TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Abbildung von Transkript- und Proteinsequenzen auf genomische Referenzdaten

Brigitte Wägele

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. I. Antes

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.-W. Mewes
2. Univ.-Prof. Dr. J. Parsch (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 12.03.2009 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 7.10.2009 angenommen.

Danksagung

An erster Stelle möchte ich Prof. Dr. Hans-Werner Mewes dafür danken, dass er es mir ermöglicht hat am Institut für Bioinformatik und Systembiologie des Helmholtz Zentrum München zu promovieren. Danke für die zahlreichen wegweisenden Gespräche und Anregungen ebenso wie den gewährten wissenschaftlichen Gestaltungsfreiraum.

Vielen Dank auch an Prof. Dr. John Parsch und Prof. Dr. Iris Antes, die - als zweiter Gutachter und Vorsitzende -, beide ohne zu zögern eingewilligt haben, meine Prüfungskommission zu vervollständigen.

Besonderen Dank möchte ich meinem Betreuer Andreas aussprechen, der mir immer mit Rat und Tat zur Seite stand. Insbesondere auch den restlichen Gruppenmitgliedern – Barbara, Irmtraud, Gisela, Goar und Corinna –, die es mir ermöglicht haben Problemstellungen dieser Dissertation neben den rein technischen Aspekten von der biologischen Seite aus zu beleuchten und zu diskutieren. Vor allem möchte ich Corinna und Irmtraud für die Zusammenarbeit an CRONOS danken. Trotz des enormen Aufwandes in die Qualitätssicherung habt Ihr nie Eure gute Laune verloren.

Ebenso möchte ich Nicole von der medizinischen Pharmakologie des Zentrums für Wirkstoffforschung und des medizinischen Zentrums der Universität Leiden, für die erfolgreiche Zusammenarbeit danken.

Ein großes Dankeschön auch an Elisabeth und Gabi, die – obwohl selbst, oder gerade weil promovierend – sich immer für fachliche und motivierende Diskussionen zur Verfügung gestellt haben.

Vielen Dank auch an Jos und Giovani für die technische Unterstützung.

Herzlichen Dank auch meinen Freunden, insbesondere Jürgen, Floh und Isa, die während meiner Doktorandenzeit nie aufgegeben haben, mir ein gesundes Maß Privatleben – neben der Wissenschaft – zu verschaffen.

Vielen herzlichen Dank auch meiner Familie für Ihre Unterstützung und Geduld, die zu guter Letzt zu dieser Dissertation geführt haben.

Vielen Dank.

Inhaltsverzeichnis

1 Einleitung

Lange war der Fortschritt der Biowissenschaften durch die Tatsache limitiert, dass nur einzelne Moleküle isoliert und charakterisiert wurden. Damals konnte eine Promotion z.B. aus der Sequenzierung eines einzelnen Gens bestehen (Venter *et al.*, 2003). Das Wissen aus den einzelnen Forschungsbereichen wurde später in spezialisierten Datenbanken organisiert, z.B. Nukleinsäuresequenzen in Genbank (Abb. 1) oder EMBL, Proteinstrukturen in PDB und Proteinsequenzen in PIR und Swiss-Prot (The UniProt Consortium, 2008; Barker *et al.*, 2001; Kulikova *et al.*, 2007; Benson *et al.*, 2008; Henrick *et al.*, 2008).

