Bachelorarbeit

Validierung der Perfekte-Phylogenie-Annahme für die Haplotypisierung

Bachelorstudiengang Molecular Life Science

Universität zu Lübeck

von

Linda Schönfeld

Gutachter:

1. **Prof. Dr. Till Tantau**
   Institut für Theoretische Informatik
   Universität zu Lübeck

2. **Dr. Steffen Möller**
   Institut für Neuro- und Bioinformatik
   Universität zu Lübeck

Datum: ……

Erklärung:

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Lübeck, den 03.09.2009 …….....

Linda Schönfeld

Zusammenfassung

Haplotypisierung als bioinformatische Methode, um aus Genotypen Haplotypen berechnen zu können, ist für Assoziationsstudien von Krankheiten und dem zugrundeliegenden genetischen Code sehr wichtig, denn Haplotypen geben die genetische Information genauer wieder als die Genotypen. Auf der Basis der Annahme der perfekten Phylogenie kann eine solche Haplotypisierung stattfinden. Diese Bachelorarbeit übersetzt reale (im Labor sequenzierte) Haplotyp-Daten aus vier verschiedenen Datensätzen in Gentoypen, sucht innerhalb der Haplotyp-Daten nach Blöcken, in denen perfekte Phylogenien vorkommen, und definiert Kenngrößen zum Vergleich der realen Haplotyp-Daten mit den berechneten Haplotypen. Diese Bewertungsparameter geben die Güte und Größe der berechneten Lösungen für einzelne Blöcke wieder.

Abstract

Validation of the perfect-phylogeny assumption for haplotyping

Haplotyping as a bioinformatic method for computing haplotypes on the basis of genotypesis important for association studies of diseases and the underlying genetic code, because haplotypes describe this information more precise than genotypes. If one takes the perfect phylogeny as an assumption it may be possible to do haplotyping this way. In this bachelor's thesis real haplotype-data (sequenced in laboratory) from four different data sets are translated into genotypes and in the real haplotype data blocks with perfect phylogeny are detected. In addition to this, parameters for analysis and comparison of the real haplotype data and computed haplotypes are defined. They stand for quality and quantity of the computed solutions in the blocks.

Inhaltsverzeichnis

1 Einleitung

Die Früherkennung von Krankheiten sowie die Erklärung der unterschiedlichen Medikamentenverträglichkeiten einer Bevölkerungsgruppe ermöglichen genauere Betrachtung von Krankheiten. Hierfür werden genaue genetische Analysen und Assoziationsstudien zwischen Krankheiten und den zugrundliegenden Genomsequenzen benötigt. Das menschliche Genom besteht aus Chromosomen, die paarweise vorliegen. Die Chromosomen selber bestehen aus DNS (Desoxyribonukleinsäure), die aus Basen aufgebaut ist, die quasi das Alphabet darstellen. Die Daten, die pro Chromosom ausgelesen werden, nennt man Haplotypen, die Daten für das Chromosomenpaar Genotypen. In den Genotypen ist aber nicht mehr die Information enthalten, welche Basen auf welchem Chromosom vorliegen, es ist nur bekannt, dass es diese Basen an der untersuchten Position gibt. Für Assoziationsstudien ist es besser, die genauesten Daten zu verwenden, die man bekommen kann, also die Haplotyp-Daten. Das Problem ist aber, dass man diese Daten schwerer bekommt als die Genotyp-Daten. Genotypen lassen sich im Labor sehr leicht und auch kostengünstig sequenzieren, während es bei Haplotypen sehr zeit- und kostenaufwändig ist [Patil et al., 2001]. Deshalb ist es von Vorteil, wenn es Algorithmen gibt, die aus Genotypen Haplotypen zuverlässig berechnen können. Es gibt verschiedene Theorien auf deren Grundlage man aus Genotypen Haplotypen berechnen kann. Eine davon ist die Perfekte-Phylogenie-Annahme. Sie enthält einige Grundvoraussetzungen, die erfüllt sein müssen und berechnet die Haplotypen auf der Grundlage eines phylogenetischen Baumes.

Diese Arbeit soll überprüfen, ob und wie gut die Perfekte-Phylogenie-Annahme auf reale Haplotyp-Daten zutrifft und ist folgendermaßen aufgebaut:

In Kapitel 2 wird der biologische Hintergrund genauer erläutert und der Ansatz der Perfekten-Phylogenie-Haplotypisierung erklärt. Anschließend wird in Kapitel 3 das Vorgehen erklärt und die zur Bewertung verwendeten Kenngrößen definiert. Kapitel 4 beinhaltet die Ergebnisse der vier untersuchten Datensätze. Zum Abschluss steht Kapitel 5, in dem die Ergebnisse diskutiert werden und ein Ausblick gegeben wird.

2 Problemstellung bei der Haplotypisierung

In diesem Abschnitt werden die biologischen Hintergründe erläutert und die Arbeit motiviert. Außerdem wird der Begriff der perfekten Phylogenie erklärt.

2.1 Biologischer Hintergrund

Die Erbinformation der Zellen liegt in Chromosomen vor, die aus Desoxyribonukleinsäure (DNS) und Proteinen, die für die kondensierte Form sorgen, bestehen. Die DNS besteht aus vier Basen, Adenin und Guanin (Purine) und Cytosin und Thymin (Pyrimidine). Diese Basen werden jeweils mit ihren Anfangsbuchstaben A, G, C und T abgekürzt. Jede Zelle, bis auf die Keimzellen, enthält 22 Chromosomenpaare sowie zwei geschlechtsspezifische Chromosomen (bei Frauen XX, bei Männern XY), sodass jede Zelle 46 Chromosomen beinhaltet. Die Keimzellen enthalten den Chromosomensatz nur einmal, dort liegen also keine Paare vor und es gibt nur 23 Chromosomen pro Keimzelle. Chromosomenpaare nennt man auch homologe Chromosomen. Im Prinzip bestehen diese Chromosomenpaare bei allen Menschen aus der gleichen Basensequenz, innerhalb einer Population können sich aber Variationen ausbilden, die für Vielfalt sorgen.

Eine Veränderung der Basensequenz entsteht durch Mutationen. Diese sind meistens Punktmutationen, treten also nur an einzelnen Stellen in der DNS-Sequenz auf. So kann die Sequenz TGC durch Punktmutation an der dritten Position zu TGA mutieren. Per Definition sind Variationen in der DNS-Sequenz erst dann SNPs (Single Nucleotide Polymorphisms, gesprochen „snips“), wenn mindestens 1% einer Population eine solche Veränderung an der untersuchten Stelle trägt, sich also diese Mutation durchgesetzt hat. Deshalb nennt man sie auch erfolgreiche Punktmutationen. Es gibt durchschnittlich 10 Millionen SNPs (ca. alle 300 Basenpaare ein SNP). SNPs bilden 90% der Genomvarianz beim Menschen [The International HapMap Consortium, 2003].

Vielfalt wird außerdem durch Rekombination erzeugt. Rekombination ist ein Vorgang, der während der Gametenbildung in der Meiose sowie bei der Bildung der Zygote während der Befruchtung stattfinden kann. Dabei paaren sich homologe Chromosomen, also die jeweiligen Chromosomenpaare, und tauschen Teile der DNS-Sequenz aus. Ein Spezialfall der Rekombination ist Crossing-Over, das während der Meiose auftreten kann. Zum Beispiel werden zwei Sequenzen mit den Basen AGGTCC und TTGAAG beim Crossing-Over in der Mitte der Sequenz zu den Sequenzen AGGAAG und TTGTCC kombiniert.

Als Haplotyp wird die Sequenz auf einem Chromosom bezeichnet. Die paarweisen Haplotypen ergeben dann das Chromosomenpaar beziehungsweise den Genotyp für das Chromosom. Der Genotyp kann an jeder einzelnen Stelle der Sequenz homozygot oder

heterozygot sei. Wenn auf beiden Chromosomen die gleiche Base vorliegt, nennt man diese Position homozygot, sonst heterozygot.

Ein Haplotyp auf einem Chromosom wäre zum Beispiel die Sequenz GTACC. Liegt auf dem homologen Chromosom die Sequenz GTGGC vor, so sieht der Genotyp wie folgt aus: GT{A/G}{C/G}C. Das Individuum ist für die ersten beiden Positionen sowie für die letzte homozygot, für die dritte und vierte Position heterozygot. Der Genotyp kombiniert die Haplotypen, enthält aber nicht mehr die Information, welche der Basen an den heterozygoten Positionen auf welchem Chromosom vorliegt. Aus zwei Haplotypen kann man den Genotyp eindeutig schließen, aus einem Genotyp aber nicht immer die zugrundeliegenden Haplotypen. Je mehr heterozygote Positionen vorkommen, desto schwieriger wird es, die Haplotypen eindeutig zuzuordnen. Das liegt daran, dass es bei n heterozygoten Positionen 2^{n-1} Haplotypenpaare gibt, die einem Genotyp zugrunde liegen können und es nicht klar ist, welches Paar nun das korrekte für diesen Genotyp ist. Ein Beispiel hierzu zeigt Abbildung 1.

A C G T **A/G** G C T **A/C** C T G	Genotyp
A C G T **A** G C T **A** C T G	Haplotyp 1
A C G T **G** G C T **C** C T G	Haplotyp 2
A C G T **A** G C T **C** C T G	Haplotyp 1
A C G T **G** G C T **A** C T G	Haplotyp 2

Abbildung 1: Ein Genotyp mit zwei heterozygoten Positionen (fett gedruckt) und den dazugehörigen möglichen Haplotypen. Einem Genotyp mit zwei heterozygoten Positionen können zwei Paare von Haplotypen zugrunde liegen.

