Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Skalierbare Indexierung humaner Mutationsprofile durch Inverted Files

Exposé zur Diplomarbeit

eingereicht von: Sascha Baese
geboren am: 07.04.1986
geboren in: Berlin

Betreuer/innen: Prof. Dr. Ulf Leser
Stefan Sprenger

eingereicht am: 5. November 2015

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit der Suche von Mutationen in menschlichen Genomen, da in diesen der Schlüssel zum Erkennen von Krankheitsrisiken gesehen wird. Ein menschliches Genom besteht aus ca. 3 Milliarden Nukleotiden [1], die auf 23 Chromosome aufgeteilt und über ihre Position adressiert werden. Vergleicht man zwei menschliche Genome miteinander, unterscheiden sie sich in etwa 0,1% der Nukleotide. Ein Großteil dieser Unterschiede besteht in Einzelnukleotid-Variationen (SNP's) [2]. Die Entschlüsselung von Genomen und deren Vergleich wird durch biotechnische Sequenzierungsmethoden ermöglicht. Als alte und etablierte Methoden gelten beispielsweise die Sanger- und die Maxim-Gilbert-Methode [3]. Durch neuere, als *next-generation sequencing* bezeichnete, Methoden, konnten die enormen Kosten der älteren Gensequenzierungsmethoden erheblich gesenkt und die Geschwindigkeit der Sequenzierung zudem stark beschleunigt [4] werden.
Durch die Verwendung immer performanterer, kostengünstiger Sequenzierungsmethoden steigt die Menge an auswertbaren Daten stetig. Die Auswertung selbst muss daher ebenfalls immer effizienter werden, was Ziel dieser Arbeit ist. Sequenzierungsdaten sind Grundlage für Mutationsforschung und die Entschlüsselung der Funktionen des Erbguts und einzelner Abschnitte (siehe Abschnitt 2) und sie sind für die aktuelle Forschung interessant, weil die enthaltenen Mutationsdaten mit diversen Krankheiten assoziiert sein bzw. auf ein höheres Risiko für eine bestimmte Krankheit weisen können [3]. Forschungsprojekte, wie das 1000 Genomes Project (siehe Abschnitt 2.1.3) erheben und vergleichen die Gensequenzen ausgewählter menschlicher Individuen.
In dieser Arbeit soll eine Suche über die Variationsdaten einer großen Anzahl an menschlichen Genomen implementiert und für die im Bereich der Variationsanalyse gängigen Abfragetypen implementiert werden. Die Variationsdaten des 1000 Genomes Projects werden exemplarisch verwendet (siehe Abschnitt 3). Für die 2.504 Samples müssen insgesamt 7,5 Milliarden Datenpunkte durchsucht werden, wenn jedes Sample im Durchschnitt 3 Millionen Mutationen besitzt.

2 Forschungsstand

Dieser Abschnitt stellt ausgewählte Forschungsprojekte sowie Werkzeuge zur Visualisierung von Genomdaten vor, die eines der Hauptanwendungsgebiete der hier betrachteten Anfragetypen sind. Anschließend wird die Gendatenbank RefSeq vorgestellt. Diese Datenbank ermöglicht, dass im implementierten Tool die Suche auch unter Angabe von Genbezeichnern erfolgen kann.

2.1 Forschungsprojekte am menschlichen Genom

Innerhalb der letzten 25 Jahre gab es große Erfolge im Bereich der Sequenzierung des menschlichen Genoms. Dieses Kapitel stellt mit dem *Human Genome Project* die Grundlage heutiger Forschungsprojekte vor und schließt mit dem *1000 Genomes Project*

ab, welches die für diese Arbeit genutzten Daten bereitstellt.

2.1.1 The Human Genome Project

1990 startete das Human Genome Project, welches sich die Sequenzierung des gesamten menschlichen Genoms zur Aufgabe machte [5]. Zu dieser Zeit herrschte die Annahme, dass das menschliche Genom 30.000 bis 100.000 Gene besitzt. Diese sollten durch das Projekt identifiziert werden. Ein weiteres Ziel des Projekts war das Auffinden der für Erbkrankheiten verantwortlichen Stellen im Genom. Dadurch erhoffte man sich Ansätze für deren Behandlung. Ebenso wollte man die evolutionäre Prozesse des menschlichen Erbguts durch den Vergleich mit Modellorganismen nachvollziehen.
Im April 2003 gaben die am Human Genome Project beteiligten Forscher die Fertigstellung der Sequenzierung bekannt [6]. Durch die vom Projekt bereitgestellten Daten wurde ermittelt, dass schätzungsweise nur 20.000 bis 25.000 Gene im menschlichen Genom existieren [7].

