Genvorhersage und vergleichende Genomanalyse am Beispiel von *Tribolium castaneum*

Doreen Werner

15. Dezember 2005

Referent : Prof. Dr. B. Morgenstern

Korreferent : Prof. Dr. E. Wimmer

Tag der Abgabe der Diplomarbeit : 15. Dezember 2005

Letzter Tag der mündlichen Prüfung : 21. Januar 2005

Verzeichnis der verwendeten Abkürzungen

ca. circa

d.h. das heißt

z.B. zu Beispiel

usw. und so weiter

u.a. unter anderem

min. minimal

max. maximal

pos. positiv

neg. negativ

WGS Whole Genome Shotgun

Gb Gigabasen

Bp Basenpaare

Kb Kilobasen

Mb Megabasen

BAC Bacterial Artificial Chromosom

YAC Yeast Artificial Chromosom

EST Expressed Sequence Tag

LINES Long Interspered Nuclear ElementS

SINES Short Interspersed Nuclear ElementS

DNA DeoxyriboNucleic Acid

RNA RiboNucleic Acid

mRNA messenger RiboNucleic Acid

cDNA complementary DeoxyriboNucleic Acid

CDS CoDing Sequence

A Adenylsäure

T Thymidylsäure

G Guanylsäure

C Cytidylsäure

UTR UnTranslated Region

GFF General Feature Format

BLAST Basic Local Alignment Search Tool

HSP High Scoring Segment Pair

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Arbeit beschreibt Methoden der Sequenzanalyse am Beispiel der Genomsequenz von *Tribolium castaneum*. Das zweite Assembly der Genomsequenz von *Tribolium* wurde im September 2005 veröffentlicht. Das Genvorhersageprogramm AUGUSTUS wurde für eine automatische Annotation dieser Genomsequenz herangezogen. Aus EST-Sequenzen von *Tribolium castaneum* wurden Genmodelle erstellt, mit denen AUGUSTUS auf die spezifische Genomsequenz von *Tribolium castaneum* trainiert wurde. Die Analyse der Qualität der Vorhersagegenauigkeit zeigt, dass die erstellte Vorhersage auf der Genomsequenz als qualitativ gute automatische Annotation angesehen werden kann. Vergleiche der Proteinsequenzen von *Tribolium castaneum*, *Drosophila melanogaster* und *Homo sapiens* wurden durchgeführt, um besondere Proteine in *Tribolium* zu finden. Für einige menschliche Proteine, für die in *Drosophila* keine homologen Proteine mit signifikanter Sequenzähnlichkeit identifiziert werden konnten, war die Suche nach signifikant ähnlichen Proteinsequenzen in *Tribolium* erfolgreich. Diese Proteine zeigen die Bedeutung der Genomsequenz von *Tribolium* in besonderem Maße. Ergebnis der vergleichenden Genomanalysen ist eine Menge an solchen Proteinen aus *Tribolium castaneum*, die mögliche Kandidaten für weitere Analysen sind.

Kapitel 1

Motivation

Seit Ende Januar 2005 steht der molekularbiologischen Forschung das erste Assembly der Genomsequenz von *Tribolium castaneum* zur Verfügung. Ende September 2005 wurde die zweite Version der assemblierten Genomsequenz von *Tribolium* veröffentlicht. Das Assembly hat eine Größe von 154 Mb. Für eine derart große Datenmenge ist die automatisierte Genannotation die einzige Möglichkeit, schnell und umfangreich einen Überblick über die Gene von *Tribolium* zu bekommen. Instrument der Wahl für diese, im Rahmen meiner Diplomarbeit zu bewältigende, Aufgabe ist das Genvorhersageprogramm AUGUSTUS [1], entwickelt von Mario Stanke. Das Programm basiert auf einem mathematischen Modell der Signale einer eukaryotischen Genstruktur und dient der Vorhersage proteinkodierender Gene in eukaryotischen Genomen. Mit annotierten Genen kann die Qualität der Genvorhersage für die spezifische Genomsequenz einer Spezies in einem Prozess, der als *Training* bezeichnet wird, verbessert werden. Die erreichte Qualität ist dabei von der Qualität und der Menge der zum Training verwendeten Annotationen abhängig. Es gilt also eine möglichst umfangreiche und korrekte Menge an Genen zusammenzustellen, um AUGUSTUS für die automatische Annotation der Genomsequenz von *Tribolium castaneum* zu trainieren. AUGUSTUS wurde bislang erfolgreich für die Genvorhersage in mindestens sechs Projekten eingesetzt und auf die Genomsequenzen der Spezies *Homo sapiens*, *Drosophila melanogaster*, *Aedes aegypti*, *Arabidopsis thaliana*, *Brugia malayi* und *Coprinus cinereus* trainiert. AUGUSTUS hat in Vergleichen mit anderen Gen-

vorhersageprogrammen eindrucksvolle Ergebnisse erzielt [1].

Die Annotation der Gene im sequenzierten Genom ist das Fundament für vergleichende Genomanalysen, welche die Notwendigkeit für die durchgeführte Sequenzierung im Fall des rotbraunen Reismehlkäfers erst untermauern. *Tribolium castaneum* gehört wie der genetische Modellorganismus *Drosophila melanogaster*, deren Genomsequenz bereits sequenziert wurde, zu den holometabolen Insekten. Es wird erwartet, dass die Genomsequenz von *Tribolium* die Identifikation von homologen Proteinen in Mensch und *Drosophila* unterstützt und entscheidend beeinflusst [2]. Vermutlich gibt es auch menschliche Proteine, zu denen in *Drosophila* keine Homologen gefunden werden, aber in *Tribolium* (Dr. Gregor Bucher, Prof. Dr. Ernst Wimmer, persönliche Mitteilung). Solche Proteine aus *Tribolium* zeigen die Bedeutung der Sequenzierung in besonderem Maße. Eine Möglichkeit zur vergleichenden Genomanalyse von *Drosophila melanogaster*, *Tribolium castaneum* und *Homo sapiens* soll im Rahmen dieser Arbeit vorgestellt werden.

Kapitel 2

Das Genomsequenzierungsprojekt des rotbraunen Reismehlkäfers *Tribolium castaneum*

Das 200 Mb große Genom des rotbraunen Reismehlkäfers *Tribolium castaneum* wurde im Human Genome Sequencing Center (HGSC) des Baylor College of Medicine in Houston, Texas durch *Whole Genome Shotgun Sequenzierung* sequenziert. Die für die Sequenzierung gewählte genomische DNA entstammt aufgereinigten Nuclei von *Tribolium castaneum* Embryos gemischten Geschlechts. Unter der URL [3] sind derzeit Projektbeschreibung und Ergebnisse der Sequenzierung und Assemblierung beziehbar. Für die vorliegende Arbeit wurden die vom HGSC unter dieser URL publizierten Assemblies Tcas_1.0 und Tcas_2.0 und ESTs verwendet sowie ESTs, die von der Universität zu Köln unter der URL [4] bezogen werden können. Diese Daten können der beiliegenden CD entnommen werden.

2.1 Das wissenschaftliche Interesse an der Genomsequenz von *Tribolium castaneum*

Das große Interesse der wissenschaftlichen Gemeinde an der Genomsequenz von *Tribolium castaneum* hat vielerlei Gründe. Einer der Wichtigsten sei aus [2] zitiert: »*Tribolium* is one of the most sophisticated genetic model organisms among all higher eukaryotes.« *Tribolium* ist einfach genetisch zu manipulieren. Viele molekularbiologische Methoden wurden für die Forschung mit *Tribolium* als Modellorganismus entwickelt. Spezifische »Forward and reverse genetic approaches« [2] zur genetischen Funktionsanalyse existieren für *Tribolium*. »Forward« meint die klassischen Methoden der Funktionsanalyse von Genen. Ungewöhnliche, seltene Phänotypen und diese Phänotypen verursachende, defekte oder ausgeschaltete, unbekannte Gene oder Allele werden gesucht. In »Reverse«-Studien wird versucht, den Phänotyp eines bekannten Gens durch Mutation oder Beeinflussung der Produkte der Genexpression zu identifizierten.

Tribolium gehört zur Ordnung der Coleoptera: Tenebrionidae, der primitivsten Ordnung der sogenannten *holometabolen* Insekten. Das sind Insekten, die während ihrer Entwicklung vom Ei zum Imagines einer vollständigen Metamorphose unterliegen [5]. Die Gene von *Tribolium* geben Aufschluss darüber, welche genetischen Veränderungen zur Entwicklung von höheren Organismen mit komplexeren Entwicklungsstadien geführt haben könnten. *Tribolium* dient der entwicklungsbiologischen Forschung als System, in dem vor allem die embryonale Entwicklung und die Evolution der Entwicklung bei Insekten studiert werden kann [2].

Die Genomsequenzierung bietet eine weitere Möglichkeit für die Entwicklung neuer Arzneistoffe und Antibiotika. Es wurden p-Benzochinone, aliphatische Kohlenwasserstoffe und andere potentiell reizend, giftig oder antibakteriell wirkende Stoffe als Produkte großer Drüsen des Käfers gefunden [6] sowie Prostaglandinsynthetaseinhibitoren in seinen Sekreten [7].

Einzelne direkte Sequenzvergleiche zwischen *Tribolium*, *Drosophila* und dem

Menschen haben gezeigt, dass es Gene in *Tribolium* gibt, die mehr Ähnlichkeit zu menschlichen Sequenzen haben als ihre Homologen in *Drosophila* (D. Beeman, nicht veröffentlicht). Die Verwandtschaft zwischen dem Gen *Zen* aus *Drosophila* und den menschlichen *HOX3* Genen konnte z.B. durch Vergleiche aufgeklärt werden, bei denen unter anderem Daten von *Tribolium* hilfreich waren [8]. So kann diese Sequenz für die Identifizierung homologer Proteine in Insekten und Vertebraten von großer Bedeutung sein, wenn direkte Sequenzvergleiche ohne Ergebnis bleiben.

Die Verfügbarkeit der Genomsequenz verstärkt auch die Hoffnung auf ein besseres Verständnis der Resistenzentwicklung gegenüber Pestiziden, da *Tribolium* weltweit in großen Lagerhallen, in denen vor allem trockene Lebensmittel gelagert werden, als Plage gefürchtet wird und eine lange Geschichte der Bekämpfung und Resistenzentwicklung, beispielsweise durch oxidative oder hydrolytische Metabolisierung, hinter sich lässt [9, 10, 11]. In Verbindung mit seiner leichten genetischen Manipulierbarkeit wird *Tribolium* hierdurch zu einem hervorragenden Kandidaten zur Identifizierung neuer Angriffspunkte für Pestizide.

2.2 Whole Genome Shotgun Sequenzierung

Mit den aktuellen Sequenzierungsapparaturen kann die Sequenz von DNA-Fragmenten mit einer Länge von 500 bis 700 Bp mit einer relativ geringen Fehlerrate ermittelt werden [12]. Bei längeren Fragmenten wird das Ergebnis zusehends ungenauer. Um nun eine ganze Genomsequenz von einigen Mb Länge zu sequenzieren, dient die Methode der *Whole Genome Shotgun Sequenzierung* (WGS). Die in mehreren Kopien vorliegende Genomsequenz wird hierfür zufällig mechanisch zerstückelt und Sequenzstücke einer bestimmten Größe, *Inserts* genannt, werden in Vektoren kloniert, deren Gesamtheit als *Genombibliothek* bezeichnet wird. Verwendete Vektoren sind *Bacterial Artificial Chromosomes* (BACs), die Inserts mit einer Länge von < 150 Kb enthalten können, *Yeast Artificial Chromosomes* (YACs), wobei hier die Inserts Längen von bis zu 3 Mb haben können und *Fosmide*[1] und *Lambda*-Phagen,

[1] *single-copy* Plasmide, welche die *cos-site* für in vitro *Lambda-Packaging* enthalten

die 20 bis 35 Kb Sequenz aufnehmen können [12]. Für die Sequenzierung der Inserts gibt es mehrere Ansätze. Ein Ansatz ist die sogenannte *Paired-End-Sequenzierung*, wobei von jedem Ende eines Inserts ein *Read* von 500 bis 600 Bp Länge gelesen wird. Für einen weiteren Ansatz wird das Insert zufällig in Reads von 500 bis 600 Kb Länge zerteilt und in universelle Vektoren subkloniert, von denen dann zufällig ausgewählte sequenziert werden. Ziel ist, die gesamte Sequenz mit Reads so abzudecken, dass möglichst wenige Lücken entstehen und die Reads ausreichend überlappen. Dazu ist eine *mittlere Coverage*[2] von 6,5 bis 8 [13] nötig. Die nun folgende Aufgabe ist die Assemblierung der Reads. Ein, aus überlappenden Reads lückenlos zusammengesetztes Sequenzstück wird ein *Contig* genannt. Aus den Contigs, deren Reihenfolge und Orientierung zueinander ermittelt werden kann, werden sogenannte *Scaffolds* konstruiert. Die Scaffolds sind untereinander durch Lücken (*physical gaps*) getrennt, deren Sequenzinhalt nicht in den Bibliotheken enthalten und deshalb nur schwer ermittelbar ist. Die größten Probleme bei der Assemblierung der Reads entstehen durch lange Abschnitte sich wiederholender Nukleotidfolgen, sogenannte *repetitive* Sequenzabschnitte, die länger als zwei Reads sind. Diese verursachen oft fehlerhafte Assemblierungen, weil nicht erkannt werden kann, wie lang die repetitiven Sequenzabschnitte in Wirklichkeit sind.

2.3 Sequenzierungsergebnisse

Das erste Assembly der durch Whole Genome Shotgun Sequenzierung generierten Reads steht seit Januar 2005 zur Verfügung. Zur Assemblierung der rund 1,8 Millionen Reads wurde das *Atlas Genome Assembly System* eingesetzt. Ergebnis dieses Prozesses sind Contigs und Scaffolds. Etliche WGS-Bibliotheken mit Inserts von 3-4 Kb und 4-6 Kb dienten der Erstellung dieser Daten. Die zur Assemblierung eingesetzten Reads repräsentieren 1,8 Gb Sequenz und 7,5-fache Coverage des gesamten klonierbaren Genoms von *Tribolium*.

[2] Aufsummierung der Längen aller Reads und Division dieser Zahl durch die Länge der Genomsequenz

Seit Ende September 2005 gibt es eine zweite Version der *Genomsequenz* von *Tribolium*. Dieser Entwurf der Genomsequenz entstand aus Genombibliotheken mit Inserts von 4-5 Kb, rund 40 Kb und rund 130 Kb durch Assemblierung von etwa 1,54 Millionen Reads mit dem Atlas Genome Assembly System. Diese Reads repräsentieren ca. 152 Mb Genomsequenz und 7,3-fache Coverage des klonierbaren Genoms von *Tribolium*. Ungefähr 70% der genomischen Sequenz konnte zu Chromosomen kartiert werden.

Da es sich bei der veröffentlichten Sequenz noch um einen Entwurf handelt sind fehlerhafte Bereiche nicht auszuschließen. Diese können unter anderem durch falsche Assemblierung sich wiederholender Sequenzen oder durch nicht vereinigte Überlappungen entstanden sein, die eine Duplikation des Abschnitts bedeuten.

2.4 Expressed sequence tags

Ein *expressed sequence tag* oder EST ist ein Teilstück (500 bis 800 Bp) einer, durch *reverse Transkription* in DNA umgeschriebenen mRNA-Sequenz, die direkt aus einer Zelle isoliert wurde und meist eine Länge von 900 bis 1500 Bp aufweist [12]. Ein EST repräsentiert einen Teil oder das komplette Transkript eines Genes und ist ein vielseitig verwendetes Hilfsmittel zur Analyse von Genen und Genexpression.

Ein großes Problem bei der Arbeit mit mRNA ist ihre Vergänglichkeit und geringe Stabilität, begründet vor allem in ihrer Einzelstrangstruktur [14]. So ist es kaum möglich, direkt aus der mRNA den Sequenzinhalt zu beziehen. Das Enzym *Reverse Transkriptase*, das die Umschreibung von mRNA in DNA (reverse Transkription) katalysiert, stammt aus einem Retrovirus. Diese Funktion ist einzigartig und wurde bisher für keine weitere Organismengruppe dokumentiert [14]. Die synthetisierte DNA-Sequenz wird cDNA (copy DNA) genannt. Die generierten cDNA-Moleküle werden kloniert (gewöhnlich in dem Bakterium *Escherichia coli*). Die Gesamtheit aller Klone wird als *c*DNA-*Bibliothek* bezeichnet. Diese kann je nach Umfang repräsentativ für das Transkriptom einer Spezies sein. Ein *Screen* dieser Bibliothek durch Se-

quenzierung der cDNAs, die hierfür gestückelt werden müssen, ergibt eine Ansammlung an EST-Sequenzen, die Indizien für Genstrukturen und Genexpression in einer Zelle zu einem bestimmten Zeitpunkt liefern. Expressed sequence tags sind hilfreiche Werkzeuge für die Genannotation und Sequenzanalyse, auch wenn ihre Qualität auf Grund technischer Grenzen bisweilen nur als sehr gering eingestuft werden kann und oft nicht feststellbar ist, ob die EST-Sequenz den Kode oder sein reverses Komplement enthält.