Wie schon erwähnt, kann die Sequenzierung eines Genotyps im Labor leicht und kostengünstig durchgeführt werden. Die Haplotyp-Information oder Phaseninformation (welches Merkmal bei heterozygoten SNPs auf welchem Chromosom vorliegt) ist dabei nicht enthalten. Gerade diese Information ist aber sehr wichtig, da angenommen wird, dass die Haplotypen eine Rolle in der Bestimmung der Medikamentenverträglichkeit einer Person spielen. Zum Beispiel kann anhand einiger Haplotypen ein höheres Risiko für einige Krankheiten angenommen werden. Deshalb ist es wichtig, sich mit bioinformatischen Methoden zu beschäftigen, die aus Genotyp-Daten zuverlässig Haplotypen berechnen können. Es gibt dort verschiedene Ansätze, einer davon ist die Perfekte-Phylogenie-Annahme.

2.2 Perfekte Phylogenie

Haplotypisierung mittels perfekter Phylogenien wurde von Gusfield [Gusfield, 2004] vorgeschlagen. Zugrunde liegen zwei Annahmen:

- Es gibt Blöcke innerhalb des Genoms, in denen keine Rekombination stattfindet.
- Die Mutationsfrequenz an einer Basenposition ist sehr gering.

Es wurde festgestellt, dass es in der Vererbungsgeschichte für Haplotypenblöcke keine Anzeichen für Crossing-Over-Ereignisse gibt [Daly et al., 2001]. Außerdem wird die Varianz in Haplotypblöcken nur durch Punktmutationen hervorgerufen und lässt sich somit durch SNPs beschreiben. Im Genom treten nur sehr wenige Mutationen auf, denn das Genom ist sehr groß und der Zeitraum, in dem die Population an Haplotypen, die betrachtet werden, entstanden ist, ist sehr kurz. Deshalb kann angenommen werden, dass pro Basenposition nur eine Mutation innerhalb dieser Zeit der Entwicklung der betrachteten Population auftritt.
Wenn diese beiden Eigenschaften erfüllt sind, so lassen sich die Haplotypen zu einem phylogenetischen Baum zusammenfassen, der eine perfekte Phylogenie ist. Graphisch dargestellt ist dies in Abbildung 2.

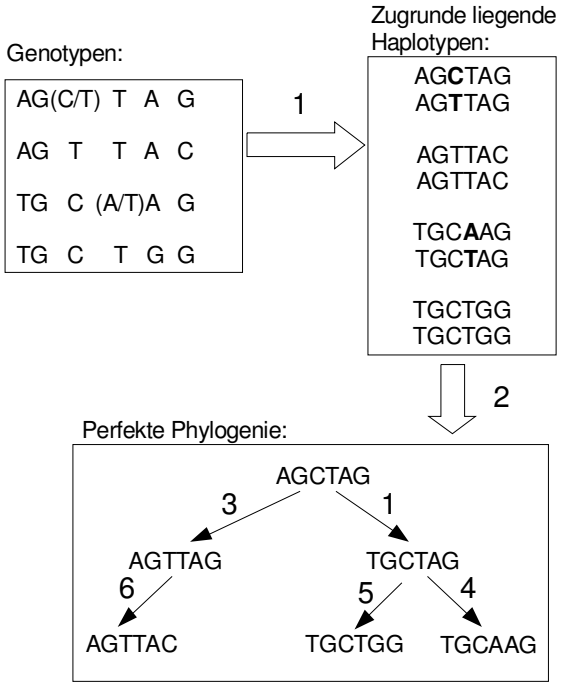


Abbildung 2: Darstellung von drei Genotypen mit fünf zugrunde liegenden unterschiedlichen Haplotypen, die sich als perfekte Phylogenie anordnen lassen. Die heterozygoten Positionen sind fett gedruckt. Jeder Pfeil im Kasten der perfekten Phylogenie zeigt eine Punktmutation an. An den Kanten der Pfeile stehen die Positionen der Sequenz, an denen eine Mutation stattgefunden hat. In Schritt 1 werden für die Genotypen die zugrunde liegenden Haplotypen berechnet. In Schritt 2 wird die perfekte Phylogenie der Haplotypen dargestellt.

Bei der Haplotypisierung mittels perfekter Phylogenien wird für die eingegebenen Genotypen die Menge an Haplotypen gesucht, die den Genotypen zugrunde liegen und sich als perfekte Phylogenie anordnen lassen (siehe Abbildung 2, Schritte 1 und 2). Zu beachten ist, dass dies nicht für jede Menge an Genotypen möglich ist. Tritt ein Crossing-Over Ereignis auf, so gibt es keine perfekte Phylogenie. Genauso verhält es sich bei mehrfachen Mutationen an einer Basenposition. Interessant ist es, dass genau dann keine perfekte Phylogenie gefunden werden kann, wenn die Kombinationen CT, TC, GC, CC in zwei SNPs vorkommen. Diese Eigenschaft bezeichnet man auch als den Vier-Gameten Test.

3 Beschreibung der Methodik und Definition der Kenngrößen

In diesem Abschnitt wird zuerst das Vorgehen beschrieben und anschließend werden die Maße und Kenngrößen definiert, die später zur Auswertung der Lösungen verwendet werden.

3.1 Methodik

Zuerst werden die vier verwendeten Datensätze (ACE, APOE, KLK13 und KLK14) so vorbereitet, dass eine Binärcodierung möglich ist. Die Datensätze bestehen alle aus Haplotyp-Sequenzen und wurden im Labor sequenziert, sind also reale Haplotyp-Daten und keine berechneten Daten. Beim Einlesen werden aus den Basen der DNS (A, G, C, T) Nullen und Einsen. Die erste Base, die pro SNP vorkommt, wird auf 0 gesetzt und wenn an dem SNP die andere Base vorkommt, dann wird diese mit 1 codiert. Dies ist möglich, weil an einer Position nur zwei verschiedene Basen in den Datensätzen vorkommen. Bei den unvollständigen Datensätzen wurden die SNPs entfernt, die unbekannte Einträge enthielten. Die nun binärcodierten Haplotypen werden anschließend in Genotypen übersetzt. Tritt in beiden Haplotypen an einem SNP das gleiche Merkmal (0 oder 1) auf, so wird dieser Wert für den Genotyp übernommen. Treten auf einem Haplotyp 0 und auf dem anderen Haplotyp dieses Haplotyppaares an der gleichen Position eine 1 auf, so wird in den Genotyp an dieser Stelle eine 2 geschrieben. Der Wert 2 repräsentiert einen heterozygoten SNP, die Werte 0 und 1 stellen homozygote SNPs dar. Zum Beispiel sehen für die Haplotypen AGC und ATG die binärcodierten Haplotypen so aus: 000 und 011. Der dazugehörige Genotyp ist dann 022. Zusätzlich werden die realen Haplotypen auf perfekte Phylogenien untersucht. Da bei Verwendung des gesamten Datensatzes keine perfekte Phylogenie auftritt, wurden aufeinanderfolgende Blöcke gesucht, in denen es eine perfekte Phylogenie gibt. Das Vorgehen dabei sieht so aus:

Es werden alle Individuen betrachtet und es wird nach Blöcken gesucht, wo perfekte Phylogenie vorkommt. Dies sind alle die Blöcke, die nicht die Kombinationen {00, 01, 10, 11} enthalten. Diese Eigenschaft nennt man die Vier-Gameten-Eigenschaft und sie zerstört die perfekte Phylogenie. Solange noch perfekte Phylogenie auf dem untersuchten Block vorhanden ist, wird ein weiterer SNP hinzugenommen und abermals auf perfekte Phylogenie überprüft. Treffen die Kriterien nun immer noch zu, so wird wieder ein weiterer SNP hinzugefügt. Ist nun aber keine perfekte Phylogenie mehr vorhanden, so wird ein neuer Block begonnen. Am Ende werden die Grenzen zurückgegeben, damit mit der Vorkenntnis dieser Grenzen die Berechnung der Haplotypen auf den Genotypen in genau diesen Blöcken arbeiten kann. Damit wird gewährleistet, dass dort auch wirklich eine perfekte Phylogenie vorliegt.

Anschließend werden die Haplotypen, die der Annahme der perfekten Phylogenie entsprechen, berechnet. Der Algorithmus, der dies übernimmt, ist im Rahmen des DFG-Projektes „Komplexität von Haplotypisierungsproblemen“ implementiert worden und wird in dieser Arbeit verwendet. Er bekommt als Eingabe die Genotypen und berechnet die zugrundeliegenden Haplotypen auf der Basis einer perfekten Phylogenie. Die Lösungen, die der Algorithmus ausgibt, werden dann mit einigen Bewertungsparametern (siehe Abschnitt 3.2 Kenngrößen) ausgewertet, dazu gehört zum Beispiel die Anzahl der Lösungen und ob einige Haplotypen in allen Lösungen vorkommen.

Abschließend werden dann die von dem Algorithmus berechneten Lösungen mit den realen Haplotypen verglichen. Die Implementierung des Vorgehens wurde mit der Programmiersprache Java durchgeführt. Das Vorgehen wird in Abbildung 3 schematisch dargestellt.

