Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung AW1 WS2011/2012

Jan-Christoph Meier

Datenanalyse mit Data Mining

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhaltsverzeichnis

1 Einleitung

Moderne Informationssysteme speichern, verwalten und analysieren sehr große Mengen an Daten, wobei das Datenaufkommen stetig steigt und permanent immense Mengen generiert werden. So ist zum Beispiel die Menge an Daten die Google, Facebook oder Amazon speichern und analysieren müssen, extrem hoch. Ziel der Datenanalyse ist es, aus den Daten Wissen zu extrahieren. Die hierfür verwendeten Techniken sind unter dem Begriff »Data Mining« zusammengefasst.
Ziel der Masterarbeit ist die Anwendung von verschiedenen Data Mining Algorithmen und Methoden der künstlichen Intelligenz auf eine Datenbank mit medizinischen Daten der Multiple Sklerose Forschung. Hierbei sollen neue Erkenntnisse über die Initiierung einer Immunreaktion im menschlichen Körper gewonnen werden.

1.1 Motivation

Im Rahmen der Bachelorarbeit wurde am Zentrum für Molekulare Neurobiologie Hamburg (ZMNH[1]), das dem Universitätsklinikum Hamburg Eppendorf (UKE[2]) zugeordnet ist, eine Software zur Unterstützung der Multiple Sklerose Forschung entwickelt.
Bei Multiple Sklerose handelt es sich um eine Autoimmunkrankheit, bei der das Immunsystem den eigenen Körper angreift. Hierbei werden körpereigene Zellen durch das Immunsystem fälschlicherweise als fremd erkannt und bekämpft. Zellen enthalten Proteine, bei denen es sich um zusammenhängende Ketten von Aminosäuren handelt. Ausgelöst wird eine Immunreaktion durch die im Blut vorhandenen T-Zellen, hierfür muss diesen ein Protein präsentiert werden und dieses jeweils an eine T-Zelle binden (vgl. Neumann (2008), Janeway u. a. (2009), Jürgen Groth (2011)). Präsentiert wird der T-Zelle das Protein durch den Haupthistokompatibilitätskomplex (Abk. MHC von engl. Major Histocompability Complex). Das Protein muss an das MHC-Komplex binden, damit es der T-Zelle präsentiert wird. Eine Immunreaktion wird somit durch eine Bindung an das MHC-Molekül sowie an die T-Zelle ausgelöst.

[1] http://www.zmnh.uni-hamburg.de
[2] http://www.uke.de

Grund für das Immun-Fehlverhalten können Strukturähnlichkeiten zwischen Viren und körpereigenen Proteinen sein, die ausreichend dafür sind, dass ein Protein an den MHC-Komplex bindet und der T-Zelle präsentiert wird. Diese Phänomen wird als »Molekulare Mimikry« bezeichnet.
In-Silico ist es möglich, für ein Protein zu bestimmen, ob es mit einer gewissen Wahrscheinlichkeit an ein MHC-Molekül bindet. Zur Auswahl stehen hierfür verschiedene Programme der Bioinformatik, mit denen eine Bindungswahrscheinlichkeit für ein Protein und ein bestimmtes MHC-Molekül bestimmt werden kann.
In der Bachelorarbeit wurde ein Webfrontend für zwei Programme zur Berechnung der Bindungswahrscheinlichkeit, das »IEDB MHC-II binding prediction«-Programm (vgl. Wang u. a. (2008)) und »SYFPEITHI« (vgl. Rammensee u. a. (1999)), entwickelt. Neben der Berechnung der Bindungswahrscheinlichkeit bietet das Webfrontend verschiedene Verbesserungen. So können Proteine in einer Datenbank gespeichert werden und zu einem späteren Zeitpunkt für die Analyse ausgewählt werden. Die rein tabellarischen Ausgaben der Programme zur Berechnung der Bindungswahrscheinlichkeit wurden verbessert, wodurch Grafiken zur Veranschaulichung generiert werden können.
In der Masterbeit soll diese Software erweitert werden und den Wissenschaftlern helfen, die Molekulare Mimikry zu untersuchen. Ziel ist es, Gemeinsamkeiten zwischen menschlichen Proteinen und viralen Proteinen zu finden und so Rückschlüsse auf die Fehlreaktion der T-Zelle zu erhalten.