Im Zuge des Sequenzierungsprojekts von *Tribolium castaneum* wurden cDNA-Bibliotheken angelegt und EST-Sequenzdaten generiert. Zum einen von der Universität zu Köln (Joel Savard) in Zusammenarbeit mit Exelixis Inc. und zum anderen vom Human Genome Sequencing Center (HGSC). Der vom HGSC durchgeführte EST-Screen umfasste 12000 Kolonien aus zwei Bibliotheken mit der Hoffnung auf Präsenz von ca. 10000 Klonen aus jeder Bibliothek (Richards, S., persönliche Mitteilung). Diese umfangreichere EST-Datensammlung ist seit Ende September 2005 verfügbar und umfasst 35649 Sequenzen.

2.5 Sequenzvalidierungen

Die Genome höherer Eukaryoten enthalten eine große Zahl von Abschnitten sich wiederholender Nukleotidfolgen, sogenannte *repetitive* Sequenzen. Diese sind intergenische, vermutlich funktionslose DNA-Bereiche [14]. Mit dem Programm RepeatMasker (A.F.A. Smit und P.Green, nicht veröffentlicht) wurden die Genomsequenzen von *Tribolium* auf spezielle repetitive Elemente, sogenannte *interspersed Repeats* und auf Regionen von geringer Komplexität (Purin-, Pyrimidin-, AT- und CG-reiche Bereiche) gefiltert. Interspersed Repeats sind charakteristisch für Pseudogene und transponierbare Elemente (DNA Transposons, LINES, SINES, Retrovirus Retrotransposons) [14]. Die Buchstaben, welche die Nukleotide dieser Sequenzabschnitte repräsentieren, wurden von dem Programm durch `N` ersetzt. Die Kennzeichnung dieser Abschnitte innerhalb der Genomsequenz mit Hilfe des Programms dient der Prävention falsch-positiver Ergebnisse, da Strukturgene fast ausschließlich in der nicht-repetitiven DNA lokalisiert sind [14]. Viele eukaryotische Ge-

nome bestehen zu einem erheblichen Teil aus transponierbaren Elementen [15], z.B. das menschliche Genom, bei dem die interspersed Repeats ca. 40% der Gesamtsequenz ausmachen [16]. Allgemein besteht die DNA tierischer Zellen durchschnittlich zu 50% aus repetitiver DNA [14]. Für *Drosophila melanogaster* ist der geschätzte Wert der repetitiven DNA ca. 30% [14]. Die Identifizierung von interspersed Repeats in der dem Programm übergebenen Genomsequenz basiert auf Sequenzvergleichen mit Bibliotheken wie Repbase [17], die eine Ansammlung bekannter repetitiver Sequenzen aus sequenzierten eukaryotischen Genomen umfassen. Das für den Sequenzvergleich integrierte Programm `cross-match` ist eine Implementierung des Smith-Waterman-Gotoh-Algorithmus [18].

Das Programm Seqclean wurde zur Validierung der expressed sequence tags eingesetzt. Störende Abschnitte aus Polyadenylsäure bzw. Polythymidylsäue an den Endbereichen der ESTs werden von dem Programm erkannt und entfernt, da diese nicht kodiert sind und posttranskriptional nach Abschluss der Synthese der RNA an diese angefügt werden [14]. Ebenso wird mit Sequenzenden von geringer Komplexität und mit solchen, die reich an nicht bestimmten Nukleotiden sind, verfahren. Sequenzen, die weniger als 100 Nukleotide enthalten, werden verworfen. Weiterhin wurde die Möglichkeit der Validierung bezüglich Kontamination der EST-Sequenzen mit Vektor- oder Adaptersequenzen aus dem Klonierungsprozess der mRNA genutzt. Identifiziert werden solche Kontaminationen durch kurze, terminale Übereinstimmungen der ESTs mit Sequenzen aus Vektor- bzw. Adapterdatenbanken. Verwendet wurde die Datenbank UniVec_Core [19]. Sie ist eine Sammlung von Oligonukleotidsequenzen (Vektoren, Adapter, Linker, Primer) bakteriellen, viralen, oder synthetischen Ursprungs sowie aus *Saccharomyces cerevisiae* und aus Phagen, die allgemein in Klonierungsprozessen eingesetzt und im Zuge dieser Prozesse an die eigentliche Sequenz angehängt werden. Der Validierungsprozess kann gegebenenfalls eine Verkürzung der Sequenzen bedeuten. Die EST-Sequenzen, die nach dem Validierungsprozess eine Mindestlänge von 100 Nukleotiden unterschreiten oder deren Prozentsatz an nicht bestimmten Nukleotiden größer als 3% ist, werden verworfen.

Kapitel 3

Annotation der Gene im Genom von *Tribolium castaneum*

Molekularbiologische Methoden der Sequenzanalyse sind zeitaufwendig und kostspielig. Daher sind sie nur für ausgewählte Gene rentabel, zu deren Identifikation in den meisten Fällen computergestützte Methoden beigetragen haben. Auch wenn automatisierte Methoden der Genannotation bislang nur recht unzuverlässige Ergebnisse liefern, geben sie doch unverzichtbare Hinweise für die Identifizierung kodierender Sequenzabschnitte.

Die bekanntesten Genvorhersageprogramme können allgemein je nach Art, der von ihnen verwendeten Daten, in drei Gruppen unterteilt werden [20]. Die erste Gruppe bilden die sogenannten *ab initio* Programme, die nur die zu annotierende Sequenz als Eingabe verwenden. Die ab initio Genvorhersage basiert auf einer mathematischen Modellierung der Merkmale von Genen und ist unabhängig von anderen Informationen bzw. Sequenzen. AUGUSTUS [1] GENSCAN [21] und GENEID [22] sind Beispiele, die diesen Ansatz verfolgen. Die zweite Gruppe bilden Programme, die mit Hilfe genomischer Sequenzen verschiedener Spezies, Gene in der zu annotierenden Sequenz vorhersagen. Dabei wird die Konservierung kodierender Abschnitte verwandter Gene ausgenutzt. Die zum Vergleich verwendeten Sequenzdaten, werden als *extrinsi-*

sche Informationen bezeichnet. Das Programm TWINSCAN [23] verfolgt z.B. einen solchen Ansatz. Die dritte Gruppe von Programmen verwenden EST- oder Proteinsequenzen zur Verbesserung der Qualität der ab initio Genvorhersage. Solche Programme sind oft Erweiterungen bekannter ab initio Programme beispielsweise GENOMESCAN, eine Erweiterung von GENSCAN [20]. AUGUSTUS+ kann in die ab initio Genvorhersage Annotationen von Exons, Exonteilen, Start- und Stoppkodons, Introns und Spleißstellen einbeziehen. Mit dem Programm AGRIPPA [24] können solche Informationen aus Sequenzvergleichen mit EST- und Proteindatenbanken, ausgeführt von dem externen Programm BLAST [25], automatisch erstellt werden.

Das ab initio Genvorhersageprogramm AUGUSTUS hat in Vergleichen mit anderen Genvorhersageprogrammen beeindruckende Ergebnisse erzielt [1] und ist aus diesem Grund eine gute Wahl für die Annotation der Genomsequenz von *Tribolium castaneum*. In mindestens sechs Projekten hat AUGUSTUS die Annotationen der Genomsequenzen eukaryotischer Spezies bereits erfolgreich unterstützt. Das Programm basiert auf einem mathematischen Modell der Merkmale eukaryotischer Gene und muss für die Vorhersage auf der Genomsequenz einer Spezies für diese optimiert werden. Die wichtigsten Merkmale eukaryotischer Gene, das mathematische Modell und das Training von AUGUSTUS für die Genomsequenz von *Tribolium castaneum* sind in den folgenden Abschnitten beschrieben.

3.1 Strukturmerkmale eukaryotischer Gene und ihre Prozessierung

Die Identifikation proteinkodierender Gene in eukaryotischen Genomen ist nicht trivial. Eukaryotische Gene bestehen aus kodierenden und nicht-kodierenden *Exons*, die durch nicht-kodierende Sequenzen, sogenannte *Introns*, voneinander getrennt sind. Ein Gen beginnt und endet immer mit einem Exon. Die Anzahl der Exons ist variabel. Es gibt Gene mit nur einem Exon und Gene mit über 100 [14]. Der größte Teil einer eukaryotischen Genomsequenz besteht vor allem bei höheren Organismen aus intergenischer Region

mit langen Abschnitten sich wiederholender Sequenzfolgen (repetitive DNA) [14]. Die Gendichte ist also oft sehr gering (ca. 10%). Die ersten und letzten zwei Nukleotide eines Introns entsprechen generell einem Konsensus und werden als *Donor-Spleißstelle* oder *5'-Spleißstelle* und *Akzeptor-Spleißstelle* oder *3'-Spleißstelle* bezeichnet. Der Konsensus für die Donor-Spleißstelle ist ein Dinukleotid aus G (Guanylsäure) und T (Thymidylsäure) und der Konsensus der Akzeptor-Spleißstelle ist ein Dinukleotid aus A (Adenylsäure) und G (Guanylsäure). Bei niederen Eukaryoten wie der Hefe ist ein weiterer Teil des Introns konserviert, die *Branch-Site*. Bei höheren Eukaryoten ist die Konservierung in diesem Bereich oft wesentlich unauffälliger [14]. Gene können an einem Lokus auf verschiedenen DNA-Strängen liegen, Überlappungen sind selten, aber dokumentiert [14].

Der Prozess, bei dem die Nukleotidsequenz eines Gens in die Aminosäuresequenz eines Proteins umgeschrieben wird, beginnt mit der Erstellung einer Kopie, der sogenannten *prä-m*RNA, des gesamten Gens, die dann durch Entfernen der Introns, dem sogenannten *Spleißen* zur fertigen mRNA (auch als *Transkript* bezeichnet) prozessiert wird. Der Teilprozess der Erzeugung einer mRNA aus der Nukleotidsequenz des Gens wird *Transkription* genannt. Die mRNA ist eine Aneinanderreihung der komplementären Nukleotide der Exonkette des DNA-Stranges. Handelt es sich um proteinkodierende mRNA folgt der Prozess der *Translation*. Der kodierende Abschnitt der mRNA wird in eine Aminosäuresequenz übersetzt, wobei je drei Nukleotide ein *Kodon* bilden und eine Aminosäure repräsentieren. Die Länge dieses Abschnitts ist also ein Vielfaches von drei. Die Länge der einzelnen Exons muss aber kein Vielfaches von drei sein, da die Sequenz eines Introns ein Kodon trennen kann.

Für die Proteinsynthese stehen 20 Aminosäuren zur Verfügung. Für eine Aminosäure gibt es, bis auf zwei Ausnahmen, mehrere (bis zu sechs) verschiedene Kodons. Die Bevorzugung bestimmter Kodons für die Aminosäuren kann artspezifisch sein. Der proteinkodierende Abschnitt eines Gens beginnt gewöhnlich immer mit dem Kodon ATG und endet mit einem von drei möglichen Stoppkodons: TAG, TGA, TAA. Eine Folge von Kodons ohne Stoppkodon

in der Sequenz eines Exons wird als *offener Leserahmen* (ORF) bezeichnet. Nicht-kodierende Leserahmen enthalten oft viele Stoppkodons, so dass eine Verschiebung des Leserahmens im Verlauf der Translation zur Termination führt und fehlerhafte Proteine nicht synthetisiert werden. Ein Gen kann mehrere Proteine kodieren. Möglich wird dies durch alternatives Spleißen. Aus einer prä-mRNA, dem sogenannten *Primärtranskript*, können verschiedene Transkripte und somit verschiedene Proteine z.B. durch »überspringen« von Exons entstehen. Alternative Transkripte bestehen aus verschiedenen Exonketten. Wie das Primärtranskript gespleißt wird, ist oft gewebeabhängig. Die *P-Elemente* von *Drosophila melanogaster* sind ein Beispiel [14]. Die Primärtranskripte dieser Gene haben in somatischen Zellen ein anderes »Spleißmuster« als in Zellen der Keimbahn.

3.2 Das Programm AUGUSTUS

AUGUSTUS basiert auf einem wahrscheinlichkeitstheoretischen Ansatz, einem sogenannten *Generalisierten Hidden-Markow-Modell*. Modelliert werden die Eigenschaften eukaryotischer Genstrukturen, deren Nukleotidsequenzfolgen als zufällig und vom Modell erzeugt (*emittiert*) betrachtet werden. Einfach ausgedrückt ist eine Genstruktur eine Folge von Merkmalen, wobei die Merkmale nur in bestimmter Reihenfolge auftreten. Eine solche Folge ist z.B. die folgende: Donor-Spleißstelle; Intron; Akzeptor-Spleißstelle; internes Exon; Donor-Spleißstelle; Intron; Akzeptor-Spleißstelle; terminales Exon. Grundsätzlich ist dabei z. B., dass auf ein Intron immer eine Akzeptor-Spleißstelle folgt und auf diese dann ein internes oder terminales Exon. Also nur eine solche Folge von Intron, Akzeptor-Spleißstelle und Exon ist biologisch sinnvoll. Das Modell besteht aus sogenannten *Zuständen*, die bestimmte der betrachteten Merkmale einer Genstruktur, wie z.B. eine Spleißstelle oder ein terminales Exon repräsentieren. Das Modell einer Genstruktur ist also eine Folge von Zuständen, wobei die Übergänge von einem Zustand in einen anderen nur in einer biologisch sinnvollen Weise erlaubt sind.

Für einige Zustände gibt es mehrere Möglichkeiten für den folgenden Zustand. Die möglichen Übergänge von einem Zustand in einen anderen haben

eine definierte Wahrscheinlichkeit. Gibt es nur einen möglichen Folgezustand so ist die Übergangswahrscheinlichkeit in diesen gleich eins. Die Folge der Zustände ist mathematisch ausgedrückt eine *homogene Markow Kette* mit definiertem Zustandsraum und Übergangsmatrix (für die genaue mathematische Definition wird auf [13, 26] verwiesen). Die Zustände emittieren eine Nukleotidsequenz zufälliger Länge und Sequenzfolge. Diese Folge, die sogenannte *Emissionsfolge*, kann beobachtet werden und ist mathematisch betrachtet eine Folge von Zufallsvariablen mit Werten aus einem definierten Alphabet. Die Folge der Zustände ist unbekannt (*hidden*) und soll unter Verwendung der Beobachtung aufgedeckt werden. Die Verteilung der Sequenzfolge, die ein Zustand emittiert und die Übergangswahrscheinlichkeiten von einem Zustand in einen anderen sind charakteristisch für eine Spezies und werden durch eine Menge an Genen mit annotierter Struktur in einem Prozess, der als *Training* des Programms bezeichnet wird, ermittelt. Für eine gegebene Sequenz gibt es oft sehr viele mögliche Kombinationen von Zustands- und Emissionsfolgen, die mit der gegebenen Sequenz konsistent sind. AUGUSTUS findet mit dem Viterbi-Algorithmus für eine gegebene Nukleotidsequenz die wahrscheinlichste Genstruktur. Alternative Transkripte werden dabei ignoriert.

Mit dem Programm `etraining`, das im Programmpaket von AUGUSTUS enthalten ist, können die spezifischen Verteilungen der Sequenzfolgen der Zustände und die Übergangswahrscheinlichkeiten aus einer Menge an annotierten Genen im GenBank-Format [27] ermittelt werden. Die ermittelten Werte (*Parameter*) werden gespeichert und bei der Vorhersage auf dem Genom dieser Spezies von AUGUSTUS benutzt. AUGUSTUS verwendet noch weitere Parameter, sogenannte *Metaparameter*, für die Vorhersage. Die Metaparameter sind auch für die Ermittlung der Übergangswahrscheinlichkeiten und der Verteilungen der Sequenzfolgen nötig. Es handelt sich um Werte, die z.B. die Größe des Fensters der Spleißstellen oder die Ordnung des Markow-Modells definieren. Auch diese Parameter sind spezifisch für die Genomsequenz einer Spezies und müssen durch Variation und Beobachtung der resultierenden Vorhersagegenauigkeit auf annotierten Genen optimiert werden.