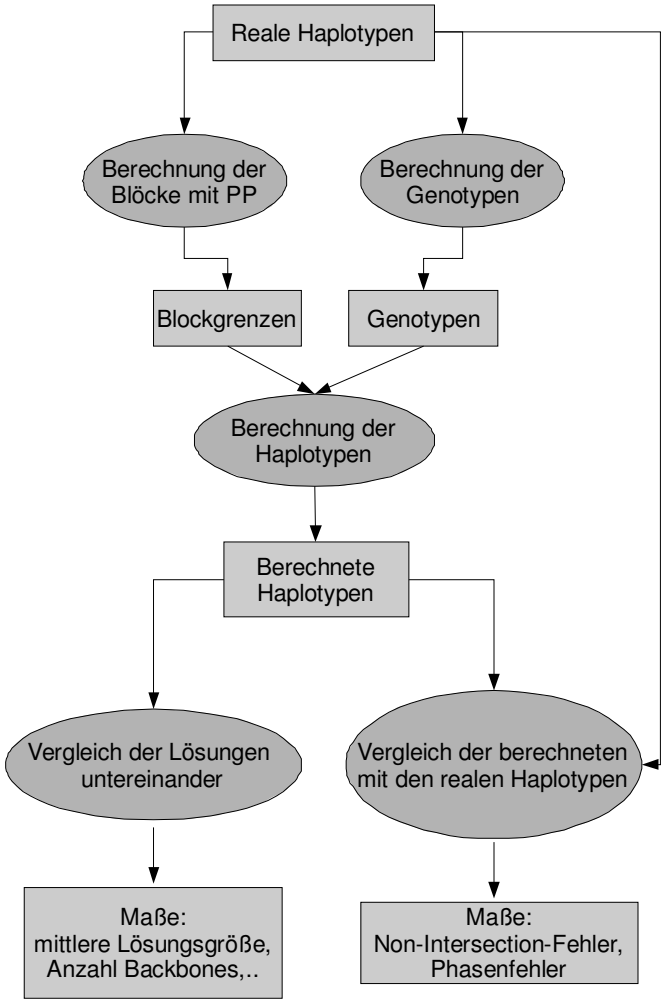


Abbildung 3: Vorgehen bei der Bachelorarbeit. Rechteckige Kästen stellen Eingaben bzw. Ausgaben dar. Ellipsen stellen Funktionen des programmierten Codes dar. Die Abkürzung PP steht für perfekte Phylogenie.

3.2 Kenngrößen und Maße

Um die Genauigkeit der von dem Algorithmus berechneten Lösungen zu bestimmen, werden verschiedene Kenngrößen verwendet. Es wird unterschieden in datensatzspezifische Maße und blockspezifische Maße. Untersucht werden die Lösungen untereinander und anschließend werden die Lösungen mit den realen Haplotypen verglichen.

3.2.1 Datensatzspezifische Maße

Unter diese Kategorie fallen alle Kenngrößen, die eine Aussage für den gesamten Datensatz liefern. Dazu gehören die Anzahl der SNPs (gegebenenfalls die Anzahl der SNPs nach Löschen der SNPs mit zu vielen unbekannten Einträgen), die Anzahl der Individuen, die Anzahl der Blöcke, für die eine perfekte Phylogenie gefunden werden kann und deren Größe, sowie die Anfangspositionen für einen Block. Ein weiteres Maß ist der Anteil der heterozygoten SNP-Positionen. Zusätzlich wurde noch der Anteil des längsten Blocks an der Gesamtlänge des Datensatzes berechnet, die Anzahl der unterschiedlichen Haplotypen pro Datensatz und die durchschnittliche Blocklänge.

3.2.2 Blockspezifische Maße

Für die Blöcke, in denen es eine perfekte Phylogenie gibt, werden einige Maße betrachtet. Dazu gehören die Anzahl der Lösungen, die es für diesen Block gibt und der prozentuale Anteil der heterozygoten Positionen. Ein weiteres Maß ist die mittlere Lösungsgröße, die betrachtet wird, wenn der Algorithmus mehr als eine Lösung liefert. Die Lösungsgröße entspricht der Anzahl an unterschiedlichen Haplotypen, die der Algorithmus benötigt, um für die eingegebenen Genotypen Haplotypen zu berechnen. Hier wird über alle Lösungen summiert und anschließend durch die Anzahl der Lösungen geteilt.

Backbone-Haplotypen

Für die Lösungen, die der Algorithmus für die eingegebenen Genotypen findet, werden wie auch bei einer anderen Veröffentlichung [Climer et al., 2009] die Backbone-Haplotypen und deren Anzahl bestimmt. Backbone-Haplotypen sind Haplotypen, die in jeder der Lösungen vorkommen, bilden also die Schnittmenge der Lösungen (siehe Abbildung 4). Je größer die Anzahl der Backbones ist, desto mehr Haplotypen kommen in allen Lösungen vor, desto ähnlicher sind sich also die Lösungen.

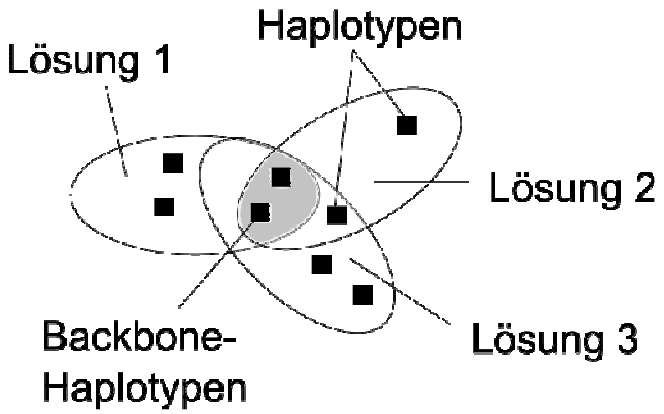


Abbildung 4: Darstellung der Backbone-Haplotypen für drei Lösungen. Die kleinen Quadrate stellen Haplotypen dar. Lösung 1 beinhaltet 4 Haplotypen, Lösung 2 enthält ebenfalls 4 Haplotypen und Lösung 3 enthält 5 Haplotypen. Lösung 1 und Lösung 2 haben zwei gemeinsame Haplotypen, Lösung 2 und Lösung 3 haben drei gemeinsame Haplotypen. Alle drei Lösungen zusammen haben zwei gemeinsame Haplotypen. Dies sind die Backbone-Haplotypen (dunkle Fläche).

Non-Intersection-Fehler

Der Non-Intersection-Fehler wird wie auch bei einer anderen Veröffentlichung [Climer et al., 2009] für jedes Paar von realen Daten mit berechneter Lösung gebildet. Er nimmt genau dann den Wert 0 an, wenn die beiden Mengen identisch sind. Je größer der Non-Intersection-Fehler wird, desto unterschiedlicher sind die beiden untersuchten Mengen. Maximal kann der Wert 1 betragen, dies bedeutet zwei komplett disjunkte Lösungen. Ein Beispiel hierfür zeigt Abbildung 5, Teil 2.

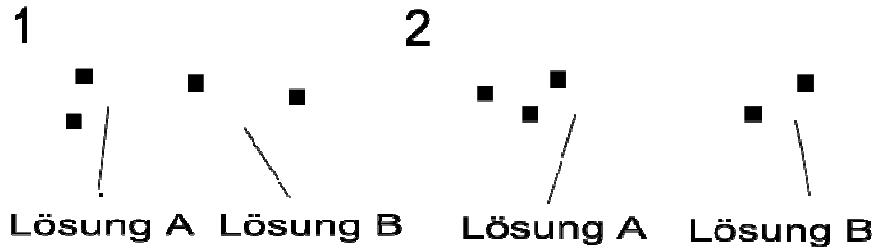


Abbildung 5: Graphische Veranschaulichung des Non-Intersection-Fehlers mit zwei Lösungen A und B. Die Quadrate stellen Haplotypen dar. Teil 1 des Bildes zeigt zwei Lösungen, die überlappen, also gemeinsame Haplotypen aufweisen, Teil 2 stellt zwei Lösungen A und B dar, die komplett unterschiedlich sind. Der Non-Intersection-Fehler nimmt bei Teil 1 den Wert $\frac{2}{3}$ an, in Teil 2 beträgt er 1.

Formal wird der Non-Intersection- Fehler so berechnet:

$$\mathbf{NonInErr} = \max\left\{\frac{|A \backslash B|}{|A|}, \frac{|B \backslash A|}{|B|}\right\}$$

Aus der Menge A werden alle die Haplotypen entfernt, die auch in B enthalten sind. So erhält man A\B. Den Wert B\A erhält man, indem man aus der Menge B alle Haplotypen entfernt, die in A sind. Liegen zum Beispiel in A drei verschiedene Haplotypen 010, 110 und 111 und in B zwei verschiedene Haplotypen 000 und 111, so besitzt A\B die Größe 2, weil der Haplotyp 111 in beiden vorkommt. B\A enthält nur noch den Haplotyp 000, hat also die Größe 1. Der Wert $\frac{|A \backslash B|}{|A|}$ ist in diesem Beispiel $\frac{2}{3}$, $\frac{|B \backslash A|}{|B|}$ nimmt den Wert $\frac{1}{2}$ an. Das Maximum dieser beiden Zahlen ist der Wert $\frac{2}{3}$, also ist der Non-Intersection-Fehler bei diesem Beispiel $\frac{2}{3}$ = 0,667. Dies bedeutet, dass die beiden Lösungen relativ unterschiedlich sind. Veranschaulicht ist dieses Beispiel in Abbildung 5, Teil 1.