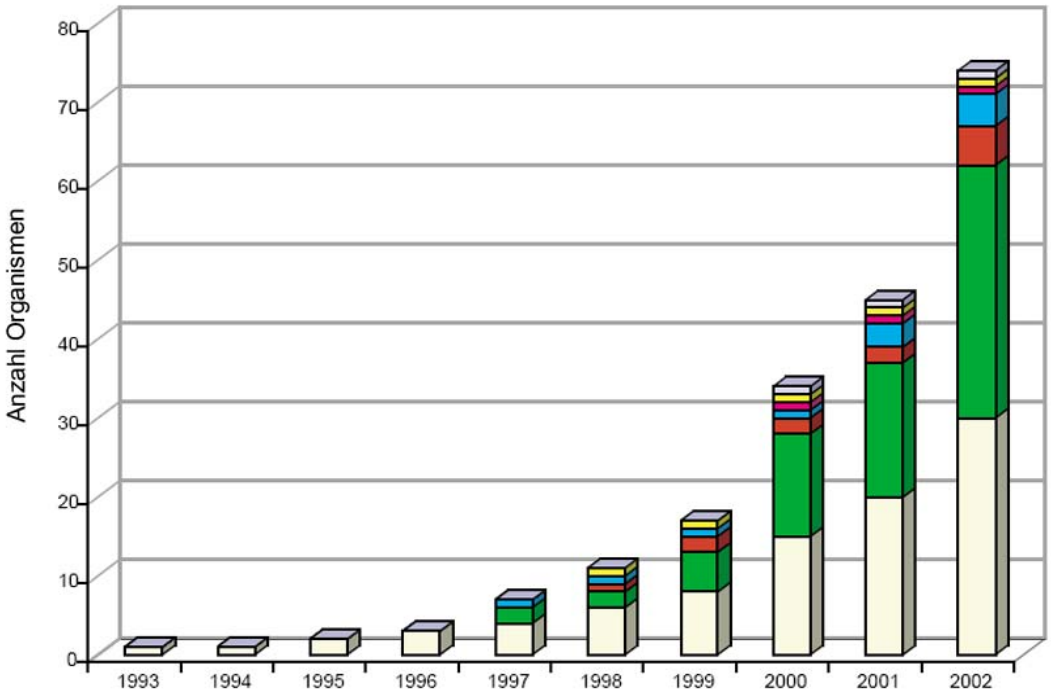


Abb. 1: Die Zunahme der EST-Sequenzierung 1992 bis 2003. Die Säulen geben die kumulative Anzahl derjenigen Organismen mit mehr als 10.000 ESTs in GenBank. Gezeigt werden die Kategorien der Tiere (weiß), Pflanzen (grün), Pilze (rot), Alveolaten (blau), Euglenozoa (pink), Mycetozoa (Schleimpilze) (gelb) und Rotalgen (grau). Verändert nach (Venter *et al.*, 2003).

Dies führte zu einem umfangreichen Faktenwissen über einzelne Gene oder Proteine, das Zusammenspiel der einzelnen Komponenten, wie es für den Ablauf komplexer zellulärer Prozesse in verschiedenen Geweben oder in komplexen Organismen notwendig ist, konnte hiermit noch nicht erklärt werden.

Die Grundlage für das Verständnis komplexer zellulärer Vorgänge sind Genomprojekte, die den gesamten Bauplan des Lebens mit allen Genen und regulatorischen Elementen eines Lebewesens enthalten. Mit der Bäckerhefe

(*Saccharomyces cerevisiae*) wurde 1997 das erste eukaryontische Genom, 2001 mit dem Humangenomprojekt das menschliche Genom sequenziert (The yeast genome directory, 1997; Lander *et al.*, 2001). Theoretisch können aus der Genomsequenz eines Organismus alle Genprodukte (RNAs und Proteine) abgeleitet werden, *de facto* sind wir aber noch nicht in der Lage, die Daten vollständig zu interpretieren. Um die Prozesse innerhalb einer Zelle oder auch die Interaktionen zwischen Zellen verstehen zu können, bedarf es weiterer Informationen wie die der Zusammensetzung des Transkriptoms, Proteoms und Metaboloms (Hieter *et al.*, 1997). Vor allem für Untersuchungen in höheren Eukaryonten müssen diese Daten noch zusätzlich räumlich und zeitlich aufgelöst sein, da die biochemische Zusammensetzung einzelner Zellen unter anderem vom Zelltyp, dem Entwicklungsstadium als auch von Umwelteinflüssen abhängt. Daten, die unter definierten experimentellen Bedingungen entstehen, bilden die Grundlage für die Beantwortung von Fragen nach der Funktion der verschiedenen Gene und an welchen zellulären Prozessen diese teilhaben, wie Gene reguliert werden und wie Gene und Genprodukte miteinander interagieren. Genauso können solche Daten Aufschluss über die Genexpression in verschiedenen Zell-Typen und die Expression spezieller Gene in Abhängigkeit von Krankheiten oder Umwelteinflüssen wie Ernährung und sportlichen Aktivitäten geben.

