liegen und wir nehmen o.B.d.A. an, dass h auf dem Weg von der Wurzel zu h' liegt. Da g ein $\{0,2\}$-Genotyp ist, existiert keine Spalte in der h und h' den Eintrag 1 enthalten. Somit folgt, dass h gleich dem Haplotyp $0 \ldots 0$ ist.

(4 ⇒ 1) Es gelte Aussage 4 und g sei ein beliebiger Genotyp in G. Wir zeigen in einem ersten Schritt, dass keiner der erklärenden Haplotypen zu g der Haplotyp $0 \ldots 0$ ist und unterscheiden dazu zwischen $\{0,2\}$-Genotypen und 1-Genotypen. Falls g ein $\{0,2\}$-Genotyp ist, dann umfasst g nach Voraussetzung Spalten in beiden Zweigen von B_H und die erklärenden Haplotypen zu g enthalten jeweils min-

destens einmal den Eintrag 1. Falls g ein 1-Genotyp ist, dann enthalten die erklärenden Haplotypen an mindestens einer gemeinsamen Stelle den Eintrag 1. Die Haplotypmatrix H, welche G erklärt, enthält somit nur Haplotypen, die den Eintrag 1 enthalten. Eine gerichtete perfekte Pfadphylogenie mit m Kanten enthält maximal m Haplotypen, die eine 1 enthalten und somit enthält H maximal m paarweise verschiedene Haplotypen, also gilt $d \leq m$. Mit obiger Argumentation gilt außerdem $m \leq d$ und folglich enthält H genau m paarweise verschiedenen Haplotypen.

Abbildung 7: Die Abbildung zeigt die beiden Fälle, die sich ergeben, wenn man die Lage von Knoten in einer perfekte Pfadphylogenie betrachtet. Der Knoten w ist die Wurzel und der Knoten v ist ein Knoten, der nicht die Wurzel ist. Die Spalten s und s' markieren Kanten. Der Knoten v ist entweder ein Blattknoten (a) oder ein innerer Knoten (b).



□

Im Beweis zu Lemma 3.24 wird bei der Fallunterscheidungen nicht der Wurzelknoten einer gerichteten perfekten Pfadphylogenie betrachtet. So ergibt sich, dass die Größe jeder gerichteten perfekten Pfadphylogenie für eine vorverarbeitete Genotypmatrix mit m Spalten nicht kleiner als m und größer als $m+1$ ist. Es lassen sich Genotypmatrizen konstruieren, für die eine gerichtete PPP-Lösung der Größe m und eine gerichtete PPP-Lösung der Größe $m+1$ existiert. Es lassen sich aber auch Genotypmatrizen konstruieren, bei denen die Größe einer minimalen gerichteten PPP-Lösung $m+1$ beträgt. Beispiele hierzu zeigt Abbildung 8.

Algorithmus für minimale gerichtete perfekte Pfadphylogenien. Nun wird untersucht, unter welchen Bedingungen eine vorverarbeitete Genotypmatrix eine gerichtete PPP-Lösung der Größe m besitzt und daraus eine logarithmisch platzbeschränkte Turingmaschine für MPPPH abgeleitet. Hierzu werden Spalten nicht durch Indizes referenziert sondern, wie im Abschnitt 3.3.4, durch Spaltenvektoren dargestellt. Wenn wir also im Folgenden von einer Spalte s aus einer Genotypmatrix G sprechen, dann kann man die einzelnen Einträge der Spalte durch $s[i]$ erhalten. Außerdem werden auch die Ordnung auf Spaltenvektoren $\succeq$ und Begriffe aus der Ordnungtheorie, die in Abschnitt 3.3.4 eingeführt wurden, weiter verwendet. Analog zur entsprechenden Definition in Abschnitt 2.3 umfasst die Zeile i in einer Genotyp- oder Haplotypmatrix die Spalte s, falls $s[i] \neq 0$ gilt. Anstatt von Zeilen sprechen wir bei Genotypmatrizen auch von Genotypen und bei Haplotypmatrizen auch von Haplotypen. Sei G eine Genotypmatrix, die eine gerichtete perfekte Pfadphylogenie zulässt und S die Menge der Spaltenvektoren in G. Aus Abschnitt 3.3.4 ist bekannt, dass dann $(S, \succeq)$ maximal die Weite 2 besitzt und für jedes $k \geq 3$ die Menge $\mathrm{hma}_k(S)$ leer ist. Für die Mengen $\mathrm{hma}_1(S)$ und $\mathrm{hma}_2(S)$ ergeben sich

Abbildung 8: Die Abbildung zeigt zwei vorverarbeitete Genotypmatrizen mit gerichteten PPP-Lösungen. Für die Genotypmatrix G ist (H_1, B_{H_1}) eine gerichtete PPP-Lösung der Größe 3 und (H_2, B_{H_2}) eine gerichtete PPP-Lösung der Größe 2. Eine minimale Lösung für G hat somit die Größe 2. Für die Genotypmatrix G' ist $(H', B_{H'})$ eine gerichtete PPP-Lösung der Größe 3. Für G' existiert keine gerichtete PPP-Lösung der Größe 2 und somit hat eine minimale Lösung für G die Größe 3. In den PPP-Lösungen der Größe 3 sind alle Knoten markiert und für die PPP-Lösung der Größe 2 gelten die Aussagen 1 bis 4 von Lemma 3.24. Die Wurzeln der perfekten Phylogenien werden in der Abbildung durch senkrechte Balken angezeigt.

$$G = \begin{matrix} & \textit{1} & \textit{2} \\ & 0 & 1 \\ & 2 & 2 \end{matrix} \qquad G' = \begin{matrix} & \textit{1} & \textit{2} \\ & 0 & 2 \\ & 2 & 0 \end{matrix}$$

$$H_1 = \begin{matrix} & \textit{1} & \textit{2} \\ & 0 & 1 \\ & 0 & 0 \\ & 1 & 1 \end{matrix} \qquad B_{H_1}: \quad \begin{matrix} & & \textit{1} & & \textit{2} & \\ 11 & & 01 & & & 00 \end{matrix}$$

$$H' = \begin{matrix} & \textit{1} & \textit{2} \\ & 0 & 1 \\ & 0 & 0 \\ & 1 & 0 \end{matrix} \qquad B_{H'}: \quad \begin{matrix} & \textit{1} & & \textit{2} & \\ 10 & & 00 & & 01 \end{matrix}$$

$$H_2 = \begin{matrix} & \textit{1} & \textit{2} \\ & 0 & 1 \\ & 1 & 0 \end{matrix} \qquad B_{H_2}: \quad \begin{matrix} & \textit{1} & & \textit{2} & \\ 10 & & & & 01 \end{matrix}$$

folgendermaßen drei Fälle: Entweder genau eine der beiden Mengen ist leer oder beide Menge sind nicht leer. Der Aufbau von $(S, \succeq)$ ist für diese Fälle schematisch in Abbildung 9 dargestellt. In den folgenden drei Lemmata wird für jeden Fall eine Aussage angegeben, die genau dann wahr ist, wenn für G eine PPP-Lösung der Größe m existiert.