2 Datenanalyse

In diesem Abschnitt wird ein Einblick in die Grundlagen von Data Mining gegeben und erläutert, wie Data Mining Algorithmen für die in der Motivation beschriebenen medizinischen Datenbanken angewendet werden können.

2.1 Data Mining Grundlagen

Als Basis für die Anwendung von Data Mining Algorithmen dienen verschiedene Datenquellen, diese können unter anderem relationale Datenbanken sein, aber auch unstrukturierte Excel-Dateien. Bevor Data Mining Algorithmen angewendet werden können müssen die Datenquellen in einem Data Warehouse zusammengefasst werden. Beim Data Mining werden dann Datensätze aus dem Data Warehouse ausgewählt und analysiert.
Abbildung 2.1 zeigt die Schritte, die für das Data Mining nötig sind.

2.1.1 Data Warehouse

Bei einem Data Warehouse handelt es sich um eine Datenbank, auf die nur lesend zugegriffen wird (vgl. Bauer und Günzel (2004)). Lediglich beim Importieren der Daten aus den Produktivsystemen wird in die Datenbank geschrieben. Das Datenbankmodell eines Data Warehouses ist denormalisiert, was zur Folge hat, dass Daten redundant gespeichert werden. Hierdurch werden wesentlich kompaktere Abfragen möglich, bei denen nicht mehrere Tabellen abgefragt werden müssen.
Als Datenbankschema einer Data Warehouse Datenbank wird das in Abbildung 2.2 aufgeführte Star-Schema verwendet. Die Abbildung zeigt die Datenbank eines Data Warehouses für einen Webshop. Das Star-Schema enthält eine zentrale Faktentabelle, in dem Beispiel ist es die Tabelle »Verkauf«, die die einzelnen Verkaufsdatensätze enthält. Die Tabelle »Verkauf« referenziert über mehrere Fremdschlüssel die einzelnen Dimensionstabellen. Im Beispiel die Tabellen »Zeit«, »Geografie« und »Produkt«. Diese enthalten die konkreten Ausprägungen wie z.B. konkrete Datumswerte. In den Dimensionstabllen werden Informationen redundant gespeichert, um die Abfragen zu vereinfachen. In der Dimensionstablle »Zeit«

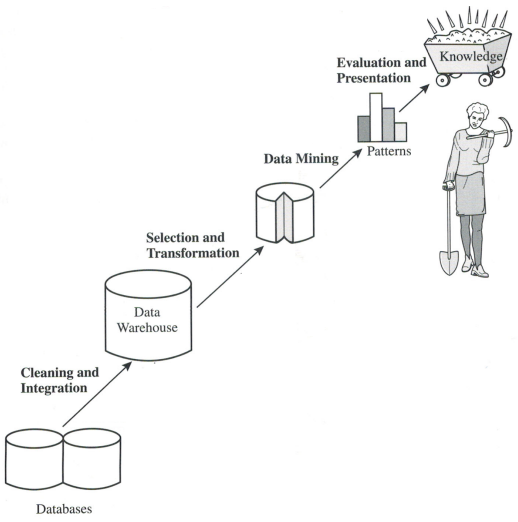


Abbildung 2.1: Benötigte Schritte bis Data Mining Algorithmen angewendet werden können (aus Han und Kamber (2006) Seite 6)

wird das Attribut Quartal zusätzlich gespeichert, obwohl das Quartal aus dem Monat abgeleitet werden kann. Das hat jedoch den Vorteil, dass mit einer einzelnen Abfrage alle Datensätze aus dem Quartal 2 im Jahre 2011 abgefragt werden können.

2.1.2 Data Mining Algorithmen

Im folgendem Abschnitt wird ein Einblick in verschiedene Data Mining Algorithmen und Ansätze gegeben (vgl. Han und Kamber (2006), Wu u. a. (2008)).