AUGUSTUS erstellt bei Übergabe einer Sequenz im GenBank-Format eine Statistik aus Vergleichen der Annotationen mit den Vorhersagen. Die Statistik kann zur Bewertung der Qualität der Vorhersagen herangezogen werden (Erläuterungen in 5.1.2). Mit dem Perl-Skript `optimize_augustus.pl` aus dem Programmpaket von AUGUSTUS ist die Bestimmung geeigneter Metaparameter durch zehnfache Kreuzvalidierung möglich. Die Trainingsmenge aus annotierten Genen im GenBank-Format wird in zehn Teile nahezu gleicher Größe gespalten. Neun Teile werden `etraining` zusammen mit den Werten für die Metaparameter übergeben. Der zehnte Teil wird für die Evaluierung der Vorhersage mit diesen Metaparametern und den von `etraining` ermittelten Parametern verwendet. Der Prozess wird ohne Veränderung der Metaparameter zehn Mal wiederholt, wobei für die Evaluierung je ein anderer Teil herangezogen wird. Dieser Schritt wird für alle zu testenden Werte eines Metaparameters durchgeführt.

Für die Spezies: *Aedes aegypti*, *Drosophila melanogaster*, *Homo sapiens*, *Arabidopsis thaliana*, *Brugia malayi* und *Coprinus cinereus* wurden die Parameter und die Metaparameter bereits optimiert. Für *Tribolium* war die Optimierung Teil der vorliegenden Arbeit.

Bei dem Prozess der Annotation der Gene im Genom von *Tribolium* mit AUGUSTUS ist das Training des Programms auf die Genomsequenz die eigentliche Aufgabe. Im optimalen Fall steht für das Training eine ausreichend große Menge an bereits annotierten Genen zur Verfügung. Dies ist aber für laufende oder gerade abgeschlossene Sequenzierungsprojekte oft nicht gegeben. Für die Genomsequenzen solcher Projekte existieren, wie auch für *Tribolium*, oft nur wenige bekannte und annotierte Genstrukturen.

3.3 Möglichkeiten der Erstellung von Genmodellen für das Training von AUGUSTUS

Für *Tribolium castaneum* gibt es nur sehr wenige annotierte Gene. Die meisten davon enthalten nur ungenaue Angaben über kodierende Sequenzab-

schnitte und können für das Training von AUGUSTUS nicht verwendet werden. Für das Training von AUGUSTUS sollte eine Menge aus mindestens 200 Genen verwendet werden (Dr. Mario Stanke, persönliche Mitteilung).

Eine Möglichkeit diese Zahl an Genmodellen zu erhalten, bieten Sequenzvergleiche mit bereits bekannten Proteinsequenzen eng verwandter Spezies, da kodierende Exons verwandter Gene in ihrer Sequenz ähnlich sind und je näher die Spezies miteinander verwandt sind, um so größer ist oft diese Ähnlichkeit [14]. Mit dem Programm BLAST [28] können effizient und schnell Ähnlichkeiten einer Sequenz zu Sequenzen in einer Datenbank gefunden werden. Durch Suche nach offenen Leserahmen können dann Genmodelle erstellt werden. *Tribolium* gehört innerhalb der Tierklasse der Insekten zu den Fluginsekten (Pterygota) und dort zur niedrigsten Ordnung der Holometabola, den Käfern (Coleoptera). Interessant für eine engere Auswahl bei diesem Ansatz sind die annotierten Gene von *Drosophila melanogaster*. *Drosophila* gehört wie *Tribolium* zu den holometabolen Insekten. Innerhalb dieses Ranges aber zur höheren Ordnung Diptera. *Drosophila* dient schon seit Jahrzehnten als genetischer Modellorganismus und die Genomsequenz dieser Spezies ist seit 2000 bekannt. Demnach müsste auch die Annotation qualitativ besser sein als bei Insekten mit ähnlichem Verwandtschaftsgrad, deren Genom aber erst viel später sequenziert wurde, wie z.B. *Apis melifera* oder *Aedes aegypti*. Für *Drosophila* existieren umfangreiche Datensammlungen annotierter Gene aus vielen bekannten Datenbankprojekten, wie z.B. FlyBase [29] oder Ensembl [30].

Eine weitere Möglichkeit ist die Vorhersage mit Parametern einer anderen Spezies. Dabei können die Parameter einer anderen als der mit *Tribolium* am engsten verwandten Spezies zufällig am besten sein (siehe Abschnitt 5.1.2). Die Vorhersage mit übernommenen Parametern kann aber von sehr schlechter Qualität sein. Auch wenn die Spezies enger verwandt sind, bedeutet das nicht, dass die optimalen Parameter die gleichen Werte haben, denn die trainierte Verteilung der Sequenzfolgen betrifft nicht nur die kodierenden Exons sondern auch nicht-kodierende Bereiche wie z.B. die Sequenzen um die Spleißstellen oder allgemein die Sequenzen der Introns. Solche Bereiche sind

bei verwandten Arten weit weniger konserviert als die kodierenden Bereiche der Gene [14] und können ganz andere Verteilungsmuster haben. Es kann also nicht gefolgert werden, welche Parameter das beste Ergebnis liefern würden. Die vorhergesagten Gene können wiederum zum Trainieren benutzt werden. Wiederholungen dieses Vorgangs könnten die Parametereinstellungen für das Genom von *Tribolium* spezifischer werden lassen.

Die beste Möglichkeit um qualitativ gute Genmodelle zu erhalten bieten die generierten ESTs. Für eine komplette Annotation der Genomsequenz ist ihre Zahl viel zu gering, für das Training von AUGUSTUS aber ausreichend. Die EST-Sequenzen enthalten spezifische Informationen über transkribierte Genabschnitte und durch Alignment mit der genomischen Sequenz können aus den ESTs Genstrukturen von *Tribolium* rekonstruiert werden. Ein EST repräsentiert meist nur einen Teil der Exons eines Gens. Für ein Gen können aber mehrere verschiedene ESTs mit Redundanz existieren, so dass im besten Fall der komplette kodierende Teil durch EST-Sequenzen abgedeckt werden kann. Überlappende ESTs gehören mit großer Wahrscheinlichkeit zu dem selben Transkript und können durch Assemblierung zu einer Sequenz (Assembly) vereinigt werden. Die Gruppierung überlappender ESTs zu einem sogenannten *Cluster* ist deshalb sinnvoll. Durch Zuordnung revers komplementärer ESTs eines Transkripts zu anderen Clustern entstehen chimäre Assemblies. Um diese Fehler zu vermeiden, ist vor dem Clustern und Assemblieren der ESTs ein Sequenzvergleich mit der genomischen Sequenz vorteilhaft [31]. Bei ESTs, deren Orientierung durch die Sequenz selbst anhand von Poly-Adenylsäureenden nicht determinierbar ist, kann die Intronsequenz mit dem Konsensus der Donor- und Akzeptor-Spleißstellen Aufschluss geben. Wenn die Genomsequenz sehr genau ist, wird auch die co-Assemblierung von ESTs unterbunden, die aus den Transkripten eng verwandter Gene generiert wurden. Die Gruppierung der ESTs ist zuverlässiger, wenn die Zuordnung zu einem Cluster durch Alignment der gesamten EST-Sequenz mit einem Abschnitt im Genom und Überlappung mit anderen EST-Sequenzen (solche, die ebenfalls mit diesem Abschnitt im Genom alignieren) definiert wird und nicht nur aus Alignments der ESTs untereinander resultiert, da schlechte Sequenzqualität, vor allem an den Randbereichen der ESTs, keine Seltenheit ist [12].

Aus diesen Gründen sind durch die Kombination aus EST-Genom-Alignment mit Clustering und Assemblierung die besten Ergebnisse für die Rekonstruktion der Transkripte der Gene und der Genstrukturen zu erwarten. Populäre Programme, entwickelt um EST- oder cDNA-Sequenzen mit genomischer DNA zu alignieren, sind BLAT [32], SIM4 [33], GAP2 [34], SPIDEY [35] und GENESEQER [36]. Es resultieren *Spliced-Alignments*. Die ESTs repräsentieren nur die transkribierten Abschnitte der Gene. An der Position, wo ein Intron in der genomischen Sequenz zwei Exons trennt, die mit einem EST alignieren, wird eine Lücke vom Spliced-Alignment-Programm in die Sequenz des ESTs eingefügt. Dadurch kann die Genstruktur identifiziert werden. Überlappende Alignments können zu Clustern gruppiert werden. Durch Assemblierung überlappender, konsistenter EST-Sequenzen eines Clusters können die Transkriptsequenzen und ihre Intron-Exon-Strukturen erhalten werden. Dieser Ansatz ist in dem Programm PASA [31] realisiert, dass im Folgenden beschrieben ist.

3.3.1 Program to Assemble Spliced Alignments

PASA wurde von TIGR (The Institute for Genomic Research) u.a. zur Verbesserung der Genannotation von *Arabidopsis thaliana* implementiert und für dieses Projekt erfolgreich [31] eingesetzt. Wenn im Folgenden von einem Alignment gesprochen wird, ist immer das Spliced-Alignment eines ESTs mit der genomischen Sequenz gemeint. Die ESTs überlappender Alignments können zu unterschiedlichen Transkripten desselben Gens gehören. Der PASA-Algorithmus [31] assembliert kompatible, überlappende Alignments eines Clusters. Die folgende Abbildung zeigt ein Cluster überlappender Alignments. Die Alignments `a` und `b` sowie `b` und `c` dieses Clusters sind kompatibel aber die Alignments `a` und `c` sind es nicht.

```
a: -----| |-----| |-------
b:    --| |-----| |--
c:    --| |-----| |---| |-----|  |-------
```

Alignments sind kompatibel, wenn sie die gleiche Orientierung haben und wenn im überlappenden Bereich die Exons und Introns identische Positionen haben. Ein Assembly aus kompatiblen Alignments könnte somit einem Teil oder der kompletten Sequenz eines Transkripts eines Gens entsprechen. Die Assemblies werden durch *Dynamisches Programmieren* erhalten.

An dem folgenden Beispiel, das ein Clusters aus überlappenden Alignments darstellt, wird der Algorithmus erläutert.

```
a: -------| |------------
b:    --| |------------
c:                   -------------|  |---------
d:                   -------------|      |--------
```

Für ein Cluster überlappender Alignments werden zunächst alle Paare kompatibler Alignments gebildet und die Alignments werden nach Anfangspositionen entlang der genomischen Sequenz sortiert. Konsistent sind Alignment `a` und `c`, `a` und `d`, `b` und `c` sowie `b` und `d`. Die Alignments werden von links nach rechts assembliert. Nur kompatible Alignments werden dabei berücksichtigt. Für das Beispiel bedeutet das, die Alignments `a` und `c` werden assembliert. Für die verbleibenden Alignments, im Beispiel sind das Alignment `b` und `d`, wird je von links nach rechts (beginnend bei dem Alignment, für welches das Assembly gesucht wird) und von rechts nach links, ein Assembly gesucht, welches das Alignment zusammen mit der maximalen Anzahl kompatibler Alignments enthält. Im Beispiel werden für das verbliebene Alignment `b` die Alignments `b` und `c` sowie die Alignments `b` und `d` assembliert. Für das verbliebene Alignment `d` werden die Alignments `b` und `d` sowie die Alignments `a` und `d` assembliert. Diese Assemblies werden nach Anzahl an enthaltenen Alignments sortiert. Enthalten die gefundenen Assemblies für ein verbliebenes Alignment nun die gleiche Anzahl an Alignments, so wird beliebig eines gewählt (Brian Haas, persönliche Mitteilung). Für das Beispiel wird für das Alignment `b` das Assembly aus `b` und `c` beliebig erwählt. Für Alignment `d`, das Assembly aus `b` und `d`. Im Beispiel enthalten beide der gewählten Assemblies die gleiche Anzahl an Alignments. Jetzt wird in den erneut sortierten Assemblies nach den verbliebenen Alignments gesucht.

Das Assembly aus `b` und `c` enthält das verbliebene Alignment `b` und das Assembly aus `b` und `d` enthält das verbliebene Alignment `d`. Somit werden für das Beispiel mit dem PASA-Algorithmus drei Assemblies erhalten. Möglich ist auch, dass für das verbliebene Alignment `b` das Assembly aus `b` und `d` zufällig erwählt wird. Dieses Assembly enthält dann beide verbliebene Alignments und somit erhält man insgesamt nur zwei Assemblies. Theoretisch gibt es vier alternative Varianten. Gefunden werden aber nur maximal drei davon. Für das Training von AUGUSTUS ist nur ein Transkript eines Gens von Bedeutung.

Der Algorithmus zur Assemblierung der Alignments ist Teil einer sogenannten *Pipeline*, einer Folge von automatischen Programmaufrufen, wobei die Ausgabe eines Programms die Eingabe des nächsten ist. Es handelt sich um Perl-Skripte, die externe Programme wie BLAT oder SIM4 aufrufen und deren Ausgabe verarbeiten. Die wichtigsten Schritte und Validierungen, die durchgeführt werden müssen, um aus den EST-Sequenzen und der genomischen Sequenz Genmodelle mit guter Qualität zu erhalten, sind kombiniert und automatisiert. Die Spliced-Alignment-Programme BLAT und SIM4 werden eingesetzt um die ESTs mit der genomischen Sequenz zu alignieren. BLAT wird zuerst verwendet. Die Ergebnisse werden einer Validierung unterzogen. Gefiltert werden die Alignments, die den [GT,GC]/AG Konsensus der Donor/Akzeptor-Spleißstellen bei allen Introns erfüllen und deren Sequenz zu min. 90% mit 95% Identität aligniert. Alignments, die dieser Validierung nicht Stand halten, werden SIM4 übergeben und erneut validiert. Eingabe ist hier der Sequenzabschnitt des Alignments und fünf Kb flankierende genomische Sequenz. Nach diesem Prozess werden Cluster überlappender Alignments durch *Single Linkage Clustering* [37] gebildet und jedes Cluster wird jetzt dem PASA Algorithmus übergeben. Die resultierenden Assemblies enthalten die, aus den ESTs und der genomischen Sequenz beziehbaren Informationen über Transkriptsequenzen und Genstrukturen, die den Validierungskriterien nach sehr wahrscheinlich sind. Die Pipeline arbeitet mit einer MySQL-Datenbank zur Speicherung der Zwischen- und Endergebnisse. Die Ergebnisse der Assemblierung werden in HTML-Dokumenten eingebettet visualisiert. Die Annotationen beschränken sich auf die Exon-Intron-

Strukturen der assemblierten ESTs und enthalten keine Informationen über offene Leserahmen.

Das PASA-Programmpaket enthält Perl-Module, mit deren Hilfe ein offener Leserahmen für ein Assembly gesucht und die Genstruktur in GFF3-Format [38] formuliert werden kann. Ein Perl-Programm, das von diesen Modulen Gebrauch macht, ist nicht im Paket enthalten, wurde aber von Brian Haas zur Verfügung gestellt. Ausgabe dieses Programms ist eine Datei im GFF3-Format, die teilweise unvollständige Gene, deren Aminosäuresequenzen eine definierbare Mindestlänge haben, enthält.

Die PASA-Pipeline wurde implementiert, um eine bereits existierende Annotation einer Genomsequenz durch Vergleiche mit den Ergebnissen der Assemblierung zu verbessern. Dazu muss diese Annotation in einer definierten Form in die von PASA benutzte MySQL-Datenbank eingelesen werden. Das Ergebnis sind Validierungen der Exon-Intron-Strukturen und Ergänzungen wie UTRs (untranslated Regions) und alternative Transkripte. Es besteht die Möglichkeit, eine mit AUGUSTUS erstellte Genannotation mit dieser Option zu verbessern. Das Ausmaß der Qualitätssteigerung ist dabei von der Anzahl der Assemblies abhängig.

3.3.2 Realisierung der Datengenerierung

Für Alignment, Clustering und Assemblierung der Sequenzdaten von *Tribolium castaneum* wurde die PASA-Pipeline verwendet. Die Sequenzdaten beliefen sich zunächst auf das erste Assembly der Genomsequenz und die EST-Sequenzen der Universität zu Köln. Aus diesen Daten wurde eine erste Trainingsmenge an Genmodellen im Wesentlichen nach den im Folgenden beschrieben Kriterien erstellt. Seit Ende September 2005 ist das zweite Assembly und eine umfangreichere Sammlung an EST-Sequenzen, produziert vom Human Genome Sequencing Center, zugänglich. Mit diesen Daten wurde eine zweite Trainingsmenge generiert.