Phasenfehler

Für jedes Paar an realen Daten und berechneter Lösung wird für jedes Genotyp-Paar der Phasenfehler bestimmt. Er sagt aus, wie unterschiedlich ein Haplotyp-Paar gelöst wurde. Es werden pro Haplotyp nur die heterozygoten SNPs betrachtet und dann für einen Genotyp die Haplotypen verglichen. Die Betrachtung kann man deshalb auf die heterozygoten SNPs beschränken, weil es an den homozygoten Positionen keine anderen Möglichkeiten gibt, den Haplotyp zu berechnen. Der erste Haplotyp des ersten Haplotyppaares wird mit den beiden Haplotypen des zweiten Haplotyppaares verglichen und der Haplotyp, der mit dem gleichen Wert (im Beispiel in Abbildung 6 der Wert 0) anfängt, wird zum Vergleichen benutzt. Dann wird gezählt, an wie vielen Positionen sich die beiden Haplotypen unterscheiden. Dieser Wert wird für alle Genotypen pro Lösungspaar aufaddiert und anschließend wird diese Summe durch die Anzahl der heterozygoten Positionen abzüglich jeweils 1 für jeden Genotyp, der heterozygote Einträge enthält. Der Phasenfehler nimmt einen Wert zwischen 0 und 1 an. Der Wert 0 bedeutet, dass alle heterozygoten Positionen korrekt gelöst worden sind, der Wert 1 hingegen sagt aus, dass alle heterozygoten Positionen falsch gelöst worden sind. Je kleiner der Wert hierbei ist, desto besser stimmen also die berechneten Haplotypen mit den realen Haplotypen überein. Abbildung 6 veranschaulicht diesen Vorgang für ein Lösungspaar. Interessant sind bei der Auswertung sowohl der durchschnittliche Phasenfehler als auch der maximale Phasenfehler. Der maximale Phasenfehler ist wichtig, weil er angibt, wie schlecht die Lösung maximal sein kann, wenn man auf Genotypen arbeitet und die Haplotypen nicht kennt. Zu diesem Zeitpunkt sind ja die korrekten Haplotypen unbekannt.

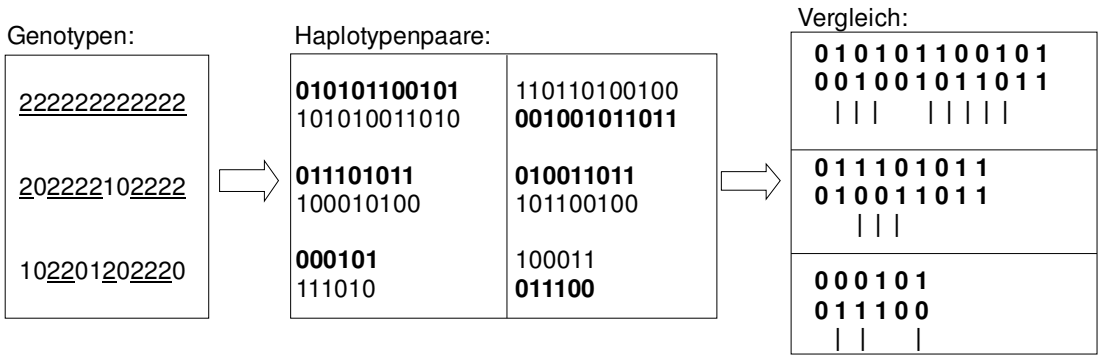


Abbildung 6: Veranschaulichung der Berechnung des Phasenfehlers. Zuerst werden aus den Genotypen die Positionen genommen, die heterozygot sind (in der Abbildung sind diese Positionen unterstrichen). Dann werden die paarweisen Haplotypenpaare betrachtet, fett gedruckt sind diejenigen Haplotypen, die für den Vergleich verwendet werden. Beim Vergleich werden die Positionen gezählt (durch senkrechte Striche angedeutet), die sich unterscheiden. In diesem Beispiel werden 14 Unterschiede gezählt (8 aus dem ersten Paar, 3 aus dem zweiten Paar und 3 aus dem letzten Paar). Geteilt wird nun durch die Summe der Anzahl der heterozygoten Positionen (12+9+6) abzüglich 3, weil jeweils die erste Position eines Paares identisch ist. Somit ist der durchschnittliche Phasenfehler mit einem Wert von 14/24 = 0,583 recht hoch, das heißt mehr als jede zweite heterozygote Position wurde falsch berechnet.

4 Die Datensätze und die berechneten Ergebnisse

Alle vier verwendeten Datensätze wurden aus Veröffentlichungen entnommen. Für einen Überblick über die Datensätze sorgt Tabelle 1. Zwei der Datensätze enthalten zum Teil fehlende Einträge, das heißt, dort fehlt Information über die tatsächliche Basensequenz. Dies kann durch Fehler bei der Sequenzierung der DNS-Sequenz erfolgt sein. Bei der Arbeit mit den zwei unvollständigen Datensätzen wurde die SNPs entfernt, die unbekannte Einträge aufweisen, denn dort wäre die Aussagekraft nicht groß. Nach dem Löschen enthielten die Datensätze also keine unbekannten Einträge mehr. Alle Datensätze bestehen aus Sequenzen von realen Haplotypen, sie wurden im Labor sequenziert. In den nächsten Abschnitten werden die Ergebnisse der Kenngrößen für jeden Datensatz einzeln dargestellt und diskutiert.

Tabelle 1: Übersicht über die verwendeten Datensätze des Apolipoprotein E-Gens (APOE), Angiotensin Converting Enzyme- Gens (ACE) und der Kallikrein-Gene (KLK13 und KLK14).

	APOE	**ACE**	**KLK13**	**KLK14**
Quelle	[Orzack et al., 2003]	[Rieder et al., 1999]	[Andrés et al., 2007]	[Andrés et al., 2007]
Anzahl Genotypen	80	11	39	39
Anzahl SNPs	9	52	190	156

4.1 Apolipoprotein E-Gen

Allgemeines

Dieser Datensatz enthält 80 Individuen und 52 SNPs. Die Sequenz-Daten wurden von Orzack et al. [Orzack et al., 2003] erhoben. Dort wurden 80 Individuen untersucht, die alle nicht miteinander verwandt sind und aus unterschiedlichen geographischen Zonen stammen. Die Sequenz wurde durch den ABI 377 DNASequencer ermittelt, die Haplotypen wurden mit allelspezifischer PCR bestimmt. Die Anzahl der SNPs ist im Vergleich zu den anderen verwendeten Datensätzen sehr gering, der Anteil der heterozygoten SNPs liegt im Mittelfeld. Er liegt zwischen den Kallikrein-Datensätzen mit einem Anteil von ca. 20% und dem Angiotensin-Converting-Enzyme-Datensatz mit ca. 34%. Die Anzahl der unterschiedlichen Haplotypen für den gesamten Datensatz ist geringer als bei den Kallikrein-Datensätzen, aber größer als beim Angiotensin-Converting-Enzyme-Datensatz. Der Anteil des größten Blocks an der Gesamtlänge ist mit ca. 67% am größten, die durchschnittliche Blocklänge mit 3 SNPs eher gering.

Tabelle 2: Übersicht über den Datensatz des Apolipoprotein E-Gens mit den datensatzspezifischen Maßen

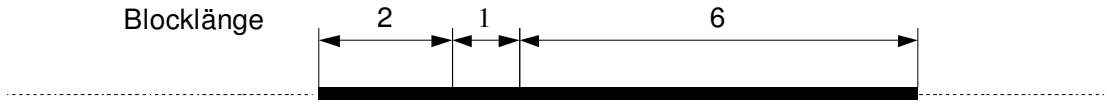
Quelle	[Orzack et al., 2003]
Anzahl Genotypen	80
Anzahl SNPs	9
Anzahl SNPs nach Löschen	9
Anteil heterozygote SNPs in %	21,4
Anzahl Blöcke mit PP	3
Größter Block mit PP	6
Anzahl unterschiedlicher Haplotypen	17
Ø Blocklänge	3
Anteil des längsten Blocks an der Gesamtlänge in %	67

Biologische Relevanz

Das Apolipoprotein E ist ein wichtiges Apolipoprotein der Darmschleimhaut und hat die Aufgabe, triglyceridreiche Lipoproteinbestandteile zu verstoffwechseln. Es besteht aus 299 Aminosäuren und transportiert Cholesterin, Triglyceride und fettlösliche Vitamine zuerst in die Lymphbahn und anschließend ins Blut. Das zugehörige Gen APOE befindet sich auf Chromosom 19. Mutationen in diesem Gen können zu familiär-bedingten, also erblichen Krankheiten führen, wie erhöhte Triglycerid- oder Cholesterinspiegel im Blut. Die bisher entdeckten Polymorphismen des Gens haben Einfluss auf die Krankheiten Hypolipoproteinämie Typ III (Polymorphismus e-2) [Wikipedia, 2009] und der Polymorphismus e-4 wird mit einem höheren Risiko für Artheriosklerose sowie Alzheimer in Verbindung gebracht [Schmidt, 2004].

Ergebnisse

Für diesen Datensatz mit 80 Individuen und 9 SNPs lassen sich drei Blöcke unterschiedlicher Länge mit perfekter Phylogenie finden. Die durchschnittliche Blocklänge beträgt 3 SNPs. Der längste Block ist 6 SNPs groß und macht 67% an der Gesamtanzahl der SNPs aus. Für jeden Block findet der Algorithmus nur eine Lösung, die mit den realen Daten übereinstimmt. Daher haben Non-Intersection-Fehler und Phasenfehler den Wert 0. Auffällig ist, dass die Lösungen für den gesamten Datensatz eindeutig sind, auch bei dem für diesen Datensatz großen Abschnitt mit 6 SNPs. Es gibt für jeden Block nur eine Lösung, die dann auch den realen Haplotyp-Daten entspricht. Dies lässt sich durch die große Anzahl an Individuen erklären. Eine Übersicht über die erhobenen Daten befindet sich in Abbildung 7.



% heterozygot	23,8	52,5	15,4
# Lösungen	1	1	1
Ø Lösungsgröße	3	2	7
# Backbones	3	2	7
	Vergleich mit realen Daten		
Ø NonInErr	0	0	0
Ø Phasenfehler	0	0	0
Max Phasenfehler	0	0	0

Abbildung 7: Darstellung der Ergebnisse von APOE. Der dicke Balken steht für die Sequenz, darüber sind die Blocklängen eingetragen. Jeweils unterhalb der Blöcke stehen die berechneten Werte für den jeweiligen Block. Anteil der heterozygoten Positionen an der Gesamtanzahl der Positionen (% heterozygot) in %, Anzahl der Lösungen (# Lösungen), die mittlere Lösungsgröße (Ø Lösungsgröße), die Anzahl der Backbones (# Backbones) und beim Vergleich der Lösungen mit den realen Daten der Non-Intersection-Fehler (Ø NonInErr), der Phasenfehler (Ø Phasenfehler) und der größte Phasenfehler (Max Phasenfehler).