Support und Confidence

In großen Datenmengen ist es nützlich, Korrelationen einzelner Datensätze zu bestimmen. Zum Beispiel bei einem Webshop könnte aus den Daten geschlussfolgert werden, welche Artikel häufig zusammen gekauft werden, wodurch einem Kunden ein weiterer Artikel vorgeschlagen werden kann, der zu seinem aktuellen Warenkorb passst.
Zur Bestimmung einer Korrelation werden Werte für den »Support« und die »Confidence« berechnet. Als Ausgangsbasis dienen hierfür beliebig viele Transaktionen, die bestimmte Elemente enthalten. Der »Support« repräsentiert die Wahrscheinlichkeit, mit der ein oder mehrere Elemente in einer der Transaktionen enthalten sind.

Mathematisch wird die Berechnund des Support folgendermaßen ausgedrückt:

$$support(A) = P(A)$$

Die Variable **A** stellt ein Element dar, das in mindestens einer der Transaktionen enthalten ist.
Durch Berechnung des »Confidence« kann eine Wahrscheinlichkeit bestimmt werden, mit der mehrere Elemente gemeinsam in einer der Transaktionen enthalten sein können.
Die Berechnung erfolgt folgendermaßen:

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)}$$

Apriori Algorithmus

Der Apriori Algorithmus (vgl. Agrawal und Srikant (1994)) wird eingesetzt, um aus einer großen Menge an Transaktionen Elemente mit einer hohen Korrelation zu bestimmten.
Der Algorithmus arbeitet in mehreren Iterationen, wobei in jeder Iteration »Itemsets« gebildet werden. In der Abbildung 2.3 wird der Ablauf des Algorithmus an einem Beispiel skizziert.
Angenommen in einer Datenbank sind beliebig viele Transaktionen, die **n** verschiedene Elemente enthalten, dann wählt der Apriori Algorithmus in der ersten Iteration **n** Elemente aus den Transkationen aus. Somit ergeben sich insgesamt **n** einstellige Itemsets. Für jedes Itemset wird der Supportwert berechnet.
In der darauffolgenden Iteration werden alle zweistelligen Kombinationen aus den Elementen gebildet und alle Tupel ausgewählt, deren Support einen festgelegeten Wert überschreitet. In dem in Abbildung 2.3 aufgeführten Beispiel ist der Schwellenwert drei. In der nächsten Iteration werden dreistellige Itemsets gebildet. Der Algorithmus beendet, sobald nur noch ein Itemset übrig ist oder eine Abbruchbedingung festgelegt wurde. Diese könnte sein, dass nur dreistellige Itemsets gefunden werden sollen.
Bei sehr vielen Transaktionen in der Datenbank erhöht sich die Laufzeit des Apriori Algorithmus, da mehrfach über die Transaktionen zur Berechnung der Support-Werte iteriert werden muss. Sofern sehr viele verschiedene Elemente in den Transaktionen vorhanden sind erhöht sich ebenfalls die Laufzeit, da eine Vielzahl an Kombinationen gebildet werden muss.

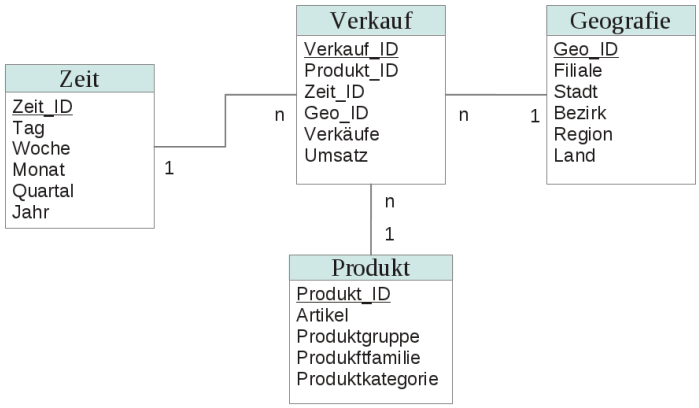


Abbildung 2.2: Star-Schema Datenbankmodell eines Data Warehouse für einen Webshop