Für das Traininging von AUGUSTUS müssen die Daten im GenBank-Format

vorliegen. Die Annotationen der zum Training verwendeten Genmodelle sollten mit großer Wahrscheinlichkeit korrekt sein und eine ausreichende Menge an Modellen ist nötig. Nur die kodierenden Teile eines Gens und ein kleiner Sequenzabschnitt um den Translationsstart sind für das Training von Bedeutung. Eine korrekte Annotation des ersten Exons ist für den Prozess des Trainings wesentlich wichtiger als ein korrekt annotiertes terminales Exon. So können auch Gene verwendet werden, deren Start bekannt ist, aber deren Stopp nur vermutet werden kann. Nach diesen Anforderungen wurden die Assemblies gefiltert und die Genmodelle erstellt.

Die mit PASA erstellten Annotationen über die Lage von Exons und Introns wurden durch Suche nach offenen Leserahmen in der assemblierten Transkriptsequenz vervollständigt. Dazu bieten einige der in Perl implementierten Module des PASA-Programmpaketes geeignete Funktionen. Die implementierten Perl-Skripte, die von diesen Modulen Gebrauch machen und welche die im Folgenden beschriebenen Schritte ausführen, können der beiliegenden CD entnommen werden. Die nötigen Informationen über die Assemblies wurden aus der von PASA benutzten MySQL-Datenbank entnommen. In Anlehnung an die, von den Modulen gebotenen Möglichkeiten wurde für jedes Assembly ein Objekt erstellt. So können die Informationen über Exon-Intron-Strukturen und Nukleotid- und Proteinsequenzen auf einfache Weise verändert, ergänzt oder abgefragt werden. Für die Sequenzen der Assemblies wurde der längste offene Leserahmen gesucht. Dieser erfüllt die folgenden Kriterien:

1. Ein Stoppkodon im gleichen Leserahmen innerhalb der Sequenz des Assemblies vor dem Startkodon.

2. Eine Mindestlänge von 100 Kodons.

Genmodelle, die ein Stoppkodon in der Sequenz des Assemblies enthalten, werden als komplett aufgefasst. Für andere Modelle wurde ein Stoppkodon in der, an die Sequenz des Assemblies angrenzenden, genomischen Sequenz gesucht. Dieses Stoppkodon ist aber nur eine Mutmaßung. Es ist möglich, dass noch ein weiteres Exon folgt und das gefundene Kodon in der Sequenz eines Introns liegt.

Die Genmodelle müssen in GenBank-Format an AUGUSTUS übergeben werden. Eine detaillierte Übersicht über die Definitionen dieses Formats ist unter der URL [27] gegeben. Zur Formulierung dieses Sequenzformats wurden Funktionen der Perl-Module *Bio::SeqFeature::Generic* und *Bio::SeqIO* aus dem Programmpaket BIOPERL verwendet. Die Dokumentationen und das gesamte Programmpaket oder einzelne Module sind unter der URL [39] veröffentlicht. Eine Sequenz der GenBank-Datei kann mehrere Gene enthalten und diese können auf dem *Plus-* oder *Minus*-Strang annotiert sein. Die Gene einer Sequenz dürfen sich aber nicht überlappen. Pro Gen ist somit auch nur ein Transkript zulässig. Aus alternativen Transkripten oder überlappenden Genen wurde eines nach den folgenden Regeln erwählt.

1. Sind unter den überlappenden Modellen komplette Gene, werden alle anderen verworfen und es wird nach den weiteren Regeln nur zwischen diesen gewählt.

2. Die größte Anzahl an Exons wird bevorzugt, da dies zum Training der Intron-Modelle von AUGUSTUS von Vorteil ist.

3. Bei gleicher Anzahl an Exons wird die längste kodierende Sequenz bevorzugt.

Die Trainingsmenge sollte nicht-redundant sein. Die kodierten Aminosäuresequenzen sollten nur zu maximal 70% identisch sein, da sonst eine bestimmte Verteilung von Nukleotiden zu große Präferenz erlangt (Dr. Mario Stanke, persönliche Mitteilung). Um dies zu überprüfen wurden die Aminosäuresequenzen der gewählten Genmodelle in einer Datei im Fasta-Format [42] zusammengefasst und ihre Ähnlichkeiten mit einer BLAST-Suche (Das Programm BLAST ist in Abschnitt 4.2 beschreiben) festgestellt. Verwendet wurde das Programm BLASTP, für die Suche mit einer Sequenz aus Aminosäuren in einer Datenbank aus Aminosäuresequenzen. Die Ausgabe wurde nach Treffern, sogenannten *High Scoring Segment Pairs* (HSPs) (siehe Abschnitt 4.2) durchsucht, die min. 70% Identität mit dem Suchmuster haben. Die entsprechenden GenBank-Sequenzen mit diesen Genmodellen wurden verworfen. Für das Training der Splice-Site-Modelle von AUGUSTUS wird

eine Datei erstellt, welche die Spleißstellen der Genmodelle mit jeweils 40 Nukleotiden flankierender Sequenz enthält. Diese kann beim Training von AUGUSTUS mit übergeben werden.

3.4 Training von AUGUSTUS

Der Prozess des Trainings von AUGUSTUS ist mit dem Skript `optimize_augustus.pl` automatisiert. Die folgenden Daten wurden für den Prozess erstellt:

- Eine Menge wahrscheinlich vollständiger Genmodelle im GenBank-Format, die zur Evaluierung und Ermittlung der Parameter während des Trainings eingesetzt werden. Die GenBank-Sequenzen enthalten nur solche Gene, deren Stoppkodon in der Sequenz des Assemblies enthalten ist.
- Eine Menge an Genmodellen im GenBank-Format, die nur zur Ermittlung der Parameter eingesetzt werden. Diese GenBank-Sequenzen enthalten mindestens ein Modell, dessen Stoppkodon durch Verlängerung in die genomische Sequenz gefunden wurde.
- Eine Menge an Genmodellen im GenBank-Format, die zufällig aus allen wahrscheinlich vollständigen Genmodellen ausgewählt wurden. Diese dienten nach Abschluss des Prozesses zum Testen der erreichten Qualität.
- Eine Zusammenstellung der Spleißstellen aller Genmodelle, die jeweils 40 Nukleotide flankierende Sequenz von beiden Seiten der Spleißstellen enthält.

Die unvollständigen Genmodelle, deren Stoppkodon nicht in der Sequenz des Assemblies gefunden wurde, werden nur zur Ermittlung der Parameter herangezogen und nicht für ihre Evaluierung. Ein unvollständig annotiertes Gen könnte vollständig und korrekt vorhergesagt werden, wird aber als falsch positives Ergebnis gewertet. Das könnte günstige Parametereinstellungen benachteiligen. Das Optimierungsskript liest aus einer Datei die Metaparameter, welche optimiert werden sollen und die Spanne ihrer möglichen Werte

aus. Für das Training wurde die Datei `generic_metapars.cfg` übergeben.

Die erste für *Tribolium* optimierte Version von AUGUSTUS wurde nicht mit dem Skript `optimize_augustus.pl` erstellt. Es wurde ein vorhandenes Skript verwendet, das ohne die beschriebene Kreuzvalidierung arbeitet und nicht zwischen kompletten und anderen Modellen unterscheidet. Die Menge an Genmodellen für das erste Training entstand aus dem ersten Assembly der Genomsequenz und den EST-Daten der Universität zu Köln. Die Daten können der beiliegenden CD entnommen werden.

Kapitel 4

Vergleichende Genomanalyse zwischen *Tribolium castaneum*, *Drosophila melanogaster* und *Homo sapiens*

Die vergleichenden Analysen der Genome von *Tribolium castaneum*, *Drosophila melanogaster* und *Homo sapiens* beziehen sich auf die Proteine, die bei *Drosophila melanogaster* und im menschlichen Genom bereits identifiziert sind und für *Tribolium castaneum* mit AUGUSTUS vorhergesagt wurden. Die Bedeutung der Genomsequenz von *Tribolium castaneum* für die Identifizierung *homologer* Proteine in Mensch und *Drosophila melanogaster* soll untersucht werden. Dabei sind mit homologen Proteinen solche gemeint, die von einem gemeinsamen *Vorläufer* abstammen. Homologe Proteine zeigen oft divergente Entwicklung in verschiedenen Spezies [14]. Die durchgeführten Analysen sollen weiterhin zeigen, ob menschliche Proteine gefunden werden können, zu denen in *Drosophila* keine Homologen identifizierbar sind, die aber signifikante Ähnlichkeit zu Proteinen aus *Tribolium* haben.

Besonderes Interesse gilt im Allgemeinen der Aufklärung der Funktionen menschlicher Proteine. Für die Klassifizierung der Funktion eines Proteins können homologe Proteine genetisch leicht manipulierbarer Organismen hilf-

reich sein, da viele Funktionen durch die Auswirkung ihres Fehlens auf den Phänotyp (*knock-out Mutationen*) eingeordnet werden können und dies nur in genetisch manipulierbaren Modellorganismen untersucht werden kann.

Ein solcher Modellorganismus ist *Drosophila melanogaster*. Innerhalb des Tierstammes der Arthropoden (Gliederfüßer) wird *Drosophila* in Bezug auf Flexibilität und Möglichkeiten der genetischen Manipulation bislang von keiner anderen Spezies übertroffen [2]. Die Suche nach homologen Proteinen für Funktionsanalysen in *Drosophila melanogaster* ist daher keine Seltenheit. Homologe Proteine können meist an ihrer wesentlichen Sequenzähnlichkeit und an konservierten Aminosäureresten erkannt werden [14]. Oft haben homologe Proteine gleiche oder noch ähnliche Funktionen. Die Ähnlichkeit von Aminosäuresequenzen kann ähnliche Proteinstrukturen und somit ähnliche Funktionen bedeuten. Ähnliche Funktionen der Proteine bedeuten aber nicht unbedingt eine Ähnlichkeit der Aminosäuresequenzen. Die divergente Entwicklung kann zu drastischen Veränderungen der Aminosäuresequenz führen, denn für den Erhalt der Funktion eines Proteins können nur wenige Aminosäuren essentiell sein [14]. Aus diesem Grund ist es vorteilhaft, mehrere Genome genetischer Modellorganismen für die Suche nach Homologien zur Verfügung zu haben. Es ist beispielsweise möglich, dass für ein menschliches Protein kein homologes Protein in *Drosophila melanogaster* über direkten Vergleich identifizierbar ist. Ein Vergleich des menschlichen Proteins mit Sequenzen von *Tribolium castaneum* kann aber ein positives Ergebnis haben. Dieses Protein aus *Tribolium* bietet eine neue Chance, das gesuchte Protein in *Drosophila* zu finden oder aber mit *Tribolium* als Modellorganismus zu arbeiten. Denn ein Experiment mit *Tribolium* als Modellorganismus kann Erfolg haben, wenn sich *Drosophila* nicht eignet.

Die Bedeutung der Genomsequenz von *Tribolium castaneum* für die Identifizierung homologer Proteine in *Drosophila melanogaster* und Mensch kann durch Betrachtung bestimmter Mengen homologer Proteine aus den drei Spezies eingeschätzt werden. Signifikante Sequenzähnlichkeit ist dabei das Kriterium, an dem eine Homologie erkannt wird. Interessant ist ein Sequenzvergleich der Proteine von *Tribolium* mit den menschlichen Proteinen, für

die in *Drosophila* keine homologen Proteine gefunden wurden. Für Proteine aus *Tribolium*, die bei diesem Vergleich durch signifikante Sequenzähnlichkeit auf eine mögliche Homologie schließen lassen, ist ein Sequenzvergleich mit Proteinen von *Drosophila* wiederum interessant.

Werden in *Drosophila* Proteine mit signifikanter Sequenzähnlichkeit zu diesen Proteinen aus *Tribolium* gefunden, kann das bedeuten, dass die menschlichen Proteine, die ebenfalls signifikante Sequenzähnlichkeit mit diesen Proteinen aus *Tribolium* haben, zu den entsprechenden Proteinen aus *Drosophila* homolog sind. Solche Proteine aus *Tribolium* könnten die Suche nach Homologien zwischen Mensch und *Drosophila* entscheidend unterstützen, denn der direkte Vergleich zwischen Mensch und *Drosophila* liefert kein Ergebnis.

Werden in *Drosophila* keine Proteine mit signifikanter Sequenzähnlichkeit zu diesen Proteinen aus *Tribolium* gefunden, gibt es für die betrachteten menschlichen Proteine vermutlich nur Homologe in *Tribolium*. Auch der Sequenzvergleich mit Proteinen aus *Tribolium* konnte eine eventuell vorhandene Homologie nicht »aufdecken«. Solche Proteine zeigen die Bedeutung der Genomsequenz von *Tribolium castaneum* in besonderem Maße.

Allgemein kann aber auch die Bestätigung einer identifizierbaren oder bekannten Homologie zwischen Mensch und *Drosophila* durch Vergleich mit Sequenzen von *Tribolium* bedeutsam sein. Um die Bedeutung der Genomsequenz von *Tribolium castaneum* für die Suche nach Homologien zu menschlichen Proteinen zu betrachten, müssen die Ähnlichkeiten der Proteine zwischen den drei Spezies identifiziert und klassifiziert werden.

4.1 Datenquellen

Die verwendeten Proteinsequenzen aus *Tribolium castaneum* wurden durch Genvorhersage mit AUGUSTUS erhalten. Vorhergesagt wurden 15309 Gene. Die Ausgabe von AUGUSTUS enthält eine Übersetzung der vorhergesagten kodierenden Abschnitte in Aminosäuresequenzen, die übernommen wurde.

Für *Drosophila melanogaster* und Mensch wurden die vom Ensembl-Projekt[1] erstellten Proteinsammlungen für diese Spezies verwendet. Informationen über das Projekt können unter der URL [30] recherchiert werden. Die Proteinsequenzen sind Übersetzungen der von Ensembl annotierten Gene. Verwendet wurden nur Proteinsequenzen, die auch in den Datenbanken Swiss-Prot, RefSeq oder SPTrEMBL enthalten sind (hier kann man von bekannten Genen ausgehen [30]) und solche, die auf der Basis von Vergleichen mit nah verwandten Spezies vorhergesagt wurden. Es sind keine Sequenzen enthalten, die durch ab initio Genvorhersage annotiert wurden. Die Sammlung an Proteinen aus *Drosophila melanogaster* besteht aus 19369 Proteinen und die Sammlung an menschlichen Proteinen enthält 33869 Sequenzen. Die Daten können der beiliegenden CD entnommen werden.

4.2 Identifizierung homologer Proteine

Homologe Proteine können auf Grund ihrer Sequenzähnlichkeit identifiziert werden. Zur Klassifizierung der Proteine in Homologe und Andere wurden die Aminosäuresequenzen untereinander verglichen. Eingesetzt wurde dafür das populäre Alignment-Programm BLAST (Basic Local Alignment Search Tool) [25].

Das Programm BLAST

BLAST ist ein Programmpaket für verschiedene Sequenzvergleiche zur Suche nach Ähnlichkeiten einer Sequenz (*Query*) zu Sequenzen aus einer Datenbank. Der Algorithmus ist aus Gründen der Schnelligkeit heuristisch und versucht, optimale lokale Alignments zwischen dem übergebenen Muster und den Sequenzen der Datenbank zu finden. Das bedeutet, ein gefundenes lokales Alignment muss nicht das Optimale sein. Ein lokales Alignment ist ein Alignment zwischen zwei Sequenzen, das sich nur auf einen Teil der Gesamtsequenz beschränkt. Für die Funktion eines Proteins sind oft nur die Aminosäuren einer oder einiger Domänen essentiell, vor allem die der *Active-Site* des Proteins, des Zentrums der katalytischen Aktivität [14]. Die

[1] Gemeinschaftsprojekt zwischen EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institut).

Erhaltung der Funktion ist bei der divergenten Entwicklung homologer Proteine oft gegeben und somit sind die dafür essentiellen Aminosäurereste meist konserviert. Die Suche nach lokalen Alignments ist für die Identifikation homologer Proteine also gerechtfertigt.