2.1.2 The International HapMap Project

Das International HapMap Project startete im Jahr 2002 [2]. Ziel war das Auffinden der Stellen im Genom, an denen Mutationen für häufige komplexe Krankheiten wie Schlaganfälle, Diabetes, psychische Erkrankungen oder Adipositas verantwortlich sind. Das Risiko für diese Krankheiten ist von verschiedenen Mutationen abhängig und daher schwer abzuschätzen.
Das International HapMap Project stellt Kopien der gesammelten Zellkulturen zur freien Verfügung [8], weshalb diese auch von anderen Projekten wie dem 1000 Genomes Project genutzt wurden.

2.1.3 The 1000 Genomes Project

Das 1000 Genomes Project, welches 2008 gestartet wurde, hatte das Ziel 2.500 menschliche Genome zu sequenzieren [9]. Bis 2015 wurden insgesamt 2.504 Genome aus Nord- und Südamerika, Europa, Ostasien, Südasien und Afrika sequenziert [10]. Laut [9] entstammen 681 der sequenzierten Genome aus der 3. Phase des International HapMap Project (siehe Abschnitt 2.1.2). Insgesamt wurden hierbei allerdings mehr Genome gesammelt, als sequenziert, um die Anonymität der Testpersonen zu wahren.
Die vorliegende Arbeit nutzt einen Teil der erhobenen Daten dieses Projekts (siehe Abschnitt 3). Diese Daten sind im Variant Call Format (VCF) gespeichert [11]. Die Mutationsdaten werden nach Chromosom getrennt in separaten Dateien gespeichert. Jede Datei enthält die Informationen für alle untersuchten Individuen. Es ist jedoch nicht das gesamte, sequenzierte Chromosom je Individuum abgelegt, sondern ausschließlich Variationen. Die Variationen werden anhand eines Referenzgenoms ermittelt. Dieses Referenzgenom wurde zuvor durch Mitglieder des 1000 Genomes Project Consortiums bestimmt.
Je Polymorphismus erfolgt unter anderem eine Angabe für die Position im Genom, die

Sequenz im Referenzgenom und die Auflistung aller möglichen auftretenden Alternativen in den Individuen [12]. Die Häufigkeiten der Alternativen werden nicht aufgeführt. Sie können aber mit Hilfe bereitgestellter Tools errechnet werden.
Zusätzlich zu den VCF-Daten liegen Informationen zu Geschlecht, Nationalität und verwandtschaftlichen Beziehungen der Individuen vor.

2.2 Tools zur Visualisierung von Genominformationen

Die im vorherigen Abschnitt vorgestellten Forschungsprojekte haben mehrere hunderte Gigabyte an Daten produziert. Die Verarbeitung und Visualisierung der Daten ist eine komplexe Aufgabe. Zwei Projekte, die große Erfolge auf diesem Gebiet verzeichnen, werden im Folgenden vorgestellt.

2.2.1 Tabix

Tabix ist ein Tool, welches Tab-getrennte, positionssortierte Dateien mit biologischen Daten indexiert [13]. Hierzu gehören unter anderem VCF-Dateien. Das Tool wurde im Jahr 2011 von Heng Li vorgestellt und greift auf lokale oder externe Dateien zu. Die Daten werden sortiert, im BGZF-Format komprimiert und anschließend als gzip-Datei gespeichert. Im Verhältnis zu den unkomprimierten Daten verringert sich der Speicherbedarf der komprimierten Dateien um einen Faktor von 3 – 5. Zusätzlich wird eine Indexdatei erstellt. Sobald ein Index erstellt wurde, kann man in der komprimierten Datei nach Positionsbereichen suchen. Eine Einschränkung auf bestimmte Samples ist allerdings nicht möglich. Tabix benötigt für die Suche die Angabe des Chromosoms und des gewünschten Positionsbereichs. Anschließend werden alle Zeilen des VCF-Dokuments ausgegeben, welche die Query erfüllen.