4.2 Angiotensin Converting Enzyme-Gen

Allgemeines

Der Datensatz besteht aus 11 Individuen mit jeweils 52 SNPs, die von 5 Individuen afroamerikanischer Herkunft und 6 Individuen europäisch-amerikanischer Herkunft stammen. Die Sequenz wurde mittels allelspezifischer PCR ermittelt. Im Datensatz des ACE-Gens wurde das Alu-Indel in X (Alu-Insertion vorhanden) und Y (Insertion nicht vorhanden) codiert. Zwei weitere Teile des Datensatzes sind unterschiedlich häufige Wiederholungen von G und CT und kamen ebenfalls nur in zwei Zuständen vor. Auch hier wurde mit X (eine Form) und Y (zweite Form) codiert. Dadurch wurde wie in den anderen Datensätzen eine binäre Codierung mit 0 und 1 möglich.

Dieser Datensatz hat mit 11 Individuen die geringste Anzahl an Individuen. Der Anteil der heterozygoten SNPs ist mit 34% am größten. Auch kommt in diesem Datensatz der größte Block vor, der aber nur einen Anteil von 38% an der Gesamtlänge hat. Die durchschnittliche Blocklänge ist hier mit 6,5 SNPs am größten, die Anzahl der unterschiedlichen Haplotypen im gesamten Datensatz am geringsten.

Tabelle 3: Übersicht über den Datensatz ACE mit den datensatzspezifischen Maßen

Quelle	[Rieder et al., 1999]
Anzahl Genotypen	11
Anzahl SNPs	52
Anzahl SNPs nach Löschen	52
Anteil heterozygote SNPs in %	34
Anzahl Blöcke mit PP	8
Größter Block mit PP	20
Anzahl unterschiedlicher Haplotypen	14
Ø Blocklänge	6,5
Anteil des längsten Blocks an der Gesamtlänge in %	38

Biologische Relevanz

Das Angiotensin konvertierende Enzym ist ein Enzym, das eine wichtige Rolle bei der Regelung des Wasser-Elektrolythaushaltes und der Aufrechterhaltung des Blutdrucks spielt. Es wirkt dabei sowohl auf das Renin-Angiotensin-System wie auch auf das Kinin-Kininogen-System ein. Das Glykoprotein, das als Peptidase wirkt, katalysiert die Umwandlung von Angiotensin I in Angiotensin II durch die Abspaltung der zwei C-terminalen Aminosäuren. Angiotensin II wirkt gefäßverengend (vasokonstriktorisch) und führt so indirekt zu einem erhöhten Blutdruck. Die zweite Funktion des Enzyms wird bei der Abspaltung von zwei C-terminalen Aminosäuren des Hormons Bradykinin benötigt. Auch dieser Weg führt zu einer Erhöhung des Blutdruckes [Silbernagl, 2005].

Für das Angiotensin-Converting-Enzyme kodiert das Gen DCP1, welches auf dem Chromosom 17 liegt. Es enthält ein 287-kb-Indel (Insertion/Deletion: es ist entweder vorhanden oder nicht), das sogenannte „Alu“-Indel in Intron 16. Dies wird von einigen Forschern als Merkmal für eine höhere Empfindlichkeit für kardiovaskuläre Erkrankungen benannt [Rigat et al., 1990], [Mayer und Schunkert, 2000], [Soubrier et al., 1994]. Außerdem wird teilweise angenommen, dass das Auftreten dieses Indels mit dem Risiko für Alzheimer assoziiert ist [Arregui et al., 2006], andere Studien bestätigen dies allerdings nicht [Zhang et al., 2005).

Ergebnisse

Dieser Datensatz weist acht Blöcke unterschiedlicher Größe mit perfekter Phylogenie auf. Im längsten Block mit 20 SNPs gibt es auch die meisten Lösungen für diesen Datensatz, es sind aber nur zwei Lösungen. Sechs der acht Blöcke lassen sich vom Algorithmus eindeutig lösen, er findet selbst für den Block mit 11 SNPs nur eine Lösung und diese ist mit den realen Haplotyp-Daten identisch (Non-Intersection-Fehler und Phasenfehler sind 0). Nur bei zwei der acht Blöcke gibt der Algorithmus zwei mögliche Lösungen zurück.

Beim Vergleich der vom Algorithmus berechneten Lösungen und den realen Daten gibt es einen durchschnittlichen Non-Intersection-Fehler von 0,167 bzw. 0,083 bei den Blöcken 2 und 3. Der Wert von 0,083 zeigt, dass die vom Algorithmus berechneten Lösungen sehr nah an den realen Daten liegen und dass nur sehr wenige Haplotypen der berechneten Lösung nicht in den realen Daten vorliegen oder andersherum. Die Phasenfehler sind mit 0,143 und 0,048 eher gering, das heißt, dass nur sehr wenige heterozygote Positionen falsch gelöst wurden. Betrachtet man die Werte des Vergleiches genauer, so stellt man fest, dass es jeweils eine Lösung gibt, die die realen Daten wiedergibt und eine, die unterschiedlich dazu ist. Dargestellt ist dies in Tabelle 4.

Überraschend ist, dass trotz wesentlich weniger Individuen als beim Apolipoprotein E-Datensatz auch hier bis auf bei zwei Blöcken die Lösungen eindeutig sind, obwohl weniger Individuen die perfekte Phylogenie nicht so genau vorgeben wie viele Individuen.

Tabelle 4: Darstellung der Non-Intersection-Fehlers und des Phasenfehlers der zwei Lösungen für die Blöcke 2 und 3

Blocknummer		**2**	**3**
Lösung 1	Non-Intersection-Fehler	0	0
	Phasenfehler	0	0
Lösung 2	Non-Intersection-Fehler	0,333	0,167
	Phasenfehler	0,143	0,048

ACE

Blocklänge	11	4
% heterozygot	20,66	29,55
# Lösungen	1	2
Ø Lösungsgröße	8	5
# Backbones	8	4
	Vergleich mit realen Daten	
Ø NonInErr	0	0,167
Ø Phasenfehler	0	0,071
Max Phasenfehler	0	0,143

Blocklänge	20
% heterozygot	41,36
# Lösungen	2
Ø Lösungsgröße	6
# Backbones	5
	Vergleich mit realen Daten
Ø NonInErr	0,083
Ø Phasenfehler	0,024
Max Phasenfehler	0,048

Blocklänge	6	1	7	2	1
% heterozygot	36,36	27,27	32,47	40,91	54,55
# Lösungen	1	1	1	1	1
Ø Lösungsgröße	5	2	5	3	2
# Backbones	5	2	5	3	2
	Vergleich mit realen Daten				
Ø NonInErr	0	0	0	0	0
Ø Phasenfehler	0	0	0	0	0
Max Phasenfehler	0	0	0	0	0

Abbildung 8: Darstellung der Ergebnisse von ACE. Zur Erklärung der verwendeten Abkürzungen siehe Abbildungsunterschrift von Abbildung 7.

4.3 Kallikrein-Gene

Allgemeines

Diese beiden Datensätze enthalten fehlende Einträge, das heißt, an einigen Positionen ist unbekannt, welche Basen an diesen Positionen vorkommen. Zusätzlich gibt es Insertionen und Deletionen. Um mit diesen Datensätzen arbeiten zu können, wurden die darin enthaltenen „+“ und „-“ gelöscht. Diese stehen für Insertionen und Deletionen, stellen also keine biallelischen SNPs dar und werden nicht weiter mit betrachtet. Zusätzlich wurden die SNPs, die unbekannte Einträge enthalten, gelöscht, da diese nicht genug Informationen enthalten. Beide Datensätze wurden von 39 Individuen (20 afroamerikanische und 19 europäisch-amerikanische) durch die Erstellung von Hybrid-Zelllinien gewonnen [Andrés et al., 2007]. Tabelle 5 gibt eine Übersicht über die beiden Datensätze. Die Anzahl der SNPs ist hier am größten, auch nach dem Löschen der unbekannten Einträge. Die Anzahl der unterschiedlichen Haplotypen ist ebenfalls hier am größten. Mit 20% heterozygoten SNPs haben diese Datensätze den geringsten Anteil an heterozygoten Einträgen. Die längsten Blöcke haben hier nur einen Anteil von 11% bzw. 12% an der Gesamtlänge. Die durchschnittliche Blocklänge ist bei im KLK13-Datensatz mit 5,7 SNPs größer als im KLK14-Datensatz, der nur eine durchschnittliche Blocklänge von 3,8 SNPs aufweisen kann.

Tabelle 5: Übersicht über die Datensätze KLK13 und KLK14 mit den datensatzspezifischen Maßen

	KLK13	**KLK14**
Quelle	[Andrés et al., 2007]	[Andrés et al., 2007]
Anzahl Genotypen	39	39
Anzahl SNPs	190	156
Anzahl SNPs nach Löschen	108	73
Anteil heterozygote SNPs in %	20	20
Anzahl Blöcke mit PP	19	19
Größter Block mit PP	12	9
Anzahl unterschiedlicher Haplotypen	70	59
Ø Blocklänge	5,7	3,8
Anteil des längsten Blocks an der Gesamtlänge	0,11	0,12

Biologische Relevanz

Kallikrein ist eine Serinprotease, das heißt es enthält Serin in seinem aktiven Zentrum und kann Proteine spalten. Es überführt inaktive Vorläufer der Gewebshormone (Kininogene) in ihre aktive Form (Kinine). Das System aus Kallikrein und Kininogenen arbeitet ähnlich wie das Renin-Angiotensin-System (siehe biologische Relevanz von Angiotensin Converting Enzyme) und hat eine Funktion in der Blutdruckregulation und der Elektrolyt- und Wasserhomöostase [Wikipedia, 2008]. Die untersuchten Gene KLK13 und KLK14 liegen zusammen mit dreizehn anderen KLK-Genen auf Chromosom 19 in einem Cluster vor. Die Expression von KLK13 wird von Steroiden reguliert und das Gen wird als möglicher Marker für Brustkrebs in Betracht gezogen. Die Kallikreine werden immer häufiger als Marker für Tumorentstehung und als andere Biomarker für Krankheiten in Betracht gezogen [NCBI Entrez Gene, 2009].