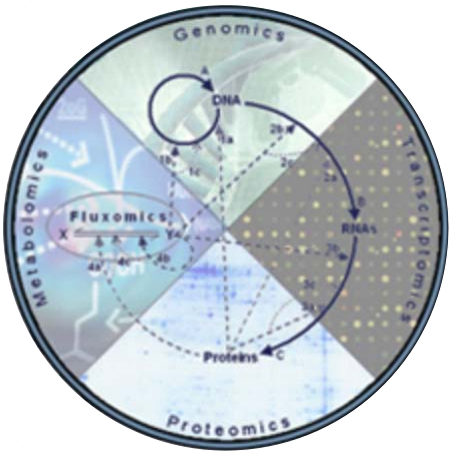


Abb. 2: Zusammenspiel der einzelnen „-omics"-Technologien: Genomics, Transcriptomics, Proteomics und Metabolomics.
Verändert nach:
http://www.ruf.rice.edu/~metabol/images/genotype.jpg

Neuere Entwicklungen in der Apparatetechnik ermöglichen es heutzutage, Organismen mit den so genannten "-omics"-Technologien wie Transcriptomics (Microarrays), Proteomics (2-D-Gelelektrophorese kombiniert mit Massenspektrometrie) und Metabolomics (Triple Quadrupol Massenspektrometer) umfassend zu untersuchen (Hollywood *et al.*, 2006). Dabei zeigt sich, dass die

Anzahl der Gene des Menschen sehr viel geringer ist als noch vor einigen Jahren vermutet. Auf der anderen Seite zeigen Ergebnisse aus neueren Untersuchungen, dass die Vielfalt der Genprodukte und der regulatorischen Mechanismen viel komplexer ist als zunächst angenommen.

Im Anschluss an die ersten Genomprojekte wurde eine Reihe von Projekten gestartet mit denen die Transkriptome von Modellorganismen untersucht wurden. Zunächst konzentrierte man sich dabei auf kurze Sequenzbereiche der mRNAs, "expressed sequence tags" (ESTs), später wurden die Transkripte auch vollständig sequenziert (Wiemann *et al.*, 2001; Gerhard *et al.*, 2004; Imanishi *et al.*, 2004). Bei der Analyse des Maus-Transkriptoms im Rahmen der RIKEN Projekte (Kawai *et al.*, 2001; Carninci *et al.*, 2005) stellte sich heraus, dass ein signifikanter Anteil der gefundenen RNAs nicht Protein-codierend war.

Zum gleichen Ergebnis kam man auch im ENCODE Projekt, in dem 1% des menschlichen Genoms hinsichtlich regulativer Elemente, Variabilität innerhalb einer Population und Chromatinstruktur und Replikation untersucht wurde (Birney *et al.*, 2007). Ergebnisse aus diesen Untersuchungen sowie Analysen mit neuen Technologien wie "tiling arrays" und DNA-Sequenzierungen mit so genanntem "second generation sequencing" (Mockler *et al.*, 2005; Mardis, 2008; Shendure *et al.*, 2008) führen zu dem Schluss, dass nicht nur die Protein-codierenden Bereiche des Genoms transkribiert werden, sondern der größte Teil des Genoms. Der Bereich der nicht-codierenden RNAs umfasst Transkripte von etwa 22 Nukleotiden Länge bei den miRNAs bis hin zu Transkripten mit einer Länge von 2 Megabasen (Kawai *et al.*, 2001). Die regulatorische Funktion von miRNAs wurde zuerst bei *C. elegans* erforscht (Johnson *et al.*, 2003; Johnston *et al.*, 2003; Lin *et al.*, 2003). Mittlerweile stellt sich heraus, dass miRNAs eine Vielzahl zellulärer Prozesse regulieren, und fehlerhaft regulierte miRNAs ein diagnostisch signifikantes Merkmal von Krankheiten wie Lungenkrebs, Brustkrebs, Leukämie und Alzheimer sind (Nelson *et al.*, 2008; Sassen *et al.*, 2008). Es gibt in der Zwischenzeit auch eine steigende Anzahl von Befunden, dass längere nicht-codierende RNAs nicht, wie früher vermutet, Artefakte sind, sondern regulatorische Funktionen ausüben (Costa, 2007; Mercer *et al.*, 2009).