Lemma 3.25. *Sei G eine vorverarbeitete $n \times m$ Genotypmatrix, die eine gerichtete perfekte Pfadphylogenie zulässt und $(S, \succeq)$ die partiell geordnete Menge der Spalten in G, für die $\text{hma}_1(S) = \{s^*\}$ und $\text{hma}_2(S) = \emptyset$ gilt. Es existiert genau dann eine* PPP*-Lösung der Größe m für G, wenn eine der folgenden Bedingungen erfüllt ist:*

1. *Die Genotypmatrix G enthält keinen $\{0,2\}$-Genotyp.*
2. *Es existiert eine Spalte, die jeder $\{0,2\}$-Genotyp umfasst und kein 1-Genotyp umfasst.*

Beweis. Wie schon im Beweis zu Satz 3.18 festgestellt, ist $(S, \succeq)$ in diesem Fall eine Kette. Abbildung 9b zeigt diesen Fall.

Nur-wenn-Teil: Nun nehmen wir an, dass für G eine gerichtete PPP-Lösung (H, B_H) der Größe m existiert. Falls ein $\{0,2\}$-Genotyp g in G vorkommt, dann umfasst g mit Punkt 4 von Lemma 3.24 Spalten in beiden Zweigen von B_H und somit ist kein Zweig leer. Wie nehmen nun o.B.d.A. an, dass die Spalte s^* in $B_{H,l}$ liegt und sehen, dass sie die oberste Spalte in $B_{H,l}$ ist. Sei nun g ein beliebig gewählter 1-Genotyp und s eine Spalte, in der g den Eintrag 1 enthält. Da $\text{hma}_1(S) = \{s^*\}$ und somit $s^* \succeq s$ gilt, hat g auch in Spalte s^* den Eintrag 1. Also hat jeder 1-Genotyp in Spalte s^* den Eintrag 1. Zusammen folgt, dass jeder Haplotyp, der für einen 1-Genotyp zur Erklärung verwendet wird, in Zweig $B_{H,l}$ liegt und daher auch jede Spalte, die von einem 1-Genotyp umfasst wird, in Zweig $B_{H,l}$ liegt. Sei nun s' das oberste Element des nichtleeren Zweiges $B_{H,r}$. Jeder 1-Genotyp umfasst nicht die Spalte s' und jeder $\{0,2\}$-Genotyp umfasst die Spalte s', der er Spalten aus beiden Zweigen umfasst.

Wenn-Teil: Für diese Beweisrichtung unterscheiden wir, ob eine Genotypmatrix nur 1-Genotypen oder auch $\{0,2\}$-Genotypen enthält. Falls einerseits G keinen $\{0,2\}$-Genotyp enthält, dann gilt mit Lemma 3.24, dass eine gerichtete PPP-Lösung der Größe m für G existiert. Falls andererseits G einen $\{0,2\}$-Genotyp enthält, dann sei s' die größte Spalte in $(S, \succeq)$, in der jeder 1-Genotyp den Eintrag 0 und jeder $\{0,2\}$-Genotyp den Eintrag 2 enthält. Da $\text{hma}_1(S) = \{s^*\}$ und somit $s^* \succeq s'$ gilt, enthält jeder $\{0,2\}$-Genotyp in der Spalte s^* den Eintrag 2. Nun betrachten wir eine beliebige PPP-Lösung (H, B_H) für G. Falls einerseits s' und s^* in verschiedenen Zweigen von B_H liegen, dann umfasst jeder $\{0,2\}$-Genotyp Spalten in beiden Zeigen und (H, B_H) hat die Größe m. Falls andererseits s' und s^* im Zweig $B_{H,l}$ liegen, lässt sich eine PPP-Lösung $(H', B_{H'})$ der Größe m aus (H, B_H) folgendermaßen konstruieren: Die Spalte s' wird zur obersten Spalte von $B_{H,r}$ gesetzt und für jeden $\{0,2\}$-Genotyp werden die beiden erklärenden Haplotypen an der Stelle s' invertiert. Dadurch wird jeder $\{0,2\}$-Genotyp weiterhin von zwei Haplotypen erklärt aber der Haplotyp $0 \ldots 0$ verschwindet. Die erklärenden Haplotypen

für einen 1-Genotyp verändern sich nicht, da sie im Zweig $B_{H,l}$ und oberhalb von s' liegen. Die so konstruierte gerichtete PPP-Lösung $(H', B_{H'})$ enthält nicht den Haplotyp $0 \ldots 0$ in H und hat daher mit Lemma 3.24 die Größe m. □

Lemma 3.26. *Sei G eine vorverarbeitete $n \times m$ Genotypmatrix, die eine gerichtete perfekte Pfadphylogenie zulässt und $(S, \succeq)$ die partiell geordnete Menge der Spalten in G, für die $\mathrm{hma}_1(S) = \emptyset$ und $\mathrm{hma}_2(S) = \{s_1, s_2\}$ gilt. Es existiert genau dann eine* PPP*-Lösung der Größe m für G, wenn jeder $\{0,2\}$-Genotyp die Spalten s_1 und s_2 umfasst.*

Beweis. Nach Voraussetzung sind die Spalten s_1 und s_2 nicht vergleichbar. Folglich liegen sie in verschiedenen Zweigen jeder PPP-Lösung (H, B_H) für G. Sei s_1 in $B_{H,l}$ und s_2 in $B_{H,r}$ (der umgekehrte Fall ist symmetrisch). Da $\mathrm{hma}_2(S) = \{s_1, s_2\}$ wissen wir, dass s_1 die oberste Spalte von $B_{H,l}$ ist und ebenso, dass s_2 die oberste Spalte von $B_{H,r}$ ist. Abbildung 9b zeigt diesen Fall der partiell geordneten Menge $(S, \succeq)$.