BLAST-Algorithmus

Für die Interpretation der Ergebnisse einer BLAST-Suche ist es vorteilhaft zu wissen, wie diese generiert werden. Die »Suchsequenz« wird in kurze Abschnitte (*Worte*) zerlegt. Für Proteinsequenzen ist die Länge der Worte drei. Für jedes Wort wird eine Liste mit ähnlichen Wörtern gleicher Länge angelegt und die Ähnlichkeit wird mit einem *Score* bewertet. Für Proteinsequenzen wird für die Bewertung der Ähnlichkeit eine sogenannte *Scoring Matrix* benutzt. Diese Matrix enthält für je zwei Aminosäuren einen Wert, der die Wahrscheinlichkeit repräsentiert, mit der diese beiden Aminosäuren im Lauf der Evolution gegeneinander ausgetauscht wurden. Dabei werden u.a. die chemischen Eigenschaften und die Molekülstrukturen der Aminosäuren berücksichtigt. Der Austausch einer Aminosäure mit essentieller Bedeutung für die Funktion des Proteins gegen eine Aminosäure mit ähnlichen Eigenschaften ist wahrscheinlicher als der Tausch gegen eine Aminosäure ganz anderer Art. Auch die Werte für zwei Paare jeweils gleicher Aminosäuren können sich unterscheiden. Der Score ergibt sich durch Aufsummierung der Bewertungen jeder einzelnen alignierten Position. Die Einfügung einer Lücke (*gap*) zwischen den alignierten Positionen, die eine Insertion oder Deletion widerspiegelt, wird dabei negativ bewertet. Mit der Liste der Wörter wird nach Übereinstimmungen in den Sequenzen der Datenbank gesucht. Treffer mit Wörtern, deren *Score* einen festgelegten Wert überschreitet, werden zu beiden Seiten solange verlängert, bis der Score des Alignments, der durch die Verlängerung erreicht wird, den bislang erreichten maximalen Score um einen festgesetzten Wert unterschreitet. Diese Alignments werden als *High Scoring Pairs* (HSPs) bezeichnet. HSPs, die eine bestimmte statistische Signifikanz haben oder deren Score einen festgesetzten Wert überschreitet, werden dann vom Programm als Ergebnis ausgegeben.

Um die Ergebnisse beurteilen zu können, werden zwei Werte für jedes Alignment errechnet. Zum einen der *E-Wert*, der die statistische Signifikanz des Treffers widerspiegelt und zum anderen der *Bitscore*, eine Normalisierung des errechneten Scores, mit dem Ziel diesen vergleichbar zu machen, ohne das verwendete Bewertungssystem berücksichtigen zu müssen. Bei der Betrachtung des Bitscores muss nur die Größe des Suchraumes berücksichtigt werden. Der von *Blast* berechnete E-Wert ist eine Schätzung der erwarteten Anzahl der Treffer mit einem Score größer oder gleich dem Score des betrachteten Treffers, wenn die Eingabesequenz und die Datenbank zufällig erzeugt wurden.

Für die Bewertung der Ergebnisse der BLAST-Suche wird der E-Wert der Hits herangezogen, denn der E-Wert setzt den Bitscore des Treffers in Bezug zur Größe des Suchraumes. Je kleiner der E-Wert, umso wahrscheinlicher ist eine Homologie zwischen Suchmuster und Treffer.

BLASTP-Ausgabe

Für die Suche nach Homologien wurde das Programm BLASTP, Version 2.2.8 [28], für Vergleiche von Aminosäuresequenzen verwendet. Die Ausgabe des Programms enthält detaillierte Informationen über die erhaltenen Treffer, nach absteigender Signifikanz sortiert. Eine Sequenz kann mehrere lokale Alignments (HSPs) enthalten, deren Bitscore und E-Wert zu einem Treffer (*Hit*) zusammengefasst werden. Jede Sequenz einer Datenbank ist im Allgemeinen mit einem eindeutigen Identifikator versehen (*Accession-Nummer, gi-Nummer*). In der Ausgabe folgt die Auflistung der Bezeichner zusammen mit dem Bitscore und dem E-Wert der Hits nach Angabe der Query und der Datenbank, die durchsucht wurde:

```
Query= gi|73486646|gb|AAJJ01000001.1|:g1.t1
         (322 letters)

Database: Homo_sapiens.NCBI35.nov.pep.fa
           33,869 sequences; 16,881,503 total letters

Searching.....................................................done
```

```
                                                                Score    E
Sequences producing significant alignments:                     (bits) Value

ENSP00000318351 pep:known-ccds chromosome:NCBI35:6:80873083:8111...   486   e-137
ENSP00000348880 pep:known-ccds chromosome:NCBI35:6:80873083:8111...   486   e-137
```

Nach dieser Auflistung erfolgen für jeden Hit die Angaben aller lokalen Teiltreffer (HSPs) mit dem Score und dem E-Wert:

```
 Score =  486 bits (1250), Expect = e-137
 Identities = 219/322 (68%), Positives = 275/322 (85%)

Query: 1   MNMFQAINNALDLALKQDESALIFGEDVAFGGVFRCTMGLQSKYGPGRVFNTPLCEQGIV 60
           MN+FQ++ +ALD +L +D +A+IFGEDVAFGGVFRCT+GL+ KYG  RVFNTPLCEQGIV
Sbjct: 71  MNLFQSVTSALDNSLAKDPTAVIFGEDVAFGGVFRCTVGLRDKYGKDRVFNTPLCEQGIV 130
```

Ein E-Wert von 0 bedeutet, dass der Wert kleiner als 10^{-180} ist.

Durchführung der Suche nach Homologien

Die Proteine, welche die Bedeutung der Genomsequenz von *Tribolium castaneum* andeuten, wurden aus den Ausgaben der drei, im Folgenden beschriebenen BLAST-Suchen extrahiert:

1. Alle Sequenzen der Sammlung an menschlichen Proteinen wurden als Query verwendet, um in der Sammlung der Proteine von *Drosophila melanogaster* Homologien zu finden.

2. Alle Sequenzen, die im Genom von *Tribolium castaneum* mit AUGUSTUS vorhergesagt wurden, wurden als Query verwendet, um in der Sammlung der menschlichen Proteine Homologien zu finden.

3. Alle Sequenzen, die im Genom von *Tribolium castaneum* mit AUGUSTUS vorhergesagt wurden, wurden als Query verwendet, um in der Sammlung der Proteine von *Drosophila melanogaster* Homologien zu finden.

Aus der ersten Suche sind nur solche menschlichen Proteine interessant, für die in *Drosophila* keine Proteine mit signifikanter Sequenzähnlichkeit gefunden werden. Diese menschlichen Proteine werden als (*negative* Treffer) dieser BLAST-Suche aus der Ausgabe selektiert. Die Proteine aus *Tribolium* mit signifikanter Ähnlichkeit zu einem negativen Treffer der ersten Suche, sind von Bedeutung. Für die zweite BLAST-Suche werden solche Proteine aus *Tribolium* als positive Treffer aufgefasst. Aus dem Ergebnis der dritten BLAST-Suche werden für diese Proteine Treffer mit Proteinen aus *Drosophila* gesucht, die signifikante oder sehr geringe Sequenzähnlichkeit andeuten. Für die dritte BLAST-Suche sind diese Proteine von *Tribolium* die positiven bzw. negativen Treffer.

Selektion und Klassifizierung der relevanten Treffer

Der E-Wert der Treffer der BLAST-Suche dient der Klassifizierung in positive oder negative Treffer. »Positive« und »negative« Grenzen für den E-Wert werden definiert. Eine Query, die einen Treffer mit einer Sequenz der Datenbank hervorbringt, dessen E-Wert kleiner oder gleich der definierten positiven Grenze für den E-Wert ist, wird als positiver Treffer bezeichnet. Eine Query, die nur Treffer mit Sequenzen der Datenbank hervorbringt, deren E-Werte größer oder gleich der definierten negativen Grenze für den E-Wert sind, wird als negativer Treffer bezeichnet.

Proteine mit signifikanter Sequenzähnlichkeit, die deshalb wahrscheinlich Homologe sind, haben im Allgemeinen einen kleinen E-Wert. Bis zu einem E-Wert von maximal 10^{-9} könnte wahrscheinlich eine Homologie vorliegen. Bei größeren E-Werten wird die Wahrscheinlichkeit einer Homologie immer geringer, denn die meisten homologen Proteine haben sehr ähnliche Sequenzfolgen. Ab welchem Wert eine Homologie mit genügend großer Sicherheit vorliegt muss vorsichtig eingeschätzt werden und ist eine sehr wage Mutmaßung. Im Zweifelsfall kann eine Betrachtung der HSPs hilfreich sein. Es bietet sich deshalb an, die relevanten Informationen aus den Ergebnissen der BLAST-Suchen in einer Datenbank bereitzustellen. Mit einer erstellten Suchmaske kann die Datenbank benutzerfreundlich mit einem Webbrowser nach

den Treffern durchsucht werden, die gewünschte Kriterien erfüllen. Grenzen für die E-Werte positiver und negativer Treffer sind definierbar und es ist auswählbar, ob positive oder negative Treffer einer bestimmten Suche gezeigt werden sollen. Für jeden Treffer können die HSPs über einen einen Link (den E-Wert des positiven oder negativen Treffers) aufgerufen werden.

Kapitel 5

Ergebnisse

Im Folgenden sollen die ermittelten Ergebnisse dargestellt und diskutiert werden. Aufgabe war, AUGUSTUS auf die spezifische Genomsequenz von *Tribolium castaneum* zu trainieren, um die anstehende Annotation dieser Sequenz mit einer qualitativ guten Genvorhersage zu unterstützen. Das Ergebnis der Genvorhersage ist ein erster Ausgangspunkt für eine vergleichende Genomanalyse mit *Drosophila melanogaster* und *Homo sapiens* mit dem Ziel, die Bedeutung der Genomsequenz von *Tribolium* für die Identifikation homologer Proteine in Mensch und *Drosophila* zu analysieren. Weiterhin wird angenommen, dass in *Tribolium* Proteine gefunden werden können, die signifikante Ähnlichkeit zu Proteinen des Menschen aufweisen, so dass eine Homologie vermutet werden kann, obwohl in *Drosophila* keine homologen Proteine zu diesen menschlichen Sequenzen identifiziert werden können. Diese Annahme soll mit Argumenten belegt werden.

5.1 Genannotation

Ein gutes Ergebnis für das Training von AUGUSTUS auf die Genomsequenz von *Tribolium castaneum* ist zu erwarten, wenn die dazu benötigten Genmodelle aus EST-Sequenzdaten erstellt werden. Diese Daten sind direkte Hinweise für exprimierte Genstrukturen. Die vom Human Genome Sequencing Center erstellten EST-Sequenzen wurden verwendet, um eine Menge an Genmodellen zu erhalten. AUGUSTUS konnte mit diesen Genmodellen erfolgreich

trainiert werden. Im Folgenden werden die Ergebnisse der durchgeführten Schritte analysiert.

5.1.1 Analyse der Trainingsdaten

Die Ergebnisse des Trainings sind entscheidend von der Menge und der Qualität, der verwendeten Genmodelle abhängig. Intuitiv klar ist die Relation: je größer die Menge und je besser die Qualität desto größer ist der zu erwartende Erfolg.

Eine Verbesserung der Qualität der Ausgangsdaten ist mit den beschriebenen Sequenzvalidierungen zu erwarten, wenn die Daten Verunreinigungen durch fremde Sequenzen aufweisen oder andere, von den Validierungsprogrammen erkannte Fehler beinhalten. Der EST-Screen des Human Genome Sequencing Center ergab 35649 Sequenzen. Nur 733 dieser Sequenzen wurden von dem Programm seqclean nach den in Abschnitt 2.5 beschriebenen Kriterien als qualitativ zu schlecht bewertet und verworfen. Die Betrachtung der Längen der EST-Sequenzen kann die Einschätzung der Qualität erleichtern. Kurze Sequenzen (weniger als 400 Bp) sind zwar von guter Qualität, decken aber mit großer Wahrscheinlichkeit nur die Enden der exprimierten Genbereiche ab. Zu lange EST-Sequenzen (länger als 1000 Bp) sind wegen der technischen Grenzen oft von schlechter Qualität, vor allem an den Randbereichen. Von den 34916 EST-Sequenzen sind nur 4544 kürzer als 700 Bp und 8782 kürzer als 800 bp. 1785 der Sequenzen sind länger als 1000 Bp und 562 sind länger als 1100 Bp. Die ESTs haben im Durchschnitt eine Länge von 837 Bp. Das ist ein guter Kompromiss zwischen Qualität und Länge.

Die Genmodelle für das Training von AUGUSTUS wurden aus diesen EST-Sequenzen und dem zweiten Assembly der Genomsequenz mit der PASA-Pipeline erstellt. Mit dem Transkript-Genom Alignmentprogramm BLAT, das wie auch das Transkript-Genom Alignmentprogramm SIM4 Teil der Pipeline ist, konnten 9235 der EST-Alignments den in Abschnitt 3.3.1 beschriebenen Validierungen des Programms Stand halten. 2081 der mit BLAT generierten Alignments, die verworfen wurden, entsprachen den geforderten

Kriterien nach Bearbeitung mit SIM4. Die 11316 EST-Alignments wurden in 3424 Cluster gruppiert aus denen von PASA 3691 Assemblies erstellt wurden. Ein Cluster enthält alle EST-Sequenzen, die einem Gen zugeordnet werden können und ein Assembly ist eine wahrscheinliche, eventuell unvollständige Sequenz eines Transkripts. In Abbildung 5.2 ist ein Beispiel für ein Cluster von EST-Alignments und den daraus resultierenden Assemblies dargestellt. Aus diesem Cluster wurden drei Assemblies erhalten. Die Abbildung 5.3 zeigt die drei Gruppen kompatibler Alignments dieses Clusters und die resultierenden Assemblies. Abbildung 5.1 zeigt die Häufigkeiten der ESTs für die einzelnen Assemblies. Wenige Assemblies wurden aus vielen ESTs konstruiert und viele der Assemblies bestehen nur aus wenigen ESTs. 1824 Assemblies wurden nur aus einer Sequenz erstellt und 944 aus zwei kompatiblen Sequenzen. 923 Assemblies wurden aus mehr als zwei kompatiblen EST-Sequenzen erstellt. Je mehr ESTs pro Assembly, umso größer ist die Wahrscheinlichkeit das das komplette Transkript rekonstruiert werden kann.

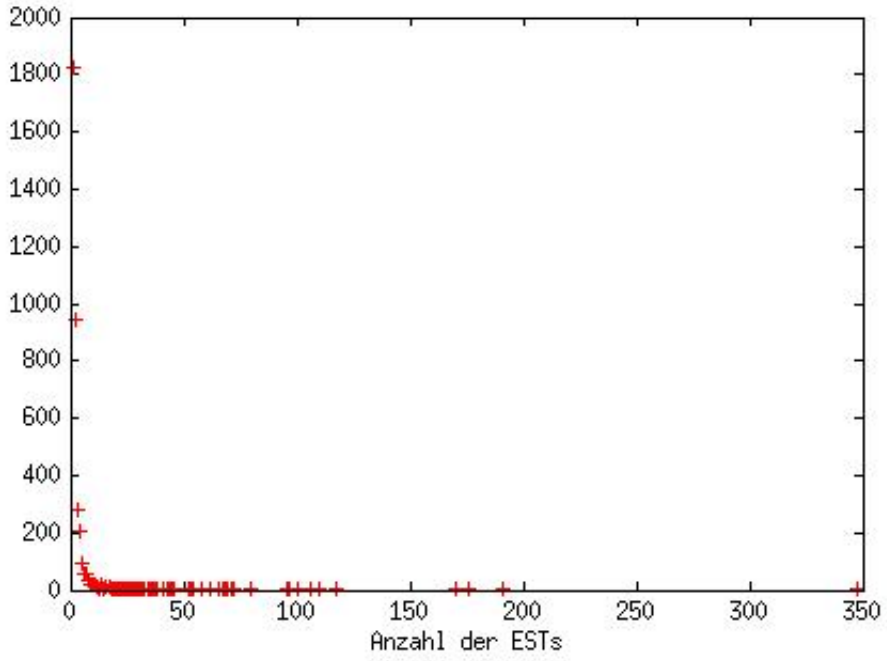

Abbildung 5.1: Die Abbildung zeigt die Häufigkeiten der EST für die, mit der PASA-Pipeline erstellten Assemblies.