Neben den Transkriptom-Projekten waren die vorliegenden Genomsequenzen auch die Grundlage für die Erforschung der Proteome. Durch massenspektrometrische Hochdurchsatzverfahren haben sich die Möglichkeiten zur Erforschung

des Proteoms (zum Beispiel das Phosphoproteom, welches Auswirkungen auf Prozesse wie Protein-Protein-Interaktionen, den Zellzyklus oder das Zytoskelett hat (Ahn *et al.*, 2001; Nita-Lazar *et al.*, 2008) oder die vergleichende Proteomanalyse bei differentiell exprimierten miRNAs (Baek *et al.*, 2008)) zwar verbessert, die Proteomik stößt jedoch auch auf technische Limits. Eines der wichtigsten ist zugleich die zentrale Eigenschaft des Proteoms: seine Komplexität. Vor allem die Modifizierung von Proteinen durch Glykosilierung, Phosphorylierung, Acetylierung etc. führt zu einer Vielfalt an unterschiedlichen Proteinen, die diejenige der Gene um ein Vielfaches übertrifft. Zudem ist zu erwarten, dass manche Proteine und insbesondere manche Modifikationen extrem selten sind. Da es sich bei der Proteomik um ein sehr zeit- und kostenintensives Forschungsgebiet handelt, wird es noch einige Zeit dauern, bis ein einigermaßen vollständiges Inventar des humanen Proteoms vorliegt.
An zellulären Prozessen sind nicht nur Proteine und RNAs sondern auch Metabolite beteiligt, z.B. für die Energiegewinnung aus den Nährstoffen oder bei der Regulierung von Proteinfunktionen durch Phosphorylierung bzw. Dephosphorylierung. Zelluläre Fehlfunktionen im Metabolitenhaushalt führen bei Menschen zu metabolischen Erkrankungen wie Diabetes und Adipositas. Fortschritte bei der umfassenden Analyse zellulärer Metabolite wurden vor allem durch Fortschritte bei der Massenspektrometrie erzielt. So ist es heute möglich, etwa 1000 Metabolite gleichzeitig zu analysieren (Altmaier *et al.*, 2008).

Die Daten aus all diesen Untersuchungen bilden die Grundlage für das Verständnis von zellulären Zusammenhängen. Isoliert betrachtet geben sie allerdings nur Einblick in Teilaspekte der Prozesse in einem Organismus. Um die komplexen biologischen Zusammenhänge der oben genannten „-omics“-Technologien untereinander zu verstehen, bedarf es bioinformatischer Lösungen, die Daten integrieren und mit Hilfe geeigneter Analysewerkzeuge neues Wissen erzeugen und weitere Experimente initiieren. Ein erster Ansatz zur integrierten Darstellung von genomischen Daten waren CYGD (Comprehensive Yeast Genome Database) und PEDANT (Protein Extraction, Description and ANalysis). In CYGD werden die genetischen, biochemischen und zellbiologischen Informationen präsentiert und zusätzliche Informationen wie die Beschreibung der Funktion genetischer Elemente und Proteine zur Verfügung gestellt. PEDANT ist

ein Tool zur automatischen Analyse von Genomen. Dies beinhaltet Motivanalysen, Sekundärstrukturanalysen oder die Strukturanalyse von Proteinen (Mewes *et al.*, 1999; Frishman *et al.*, 2001; Guldener *et al.*, 2005). Im Bereich der Mammalia werden vor allem automatische Systeme wie der UCSC Genome Browser (Abb. 3) oder Ensembl verwendet (Karolchik *et al.*, 2007; Flicek *et al.*, 2008). Diese haben die Aufgabe für bekannte Genome einen Großteil der öffentlich zugänglichen Sequenzen auf diese abzubilden.

Abb. 3: Der UCSC Genome Browser. Ansicht des menschlichen Chromosoms 16; Region 16p13.3

Mit der Weiterentwicklung der Sequenzierungstechnologien (“next generation DNA sequencing“ (Shendure *et al.*, 2008)) und Verarbeitung nicht öffentlicher Daten, entsteht vermehrter Bedarf nach Lösungen, die eine interne Datenverwaltung und Datenverarbeitung ermöglichen. Um auf individuelle oder Projekt-spezifische Anforderungen schnell und flexibel reagieren zu können, wurde innerhalb dieser Arbeit ein generisches Software-System entwickelt, das verschiedenste Sequenzdaten automatisch verarbeiten und graphisch anzeigen kann. Damit werden für Transkript- und Proteinsequenzen Abbildungen auf genomische Referenzdaten erzeugt, mit denen weiterführende Analysen durchgeführt werden können.