Mit Lemma 3.24 gelten folgende zwei Aussagen: 1) Falls jeder $\{0,2\}$-Genotyp s_1 und s_2 umfasst, hat eine PPP-Lösung für G die Größe m. 2) Falls eine PPP-Lösung (H, B_H) für G die Größe m besitzt, umfasst jeder $\{0,2\}$-Genotyp Spalten in beiden Zweigen und damit die Spalten s_1 und s_2. □

Lemma 3.27. *Sei G eine vorverarbeitete $n \times m$ Genotypmatrix, die eine gerichtete perfekte Pfadphylogenie zulässt und $(S, \succeq)$ die partiell geordnete Menge der Spalten in G, für die $\mathrm{hma}_1(S) = \{s^*\}$ und $\mathrm{hma}_2(S) = \{s_1, s_2\}$ gilt. Es existiert eine* PPP*-Lösung der Größe m für G, wenn eine der folgenden Fälle eintritt:*

1. *Ein 1-Genotyp umfasst die Spalte s_1 und jeder $\{0,2\}$-Genotyp umfasst die Spalte s_2.*
2. *Ein 1-Genotyp umfasst die Spalte s_2 und jeder $\{0,2\}$-Genotyp umfasst die Spalte s_1.*
3. *Es existiert eine Spalte s' mit $s' \succeq s_1$, $s' \succeq s_2$ und $s' \neq s^*$, so dass jeder $\{0,2\}$-Genotyp die Spalte s' umfasst und kein 1-Genotyp die Spalte s' umfasst.*

Beweis. Nach Voraussetzung gilt $\mathrm{hma}_1(S) = \{s^*\}$ und $\mathrm{hma}_2(S) = \{s_1, s_2\}$. Abbildung 9c zeigt diesen Fall der partiell geordneten Menge $(S, \succeq)$, die die Weite 2 besitzt.

Nur-wenn-Teil: Wie nehmen an, dass G eine gerichtete PPP-Lösung (H, B_H) besitzt und zeigen, dass dann eine der Aussagen 1 bis 3 gilt. Nach Voraussetzung liegen die Spalten s_1 und s_2 in verschiedenen Zweigen von B_H und somit besteht B_H aus zwei nichtleeren Zweigen, die wir mit $B_{H,l}$ und $B_{H,r}$ bezeichnen. Wir nehmen an, dass s^* und s_1 in $B_{H,l}$ liegen und dass s_2 in $B_{H,r}$ liegt (für s^* und s_2 in einem gemeinsamen Zweig lässt sich die nun folgende Argumentation analog führen). Wie im Beweis zu Lemma 3.25 lässt sich erkennen, dass s^* die oberste Spalte in $B_{H,l}$ ist, jeder Genotyp die Spalte s^* umfasst und jede Spalte, die von einem 1-Genotyp umfasst wird, im Zweig $B_{H,l}$ liegt. Im Weiteren bezeichne S_1 die Menge

der Spalten, die von einem 1-Genotyp umfasst werden. Wir unterscheiden nun, ob die Spalte s_1 in S_1 liegt oder nicht in S_1 liegt.

Falls s_1 in S_1 liegt, dann liegt jede Spalte, die größer als s_1 ist, ebenfalls in S_1. Wegen $\text{hma}_2(S) = \{s_1, s_2\}$ ist folglich s_2 die oberste Spalte im Zweig $B_{H,r}$. Da (H, B_H) die Größe m besitzt, umfasst mit Lemma 3.24 jeder $\{0,2\}$-Genotyp Spalten in beiden Zweigen und insbesondere s_2.

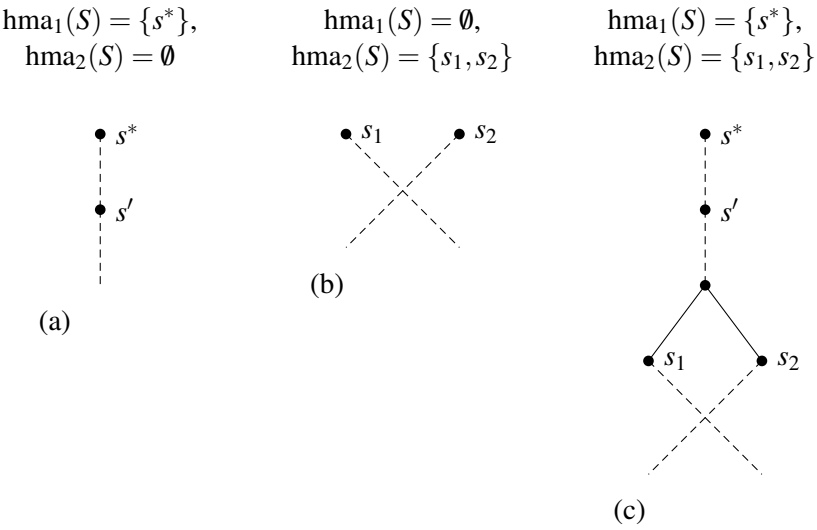
Falls andererseits s_1 nicht in S_1 liegt, nehmen wir an, dass s_2 das oberstes Element von $B_{H,r}$ ist. Wieder umfasst jeder $\{0,2\}$-Genotyp die Spalte s_2, da (H, B_H) die Größe m besitzt. Da jede Spalte aus S_1 im Zweig $B_{H,l}$ liegt, ist s_2 nicht in S_1 enthalten. Es umfasst also kein 1-Genotyp die Spalten s_1 oder s_2. Da s_1 und s_2 nach Voraussetzung nicht vergleichbar sind, muss G somit in den Spalten s_1 und s_2 die Untermatrix $\left[\begin{smallmatrix} 0 & 2 \\ 2 & 0 \end{smallmatrix}\right]$ enthalten. Dies steht im Widerspruch dazu, dass jeder $\{0,2\}$-Genotyp die Spalte s_2 umfasst. Folglich kann die Spalte s_2 in diesem Fall nicht oberstes Element von $B_{H,r}$ sein. Es existiert daher eine Spalte s' mit $s' \succeq s_2$, $s' \succeq s_1$ und $s' \neq s^*$, die oberstes Element von $B_{H,r}$ ist. Diese Spalte wird von jedem $\{0,2\}$-Genotyp umfasst und von keinem 1-Genotyp umfasst.

Wenn-Teil: Wir nehmen nun nacheinander an, dass eine der Aussagen 1 bis 3 gilt und zeigen jeweils, dass dann eine gerichtete PPP-Lösung der Größe m für G existiert.

Für die erste Aussage nehmen wir an, dass mindestens ein 1-Genotyp die Spalte s_1 umfasst und jeder $\{0,2\}$-Genotyp die Spalte s_2 umfasst. Sei nun (H, B_H) eine beliebige gerichtete PPP-Lösung für G und Spalte s^* liege im Zweig $B_{H,l}$. Nach Voraussetzung gilt $\text{hma}_1(S) = \{s^*\}$ und somit umfasst jeder Genotyp die Spalte s^*. Hieraus folgt, dass jede Spalte, die von einem 1-Genotyp umfasst wird, im Zweig $B_{H,l}$ liegt. Insbesondere gilt dies also auch für s_1. Da s_1 in $B_{H,l}$ liegt und s_1 und s_2 nicht vergleichbar sind, liegt s_2 folglich in $B_{H,r}$. Zusammen gilt, dass s^* und s_2 in verschiedenen Zweigen von B_H liegen und von jedem $\{0,2\}$-Genotyp umfasst werden. Mit Lemma 3.24 hat (H, B_H) daher Größe m. Für den Fall, dass mindestens ein 1-Genotyp die Spalte s_2 umfasst und jeder $\{0,2\}$-Genotyp die Spalte s_1 umfasst, lässt sich der Beweis analog führen.