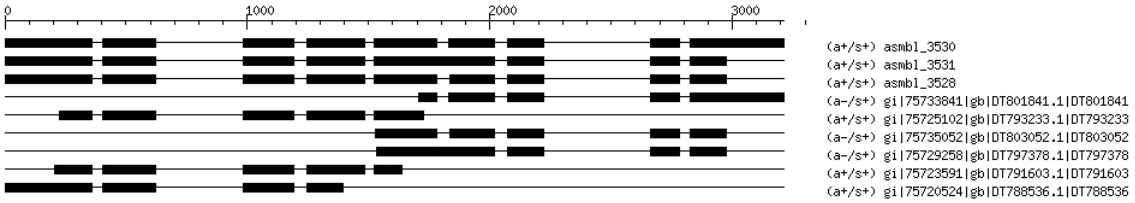

Abbildung 5.2: Dargestellt ist ein Cluster aus überlappenden Alignments und die von PASA daraus erstellten Assemblies.

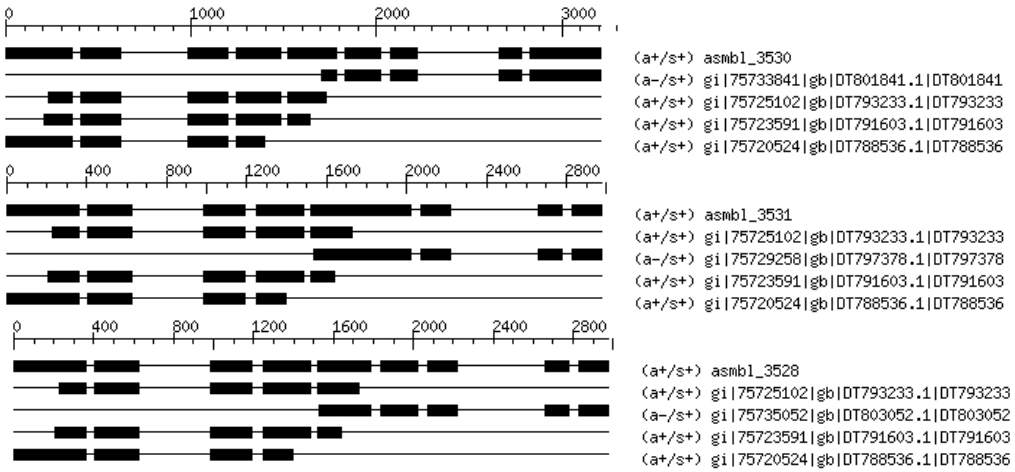

Abbildung 5.3: Die Abbildung zeigt die Gruppen konsistenter Alignments aus Abbildung 5.2 und die resultierenden Assemblies.

876 der Assemblies aus 601 Sequenzen sind Genmodelle, die bestimmte Kriterien erfüllen und für das Training von AUGUSTUS verwendet wurden. Es handelt sich um Sequenzen mit einem offenen Leserahmen, der eine Mindestlänge von 100 Aminosäuren hat. Vor dem Startkodon gibt es ein Stoppkodon im gleichen Leserahmen. Das den offenen Leserahmen begrenzende Stoppkodon wurde entweder in der Sequenz des Assemblies gefunden oder stromabwärts in der genomischen Sequenz. Abbildung 5.4 zeigt die Häufigkeitsverteilung der ESTs für diese Assemblies. 236 Assemblies entstanden aus nur einem EST und 169 durch Assemblierung von zwei kompatiblen EST-Sequenzen. 471 der Assemblies entstanden durch Assemblierung von mehr als zwei konsistenten EST-Sequenzen. Die meisten Genmodelle wurden aus mindestens zwei EST-Sequenzen erstellt. Die Redundanz der ESTs ist also

offenbar von Bedeutung für die Rekonstruktion der Genmodelle.

Für 570 Genmodelle wurde das Stoppkodon in der Sequenz des Assemblies gefunden. Sequenzen, die Gene enthalten, deren Stoppkodon nicht in der Sequenz des Assemblies gefunden wurde, können nicht zur Evaluierung im Laufe des Trainings und auch nicht für die Bewertung der Qualität der Vorhersagen nach Abschluss der Parameterbestimmung herangezogen werden. Das gefundene Stoppkodon kann durchaus in der Sequenz eines Introns liegen, wenn noch ein weiteres Exon folgt. Das nicht annotierte Exon kann von AUGUSTUS richtig erkannt werden, wird aber als falsch-positive Vorhersage gezählt und verfälscht im Prozess der Evaluierung der veränderten Parameter und bei der abschließenden Bewertung der erreichten Qualität die Spezifität der Vorhersage. Das könnte optimale Parameter benachteiligen.

In Abbildung 5.5 und 5.6 sind die Häufigkeiten der Längen der Aminosäuresequenzen dargestellt, die zum Training verwendet wurden. Modelle mit ausreichend langen Aminosäuresequenzen sind sehr wahrscheinlich komplette Gene. Die Modelle haben eine durchschnittliche Länge von rund 420 Aminosäuren. Wenige der Sequenzen sind sehr lang oder sehr kurz. Das entspricht den erwarteten Werten. Proteine aus *D. melanogaster* haben ebenfalls diese Durchschnittslänge [14] und Häufigkeitsverteilung. Um den Prozess des Trainings zu beschleunigen und um die Anzahl nicht annotierter Gene, die bei richtiger Vorhersage falsch-positive Ergebnisse ergeben, zu reduzieren, wurden die flankierenden Bereiche stromaufwärts (5'-Richtung) des ersten annotierten Gens und stromabwärts (3'-Richtung) des letzten annotierten Gens bis auf 3000 Nukleotide verworfen.

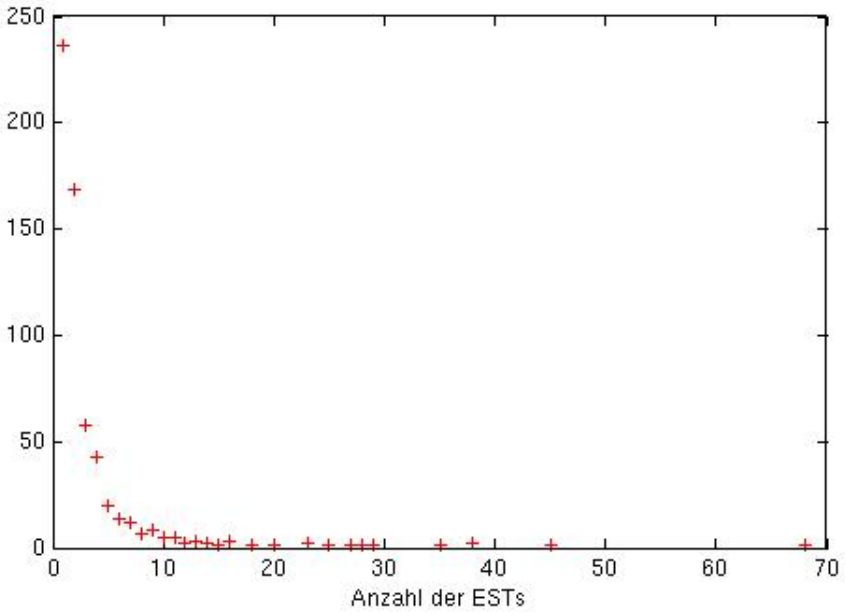

Abbildung 5.4: Dargestellt sind die Häufigkeiten der ESTs für die Assemblies, aus denen die Genmodelle für das Training von AUGUSTUS entstanden.

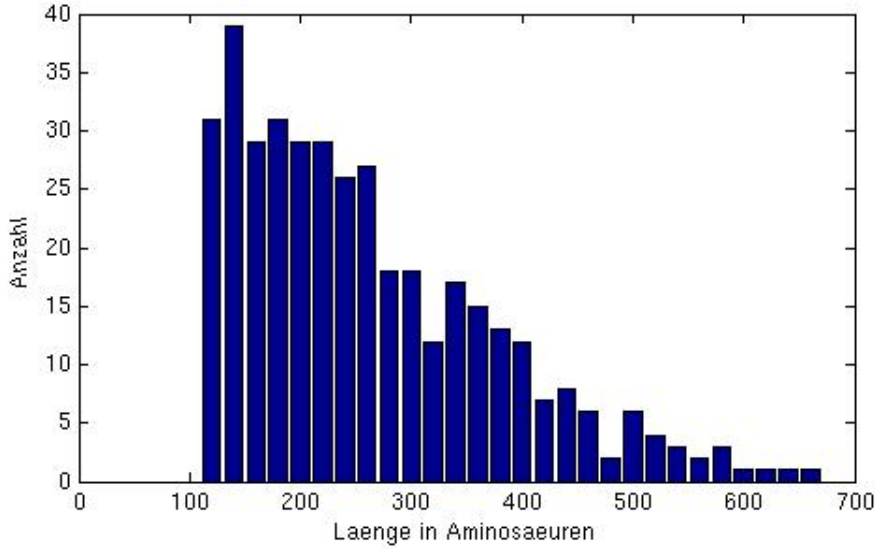

Abbildung 5.5: Die Abbildung zeigt die Häufigkeiten der Proteinlängen wahrscheinlich kompletter Genmodelle aus der Menge der zum Training von AUGUSTUS benutzten Modelle.

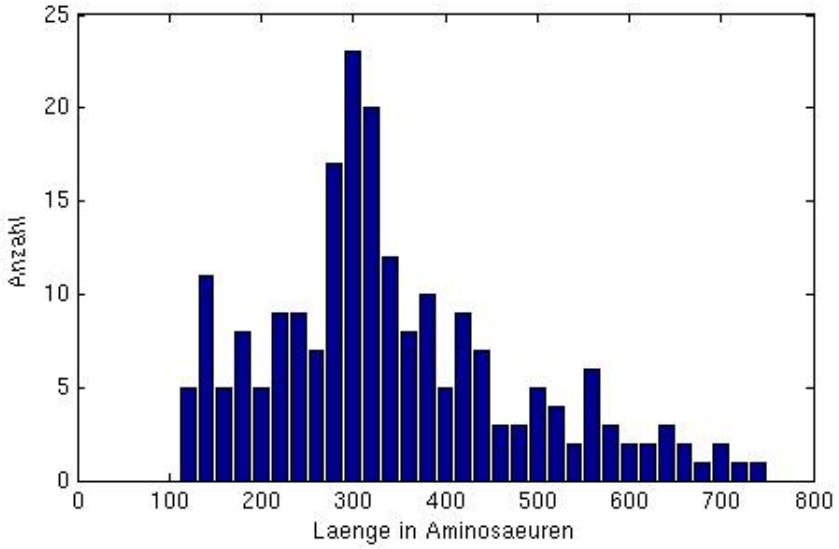

Abbildung 5.6: Die Abbildung zeigt die Häufigkeiten der Proteinlängen von Genmodellen, deren Stoppkodon nicht in der Sequenz des Assemblies gefunden wurde aus der Menge der zum Training von AUGUSTUS benutzten Modelle.

5.1.2 Ergebnisse des Trainings von AUGUSTUS

Für die Bewertung der Qualität einer Genvorhersage wird im Allgemeinen die *Sensitivität* und die *Spezifität* betrachtet. AUGUSTUS sagt in einer Nukleotidsequenz für ein Gen die Positionen der kodierenden Exons vorher. Mit Vorhersage ist im folgenden die Vorhersage der Positionen dieser Exons eines Gens gemeint.

AUGUSTUS ermittelt bei Übergabe von Sequenzen im GenBank-Format eine Statistik über Sensitivität und Spezifität der Vorhersage. TP (true positives) bezeichne die Anzahl der korrekten Vorhersagen und FN (false negatives) die Anzahl der nicht vorhergesagten Annotationen. Die Sensitivität ist folgendermaßen definiert:

$$\text{Sensitivität} := \frac{TP}{TP + FN}$$

Die Sensitivität ist die Prozentangabe der richtigen Vorhersagen aus der Menge aller Annotationen. Die Summe aus TP und FN ist gleich der Anzahl der Annotationen. FP (false positives) bezeichne die Anzahl falscher Vorhersagen. Die Spezifität ist wie folgt definiert:

$$\text{Spezifität} := \frac{TP}{TP + FP}$$

Die Spezifität ist der Prozentsatz der richtigen Vorhersagen aus der Menge aller Vorhersagen.

Die Menge an TP und FP und daraus Sensitivität und Spezifität werden drei Mal errechnet. Erstens für den *Genlevel*. Dabei wird ein vorhergesagtes Gen als korrekt betrachtet (TP), wenn die Vorhersage mit der Annotation vollständig übereinstimmt. Zweitens für den *Exonlevel*. Hier wird jedes vorhergesagte Exon einzeln mit der Annotation verglichen. Und drittens für den Nukleotidlevel. Dafür wird jedes einzelne, als kodierend vorhergesagte Nukleotid mit der Annotation verglichen.

Die von AUGUSTUS erstellte Statistik wird zur Einschätzung der Qualität der Vorhersagen verwendet. Die Sensitivität auf dem Genlevel ist dabei von besonderer Bedeutung. Sie zeigt, wie viele der annotierten Gene richtig erkannt wurden. Neben ihr ist aber auch die Spezifität dieses Levels wichtig. Sie gibt Aufschluss darüber, wie viele der vorhergesagten Gene korrekt sind. Eine hohe Sensitivität, bei gleichzeitig geringer Spezifität bedeutet, dass viele Gene richtig vorhergesagt wurden, aber auch das viele der vorhergesagten Gene falsch sind.

409 der 601 im GenBank-Format formulierten Sequenzen enthalten nur Genmodelle mit wahrscheinlich korrekt annotiertem Stoppkodon. Aus diesen Sequenzen wurden zufällig 100 erwählt und zur Bewertung der durch das Training erreichten Qualität aufbewahrt und nicht in diesen Prozess mit einbezogen. Bei 85 dieser Sequenzen ist nur ein Gen annotiert. Zwei Gene sind

bei 13 Sequenzen annotiert und zwei Sequenzen enthalten drei annotierte Gene. 501 Sequenzen mit insgesamt 759 Genen konnten für die Parameteroptimierung genutzt werden. In Tabelle 5.1 ist die, aus der Vorhersage auf diesen Sequenzen ermittelte Statistik zusammengefasst. Für den Genlevel konnte eine Sensitivität der Vorhersage von 48,7% und eine Spezifität von 19,2% erreicht werden. Im Vergleich mit einer Vorhersage auf annotierten Sequenzen von *Drosophila* (fly100, annotierte Sequenzen aus der Datenbank FlyBase [29]) unter Verwendung der Parameter für *D. melanogaster* (Statistik aus [1]) ist die für *Tribolium* erreichte Gensensitivität nur rund 3% schlechter (siehe Tabelle 5.1). Die Genspezifität der Vorhersage ist um rund 8% schlechter.

Die geringe Spezifität resultiert aus einer großen Zahl falsch-positiver Vorhersagen. In der Testmenge sind 117 Gene annotiert und 297 wurden von AUGUSTUS vorhergesagt. Wie viele der falsch-positiven Vorhersagen in Wirklichkeit nicht annotierte Gene sind kann nicht gesagt werden. Es ist aber wahrscheinlich, dass die für *Drosophila* verwendeten Testsequenzen eine qualitativ bessere und vollständigere Annotation enthalten, da die Sequenzdaten für diese Spezies schon seit längerem zur Verfügung stehen und *Drosophila* seit Jahrzehnten als Modellorganismus in der molekularbiologischen Forschung angesehen ist. Somit ist zu erwarten, dass sehr viele Gene identifiziert, annotiert und in vielen Fällen auch experimentell verifiziert sein müssten.

Die für *Tribolium* erreichte Exonsensitivität von 75,8 % ist nur um 4% schlechter als das für *Drosophila* erreichte Ergebnis. Vorteilhaft war sehr wahrscheinlich die oft große Zahl an Exons pro Gen. Tabelle 5.1 zeigt die ermittelten Sensitivitäten und Spezifitäten der Vorhersage von AUGUSTUS auf den Testdaten der zweiten Trainingsmenge mit den im ersten Trainingslauf ermittelten Parametern. Bei der Betrachtung der Werte muss berücksichtigt werden, dass die Testsequenzen Gene enthalten könnten, die auch in der Trainingsmenge des ersten Trainings enthalten waren. Diese Genmodelle sind »trainiert« und müssten demnach auch besser erkannt werden als andere.