Neben den Organismen, deren Genom bereits vollständig sequenziert ist und in UCSC oder Ensembl integriert sind, gibt es zahlreiche Modellorganismen ohne zugehöriges Genomprojekt. Die Sequenzierung von cDNA-Bibliotheken liefert einen einfachen Weg Informationen über Protein-codierende Gene solcher Organismen

zu erhalten. Hochdurchsatz-Analysen tausender kurzer cDNA-Sequenzen erfordern ein vollautomatisches Analysewerkzeug, welches die folgenden Anforderungen erfüllt: einfache Bedienbarkeit ohne spezielles bioinformatisches Wissen, die Option zur Analyse von ESTs von Organismen verschiedenster phylogenetischer Abstammung, hohe Genauigkeit bei der Erkennung von korrespondierenden Genprodukten, erste Charakterisierung des Datensatzes mit Hilfe der Annotation der Funktion und darauf aufbauenden statistischen Analysen, die über- oder unterrepräsentierte Funktionen im Datensatz detektieren. Es wurden bereits Werkzeuge wie zum Beispiel ESTExplorer (Nagaraj *et al.*, 2007), PartiGene (Parkinson *et al.*, 2004) oder EST2uni (Forment *et al.*, 2008) zur EST-Analyse entwickelt. Diese erfordern die lokale Installation der Software und eigenständige Wartung der benötigten Datenbanken. Zusätzlich erfüllen sie die oben angeführten Anforderungen nur zum Teil. Um ein leistungsfähiges Tool zur Verfügung zu haben, das zu EST-Analysen in realen Projekten eingesetzt werden kann, wurde die Web-Applikation OREST (Waegele *et al.*, 2008) entwickelt. Sie sollte die Methodik der genannten Applikationen verbessern und die oben genannten Anforderungen insbesondere die Möglichkeit zu statistischer Auswertung erfüllen. OREST ist zusätzlich in der Lage Oligonukleotide zu charakterisieren (25mere und längere Sequenzen) wie man sie für die Erstellung von Oligonukleotid Microarrays benötigt (Schena *et al.*, 1995; Schena, 1996; Shalon *et al.*, 1996; Ramsay, 1998; Kurian *et al.*, 1999; Watson *et al.*, 2000; Chittur, 2004). Diese müssen speziell auf die Modellorganismen und die zu erforschenden biologischen Prozesse zugeschnitten werden. Eine effiziente Lösung für die Analyse von ESTs aus Organismen mit fehlender Genomsequenz bietet OREST, das mit Hilfe eines Referenzdatensatzes eines möglichst nah verwandten und bereits erforschten Organismus die Auswahl der Proben für den Chip übernimmt. In dieser Arbeit wird diese Vorgehensweise zur Erstellung eines DNA-Chips (“chip design“) für einem dem Menschen ähnlichen Modellorganismus zur Erforschung der Langzeitauswirkung von pränataler Glukokortikoideinwirkung angewendet (Datson *et al.*, 2007). Microarrays werden für ein breites Spektrum von Anwendungen, wie zum Beispiel bei der Erforschung von Infektionskrankheiten, den kognitiven Neurowissenschaften, der Entwicklung von Medikamenten und deren Risikoabschätzung (“safety assessment“) (Afshari *et al.*, 1999; Cuzin, 2001; Dhiman *et al.*, 2001) oder der Stammzellforschung eingesetzt.

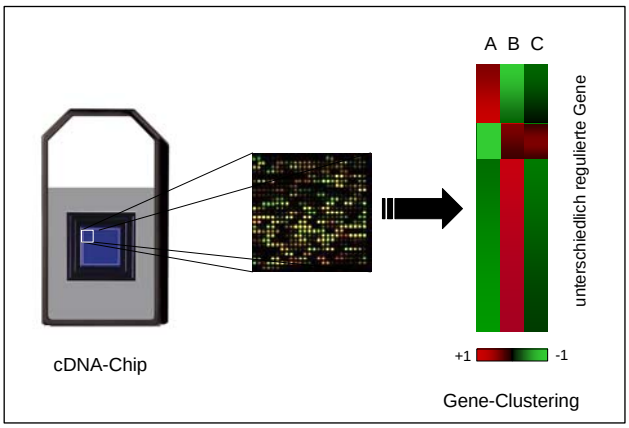


Abb. 4: Verlauf eines Transkriptomik-experiments. Vom Chip-Design bis zur Analyse der Genexpression unter verschiedenen Bedingungen A, B, und C.