Für die dritte Aussage nehmen wir an, dass eine Spalte s' mit $s' \succeq s_1$, $s' \succeq s_2$ und $s' \neq s^*$ existiert, die jeder $\{0,2\}$-Genotyp umfasst und kein 1-Genotyp umfasst. Dieser Beweisteil ist ähnlich zum Beweis von Lemma 3.25. Zuerst stellen wir, analog zu den schon geführten Beweisen fest, dass jeder $\{0,2\}$-Genotyp die Spalte s^* umfasst. Jetzt betrachten wir eine beliebige gerichtete PPP-Lösung (H, B_H) und unterscheiden zwei Fälle. Falls einerseits s^* und s' in verschiedenen Zweigen liegen, dann umfasst jeder $\{0,2\}$-Genotyp Spalten in beiden Zweigen und (H, B_H) hat die Größe m. Falls andererseits s^* und s' im gleichen Zweig $B_{H,l}$ liegen, dann lässt sich eine gerichtete PPP-Lösung $(H', B_{H'})$ der Größe m für G wie im Beweis zu Lemma 3.25 konstruieren: Die Spalte s' wird zur obersten Spalte von $B_{H,r}$ gesetzt und für jeden $\{0,2\}$-Genotypen werden die erklärenden Haplotypen in Spalte s' invertiert, wobei der $0 \ldots 0$ verschwindet. Für jeden 1-Genotyp bleiben die erklärenden Haplotypen erhalten. Die so konstruierte gerichtete PPP-Lösung enthält nicht den Haplotyp $0 \ldots 0$ und hat damit nach Lemma 3.24 die Größe m. □

Abbildung 9: Falls eine partielle geordnete Menge $(S, \succeq)$ die Weite 2 besitzt, dann ist für jedes $k \geq 3$ die Menge $\text{hma}_k(S)$ leer. Betrachtet man die Mengen $\text{hma}_1(S)$ und $\text{hma}_2(S)$, so ist entweder genau eine der Mengen leer oder beide sind nicht leer. Hieraus ergeben sich drei Möglichkeiten zum Aufbau von $(S, \succeq)$, die in dieser Abbildung schematisch dargestellt werden. Eine durchgezogene Linie zwischen einem Element t und einem darüberliegenden Element t' bedeutet, dass $t' \succeq t$ gilt und in der Ordnung kein drittes Element existiert, welches größer als t und kleiner als t' ist. Eine gestrichelte Linie bedeutet, dass sich eine Kette beliebiger Länge zwischen den Elementen befinden kann. Diese Kette kann auch leer sein, dann sind die angrenzenden Elemente identisch. Gekreuzte gestrichelte Linien zeigen eine beliebige Ordnung der Weite 2 an, die an zwei oberste Elemente angefügt ist. In den Lemmata 3.25 und 3.27 werden die Eigenschaften einer Spalte s' beschrieben, die kein 1-Genotyp aber jeder $\{0,2\}$-Genotyp umfasst. Falls eine solche Spalte s' existiert, dann liegt sie wie abgebildet in der partiellen Ordnung.



Satz 3.28. ger-MPPPH $\in$ L.

Beweis. Es wird nun gezeigt, dass das Problem ger-MPPPH, dessen Eingabe aus einer Genotypmatrix G und einem Budgetwert d besteht und das nach der Existenz einer gerichteten PPP-Lösung mit maximaler Größe d für G fragt, durch eine logarithmisch platzbeschränkte deterministische Turingmaschine gelöst werden kann. Hierzu wird im Folgende ein Algorithmus beschrieben, der das Problem ger-MPPPH löst und sich mit einer Turingmaschine implementieren lässt.

Der Algorithmus erhält als Eingabe eine $n \times m$ Genotypmatrix G und einen Budgetwert d. Die Entscheidung, ob G eine gerichtete PPP-Lösung der Größe d besitzt, teilt sich folgendermaßen in mehrere Schritte auf:

Zuerst wird getestet, ob G eine gerichtete PPP-Lösung besitzt. In Satz 3.18 wurde gezeigt, dass sich das Problem ger-PPPH durch eine Formel in Prädikatenlogik erster Stufe beschreiben lässt. Jedes Problem, das sich durch eine Formel in

Prädikatenlogik erster Stufe beschreiben lässt, lässt sich auch mit einer deterministischen logarithmisch platzbeschränkten Turingmaschine entscheiden [25, Seite 45f]. Falls G keine PPP-Lösung besitzt, wird an dieser Stelle im Algorithmus mit „nein“ geantwortet und beendet. Andernfalls wird folgendermaßen fortgefahren:

Es sei G' die $n' \times m'$ Genotypmatrix, die aus G durch Vorverarbeitung entsteht. Mit der Vorverarbeitung folgt, dass die Größe einer kleinsten PPP-Lösung für G und G' gleich ist. Die Frage, ob G eine PPP-Lösung der Größe maximal d besitzt, ist also gleichbeutend mit der Frage, ob G' eine PPP-Lösung der Größe maximal d besitzt. Es reicht also aus die Frage für G' korrekt zu beantworten. Wir wissen an dieser Stelle im Algorithmus, dass die Genotypmatrix G' eine gerichtete PPP-Lösung besitzt, da G eine gerichtete PPP-Lösung besitzt. Lemma 3.24 sagt aus, dass jede PPP-Lösung für G' maximal die Größe $m'+1$ besitzt. Falls $d > m'$ gilt, wird daher an dieser Stelle „ja“ ausgegeben und beendet. Falls $d < m'$ gilt, besitzt G' mit Lemma 3.24 keine gerichtete PPP-Lösung der Größe d und es wird an dieser Stelle „nein“ ausgegeben und beendet. Der Wert m', die Anzahl der Spalten in der Genotypmatrix G', lässt sich folgendermaßen mit logarithmischem Platzaufwand bestimmen: Es wird Schritt für Schritt jede Spalte der Genotypmatrix G betrachtet. Falls eine Spalte keine 0-Spalte ist und Spalten, die mit ihr identisch sind, nur einen kleineren Index besitzen, wird ein Zähler um den Wert 1 erhöht. Auf diese Weise wird von gleichen Spalte jeweils eine gezählt und die 0-Spalten werden ausgespart. Nachdem alle Spalten betrachtet wurden, entspricht der Zähler der Anzahl von Spalten in G'. Der Vergleich von d und m' ist ebenfalls mit logarithmischem Platzaufwand möglich und somit benötigt der gesamte Schritt nur logarithmischen Platz.