Für das erste Training wurden 405 Genmodelle aus 314 Sequenzen verwendet. Die Testmenge für die Bewertung der erreichten Qualität der Vorhersage mit bestimmten Werten der Metaparameter bestand aus 189 Genen aus 149 Sequenzen. Wie erwartet ist die Qualität der Vorhersage mit den Parametereinstellungen, die im zweiten Training ermittelt wurden, besser. Vor allem macht dies eine Verbesserung der Exonsensitivität um rund 6% deutlich. Allerdings ist ein genauer Vergleich der Werte wegen der genannten Gründe schwer. Tatsache ist aber eine Verbesserung der Qualität, das Ausmaß der Qualitätssteigerung ist nur schwer einschätzbar. Die Ergebnisse zeigen ein erfolgreiches Training von AUGUSTUS auf die Genomsequenz von *Tribolium castaneum*.

In Tabelle 5.1 sind die Statistiken der Vorhersagen auf den Testsequenzen von *Tribolium castaneum* mit den Parametern für *Drosophila melanogaster*, *Aedes aegypti*, *Homo sapiens*, *Arabidopsis thaliana*, *Brugia malayi* und *Coprinus cinereus* zum Vergleich aufgelistet. Das beste Ergebnis von 17,1% für die Gensensitivität konnte mit den Parametern für *Aedes aegypti* erreicht werden. Am schlechtesten war der Wert der Gensensitivität für *Drosophila melanogaster* mit nur 4,27%. Der beste Wert, der für die Genspezifität erreicht wurde, war 16,4% mit den Parametern für *Coprinus cinereus*. Phylogenetisch am engsten mit *Tribolium* verwandt sind *Drosophila melanogaster* und *Aedes aegypti*. Trotz dieser Tatsache sind die erreichten Ergebnisse mit den Parametern dieser Spezies sehr schlecht. Ein enger Verwandtschaftsgrad bedeutet also nicht unbedingt auch ein gutes Ergebnis bei Übernahme der Parameter.

Der Vergleich mit diesen Werten zeigt eine enorme Verbesserung der Qualität der Vorhersagen auf der Genomsequenz von *Tribolium castaneum* durch erfolgreiches Training der Parameter der Modelle von AUGUSTUS. Eine Vorhersage auf der Genomsequenz von *Tribolium castaneum* mit den ermittelten Parametern sollte ein gutes Ergebnis haben.

Fazit

Der Erfolg des Trainings von AUGUSTUS für das Genom einer Spezies ist sehr von der Qualität der, für den Prozess verwendeten Genmodelle abhängig. Die automatische Annotation einer Genomsequenz durch ab initio Genvorhersage ist einer der ersten Schritte zur Analyse neu sequenzierter Genome, für die oft nur wenige Gene bekannt und annotiert sind, wie es auch für die zu analysierende Genomsequenz von *Tribolium castaneum* der Fall ist. Die beste Lösung für dieses Dilemma bieten die, sehr oft im Zuge eines Sequenzierungsprojekts erstellten cDNA-Bibliotheken. Ein Screen dieser Bibliotheken erzeugt eine Menge an Sequenzen, die Indizien für exprimierte Gene sind und einen Teil oder sogar die komplette Sequenz der Transkripte der Gene darstellen. Durch Alignment mit der genomischen Sequenz und geeignete Validierungskriterien können aus diesen Transkriptsequenzen Trainingsmodelle erhalten werden. Dabei ist die Qualität der verfügbaren Sequenzdaten von entscheidender Bedeutung. Für *Tribolium* ist die Datenlage gut. Das zweite Assembly der Genomsequenz ist öffentlich und konnte verwendet werden. Die Anzahl der EST-Sequenzen und ihre Qualität war ausreichend um genügend Genmodelle auch nach strengeren Validierungskriterien zu erhalten. Bei viel schlechterer Datenlage wäre die Erstellung von Genmodellen aus Sequenzvergleichen annotierter Gene verwandter Arten mit guter experimenteller Datenlage, wie z.B. Vergleiche mit *Drosophila melanogaster* erfolgversprechender. So ist aber der Sequenzvergleich Transkript-Genom viel genauer und spezifischer, denn die Ähnlichkeit der Gene von *Tribolium* zu verwandten Genen anderer Genome wird geringer eingeschätzt als die Ähnlichkeit zu den gegebenen Transkriptsequenzen, die ihrer durchschnittlichen Länge nach zu urteilen auch oft die komplette Transkriptsequenz repräsentieren müssten.

Die PASA-Pipeline kombiniert alle wichtigen Validierungsschritte um aus den EST-Daten eindeutige und möglichst korrekte Transkriptsequenzen zu erhalten. Die Kriterien dieser Schritte sind ausreichend streng. PASA bietet einfache Möglichkeiten mit Perl-Skripten die relevanten Sequenzen zu selektieren und auch die vorhergesagten Gene mit den assemblierten EST-Sequenzen zu validieren. Dies kann von Nutzen sein, wenn viele EST- und vor allem

cDNA-Sequenzen für die Spezies existieren. Für die Verbesserung der Annotation der Genomsequenz von *Arabidopsis thaliana* standen 182540 validierte Transkript-Genom Alignments zur Verfügung. Die Transkripte wurden mit PASA zu 25165 Sequenzen assembliert [31]. 16542 davon zeigten mögliche Verbesserungen für 14247 annotierten Genstrukturen auf. Die für *Tribolium castaneum* zur Verfügung stehende Datenmenge von nur rund 3691 EST-Assemblies ist im Vergleich zu den für dieses Projekt verfügbaren Daten sehr klein und der zu erwartende Nutzen wird demnach als gering eingeschätzt.

AUGUSTUS hat für die Genomsequenz von *Tribolium castaneum* 15309 Genstrukturen vorhergesagt. Für die Genomsequenz von *Drosophila melanogaster* wurden mit den Parametern dieser Spezies 12357 Gene vorhergesagt. Jüngste Angaben über die Anzahl der Gene im Genom von *Drosophila melanogaster* belaufen sich auf rund 14000 Gene (Prof. Dr. Heinz Sass, Pressemitteilung der Universität zu Leipzig, 28.11.2005). Die Zahl der Gene für *Tribolium* könnte sich wegen der näheren Verwandtschaft zu *Drosophila* auch in dieser Größenordnung bewegen. Die mit AUGUSTUS automatisch erstellte Annotation der Genomsequenz von *Tribolium castaneum* bietet auf Grund der erreichten Qualität der Vorhersage eine fundierte Grundlage für weitere Analysen. Trotz der geringen Spezifität der Vorhersage ist es wahrscheinlich, das in der Menge der vorhergesagten Gene viele relevante und gesuchte Sequenzen oder Teilsequenzen enthalten sind, denn es wurde eine hohe Sensitivität erreicht. Die Zahl der vorhergesagten Gene entspricht ungefähr der vermuteten Anzahl bzw. ist gering größer. Die zu analysierende Datenmenge bei der Suche nach speziellen kodierten Funktionen hat sich erheblich verringert – von der gesamten Genomsequenz auf die erhaltene Annotation. Von besonderem Interesse ist die erstellte Annotation für die *Tribolium Gemeinde*. Das sind viele der Forscher aus aller Welt, die mit *Tribolium castaneum* als Modellorganismus arbeiten und in Kontakt zueinander stehen. Die mit dem ab initio Genvorhersageprogramm AUGUSTUS erstellte Annotation soll diese Wissenschaftler in ihrer Arbeit unterstützen.

Tabelle 5.1: In der Tabelle sind die von AUGUSTUS ermittelten Werte für die Sensitivität und die Spezifität der Genvorhersage auf den entsprechenden Testdaten unter Verwendung der Parameter der genannten Spezies aufgelistet. Die Testdaten mit der Bezeichnung 2 sind die Genmodellen, die aus dem zweiten Assembly der Genomsequenz von *Tribolium* und den ESTs des HGSC erstellt wurden. Die Testdaten mit der Bezeichnung 1 sind die Genmodelle der ersten Trainingsmenge, generiert aus dem ersten Assembly der Genomsequenz von *Tribolium* und den ESTs der Universität zu Köln. Die Testdaten mit der Bezeichnung fly100 sind annotierte Sequenzen aus der Datenbank FlyBase.

Spezies	**Testdaten**	**Nukleotidlevel**		**Exonlevel**		**Genlevel**	
		Sensitivität	**Spezitivität**	**Sensitivität**	**Spezitivität**	**Sensitivität**	**Spezitivität**
Tribolium castaneum	2	0.881	0.249	0.758	0.26	0.487	0.192
Drosophila melanogaster	fly100	0.97	0.59	0.80	0.49	0.52	0.27
Tribolium castaneum	1	0.827	0.243	0.695	0.255	0.462	0.194
Drosophila melanogaster	2	0.222	0.285	0.0938	0.185	0.0427	0.0538
Aedes aegypti	2	0.508	0.266	0.26	0.195	0.171	0.113
Homo sapiens	2	0.493	0.235	0.185	0.147	0.094	0.0553
Arabidopsis thaliana	2	0.522	0.254	0.201	0.14	0.111	0.0594
Brugia malayi	2	0.729	0.22	0.331	0.117	0.111	0.0408
Coprinus cinereus	2	0.319	0.308	0.195	0.242	0.154	0.164

Ausblick

Für die Vorhersage mit AUGUSTUS kann die Option genutzt werden, extrinsische Informationen einzubeziehen. Diese können mit dem Programm AGRIPPA [24] aus EST- und Proteinsequenzen erstellt werden. Die Zuverlässigkeit der extrinsischen Hinweise wird in Parametern bewertet, deren Werte trainiert werden müssen. Die Qualität der Vorhersage wird durch diese Option sehr wahrscheinlich positiv beeinflusst [20], [24].

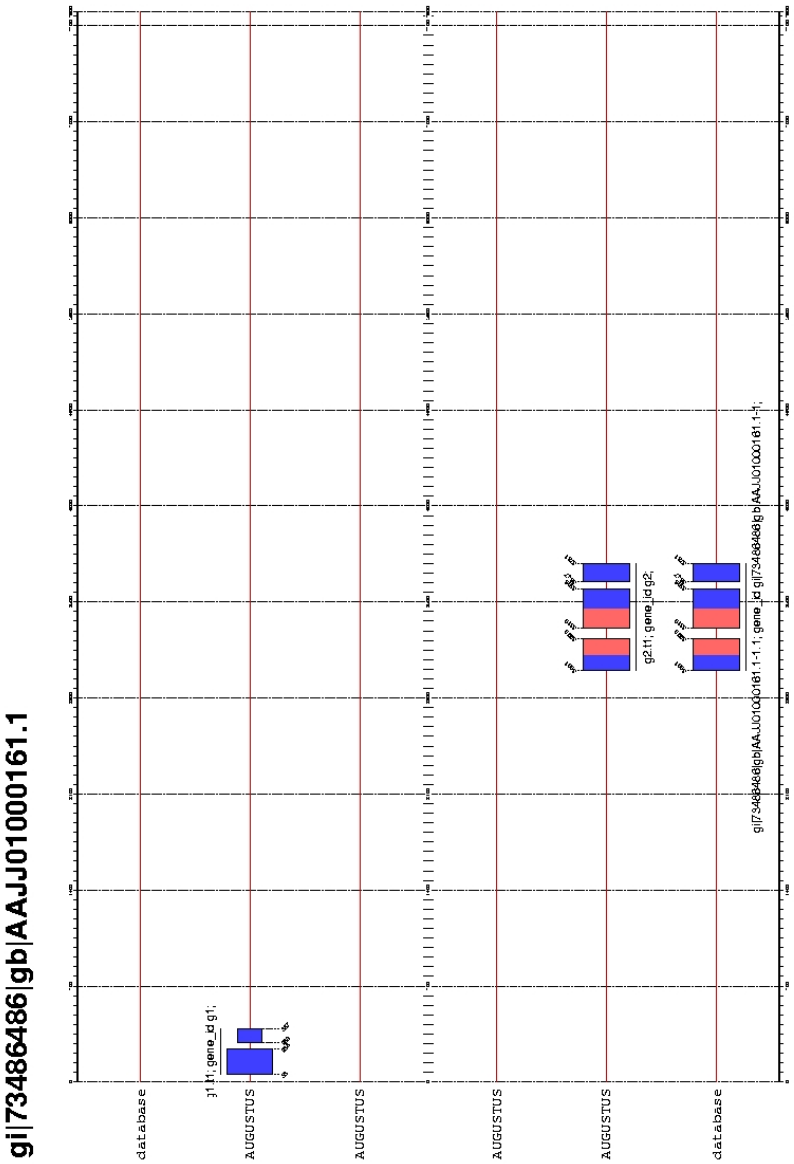

Abbildung 5.7: Die Abbildung zeigt einen Ausschnitt aus einer mit dem Programm GFF2PS erstellten Visualisierung einer mit AUGUSTUS vorhergesagten Genstruktur von *Tribolium*. Die Exons sind als dicke Balken dargestellt. Die verschiedenen Farben zeigen den jeweiligen Leserahmen an. Die Zeile »database« beizeichnet, die in diesem Sequenzabschnitt annotierten Gene. Die Annotation wurde aus den EST- und Genomdaten von *Tribolium* erstellt. Die Abbildung zeigt ein typisches Beispiel für die Vorhersage. AUGUSTUS hat sehr viele Gene richtig vorhergesagt. Unter den zusätzlich von AUGUSTUS vorhergesagten Genen, die nicht annotiert sind, sind wahrscheinlich viele falsch-negative Annotationen.

5.2 Vergleichende Genomanalyse

Die vergleichenden Genomanalysen zwischen *Tribolium castaneum*, *Drosophila melanogaster* und *Homo sapiens* beziehen sich auf die Proteine. Die Bedeutung der Genomsequenz von *Tribolium castaneum* für die Identifizierung von homologen Proteinen in *Homo sapiens* und *Drosophila melanogaster* ist dabei interessant und wird im Folgenden diskutiert. Die durchgeführten Analysen haben weiterhin das Ziel, menschliche Proteinen zu finden, die keine identifizierbaren Homologen in *Drosophila* haben, aber in *Tribolium*. Homologe Proteine können anhand der Ähnlichkeit ihrer Aminosäuresequenzen identifiziert werden. Diese Ähnlichkeit kann global oder lokal sein oder nur einzelne Aminosäurereste betreffen, die zur Ausübung der Funktion des Proteins notwendig und deshalb konserviert sind. Ein Vergleich der kodierenden Nukleotidsequenzen ist weniger eindeutig, da eine Aminosäure meist durch verschiedene Kodons repräsentiert werden kann.

Für die automatische Suche nach homologen Proteinen in ganzen Genomen eignet sich das Programm BLASTP gut. Das Programm ist in Abschnitt 4.2 beschrieben. Der E-Wert wird herangezogen, um die Treffer der BLAST-Suchen zu klassifizieren. Der von *Blast* berechnete E-Wert ist eine Schätzung der erwarteten Anzahl der Treffer mit einem Score größer oder gleich dem Score des betrachteten Treffers, wenn die Eingabesequenz und die Datenbank zufällig erzeugt wurden. Signifikante Sequenzähnlichkeiten, wie sie bei vielen homologen Proteinen zu finden sind, ergeben im Allgemeinen einen sehr kleinen E-Wert (min. 10^{-10}). Entfernt verwandte Proteine mit geringer Sequenzähnlichkeit sind nur schwer von negativen Treffern zu unterscheiden. Oft können entfernt verwandte Proteine nur durch einzelne Betrachtung der Treffer und Kenntnis konservierter Aminosäurereste erkannt werden.

Die für *Drosophila* und Mensch bislang bekannten Proteine sind zahlreich und die, in der Datenbank Ensembl zusammengestellten Annotationen können als qualitativ gut angesehen werden. Für die Proteine von *Tribolium* muss bedacht werden, dass es sich um eine automatisch erstellte ab initio Annotation handelt. Diese ist, den Analysen (siehe Abschnitt 5.1.2) und Re-

ferenzen von AUGUSTUS nach, von guter Qualität, muss aber trotzdem mit Vorsicht betrachtet werden. Gene, die benachbart sind, könnten beispielsweise als nur ein Gen vorhergesagt wurden sein. Bei einer BLAST-Suche ergibt dann das resultierende Protein wahrscheinlich mehrere Treffer in der Datenbank. Multiple Treffer für viele Proteine sind auch aus anderen Gründen zu erwarten. Innerhalb eines Organismus gibt es Familien homologer Gene, die auch oft als Gencluster organisiert sind, wie z.B. die *HOX*-Gene des Menschen [12]. Diese sind vermutlich durch Genduplikation entstanden. Wegen der vielen möglichen Trefferkombinationen homologer Proteinfamilien aus den Organismen wird bei der Ermittlung der Trefferzahl für die durchgeführten BLAST-Suchen nur der beste Treffer gezählt.