2.2.2 Integrative Genomics Viewer

Der Integrative Genomics Viewer (IGV) ist ein Tool, welches 2007 entwickelt wurde, um die Daten des *Cancer Genome Atlas* zu visualisieren [14]. Das Tool ist in Java geschrieben und kann mit verschiedenen Dateiformaten, wie dem VCF-Format (siehe Abschnitt 2.1.3), umgehen. Das Tool bietet die Möglichkeit der Betrachtung gesamter Genome bis hin zu einzelnen Basenpaaren. Ebenso können mehrere Regionen gleichzeitig betrachtet und somit gegenübergestellt werden. Um einen Wechsel der Auflösungsstufen in Echtzeit zu ermöglichen, wird das so genannte *data tiling* eingesetzt. Hierbei werden die groben Daten verschiedener Auflösungsstufen des Genoms vorberechnet. Während eine genaue Vorberechnung aller Auflösungsstufen einen hohen Speicherplatzbedarf nach sich zieht, wird dieser durch data tiling klein gehalten. IGV speichert daher eine vordefinierte Anzahl grober Auflösungsstufen abhängig von der Größe des zugrundeliegenden Datensatzes. Die Berechnung der feineren Auflösungsstufen erfolgt in Echtzeit. Dies ist möglich, da nur kleine Genabschnitte betrachtet werden.

2.3 RefSeq

RefSeq ist eine Datenbank des National Center for Biotechnology Information (NCBI) [15]. Sie beinhaltet unter anderem Gennamen aus Genomen von 3.774 Organismen, einschließlich des Menschen. Die Gennamen sind den entsprechenden Positionen im Genom zugeordnet. Diese Informationen werden im Index des in dieser Arbeit implementierten Tools hinterlegt und sind somit abrufbar.

3 Ziel der Arbeit

Durch das im Rahmen dieser Arbeit entwickelte Programm sollen Variationen in VCF-Daten gesucht werden. Als Ergebnis der Anfragen wird eine VCF-Datei erzeugt, welche dann in beispielsweise IGV ausgewertet werden kann. Für die durchgeführten Experimente werden die Daten des 1000 Genomes Projekts (siehe Abschnitt 2.1.3) genutzt und anhand verschiedener Filter durchsuchbar gemacht:

(a) *Mutationsanfragen:* Variationen werden über festgelegte Positionsintervalle gesucht. Anstatt eines Positionsbereichs können auch Gennamen angegeben werden, die aus der RefSeq-Datenbank (siehe Abschnitt 2.3) eingelesen wurden. Als Ergebnis werden alle Mutationen für die einzelnen Positionen der Intervalle zurückgegeben. Wir suchen beispielsweise alle Mutationen zwischen den Positionsbereichen 16.050.075 und 16.050.646 auf Chromosom 22.

(b) *Sampleanfragen:* Es werden Samples mit bestimmten Attributen gesucht. Die Abfragen sind durch Konjunktion, Disjunktion und Negation kombinierbar. Zur Auswahl stehen hierbei Geschlecht, Nationalität und verwandtschaftliche Beziehungen der Individuen. Beispielsweise suchen wir nach allen männlichen Individuen aus Großbritannien. Mutationen spielen bei dieser Abfrage keine Rolle.

(c) *Verknüpfung von (a) und (b):* Gesucht werden Samples mit bestimmten Attributen über festgelegte Positionsbereiche. Als Operatoren sind Konjunktionen zwischen (a) und (b) zulässig. Es lassen sich beispielsweise alle Mutationen zwischen den Positionsbereichen 16.050.075 und 16.050.646 von männlichen Individuen aus Großbritannien filtern.

Die erstellte Suchlösung wird anschließend in Experimenten evaluiert (siehe Abschnitt 4).

4 Vorgehensweise

Die VCF-Daten werden mithilfe von *Lucene* indexiert. Lucene ist eine auf Java basierte Suchbibliothek [16]. Diese wurde 1997 von Doug Cutting entwickelt und gehört seit 2005 zur Apache Foundation. Durch Lucene lassen sich diverse Daten indexieren, wobei die erzeugten Indexe als *inverted files* erzeugt werden. Diese richten sich nicht nach den

zugrundeliegenden Dokumenten, sondern nach den Termen innerhalb der Dokumente. Jeder Term verweist hierbei auf die Dokumente, in denen er vorkommt.
Das in dieser Arbeit erstellte Tool soll, wie Tabix, auf Kommandozeilenebene arbeiten und VCF-Daten erzeugen. In diesen Daten kommt jede Position einmalig vor und wird in einer separaten Zeile erfasst. Zu jeder Position werden die Mutationen der Samples aufgelistet. Daher wird für den Index jede Zeile als eigenständiges Dokument betrachtet. Jedes Dokument enthält Position und Samplemutationen als Terme. Erfolgt nun beispielsweise eine Bereichsabfrage, werden die Terme, die gemeinsam mit der Position in einem Dokument liegen, ausgegeben. Für die Erstellung des Index gibt es mehrere Überlegungen:

(a) Ein Index für jeden Chromosomendatensatz. Jeder Index soll alle allgemeinen Informationen für die Position einer Mutation enthalten. Für die folgenden Spalten sollen nur die Informationen der Individuen mit Polymorphismus gespeichert werden. Ein weiterer Index soll die vom 1000 Genomes Project zur Verfügung gestellten Attribute für alle Individuen enthalten. Es handelt sich hierbei um Geschlecht, verwandtschaftliche Beziehungen und Herkunft des jeweiligen Individuums.

(b) Ein gemeinsamer Index für alle Chromosomen, der zudem die Attribute zu den Samples enthält.

(c) Ein gemeinsamer Index für alle Chromosomen und ein separater Index für die Attribute zu den Samples.

Die Evaluation erfolgt gegen tabix und CVCI. CVCI ist ein Werkzeug, das am Lehrstuhl für Wissensmanagement in der Bioinformatik an der Humboldt-Universität zu Berlin entwickelt wurde und für die Indexierung von Mutationsdaten genutzt werden kann.
Durch Stresstests und Skalierungsabfragen sollen Stärken und Schwächen der Tools abgegrenzt werden:

(a) *Stresstests:* Mehrere Instanzen des Programms werden gestartet, um den Index parallel abzufragen. Hierbei soll festgestellt werden ab welcher Last das System einbricht.

(b) *Skalierungsabfragen:* Es werden verschiedene einfache und komplexe Suchanfragen zwischen den vorgeschlagenen Indexierungsmethoden und Tabix verglichen. Interessant ist vor allem die Geschwindigkeit des Systems in Abhängigkeit von der Indexgröße. Vor den Abfragen werden zufällig gewählte Intervallgrenzen generiert, damit die Queries auf den Tools identisch sind. Anschließend werden die Ausführungsgeschwindigkeit der Suchanfragen und die Indexgrößen verglichen.

Das 1000 Genomes Projekt gewährt Zugriff auf Mutationen von 2504 Samples. Da zu erwarten ist, dass die Anzahl Samples mit der Zeit stetig wächst, werden beide Tests zusätzlich mit wesentlich erhöhter Sampleanzahl wiederholt. Beobachtet wird hierbei

die Antwortzeit zufällig generierter VCF-Daten mit ca. 10 mal so vielen Samples wie vom 1000 Genomes Projekt zur Verfügung gestellt.

Literatur

[1] N. E. Morton, "Parameters of the human genome." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, pp. 7474–7476, 1991.

[2] The International HapMap Consortium, "The International HapMap Project." *Nature*, vol. 426, pp. 789–796, 2003.

[3] J. Graw, *Genetik*, 5th ed. Springer, 2010.

[4] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nat Biotechnol*, vol. 26, no. 10, pp. 1135–1145, 2008.

[5] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro, "Human Genome Project," *The American Journal of Surgery*, vol. 165, pp. 258–264, 1993.

[6] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: lessons from large-scale biology." *Science*, vol. 300, pp. 286–290, 2003.

[7] The International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 2, pp. 931–945, 2004.

[8] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs." *Nature*, vol. 449, pp. 851–861, 2007.

[9] The 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, 2012.

[10] P. H. Sudmant, T. Rausch, E. J. Gardner, and A. Others, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, pp. 75–81, 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature15394

[11] L. Clarke, X. Zheng-Bradley, R. Smith, *et al.*, "The 1000 Genomes Project: data management and community access," *Nature Methods*, vol. 9, no. 5, pp. 459–462, 2012.

[12] P. Danecek, A. Auton, G. Abecasis, *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[13] H. Li, "Tabix: Fast retrieval of sequence features from generic TAB-delimited files," *Bioinformatics*, vol. 27, no. 5, pp. 718–719, 2011.

[14] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, 2012.

[15] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, 2007. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl842

[16] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action*, 2nd ed. Manning Publications, 2010.