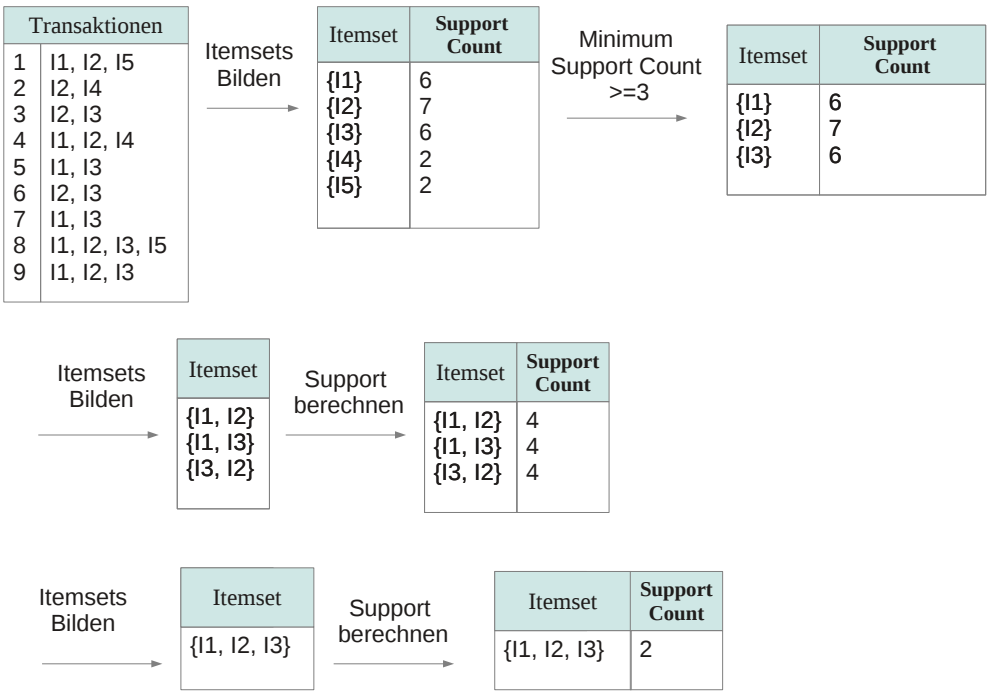


Abbildung 2.3: Beispiel Apriori Algorithmus

ID	Transaktionen	Elemente mit Support >= 3
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, l, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, k, s, p	c, b, p
5	a, f, c, e, l, p, m, n	f, c, a, m, p

Tabelle 2.1: Transaktionen auf denen der FP-Tree Algorithmus angewendet wird

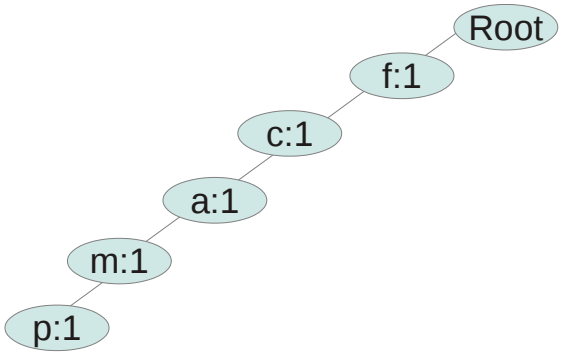
FP-Tree Algorithmus

Der Frequent-Pattern-Tree (kurz FP-Tree, vgl. Han u. a. (2004)) Algorithmus dient wie der Apriori Algorithmus zur Bestimmung von Korrelationen in beliebig vielen Transaktionen.
Der FP-Tree Algorithmus beruht auf einer Baumstruktur, wodurch die Laufzeit bei sehr vielen Transaktionen reduziert wird.
Im folgendem wird der Algorithmus vereinfacht dargestellt, wobei ein Einblick in die Arbeitsweise des Algorithmus gegeben werden soll.
In der Tabelle 2.1 sind die Transaktionen, auf denen der FP-Tree-Algorithmus angewendet wird, aufgeführt. Die zweite Spalte enthält die Elemente der einzelnen Transaktionen, in der dritten Spalte sind alle Elemente aufgeführt, deren Support größer oder gleich drei ist. Die Elemente sind nach der Häufigkeit sortiert. Der Baum wird aufgebaut, indem in mehreren Schritten die Elemente mit dem Support größer drei in den Baum eingefügt werden. Abbildung 2.4 zeigt die Bäume nach dem Einfügen der ersten und zweiten Transaktion, sowie den gesamten Baum, nachdem alle Transaktionen eingefügt wurden. Die Blätter des Baumes enthalten den Namen des Elements und den Supportwert.
Zur Berechnung der Korrelation, z.B. des Elements »p«, müssen alle Pfade in dem Baum bestimmt werden, die »p« enthalten. Diese sind »f:2, c:2, a:2, m:2« und »c:1, b:1«. Der Support für das Tupel »p,c« ergibt somit 3.