Ergebnisse

Der Datensatz des KLK13-Gens enthält 19 Blöcke mit perfekter Phylogenie. Die Größe der Blöcke variiert stark, es kommen viele kleine Blöcke mit einer Länge von 1 und 2 SNPs aber auch ein Block mit 12 SNPs (11% der Gesamtlänge), einer mit 11 SNPs und drei mit einer Länge von 10 SNPs vor. Diese größeren Blöcke haben einen Anteil von 49% an der Gesamtanzahl der SNPs in diesem Datensatz. Die mittlere Blocklänge beträgt 5,7 SNPs. Es gibt Bereiche, in denen gehäuft kleine Blöcke mit einer Länge von 1, 2 oder 3 SNPs vorkommen, dann wieder Bereiche in denen größere Blöcke auftreten. Man sieht die Tendenz, dass mit zunehmender Blocklänge die Anzahl der Lösungen größer wird. Bis zu einer Blocklänge von 3 SNPs gibt es hier nur eine Lösung, bei einer Länge von 10 SNPs variiert die Anzahl der Lösungen von 8 über 32 bis 64 Lösungen. Dies muss dazu noch von anderen Faktoren abhängen, als der Art, wie die heterozygoten Positionen innerhalb des Blockes verteilt sind. Denn der Anteil der drei 10er Blöcke ist mit 10,5%, 10,8% und 11,3% sehr ähnlich und trotzdem gibt es einen großen Unterschied in der Anzahl der Lösungen. Ein Überblick über die berechneten Maße ist in Abbildung 11 dargestellt. In diesem Datensatz kommen mehrere Blöcke vor, die viele Lösungen aufweisen, zum Beispiel Block 9 mit 16 Lösungen, Block 11 mit 64 Lösungen und Block 13 mit 32 Lösungen. Bei diesen vielen Lösungen ist die Verschiedenheit der einzelnen Lösungen interessant. Hierzu wurde für Block 11 ein Histogramm erstellt, das in Abbildung 9 zu finden ist. Es zeigt, dass der Non-Intersection-Fehler stark variiert, aber nicht größer als 0,6 wird. Außerdem ist festzustellen, dass es mehr Lösungen gibt, die einen großen Non-Intersection-Fehler haben, als Lösungen, die mit den realen Haplotypen besser übereinstimmen. Die meisten Werte für den Non-Intersection-Fehler liegen zwischen 0,4 und 0,5, dies bedeutet, dass 40% bis 50% der

Haplotypen der berechneten Lösung nicht mit den Haplotypen der realen Lösung übereinstimmen.

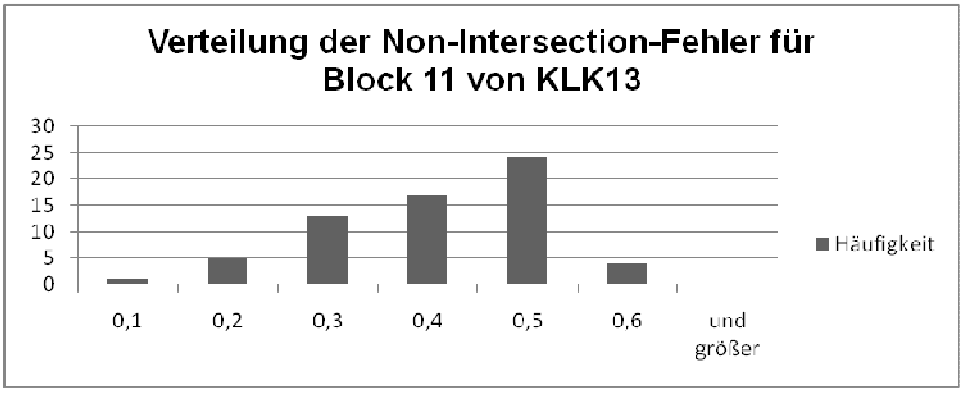


Abbildung 9: Histogramm für den Non-Intersection-Fehler für Block 11 von KLK13 mit 64 Lösungen. Auf der x-Achse sind die Intervalle in Schritten einer Größe von 0,1 aufgetragen, auf der y-Achse ist die Häufigkeit der Werte in diesem Bereich eingetragen. In der Spalte mit dem Wert 0,2 sind alle die Werte zusammengefasst, die größer als 0,1 aber kleiner oder gleich 0,2 sind.

Betrachtet man das gleiche nun für den Phasenfehler, so stellt sich dort ein etwas anderes Bild dar. Das Histogramm dazu befindet sich in Abbildung 10. Hier gibt es wenig sehr große Fehler, die größer als 0,5 sind. Aber schon ein Phasenfehler von 0,3 bedeutet, dass im Schnitt fast 30% der heterozygoten Positionen falsch gelöst worden sind. Es gibt jedoch hier immerhin 5 von 64 Lösungen, die nur einen Phasenfehler von maximal 0,1 aufweisen. Diese Lösungen sind also den realen Daten am ähnlichsten.

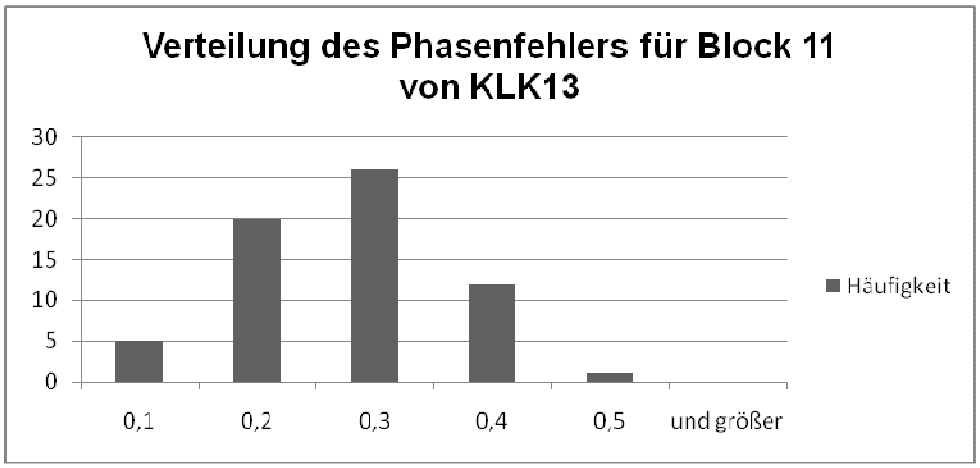


Abbildung 10: Histogramm für den Phasenfehler für Block 11 von KLK13 mit 64 Lösungen. Auf der x-Achse sind Intervalle der Größe 0,1 eingetragen, auf der y-Achse die Häufigkeit mit der die Phasenfehler in diesem Bereich auftraten. In der Spalte mit dem Wert 0,2 sind alle die Werte zusammengefasst, die größer als 0,1 aber kleiner oder gleich 0,2 sind.