Die Ergebnisse der durchgeführten Analysen sind im Folgenden beschrieben. Tabelle 5.2 zeigt die Häufigkeiten negativer Treffer der BLAST-Suche mit menschlichen Proteinen als Query in der Sammlung der Proteine von *Drosophila*. Ein negativer Treffer ist ein menschliches Protein, dessen bester Treffer bei der BLAST-Suche (der Treffer mit dem kleinsten E-Wert) den im Tabellenkopf angegebenen Wert `e` nicht unterschreitet. Ein E-Wert von 10^{-3} ist allgemein ein Zeichen geringer Sequenzähnlichkeit. Für 9240 menschliche Sequenzen gibt es keine Treffer mit einem E-Wert, der kleiner als 10^{-3} ist. Für diese menschlichen Proteine können also homologe Proteine in *Drosophila* nur schwer oder sogar nicht identifiziert werden.

Die Genomsequenz von *Tribolium* bietet für diese menschlichen Proteine eine neue Chance. Tabelle 5.3 zeigt die Häufigkeiten der positiven Treffer einer BLAST-Suche mit Proteinen von *Tribolium* als Query in der Sammlung der Proteine des Menschen. Die Sammlung der menschlichen Proteine beschränkt sich dabei auf solche Sequenzen, die bei der BLAST-Suche in den Proteinsequenzen von *Drosophila* nur Treffer hervorgebracht haben, deren E-Werte die, im Tabellenkopf aufgeführten Werte e_n, nicht unterschreiten. Als positive Treffer werden Proteine aus *Tribolium* gezählt, die mindestens einen Treffer hervorgebracht haben, mit einem E-Wert, der maximal den, in der ersten Spalte definierten Wert e_p hat. Ein E-Wert von 10^{-9} ist im Allgemeinen ein Zeichen für signifikante Sequenzähnlichkeit. Für 231 Proteine

aus *Tribolium* wurde für mindestens ein menschliches Protein (von den genannten 9240 menschlichen Proteinen) ein E-Wert von max. 10^{-9} erhalten. Die Bezeichner dieser 231 Proteine aus *Tribolium* sowie alle Bezeichner der menschlichen Proteine, für die ein E-Wert von maximal 10^{-9} erhalten wurde, sind in der Textdatei `Tribolium_Human_pos_Treffer.txt` zusammengestellt, die der beiliegenden CD entnommen werden kann. Diese 231 Proteine aus *Tribolium* sind wahrscheinlich bedeutsame Kandidaten für die Identifikation unbekannter Homologien in Mensch und *Drosophila* oder bieten neue Möglichkeiten, wenn eine Homologie zu *Drosophila* auch durch Vergleich mit *Tribolium* nicht erkannt werden kann.

Die Ergebnisse der BLAST-Suche mit Proteinen von *Tribolium* als Query in der Sammlung der Proteine von *Drosophila* sind in den Tabellen 5.4 und 5.5 zusammengefasst. In den Zellen der Tabelle 5.4 ist die Anzahl der Proteine von *Tribolium* aufgelistet, die bei der BLAST-Suche als Query mindestens einen Treffer mit einem E-Wert von maximal dem in der ersten Spalte genannten Wert (e_p) hervorgebracht haben. Der Tabellenkopf definiert die Grenzen für negative Treffer. Gemeint sind die menschlichen Proteine, die bei der BLAST-Suche in der Sammlung der Proteine von *Drosophila* nur negative Treffer, deren kleinster E-Wert den jeweiligen Wert e_n nicht unterschreitet, hervorgebracht haben.

In der Tabelle 5.5 sind die Häufigkeiten der negativen Treffer dieser BLAST-Suche zusammengefasst. Die Grenzen, die einen negativen Treffer definieren, sind in der ersten Zeile genannt. In den Zellen ist die Anzahl der Proteine von *Tribolium* aufgeführt, die nur Treffer hervorgebracht haben, die den jeweiligen Wert e_n nicht unterschreiten. Die Werte e_n definieren ebenfalls wieder die Grenzen für die Selektion der negativen Treffer aus der BLAST-Suche zwischen Mensch und *Drosophila*. Die erste Spalte definiert die Grenzen für positive Treffer aus der BLAST-Suche zwischen Mensch und *Tribolium*. Die *Blast*-Suche zwischen *Tribolium* und *Drosophila* ergab für einen positiven Grenzwert von 10^{-9} und einen negativen Grenzwert von 10^{-3} 80 positive und 121 negative Treffer. Die Bezeichner dieser Proteine von *Tribolium* sowie die Bezeichner aller positiven oder negativen Treffer können der beiliegenden

CD entnommen werden. Die Dateien haben die Namen:
`Tribolium_Drosophila_pos_Treffer.txt` und
`Tribolium_Drosophila_neg_Treffer.txt`.

Fazit

Die durchgeführten Analysen haben Argumente geliefert, um die Bedeutung der Genomsequenz von *Tribolium castaneum* zu »betonen«. Proteine von *Tribolium* mit signifikanter Sequenzähnlichkeit zu »interessanten« menschlichen Proteinen wurden identifiziert. Interessant ist, dass für diese menschlichen Proteine keine signifikanten Ähnlichkeiten zu Proteinen aus *Drosophila melanogaster* gefunden wurden. Für einige dieser Proteine aus *Tribolium* wurden in *Drosophila* mögliche Homologe identifiziert. Für einen größeren Teil konnten aber keine möglichen homologen Proteine mit signifikanter Sequenzähnlichkeit in *Drosophila* identifiziert werden. Diese Proteine von *Tribolium* zeigen die Bedeutung der Genomsequenz in besonderem Maße. Die Anzahl der gefundenen Proteine erscheint im Vergleich zu der Gesamtmenge aller Proteine gering. Doch können Proteine enthalten sein, die entscheidenden Einfluss auf wichtige Erkenntnisse haben. Die ermittelten Ergebnisse bestärken die Aussichten auf Erfolg einer tiefergehendenden vergleichenden Genonomanalyse.

Ausblick

Die identifizierten Sequenzähnlichkeiten könnten durch multiple Alignments der Proteinsequenzen genauer betrachtet und besser bewertet wertet werden. Geeignet für diesen Ansatz sind z.B. die Alignment-Programme DIALIGN [40] und CLUSTALW [41]. Für die »interessanten« menschlichen Proteine könnte nach möglichen, bereits bekannten Funktionen gesucht werden. Die Ergebnisse der durchgeführten BLAST-Suchen könnten durch wiederholte Analyse, wobei Query und Datenbank zu vertauschen sind, genauer analysiert und bestätigt werden.

Tabelle 5.2: Häufigkeiten menschlichen Proteine, die in einer BLAST-Suche in der Sammlung der Proteine von *Drosophila* nur Treffer hervorgebracht haben, deren E-Werte den jeweiligen Wert `e` nicht unterschreiten.

e	**1**	**0,1**	**0,01**	**0,001**	**0,0001**
	4188	6878	8328	9240	9968

Tabelle 5.3: Häufigkeiten der Proteine aus *Tribolium*, die in einer BLAST-Suche in der Sammlung der Proteine des Menschen mindestens einen Treffer hervorgebracht haben, dessen E-Wert maximal so groß ist, wie der jeweilige Wert für e_p. Die Sammlung der menschlichen Proteine beschränkt sich auf solche, die in einer BLAST-Suche in der Sammlung der Proteine von *Drosophila* nur Treffer hervorgebracht haben, deren E-Werte den jeweiligen Wert für e_n nicht unterschreiten.

e_p	e_n				
	1	**0,1**	**0,01**	**0,001**	**0,0001**
10^{-5}	118	188	270	410	498
10^{-6}	104	160	228	336	410
10^{-7}	93	143	202	285	348
10^{-8}	84	129	179	254	310
10^{-9}	79	122	165	231	277
10^{-10}	72	110	147	210	249
10^{-20}	47	68	81	105	125
10^{-30}	35	48	55	71	83
10^{-40}	28	39	44	57	64
10^{-50}	21	29	34	45	49
10^{-60}	19	27	32	40	44
10^{-70}	13	20	23	31	34

Tabelle 5.4: Häufigkeiten der Proteine aus *Tribolium*, die in einer BLAST-Suche in der Sammlung der Proteine aus *Drosophila* mindestens einen Treffer hervorgebracht haben, dessen E-Wert maximal so groß ist, wie der jeweilige Wert für e_p. Die Proteine von *Tribolium* beschränken sich auf die der jeweiligen Zelle aus Tabelle 5.3

e_p	e_n				
	1	**0,1**	**0,01**	**0,001**	**0,0001**
10^{-5}	46	79	126	235	305
10^{-6}	33	55	92	166	220
10^{-7}	25	43	73	122	166
10^{-8}	19	34	58	100	138
10^{-9}	14	27	45	80	107
10^{-10}	9	20	31	64	83
10^{-20}	2	3	10	12	20
10^{-30}	1	1	5	5	7
10^{-40}	1	1	4	4	5
10^{-50}	1	1	3	4	5
10^{-60}	1	1	3	3	4
10^{-70}	0	0	1	2	2

Tabelle 5.5: Häufigkeiten der Proteine aus *Tribolium*, die in einer BLAST-Suche in der Sammlung der Proteine von *Drosophila* nur Treffer hervorgebracht haben, deren E-Werte den jeweiligen Wert e_n nicht unterschreiten. Die Proteine von *Tribolium* beschränken sich auf die der jeweiligen Zelle aus Tabelle 5.3

e_p	e_n				
	1	**0,1**	**0,01**	**0,001**	**0,0001**
10^{-5}	32	84	122	155	184
10^{-6}	32	80	113	146	174
10^{-7}	31	77	108	135	161
10^{-8}	30	73	101	125	150
10^{-9}	30	72	99	121	145
10^{-10}	29	68	94	115	139
10^{-20}	19	51	61	76	90
10^{-30}	15	39	42	52	66
10^{-40}	11	31	33	40	51
10^{-50}	7	23	26	33	40
10^{-60}	6	21	24	29	36
10^{-70}	4	15	18	23	28

Literaturverzeichnis

[1] M. Stanke and S. Waack. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics*, 19(Suppl. 2):ii215–ii255, 2003.

[2] S. Brown, D.E. Denell, and R. Beeman. Beetling around the genome. *Genet. Res., Camb.*, 82:155–161, 2003.

[3] Baylor College of Medicine *Tribolium castaneum* Genome Project (TGP). http://www.hgsc.bcm.tmc.edu/projects/tribolium.

[4] *Tribolium castaneum* FTP-Server (Universität Köln). ftp://ftp.uni-koeln.de/institute/genetik/tribolium/.

[5] A. Kaestner and H.H. Dathe. *Lehrbuch der speziellen Zoologie, Bd.1/5: Wirbellose Tiere, 1. Auflage*. Spektrum Akademischer Verlag, 2002.

[6] M.S. Blum. *Chemical Defenses of Arthropodes*. New York: Academic Press, 1981.

[7] R.W. Howard, R.A. Jurenka, and G.J. Blomquist. Prostaglandin synthetase inhibitors in the defensive secretion of the red flower beetle *Tribolium castaneum* (herbst) (coleoptera:tenebrionidae). *Insect Biochemistry*, 16:757–760, 1986.

[8] F. Falciani, B. Hausdorf, R. Schröder, M. Akam, D. Tautz, R. Denell, and S. Brown. Class 3 hox genes in insects and the origin of zen. In *Proceedings of the National Academy Sciences of the USA*, volume 8479-8484, page 93, 1996.

[9] D. Andreev, T. Rocheleau, T.W. Phillips, R.W. Beeman, and R. H. R.H. FFrench-Constant. A PCR diagnostic for cyclodiene insecticide resistance in the red flower beetle , *Tribolium castaneum*. *Pesticide Science*, 41:345–349, 1994.

[10] R.W. Beeman and S.M. Nanis. Malathion resistance alleles and their fitness in the red flour beetle (coleoptera: Tenebrionidae). *J. Econ. Entomol.*, 1986.

[11] R.W. Beeman and J.J. Stuart. A gene for lindene + cyclodiene resistance in the red flour beetle (coleoptera: Tenebrionidae). *J. Econ. Entomol.*, 83:1745–1751, 1990.

[12] F. Lottspeich and H. Zorbas. *Bioanalytik*. Spektrum Akademischer Verlag Heidelberg, Berlin, 1998.

[13] M. Stanke. Algorithmen der Bioinformatik II Teile Genvorhersage und Sequenzierung und Assemblierung. 2004.

[14] B. Lewin. *Molekularbiologie der Gene*. Spektrum Akademischer Verlag Heidelberg, Berlin.

[15] N.L. Craig, R.C. Craigie, M. Gellert, and A.M. Lambowitz. *Mobile DNA II*. ASM Press, Washington D.C., 2002.

[16] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, and W. Morris *et al.* nternational Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[17] J. Jurka, V.V. Kapitonov, P. Klonowski, J. Walichiewicz, and A.F. Smit. Identification of new medium reiteration frequency repeats in the genomes of primates, rodentia and lagomorpha. *Genetica*, 98:235–247, 1996.

[18] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 27(162):705–708, 1982.

[19] P.A. Kitts, T.L. Madden, H. Sicotte, and J.A. Ostell.

[20] M. Stanke, O. Schöffmann, B. Morgenstern, and S. Waack. AUGUSTUS+: Gene prediction in eukaryotes with a Generalized Hidden Markov Model using EST and protein sequence information.

[21] C.B. Burge. Identification of genes in human genomic DNA. 1997.

[22] G. Parra, B. Enrique, and R. Guigo. Geneid in drosophila. *Genome Research*, 10:511–515, 2000.

[23] I. Korf, P. Flicek, D. Duan, and M. R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 1(Suppl. 1):S1–S9, 2001.

[24] O. Schöffmann. *Gewinnung extrinsischer Informationen zur Genvorhersage und und Einbindung in ein Hidden Markov Model.* PhD thesis, 2003.

[25] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410, 1990.

[26] R. Merkl and S. Waack. *Bioinformatik Interaktiv.* Wiley-VCH, 1 edition, 2002.

[27] GenBank. ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt.

[28] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, W. Miller Z. Zhang, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[29] R.A. Drysdale, M.A. Crosby, and The FlyBase Consortium. FlyBase: genes and gene models. *Nucleic Acids Research*, 33:D390–D395, 2005.

[30] Ensembl. http://www.ensembl.org/.

[31] B.J. Haas, A.L. Delcher, S.M. Mount, J. R. Wortman, R.K. Smith Jr, L.I. Hannick, R. Maiti, C.M. Ronning, and D.B. Rusch. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31:5654–5666, 2003.

[32] W.J. Kent. BLAT-the BLAST-like alignment tool. *Genome Res.*, 12:656–664, 2002.

[33] L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, W., and Miller. A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome Res.*, 8:967–974, 1998.

[34] X. Huang, M.D. Adams, H. Zhou, and A.R. Kerlavage. A tool for analyzing and annotating genomic sequences. *Genomics*, 47:37–45, 1997.

[35] S.J. Wheelan, D.M. Church, and J.M. Ostell. Spidey: a tool for mrna-to-genomic alignments. *Genome Res.*, 11:1952–1957, 2001.

[36] J. Usuka, W. Zhu, V., and Brendel. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, 16:203–211, 2000.

[37] L. Hubert. Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, 69:698–704, 1974.

[38] R. Durbin and G. Haussler. General Feature Format (GFF) http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml.

[39] BioPerl. http://search.cpan.org/dist/bioperl/.

[40] A.R. Subramanian, J. Weyer-Menkhoff, M. Kaufmann, and B. Morgenstern. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *Bioinformatics*, 6:6:66, 2005.

[41] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–80, 1994.

Danksagung

Ich danke meinem Betreuer Dr. Mario Stanke besonders für seine Unterstützung, seine Ideen, mit denen er mir den Weg gezeigt hat und die Korrektur meiner Arbeit.

Prof. Burkhard Morgenstern danke ich für die Möglichkeit in der Abteilung für Bioinformatik meine Diplomarbeit zu schreiben, sowie Prof. Ernst Wimmer für seine freundliche Bereitschaft, die Zweitkorrektur meiner Arbeit zu übernehmen.

Ich danke allen Mitarbeiter der Abteilung für ihre Motivation und Freundlichkeit, ganz besonders Maike Tech für das Korrekturlesen und ihre Hilfe.

Weiterhin danke ich Brian Haas für seine Unterstützung bei der Arbeit mit dem von ihm entwickelten Programm Pasa.

Ich danke meinem Vater, Nadine und Meik und meinen Freunden Kerstin, Holger und Alex sehr für ihre Unterstützung und Motivation.