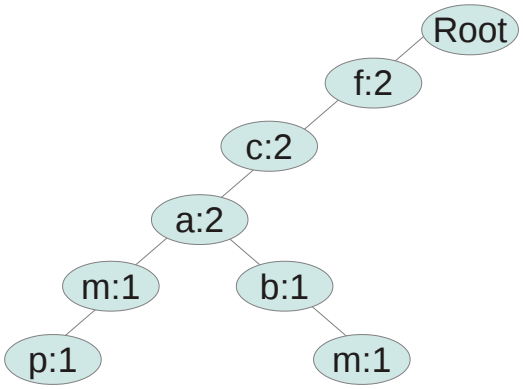
2.2 Data Mining einer medizinischen Datenbank

Im Rahmen der Masterarbeit sollen Data Mining Algorithmen in die in der Motivation (Abschnitt 1.1) beschriebene Analysesoftware und Datenbank integriert werden. Mithilfe der Algorithmen können Eigenschaften von Proteinsequenzen bestimmt werden, die zur Auslösung einer Immunreaktion beitragen.

Einfügen der 1. Transaktion: f, c, a, m, p



Einfügen der 2. Transaktion: f, c, a, b, m



Ergebnis – Alle Transaktionen wurden eingefügt

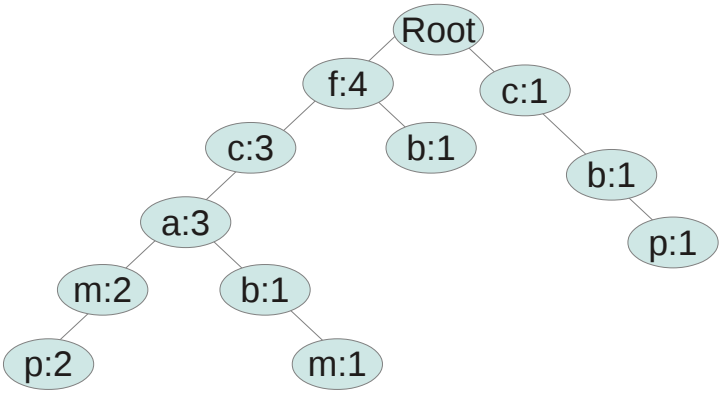


Abbildung 2.4: Einfügen mehrerer Transaktionen in den FP-Tree

Position	**1**	2	3	**4**	5	6	**7**
Aminosäure	I			I			I
	L			F			L
	V			Y			M
							F
							V

Tabelle 2.2: Anker-Motiv des MHC-II Moleküls »HLA_DRB1*1501«

Als Basis für die Analyse dienen verschiedene Virendatenbanken, die mehrere tausend Proteinsequenzen enthalten und in die Analysesoftware importiert werden müssen. Diese viralen Proteine werden In-Silico auf die Bindung an das MHC-Molekül untersucht. Hierbei wird für jede Proteinsequenz eine Bindungswahrscheinlichkeit bestimmt und, sofern diese einen bestimmten Schwellenwert überschreitet, für die weitere Analyse in der Datenbank gespeichert.
Im zweiten Schritt werden die als gut bindend klassifizierten Proteinsequenzen auf Vorhandensein von Anker-Motiven analysiert. Ein Anker-Motiv (vgl. Rammensee u. a. (1999)) ist eine Kombination von Aminosäuren, die in der Proteinsequenz vorhanden sein können. Ein Beispiel für ein Anker-Motiv ist in Tabelle 2.2 aufgeführt, dies hat insgesamt drei Anker-Elemente. Das erste Anker-Element enthält die Aminosäuren »I« (Isoleucin), »L« (Leucin) und »V« (Valin). Das Anker-Motiv passt auf eine Proteinsequenz, sofern die Aminosäure jeweils eine Aminosäure der in den Anker-Elementen enthaltenen Aminosäuren enthält.
In Abbildung 2.5 ist die Bindung einer Proteinsequenz an den T-Zell-Rezeptor und an das MHC-Molekül dargestellt. Das Anker-Motiv besteht aus den Aminosäuren »F« (Phenylalanin), »I« (Isoleucin) und »R« (Arginin). Die Aminosäuren neben und zwischen den Anker-Elementen sind für die Bindung an die T-Zelle zuständig.
Die an die T-Zelle bindenden Aminosäuren sollen untersucht werden. Ziel ist es strukturelle Eigenschaften zu finden, die ein Auslöser für die Bindung an die T-Zelle sein können. Menschliche Proteine sollen dann auf diese Eigenschaften untersucht werden und somit ein möglicher Auslöser für die Fehlfunktion des Immunsystems gefunden werden.