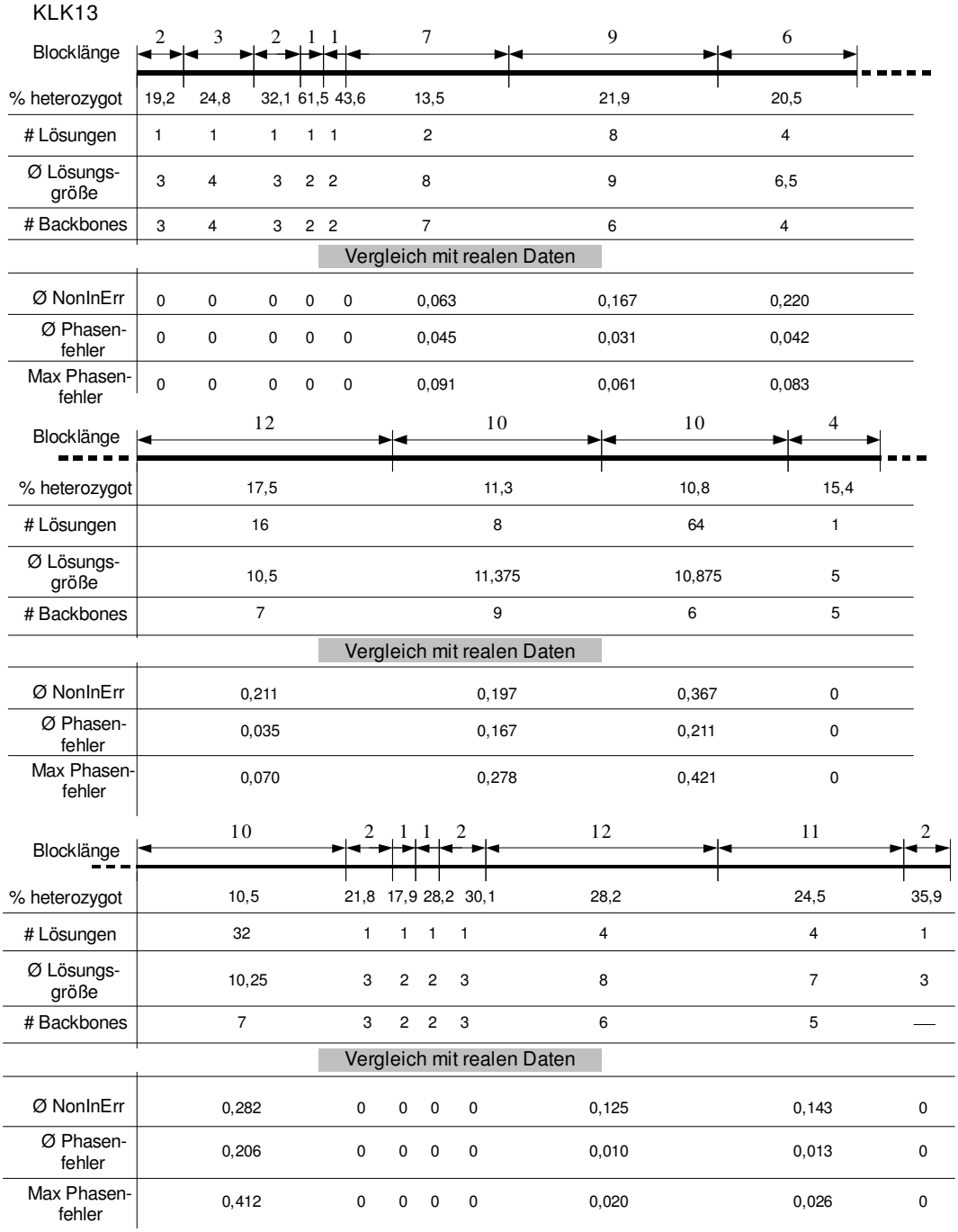


Abbildung 11: Darstellung der Ergebnisse von KLK13. Für die Erklärung der Abkürzungen siehe Abbildung 7.

Für den Datensatz KLK14, dessen Ergebnisse in Abbildung 14 dargestellt sind, wurden 19 Blöcke unterschiedlicher Länge gefunden. Der längste Block ist nur 9 SNPs lang, dies entspricht einem Anteil von 12% an der Gesamtlänge des Datensatzes. Wie in Abbildung 14 zu sehen ist, gibt es hier zwei Blöcke, bei denen der maximale Phasenfehler bei 1 liegt, bei Block 8 mit einer Länge von 2 SNPs wie auch bei Block 16 mit einer Länge von 5 SNPs.

Bei diesen Blöcken wurden mit einer von 2 bzw. 8 Lösungen alle heterozygoten Positionen komplett falsch gelöst. In Block 16 ist auch der durchschnittliche Non-Intersection-Fehler sehr groß. Betrachtet man hier auch die Verteilung der Werte des Non-Intersection-Fehlers wie schon beim Datensatz KLK13 in Block 11, so sind hier Sprünge zu beobachten. Auffällig ist, dass auch hier die meisten (4 von 8) Werte im größten Bereich von 0,6 bis 0,7 liegen, die meisten Lösungen also stark abweichen. Dargestellt ist dies in Abbildung 12.

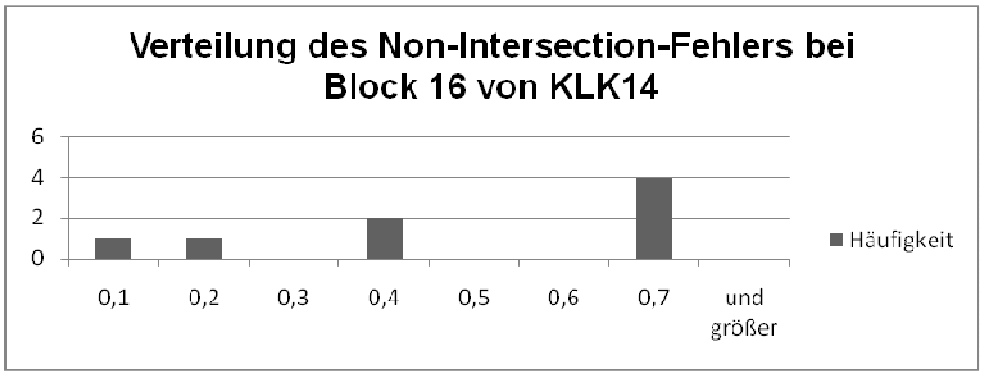


Abbildung 12: Histogramm für den Non-Intersection-Fehler für Block 16 von KLK14 mit 8 Lösungen. Auf der x-Achse sind die Intervalle in Schritten einer Größe von 0,1 aufgetragen, auf der y-Achse ist die Häufigkeit der Werte in diesem Bereich eingetragen.

Betrachtet man nun den Phasenfehler für diesen Block, so sieht man, dass auch diese Werte weit verteilt sind. Ein Viertel der Lösungen hat einen geringen Phasenfehler, aber ein Viertel der Lösungen hat auch einen sehr großen Phasenfehler mit einem Wert von fast 1, das heißt fast alle bis alle heterozygoten Positionen wurden hier falsch gelöst. Gezeigt wird dies in Abbildung 13. Bei 6 von 8 Lösungen wurden hier mindestens 50% der heterozygoten Positionen falsch berechnet.

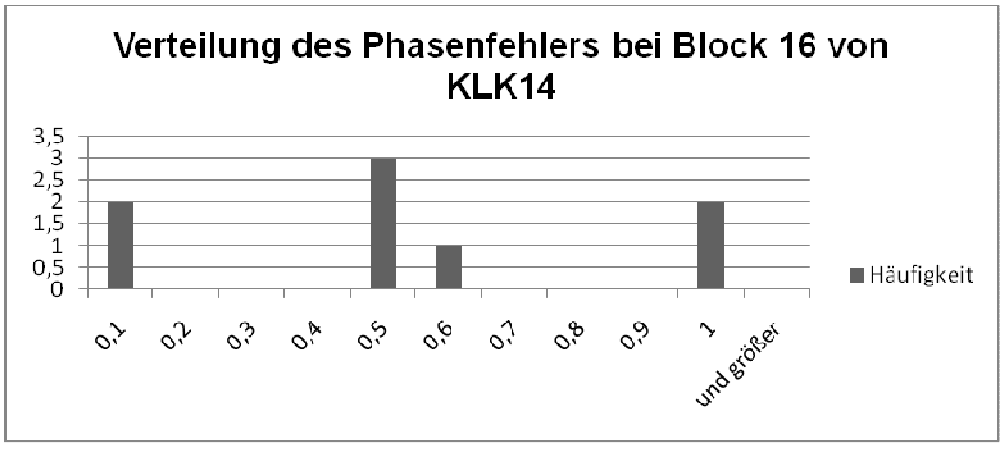


Abbildung 13: Histogramm für den Phasenfehler für Block 11 von KLK13 mit 64 Lösungen. Auf der x-Achse sind Intervalle der Größe 0,1 eingetragen, auf der y-Achse die Häufigkeit mit der die Phasenfehler in diesem Bereich auftraten.

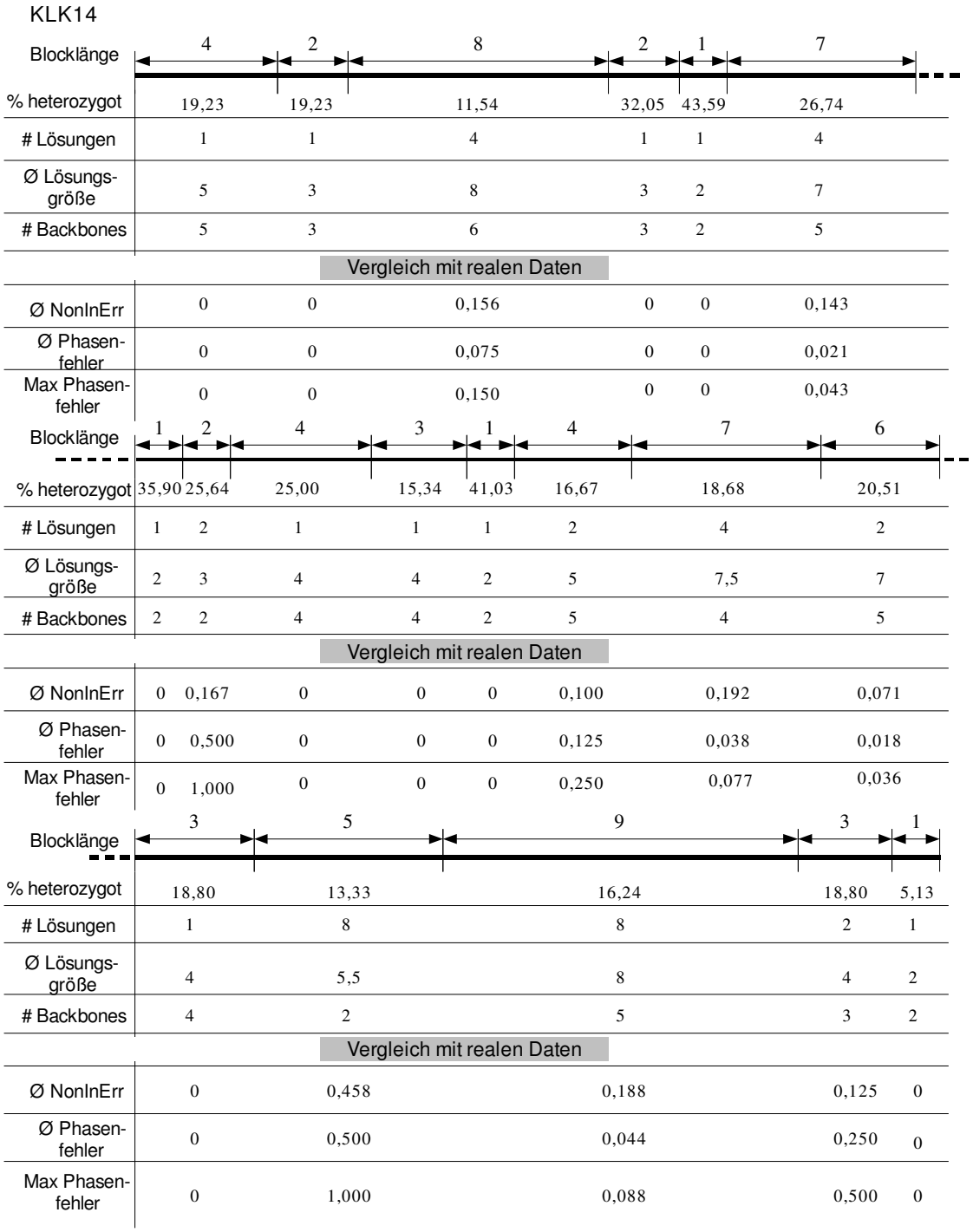


Abbildung 14: Darstellung der Ergebnisse von KLK14. Für die Erläuterungen der Abkürzungen siehe Abbildung 7.