Für den Fall, dass $d = m$ gilt, besitzt G' genau dann eine gerichtete PPP-Lösung der Größe d, wenn eines der Lemmata 3.25, 3.26 oder 3.27 zutrifft. Das heißt, es muss jeweils neben der Voraussetzung die entsprechende Aussage gelten. Wie wissen, dass G' eine gerichtete perfekte Pfadphylogenie zulässt. Die weiteren Voraussetzungen und die Aussagen für die Lemmata 3.25, 3.26 und 3.27 werden nun Schrittweise durch Formeln beschrieben. Ziel ist es, dadurch zu zeigen, dass sich die Voraussetzungen und Aussagen durch eine deterministische Turinmaschine mit logarithmischem Platz testen lassen.

Um nur die Spalten zu betrachten, die bei der Vorverarbeitung von G nach G' nicht gelöscht werden, beschreibt folgende Formel eine Spalte, die in der Vorverarbeitung nicht gelöscht wird:

$$\begin{aligned}\phi_{\text{Vor}}(s) \equiv\ & (\exists z.Z(z))[G_1(z,s) \vee G_2(z,s)] \wedge \\ & (\neg\exists s'.S(s'))[s' < s \wedge (\forall z.Z(z))[G_0(z,s') \wedge G_0(z,s) \vee \\ & \qquad G_1(z,s') \wedge G_1(z,s) \vee \\ & \qquad G_2(z,s') \wedge G_2(z,s)]].\end{aligned}$$

Die Voraussetzungen der Lemmata 3.25, 3.26 und 3.27 betreffen unter anderem die höchsten maximalen Antiketten der Größe 1 und 2. Um diese Voraussetzungen

zu testen, beschreiben folgende Formeln, dass eine Spalte in der höchsten maximalen Antikette der Größe 1 liegt bzw. dass zwei Spalten in der höchsten maximalen Antikette der Größe 2 liegen:

$$\begin{aligned}
\phi_{\text{hma}_1}(s^*) &\equiv (\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\phi_{\preceq}(s^*, s)], \\
\phi_{\text{hma}_2}(s_1, s_2) &\equiv \neg(\phi_{\preceq}(s_1, s_2) \vee \phi_{\preceq}(s_2, s_1)) \wedge \\
&\quad (\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\\
&\quad \phi_{\preceq}(s_1, s) \vee \phi_{\preceq}(s, s_1) \vee \phi_{\preceq}(s_2, s) \vee \phi_{\preceq}(s, s_2)] \wedge \\
&\quad (\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\\
&\quad ((\phi_{\preceq}(s, s_1) \wedge s \neq s_1) \vee (\phi_{\preceq}(s, s_2) \wedge s \neq s_2)) \rightarrow \\
&\quad (\forall t.S(t) \wedge \phi_{\text{Vor}}(t))[\phi_{\preceq}(s, t) \vee \phi_{\preceq}(t, s)]].
\end{aligned}$$

Nachfolgende Formel wird als Teilformel verwendet und beschreibt eine Spalte, die in G' jeder $\{0,2\}$-Genotyp umfasst und jeder 1-Genotyp nicht umfasst:

$$\begin{aligned}
\phi_{\text{jederNZ}}(s') \equiv (\forall z.Z(z))[&((\exists s.S(s) \wedge \phi_{\text{Vor}}(s))[G_1(z, s)] \rightarrow G_0(z, s')) \wedge \\
&((\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\neg G_1(z, s)] \rightarrow G_2(z, s'))].
\end{aligned}$$

An dieser Stelle im Algorithmus wissen wir, dass G' eine gerichtete perfekte Pfadphylogenie zulässt. Die folgenden drei Formeln beschreiben die restlichen Teile der Vorraussetzungen von Lemmata 3.25, 3.26 und 3.27:

$$\begin{aligned}
\phi_{\text{VorrL}_1} &\equiv (\exists s^*.S(s^*) \wedge \phi_{\text{Vor}}(s^*))[\phi_{\text{hma}_1}(s^*)] \wedge \\
&\quad (\forall s_1.S(s_1) \wedge \phi_{\text{Vor}}(s_1), s_2.S(s_2) \wedge \phi_{\text{Vor}}(s_2))[\neg\phi_{\text{hma}_2}(s_1, s_2)], \\
\phi_{\text{VorrL}_2} &\equiv (\forall s^*.S(s^*) \wedge \phi_{\text{Vor}}(s^*))[\neg\phi_{\text{hma}_1}(s^*)] \wedge \\
&\quad (\exists s_1.S(s_1) \wedge \phi_{\text{Vor}}(s_1), s_2.S(s_2) \wedge \phi_{\text{Vor}}(s_2))[\phi_{\text{hma}_2}(s_1, s_2)], \\
\phi_{\text{VorrL}_3} &\equiv (\exists s^*.S(s^*) \wedge \phi_{\text{Vor}}(s^*))[\phi_{\text{hma}_1}(s^*)] \wedge \\
&\quad (\exists s_1.S(s_1) \wedge \phi_{\text{Vor}}(s_1), s_2.S(s_2) \wedge \phi_{\text{Vor}}(s_2))[\phi_{\text{hma}_2}(s_1, s_2)].
\end{aligned}$$

Nachfolgende Formeln beschreiben die Aussagen von Lemmata 3.25, 3.26 und

3.27:

$$
\begin{aligned}
\phi_{\text{Aussage 1}} \equiv\ & (\exists s^*.S(s^*) \wedge \phi_{\text{Vor}}(s^*))[\phi_{\text{hma}_1}(s^*) \wedge \\
& ((\forall z.Z(z))(\exists s.S(s) \wedge \phi_{\text{Vor}}(s))[G_1(z,s)] \vee \\
& (\exists s.S(s) \wedge \phi_{\text{Vor}}(s))[s \neq s^* \wedge \phi_{\text{jederNZ}}(s)])], \\
\phi_{\text{Aussage 2}} \equiv\ & (\exists s_1.S(s_1) \wedge \phi_{\text{Vor}}(s_1), s_2.S(s_2) \wedge \phi_{\text{Vor}}(s_2))[\phi_{\text{hma}_2}(s_1,s_2) \wedge \\
& (\forall z.Z(z))[\\
& (\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\neg G_1(z,s)] \rightarrow (G_2(z,s_1) \wedge G_2(z,s_2))]], \\
\phi_{\text{Aussage 3}} \equiv\ & (\exists s^*.S(s^*) \wedge \phi_{\text{Vor}}(s^*), s_1.S(s_1) \wedge \phi_{\text{Vor}}(s_1), s_2.S(s_2) \wedge \phi_{\text{Vor}}(s_2))[\\
& \phi_{\text{hma}_1}(s^*) \wedge \phi_{\text{hma}_2}(s_1,s_2) \wedge \\
& (((\exists z.Z(z), s.S(s) \wedge \phi_{\text{Vor}}(s))[G_1(z,s) \wedge (G_1(z,s_1) \vee G_2(z,s_1))] \wedge \\
& (\forall z.Z(z))[(\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\neg G_1(z,s)] \rightarrow G_2(z,s_2)]) \vee \\
& ((\exists z.Z(z), s.S(s) \wedge \phi_{\text{Vor}}(s))[G_1(z,s) \wedge (G_1(z,s_2) \vee G_2(z,s_2))] \wedge \\
& (\forall z.Z(z))[(\forall s.S(s) \wedge \phi_{\text{Vor}}(s))[\neg G_1(z,s)] \rightarrow G_2(z,s_1)]) \vee \\
& (\exists s'.S(s') \wedge \phi_{\text{Vor}}(s'))[\\
& \phi_{\succeq}(s',s_1) \wedge \phi_{\succeq}(s',s_2) \wedge s' \neq s^* \wedge \phi_{\text{jederNZ}}(s')])].
\end{aligned}
$$