Statistische Analyse

Ein erster Ansatz ist die statistische Analyse der Aminosäuren zwischen den Anker-Elementen. Hierbei wird berechnet, wie häufig einzelne Aminosäuren vorkommen und eine Verteilung der Aminosäuren aufgestellt.

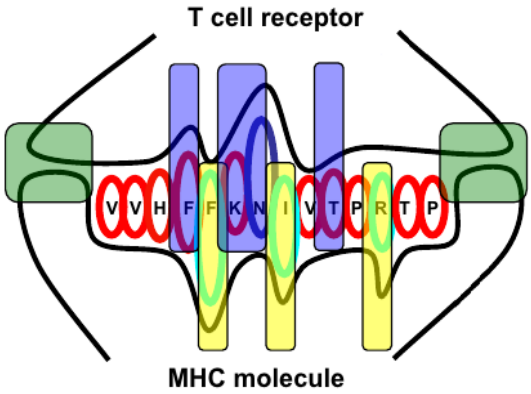


Abbildung 2.5: Bindung einer Proteinsequenz an das MHC-Molekül und an den T-Zell-Rezeptor

Data Mining

Aminosäuren verfügen über verschiedene Eigenschaften, somit kann jeder anhand ihrer Ladung, Größe und Hydrophobie klassifiziert werden.
Anhand dieser Eigenschaften können Korrelationen der zwischen den Anker-Elementen vorhandenen Aminosäuren mit Data Mining Algorithmen gefunden werden. Hierbei ist es interessant zu untersuchen, welche Aminosäuren häufig gemeinsam auftreten und welche Eigenschaften sie haben.

Maschinelles Lernen

Anhand der im Data Mining Schritt gewonnenen Informationen über die Eigenschaften der an die T-Zelle bindenden Aminosäuren kann eine Support Vector Machine trainiert werden. Diese soll eine Klassifizierung bisher nicht analysierter Proteinsequenzen durchführen und bestimmen, mit welcher Wahrscheinlichkeit die Proteinsequenz an die T-Zelle bindet.

3 Schluss

Die Masterarbeit soll einen neuen Ansatz in der Immunologie-Forschung bieten und es ermöglichen, eine Vielzahl an Proteinen auf Strukturähnlichkeiten zu untersuchen. Es ist denkbar, dass Strukturen gefunden werden, die den Proteinen des menschlichen Körpers ähneln und so ein Auslöser für die molekulare Mimikry sind. Dies wäre ein großer Fortschritt für die Immunologie-Forschung, jedoch ist es auch möglich, dass keine Strukturen in den Proteinsequenzem gefunden können.

3.1 Ausblick

Bis Data Mining Algorithmen implementiert und angewendet werden können, muss Vorarbeit geleistet werden. Es müssen Programme entwickelt werden, die die Virendatenbanken importieren und mit den verschiedenen Formaten der Virendatenbanken arbeiten können. Die Analyse einer Vielzahl von Proteinsequenzen ist sehr rechenaufwändig, daher muss ein Mechanismus entwickelt werden, der es den Wissenschaftlern ermöglicht, Proteinsequenzen auszuwählen und diese durch die Software über einen längeren Zeitraum analysieren zu lassen. Wichtig hierbei ist es zu parallelisieren, um auf die Rechenleistung mehrerer Prozessoren zurückzugreifen.
Das »Apache Hadoop«-Projekt[1] ist ein Framework das nach dem Konzept des von Google entwickelten »MapReduce«-Algorithmus (vgl. Dean und Ghemawat (2008)) arbeitet. Es ermöglicht eine sehr starke Parallelisierung von Algorithmen, die transparent über mehrere Rechner verteilt laufen können. Das auf Hadoop aufsetzende Projekt »Mahout«[2] implementiert verschiedene Data Mining Algorithmen, wie z.B. einen FP-Tree und eignet sich somit für die Analyse der Aminosäuren.
Eine wissenschaftliche Publikation der Software ist in Planung. Sobald die Software erste Ergebnisse für die Immunologie-Forschung liefern kann, werden weitere medizinische Publikationen folgen.