Bei diesem Datensatz sind die Blöcke sehr klein, im Durchschnitt nur 3,8 SNPs lang, das heißt, der Datensatz ist sehr zerteilt. Weil die Blöcke so klein sind, gibt es auch nur wenige Lösungen für einen Block.

5 Diskussion und Ausblick

In der vorliegenden Arbeit wurden für die vier verwendeten Datensätze aufeinanderfolgende Blöcke an SNPs berechnet, für die es für die Haplotypen der realen Daten eine perfekte Phylogenie gibt. Mit der Kenntnis dieser Blöcke wurden für die aus den Haplotyp-Daten berechneten Genotypen die möglichen zugrunde liegenden Haplotypen berechnet. Für diese berechneten Lösungen wurden die in Kapitel 3.2 definierten Maße berechnet. Dabei wurden für die unterschiedlichen Datensätze verschiedene Beobachtungen gemacht. So scheint die Anzahl der Individuen im Fall des Datensatzes des Apolipoprotein E-Gen die Eindeutigkeit der Lösung zu fördern. Bei Datensätzen mit weniger Individuen ist die Länge des Blockes für die Lösungsgröße anscheinend entscheidend, denn es lässt sich bis auf die Ausnahme eines 20 SNPs langen Blocks beim Datensatz des Angiotensin-Converting-Enzyme-Gens eine wachsende Tendenz der Anzahl der Lösungen in Korrelation zur Blockgröße feststellen.

Der Anteil der heterozygoten SNPs am gesamten Datensatz scheint keine große Auswirkung auf die Anzahl der Lösungen zu haben und auch nicht auf die Genauigkeit der Lösungen, denn diese Werte schwanken sehr stark und eine Korrelation lässt sich nicht erkennen. Dies müsste aber noch genauer untersucht werden, um einen möglichen Zusammenhang aufzudecken, denn tendenziell sollten mehr heterozygote SNPs in einer Population die Schwierigkeit erhöhen, die zugrunde liegenden Haplotypen für die Genotypen zu bestimmen. Bei dem Datensatz KLK13 ist dies zumindest nicht der Fall. Dieser Datensatz hat mit einem Wert von ca. 20% für den gesamten Datensatz den niedrigsten Anteil an heterozygoten Positionen, weist aber selbst in Blöcken, wo der Anteil bei ca. 11% liegt, eine große Anzahl an Lösungen auf.

Desweiteren ist festzustellen, dass bei einer großen Anzahl von Lösungen die Anzahl der Backbones im Vergleich zu der mittleren Lösungsgröße geringer ist. Dies zeigt, dass die zugrunde liegenden Haplotypen, die für die Genotypen bestimmt wurden, nicht in jeder Lösung identisch sind und auch deren Anzahl nicht. Denn auch die Lösungsgrößen variieren bei einer großen Anzahl von Lösungen stärker als bei einer kleinen Anzahl von Lösungen.

Die Genauigkeit der Haplotypisierung aufgrund der Perfekte-Phylogenie-Annahme lässt sich nicht allgemein quantitativ auswerten. Grundsätzlich lassen sich für kleine Blöcke bis zu einer Länge von 4 SNPs die korrekten zugrundeliegenden Haplotypen für die Genotypen eindeutig bestimmen, aber auch schon bei einer Größe von 2 SNPs kann es Ausnahmen geben, wie beim Block 8 des Datensatzes KLK14 mit zwei Lösungen für eine Blocklänge von zwei SNPs.

Bei großen Blöcken schwankt die Genauigkeit der Haplotypisierung sehr. Es gibt Blöcke, in denen liegt der maximale Phasenfehler nur bei 0,02 wie bei Block 7 des Datensatzes KLK13, bei anderen Blöcken ähnlicher Größe liegt der maximale Phasenfehler aber bei 0,206 (Block 13 von KLK13), also um Faktor 10 höher, obwohl die Länge der beiden Blöcke sich nur um einen SNP unterscheidet. Hier müssen also noch andere Faktoren eine Rolle spielen.

Für eine Verallgemeinerung der Aussagen müssten noch genauere Auswertungen gemacht werden und es wäre sinnvoll, diese Daten vergleichbar zu machen, denn aufgrund der unterschiedlichen Blockgröße ist ein Vergleich nicht immer sinnvoll. Außerdem wäre es interessant, welche Werte sich ändern würden, wenn bei den Kallikrein-Datensätzen die unbekannten Einträge mit betrachtet würden. Dies konnte aufgrund des eingeschränkten Zeitrahmens der Bachelorarbeit nicht geleistet werden.

Literaturverzeichnis

[Andrés et al., 2007] Aida M. Andrés, Andrew G. Clark, Lawrence Shimmin, Eric Boerwinkle, Charles F. *Understanding the accuracy of statistical haplotype inference with sequence data of known phase.* [Article] // Genetic Epidemiology. - November 2007. - 31(7):659-71.

[Arregui et al., 2006] Alberto Arregui, Elaine K. Perry, Martin Rossor and Bernard E. Tomlinson *Angiotensin-converting enzyme in Alzheimer's disease: Increased activity in caudate nucleus and cortical areas.* [Article] // Journal of Neurochemistry. - 2006. - 38.

[Climer et al., 2009] Sharlee Climer , Gerold Jäger, Alan R. Templeton 4 and Weixiong Zhang *How frugal is mother nature with haplotypes?* [Article] // Bioinformatics. - Januar 1, 2009. - pp. 25(1):68-74.

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

[Gusfield, 2004] Dan Gusfield, In Sorin Istrail, Michael S. Waterman, and Andrew G. Clark, editors *An overview of combinatorial methods for haplotype inference.* [Article] // Revised Papers, volume 2983 of Lecture Notes in Computer Science. - 2004. - pp. 9–25.

[Mayer und Schunkert, 2000] Björn Mayer und Heribert Schunkert *ACE-Gen polymorphism and cardiovascular diseases* [Article] // Herz. : Urban&Vogel, 2000. - Volume 25.

[NCBI Entrez Gene, 2009] NCBI Entrez Gene [Online]. - August 2, 2009. - August 28, 2009. - *http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=26085.*

[Orzack et al., 2003] Steven Hecht Orzack, Daniel Gusfield, Jeffrey Olson, Steven Nesbitt, Lakshman Subrahmanyan, and Vincent P. Stanton, Jr. *Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes.* [Article] // Genetics. - Oktober 2003. - pp. 165(2):915-28..

[Patil et al., 2001] 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, et al. *Blocks of limited haplotype diversity revealed by highresolution* [Article] // Science. - 2001. - pp. 1719–1723.

[Rieder et al., 1999] Mark J. Rieder, Scott L. Taylor, Andrew G. Clark & Deborah A. Nickerson *Sequence variation in the human angiotensin converting enzyme.* [Article] // Nature Genetics. - Mai 1999. - pp. 22(1):59-62..

[Rigat et al., 1990] B. Rigat, C. Hubert, F. Alhenc-Gelas, F. Cambien, P. Corvol and F. Soubrier *An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. [Article] // The Journal of Clinical Investigation.* - 1990.

[Schmidt, 2004] Robert F. Schmidt *Zelluläre und molekulare Ursachen des Alterns* [Book Section] // Physiologie des Menschen / book auth. Robert F. Schmidt Florian Lang, Gerhard Thews. - Würzburg, Tübingen : Springer Medizin Verlag, 2004.

[Silbernagl, 2005] Stefan Silbernagl *Renin und Nierenhormone [*Book Section] // Physiologie / book auth. Rainer Klinke Hans-Christian Pape, Stefan Silbernagl. - Franktfurt am Main, Münster, Würzburg : Georg Thieme Verlag, 2005.

[Soubrier et al., 1994] F. Soubrier, S. Nadaud and T. A. Williams *Angiotensin I Converting Enzyme Gene: Regulation, Polymorphism and Implications in Cardiovascular Diseases* [Article] // European Heart Journal. - 1994.

[The International HapMap Consortium, 2003] The International HapMap Consortium *The international HapMap project* [Article] // Nature. - 2003. - pp. 789–796.

[Wikipedia, 2008] Wikipedia [Online]. - Dezember 28, 2008. - September 1, 2009. - *http://de.wikipedia.org/wiki/Kallikrein.*

[Wikipedia, 2009] Wikipedia [Online]. - August 3, 2009. - September 1, 2009. - *http://en.wikipedia.org/wiki/Apolipoprotein_E.*

[Zhang et al. 2005] Jun-Wu Zhang, Xiao-Qing Li, Zhen-Xin Zhang, Deng Chen, Hua-Lu Zhao, Ya-Ning Wu, Qiu-Ming Qu *Association between angiotensin-converting enzyme gene polymorphism and Alzheimer's disease in a Chinese population.* [Article] // Dementia and Geriatric cognitive disorders. - 2005. - pp. 20(1):52-6.