Da sich die Voraussetzungen und die Aussagen von Lemmata 3.25, 3.26 und 3.27 als Formeln in Prädikatenlogik erster Stufe beschreiben lassen, können sie auch mit logarithmischem Platzaufwand von einer deterministischen Turingmaschine geprüft werden. Falls nun für ein Lemma die Voraussetzung und die Aussage erfüllt ist, so besitzt G' eine gerichtete PPP-Lösung der Größe $d = m$ und es wird „ja“ ausgegeben und beendet. Andernfalls wird „nein“ ausgegeben und beendet. Das beschriebene Verfahren lässt sich insgesamt durch eine deterministische Turingmaschine berechnen, deren Platzbedarf logarithmisch in der Eingabe ist. □

Neben dem Auswerten der verschiedenen Formeln berechnet der Algorithmus im Beweis zu Satz 3.28 nur die Anzahl der Spalten in der Genotypmatrix G', die durch Vorverarbeitung aus der Eingabe entsteht. Die Anzahl der Spalten in G' kann man auch durch einen Schaltkreis logarithmischer Tiefe berechnen und die Formeln von einem Schaltkreis logarithmischer Tiefe auswerten lassen. Durch Kombination dieser Schaltkreise erhält man einen Schaltkreis logarithmischer Tiefe für ger-MPPPH und es lässt sich erkennen, dass ger-MPPPH auch in der Schaltkreiskomplexitätsklasse NC^1 liegt.

4 Zusammenfassung und Ausblick

In dieser Arbeit wurde die Komplexität von Haplotypisierungsverfahren, die auf perfekten Phylogenien und kleinsten Haplotypmengen basieren, untersucht. Indem neben neuen Resultaten auch Resultate aus der Literatur vorgestellt wurden, ergab sich ein Überblick über die bisherige Forschung in diesem Bereich. Wie die Probleme im einzelnen bearbeitet wurden, wird nun zusammenfassend dargestellt.

4.1 Ergebnisse der Arbeit

Für das Problem MH und $\{k,l\}$-beschränkte Varianten von MH wurden in Abschnitt 3.2 Resultate aus der Literatur zusammengetragen, die aussagen, dass schon $\text{MH}(3,3)$ NP-vollständig ist, aber die Probleme $\text{MH}(2,\infty)$ und $\text{MH}(\infty,1)$ in P liegen.

In Abschnitt 3.3 wurden erste Beweise dazu gegeben, dass PPH in Mod_2L liegt und L-hart ist. Diese Beweise basieren auf mündlich überlieferten Beweisskizzen von Arfst Nickelsen und Till Tantau. Mit der Aussage $\text{PPH} \in \text{Mod}_2\text{L}$ folgte, dass PPH in der Klassen NC^2 liegt. Aus der L-Härte von PPH folgte zum Beispiel, dass sich PPH nicht durch eine Formel in Prädikatenlogik erster Stufe beschreiben lässt. Da für die Probleme $\text{PPH}(2,\infty)$ und $\text{PPH}(\infty,1)$ gezeigt wurde, dass sie in FO liegen, konnten wir feststellen, dass die Haplotypisierung mittels perfekten Phylogenien einfach ist, wenn die Eingabe nur sehr wenige heterozytische Einträge besitzt und schwerer wird, wenn die Anzahl der heterozytischen Einträge in der Eingabe zunimmt. Durch Resultate aus der Literatur hat sich gezeigt, dass man ein solches Verhalten nicht nur bei der Komplexität (k,l)-beschränkter Varianten von PPH sondern auch bei der Komplexität (k,l)-beschränkter Varianten MH und MPPH beobachten kann.

In Abschnitt 3.3 wurde neben der Haplotypisierung mittels perfekten Phylogenien auch die Haplotypisierung mittels perfekten Pfadphylogenien betrachtet. Aufbauend auf der Arbeit von Gramm et al. [17] wurde in diesem Abschnitt gezeigt, dass ger-PPPH und PPP in FO liegen und festgestellt, dass im gerichteten Fall die Komplexität der Haplotypisierung mittels perfekten Pfadphylogenien (ger-PPPH in FO) echt kleiner als die Komplexität der Haplotypisierung mittels perfekten Phylogenien (ger-PPH ist L-hart) ist.

Zu MPPH wurden in Abschnitt 3.4 Resultate aus der Literatur aufgezeigt, die aussagen, dass schon $\text{MPPH}(3,3)$ NP-vollständig ist, aber $\text{MPPH}(2,\infty)$ und $\text{MPPH}(\infty,1)$ in P liegen. An dieser Stelle zeigte sich, dass für die $\{k,l\}$-beschränkten Varianten von MH und MPPH die gleichen komplexitätstheoretischen Resultate bekannt sind.

Abschließend wurde in Abschnitt 3.4 das Hauptresultat dieser Arbeit, dass ger-MPPPH in L liegt, bewiesen. Der Beweis hierzu baute auf einer Idee von Gramm et al. [17] auf, eine partielle Ordnung über den Spalten einer Genotypmatrix so mit zwei Ketten zu überdecken, dass sich eine PPP-Lösung ergibt. Dieser Ansatz wurde in Abschnitt 3.4 um die Idee, die Genotypmatrizen bezüglich

der Struktur der zugehörigen partiellen Ordnung zu unterscheiden, erweitert. Auf diese Weise konnten genaue Aussagen über den Aufbau und die Größe perfekter Pfadphylogenien getroffen werden.

Abbildung 4.1 gibt einen Überblick über die Komplexität der vorgestellten Haplotypisierungsprobleme. In der Abbildung sind sowohl die neuen Resultate dieser Arbeit als auch die bekannten Resultate aus der Literatur dargestellt.

4.2 Ausblick

Durch die Ergebnisse dieser Arbeit kann die eingangs gestellte Frage nach der genauen Komplexität der Haplotypisierungsprobleme weiter eingeschränkt, aber noch nicht abschließend beantwortet werden. Es lassen sich neue Teilfragen zur Komplexität der Haplotypisierungprobleme stellen.