[1] http://hadoop.apache.org
[2] http://mahout.apache.org/

Literaturverzeichnis

[Agrawal und Srikant 1994] AGRAWAL, Rakesh ; SRIKANT, Ramakrishnan: Fast Algorithms for Mining Association Rules in Large Databases. In: BOCCA, Jorge B. (Hrsg.) ; JARKE, Matthias (Hrsg.) ; ZANIOLO, Carlo (Hrsg.): *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, Morgan Kaufmann, 1994, S. 487–499. – ISBN 1-55860-153-8

[Bauer und Günzel 2004] BAUER, Andreas ; GÜNZEL, Holger: *Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung*. Dpunkt.Verlag, 2004. – ISBN 978-3898642514

[Dean und Ghemawat 2008] DEAN, Jeffrey ; GHEMAWAT, Sanjay: MapReduce: simplified data processing on large clusters. In: *Commun. ACM* 51 (2008), Januar, S. 107–113. – URL `http://doi.acm.org/10.1145/1327452.1327492`. – ISSN 0001-0782

[Han und Kamber 2006] HAN, Jiawei ; KAMBER, Micheline: *Data Mining. Concepts and Techniques*. Morgan Kaufmann, 2006. – ISBN 978-1558609013

[Han u. a. 2004] HAN, Jiawei ; PEI, Jian ; YIN, Yiwen ; MAO, Runying: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. In: *Data Min. Knowl. Discov.* 8 (2004), Nr. 1, S. 53–87

[Janeway u. a. 2009] JANEWAY, Charles A. ; MURPHY, Kenneth M. ; TRAVER, Paul ; WALPORT, Mark: *Janeway Immunologie*. Spektrum Akademischer Verlag, 2009. – ISBN 3-8274-1079-7

[Jürgen Groth 2011] JÜRGEN GROTH, Dr. rer. n.: *Meine Moleküle. Deine Moleküle - Online Buch*. 2011. – URL `http://www.meine-molekuele.de/`

[Kim u. a. 2009] KIM, Yohan ; SIDNEY, John ; PINILLA, Clemencia ; SETTE, Alessandro ; PETERS, Bjoern: Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. In: *BMC Bioinformatics* 10:394 (2009), November

[Lesk 2003] LESK, Arthur M.: *Bioinformatik - Eine Einführung*. Spektrum, Akad. Verlag, 2003. – ISBN 3-8274-1371-0

[Neumann 2008] NEUMANN, Jürgen: *Immunbiologie*. Springer Berlin Heidelberg, 2008 (Springer-Lehrbuch). – ISBN 978-3-540-72569-5

[Rammensee u. a. 1999] RAMMENSEE, Hans-Georg ; BACHMANN, Jutta ; EMMERICH, Niels Philipp N. ; BACHOR, Oskar A. ; STEVANOVIĆ, Stefan: SYFPEITHI: database for MHC ligands and peptide motifs. In: *Immunogenetics* (1999), Nr. 50, S. 213–219

[Wang u. a. 2008] WANG, Peng ; SIDNEY, John ; DOW, Courtney ; MOTHÉ, Bianca ; SETTE, Alessandro ; PETERS, Bjoern: A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. In: *PLoS Computational Biology* volume 4 issue 4 (2008), April

[Wu u. a. 2008] WU, Xindong ; KUMAR, Vipin ; J. ROSS QUINLAN, et a.: Top 10 algorithms in data mining. In: *Knowledge and Information Systems* 14 (2008), January, S. 1–37