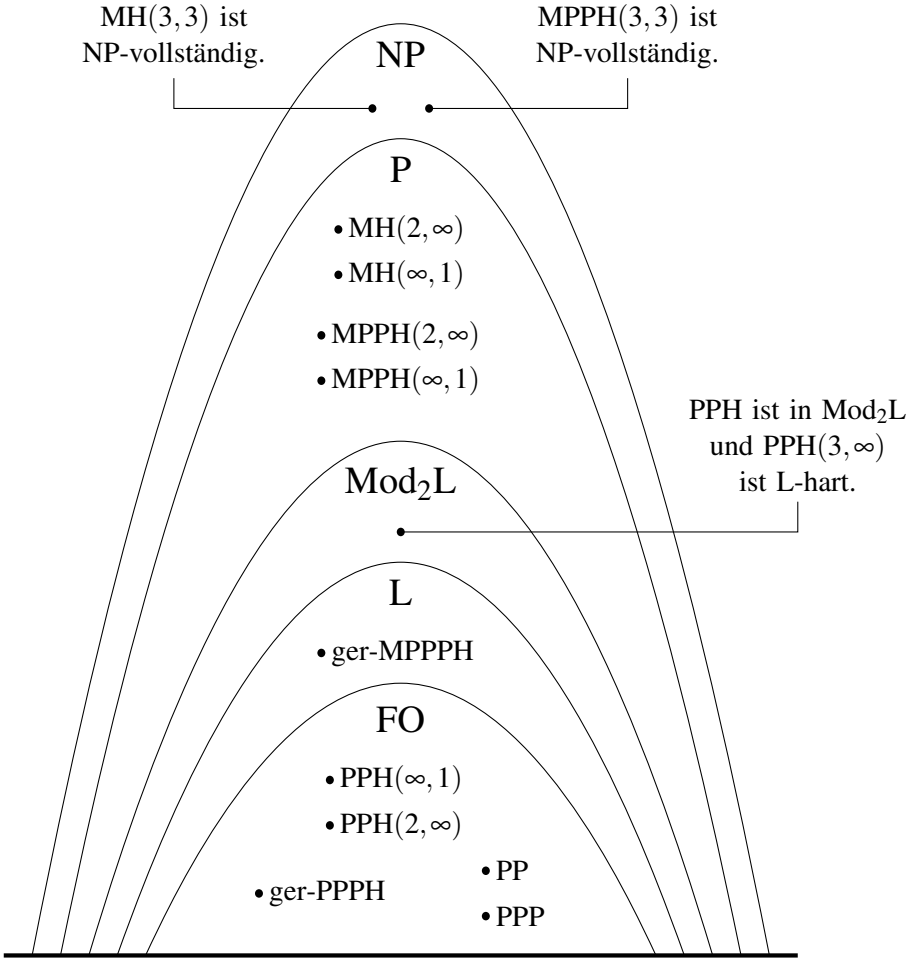
Als eines der Resultate dieser Arbeit wurde gezeigt, dass PPH in $\mathrm{Mod}_2\mathrm{L}$ liegt und hart für die Klasse L ist. Dies wirft die Frage auf, ob PPH auch hart für $\mathrm{Mod}_2\mathrm{L}$ ist oder vielleicht in L liegt.

Als ein weiteres Resultat wurde gezeigt, dass ger-MPPPH in L liegt. Es ist aber nicht klar, ob auch die ungerichtete Variante MPPPH in L liegt. Eine ähnliche Frage stellt sich für die Haplotypisierung mittels perfekten Pfadphylogenien. Die gerichtete Variante ger-PPPH ist zwar in FO, aber es ist nicht klar, ob auch die ungerichtete Variante PPPH in FO liegt. Durch eine Reduktion von PPPH auf PPH (Hinzufügen von Genotyp $2 \ldots 2$ zur Genotypmatrix) wissen wir bisher nur, dass PPPH in $\mathrm{Mod}_2\mathrm{L}$ liegt. Die Frage ist nun: Lässt sich für die Haplotypisierung mittels perfekten Pfadphylogenien beweisen, dass die ungerichtete Variante in FO liegt (zum Beispiel durch eine Reduktion auf die gerichtete Problemvariante) oder ist die ungerichtete Variante echt schwerer (zum Beispiel L-hart).

Durch die Resultate, die aus der Literatur angegeben wurden, lassen sich weitere Fragestellungen ableiten. Zwar sind MH und MPPH NP-vollständig aber vielleicht lassen sich die Probleme auch durch Approximationsalgorithmen oder Festparameteralgorithmen (zusammen mit einer sinnvollen Parametrisierung der Probleme) geeignet lösen. Erste Ergebnisse in diese Richtungen sind bereits vorhanden (für das Problem MH in [28, 29, 34, 35] und für das Problem MPPH in [35]).

Ein weiteres Problem, das nicht in dieser Arbeit behandelt wurde, ergibt sich, wenn die Eingaben für die Haplotypisierung unvollständig sind. Dies kann in der Praxis auftreten, wenn die Methoden, die im Labor zum Auslesen der Erbinformation verwendet werden, nicht die Allele an jeder Basenposition bestimmen können. Für die Eingabe der Haplotypisierungsprobleme bedeutet dies, dass ein Eintrag nicht nur den Wert 0, 1 oder 2 besitzen kann sondern zusätzlich einen vierten Wert annehmen kann, welcher angibt, dass für diesen Eintrag keine Information vorhanden ist. Betrachtet man beispielsweise die Haplotypisierung mittels perfekten Phylogenien und lässt unvollständige Eingaben zu, dann lässt sich die Problemstellung folgendermaßen definieren: Lassen sich die unbekannten Einträge in einer unvollständigen Genotypmatrix so mit 0, 1 oder 2 auffüllen, dass die aufgefüllte Genotypmatrix eine perfekte Phylogenie zulässt? Es wurde bisher gezeigt, dass

Abbildung 10: Die Abbildung zeigt eine Übersicht über die Komplexität der Haplotypisierungsprobleme, die in dieser Arbeit behandelt wurden. Die verschiedenen Komplexitätsklassen FO, L, Mod_2L, P und NP wurden in Abschnitt 3.1 beschrieben. Es lässt sich erkennen, dass sich die Komplexität (k,l)-beschränkter Varianten von MH, PPH und MPPH unterscheidet. Beispielsweise ist PPH$(3,3)$ L-hart, aber PPH$(2,\infty)$ und PPH$(\infty,1)$ liegen in FO. Die Komplexität nimmt also zu, wenn heterozytische Einträge in der Eingabe vorkommen. Analog lässt sich dies für MH und MPPH erkennen, wobei die dargestellten (k,l)-beschränkten Varianten hier NP-vollständig sind oder in P liegen.



dieses Problem NP-vollständig ist [26] und dass die entsprechende Pfadvariante ebenfalls NP-vollständig ist [17]. In Bezug auf diese Arbeit stellt sich nun die Frage, wie sich die Komplexität der verschiedenen Haplotypisierungprobleme und der (k,l)-beschränkten Varianten im Detail verändert, wenn man unvollständige Daten zulässt.

Literatur

[1] Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, Viggo Kann, Alberto Marchetti-Spaccamela, and Marco Protasi. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer, 1999.

[2] Vineet Bafna, Dan Gusfield, Sridhar Hannenhalli, and Shibu Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. *Journal of Computational Biology*, 11(5):858–866, 2004.

[3] Vineet Bafna, Dan Gusfield, Giuseppe Lancia, and Shibu Yooseph. Haplotyping as perfect phylogeny: A direct approach. *Journal of Computational Biology*, 10(3–4):323–340, 2003.

[4] Paola Bonizzoni, Gianluca Della Vedova, Riccardo Dondi, and Jing Li. The haplotyping problem: an overview of computational models and solutions. *Journal of Computer Science Technology*, 18(6):675–688, 2003.

[5] Daniel G. Brown and Ian M. Harrower. Integer programming approaches to haplotype inference by pure parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):141–154, 2006.

[6] Gerhard Buntrock, Carsten Damm, Ulrich Hertrampf, and Christoph Meinel. Structure and importance of logspace-MOD-classes. *Mathematical Systems Theory*, 25(3):223–237, 1992.

[7] Rudi Cilibrasi, Leo van Iersel, Steven Kelk, and John Tromp. On the complexity of several haplotyping problems. In *Proceedings of the 5th International Workshop on Algorithms in Bioinformatics (WABI 2005)*, volume 3692 of *Lecture Notes in Computer Science*, pages 128–139. Springer, 2005.

[8] Francis S. Collins, Mark S. Guyer, and Aravinda Chakravarti. Variations on a theme: Cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581, 1997.

[9] The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003.

[10] Mark J. Daly, John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson, and Eric S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

[11] Zhihong Ding, Vladmimir Filkov, and Dan Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping (PPH) problem. *Journal of Computational Biology*, 13(2):522–553, 2006.

[12] Eleazar Eskin, Eran Halperin, and Richard M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, 1(1):1–20, 2003.

[13] Stacey B. Gabriel, Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly, and David Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

[14] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York, 1979.

[15] Jens Gramm, Arfst Nickelsen, and Till Tantau. Fixed-parameter algorithms in phylogenetics. *The Computer Journal*, 2007. Zur Veröffentlichung akzeptiert. DOI 10.1093/comjnl/bxm049.

[16] Jens Gramm, Arfst Nickelsen, and Till Tantau. Fixed-parameter algorithms in phylogenetics. In Jonathan Keith, editor, *Bioinformatics (tentative)*, Methods in Molecular Biology. The Humana Press, 2007. Zur Veröffentlichung akzeptiert.

[17] Jens Gramm, Till Nierhoff, Roded Sharan, and Till Tantau. Haplotyping with missing data via perfect path phylogenies. *Discrete and Applied Mathematics*, 155(6–7):788–805, 2007.

[18] Dan Gusfield. Efficient algorithms for inferring evolutionary history. *Networks*, 21:19–28, 1991.

[19] Dan Gusfield. *Algorithm on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge Press, 1997.

[20] Dan Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 166–175. ACM Press, 2002.

[21] Dan Gusfield. Haplotype inference by pure parsimony. In *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching (CPM 2003)*, volume 2676 of *Lecture Notes in Computer Science*, pages 144–155. Springer, 2003.

[22] Dan Gusfield. An overview of combinatorial methods for haplotype inference. In Sorin Istrail, Michael S. Waterman, and Andrew G. Clark, editors, *Computational Methods for SNPs and Haplotype Inference, DIMACS/RECOMB Satellite Workshop, Piscataway, NJ, USA, November 21-22,*

2002, Revised Papers, volume 2983 of *Lecture Notes in Computer Science*, pages 9–25. Springer, 2004.

[23] Bjarni V. Halldórsson, Vineet Bafna, Nathan Edwards, Ross Lippert, Shibu Yooseph, and Sorin Istrail. A survey of computational methods for determining haplotypes. In Sorin Istrail, Michael S. Waterman, and Andrew G. Clark, editors, *Computational Methods for SNPs and Haplotype Inference, DIMACS/RECOMB Satellite Workshop, Piscataway, NJ, USA, November 21-22, 2002, Revised Papers*, volume 2983 of *Lecture Notes in Computer Science*, pages 26–47. Springer, 2004.

[24] Yao-Ting Huang, Kun-Mao Chao, and Ting Chen. An approximation algorithm for haplotype inference by maximum parsimony. *Journal of Computational Biology*, 12(10):1261–1274, 2005.

[25] Neil Immerman. *Descriptive Complexity*. Springer-Verlag, New York, 1999.

[26] Gad Kimmel and Ron Shamir. The incomplete perfect phylogeny haplotype problem. *Journal of Bioinformatics and Computational Biology*, 3(2):359–384, 2005.

[27] Eric Lai. Application of snp technologies in medicine: Lessons learned and future challenges. *Genome Research*, 11(6):927–929, 2001.

[28] Giuseppe Lancia, Maria Cristina Pinotti, and Romeo Rizzi. Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. *INFORMS Journal on Computing*, 16(4):348–359, 2004.

[29] Giuseppe Lancia and Romeo Rizzi. A polynomial case of the parsimony haplotyping problem. *Operations Research Letters*, 34(3):289–295, 2006.

[30] Yunkai Liu and Cun-Quan Zhang. A linear solution for haplotype perfect phylogeny problem. In *Proceedings of the International Conference Advances in Bioinformatics and its Applications*, pages 173–184. World Scientific, 2005.

[31] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[32] Nila Patil, Anthony J. Berno, David A. Hinds, Wade A. Barrett, Jigna M. Doshi, Coleen R. Hacker, Curtis R. Kautzer, Danny H. Lee, Claire Marjoribanks, David P. McDonough, Bich T. N. Nguyen, Michael C. Norris, John B. Sheehan, Naiping Shen, David Stern, Renee P. Stokowski, Daryl J. Thomas, Mark O. Trulson, Kanan R. Vyas, Kelly A. Frazer, Stephen P. A. Fodor, and David R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[33] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[34] Roded Sharan, Bjarni V. Halldórsson, and Sorin Istrail. Islands of tractability for parsimony haplotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3):303–311, 2006.

[35] Leo van Iersel, Judith Keijsper, Steven Kelk, and Leen Stougie. Beaches of islands of tractability: Algorithms for parsimony and minimum perfect phylogeny haplotyping problems. In *Proceedings of the 6th International Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lecture Notes in Computer Science*, pages 80–91. Springer, 2006.

[36] Ravi Vijayasatya and Amar Mukherjee. An optimal algorithm for perfect phylogeny haplotyping. *Journal of Computational Biology*, 13(4):897–928, 2006.

[37] Lusheng Wang and Ying Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780, 2003.