Die interdisziplinäre Systembiologie vereint Biologie, Mathematik und Physik. In dieser *„geht es darum, die komplexen und dynamischen Abläufe einer Zelle oder eines Organs z.B. bei Umweltanpassung, Alterung oder Immunabwehr zu verstehen und abzubilden. Die große Fülle von Daten über einzelne Zellbestandteile bzw. -funktionen, die auf verschiedenen Ebenen der Lebensprozesse gewonnen wurde (Genom, Proteom, Metabolom), muss in einen sinnvollen Gesamtzusammenhang gebracht und im Computer nachgebildet werden, so dass Simulationen und Vorhersagen auch ohne Experimente im Labor möglich werden“* (Bundesministerium für Bildung und Forschung; BMBF; http://www.bmbf.de/de/1140.php). Für solche systembiologischen Analysen müssen Daten aus unterschiedlichen Projekten und Datenbanken integriert werden. Jede dieser Datenbanken hat eigene Formate und eigene Regeln bei der Nomenklatur eingeführt, was dazu geführt hat, dass die Daten nicht ohne weiteres kompatibel sind.

Protein-Protein Interaktionsdaten aus MPPI - “MIPS Mammalian Protein-Protein Interaction Database“ (Pagel *et al.*, 2005) und HPRD – “Human Protein Reference Database“ (Keshava Prasad *et al.*, 2009) werden beispielsweise mit UniProt-Identifiern beziehungsweise EntrezGene- Identifiern annotiert. Soll aus diesen beiden Daten-Ressourcen ein nicht-redundanter Datensatz erzeugt werden, ist es nötig sie in ein einheitliches Format umzuwandeln und doppelte Einträge zu löschen. Die Analyse der redundanten Einträge scheitert dabei häufig an der Verwendung von mehrdeutigen biologischen Termen wie Gen- oder Proteinnamen, oder an der unterschiedlichen Prozessierung der Proteinsequenzen der einzelnen Datenbanken. Die manuelle Auflösung dieser Redundanzen ist zudem sehr zeitaufwändig.

In der Vergangenheit wurden Applikationen wie MatchMiner (Bussey *et al.*, 2003) oder PICR (Cote *et al.*, 2007) entwickelt, um korrespondierende Sequenzen einander zuzuordnen. Dies erfolgt über Gennamen oder identische Sequenzen. Analysen der Gennamen und Proteinnamen ergeben einen hohen Anteil mehrdeutiger Terme, die in MatchMiner nicht berücksichtigt werden, und damit falsche Zuordungen verursachen. Der Ansatz von PICR - über identische Sequenzen Verknüpfungen zu erstellen - liefert in der Regel korrekte Beziehungen. Da aber korrespondierende Sequenzen sich in verschiedenen Datenbanken minimal unterscheiden können (zum Beispiel am N-Terminus der Proteinsequenz), werden viele Zuordnungen nicht erstellt. In dieser Arbeit wird daher die Entwicklung der Web-Applikation CRONOS (Waegele *et al.*, 2008) beschrieben, mit der mehrdeutige Terme detektiert und bei der Erzeugung der Zuordnungen berücksichtigt werden sollen.

Ein weiterer wichtiger Aspekt der Wissensgenerierung in der Biomedizin und Molekularbiologie ist das Text Mining, nicht zuletzt wegen der immer größeren Anzahl wissenschaftlicher Publikationen in PubMed. Mit dem Text Mining wird versucht, aus Textdaten relevante Informationen zu extrahieren und das bestehende Wissen zu erweitern. In den Biowissenschaften wird das Text Mining beispielsweise in den Bereichen der Funktionsannotation und zur Erkennung von Interaktionen wie Inhibition oder Aktivierung, und Protein-Protein-Interaktionen eingesetzt (Abb. 5) (Hoffmann *et al.*, 2005; Krallinger *et al.*, 2005; Krallinger *et al.*, 2005; Altman *et al.*, 2008). Kernproblematik des Text Minings ist die Erkennung von biologischen Termen im Text und deren Eindeutigkeit, insbesondere von Gen- und Proteinnamen. Im Rahmen der Entwicklung von CRONOS werden Listen von mehrdeutigen Termen generiert, die neben dem Text Mining auch andere Anwendungen der Bioinformatik und Systembiologie, die mit diesen Annotationen arbeiten, erheblich verbessern werden. Zusätzlich unterstützen sie die Integration von Daten für systembiologische Untersuchungen.

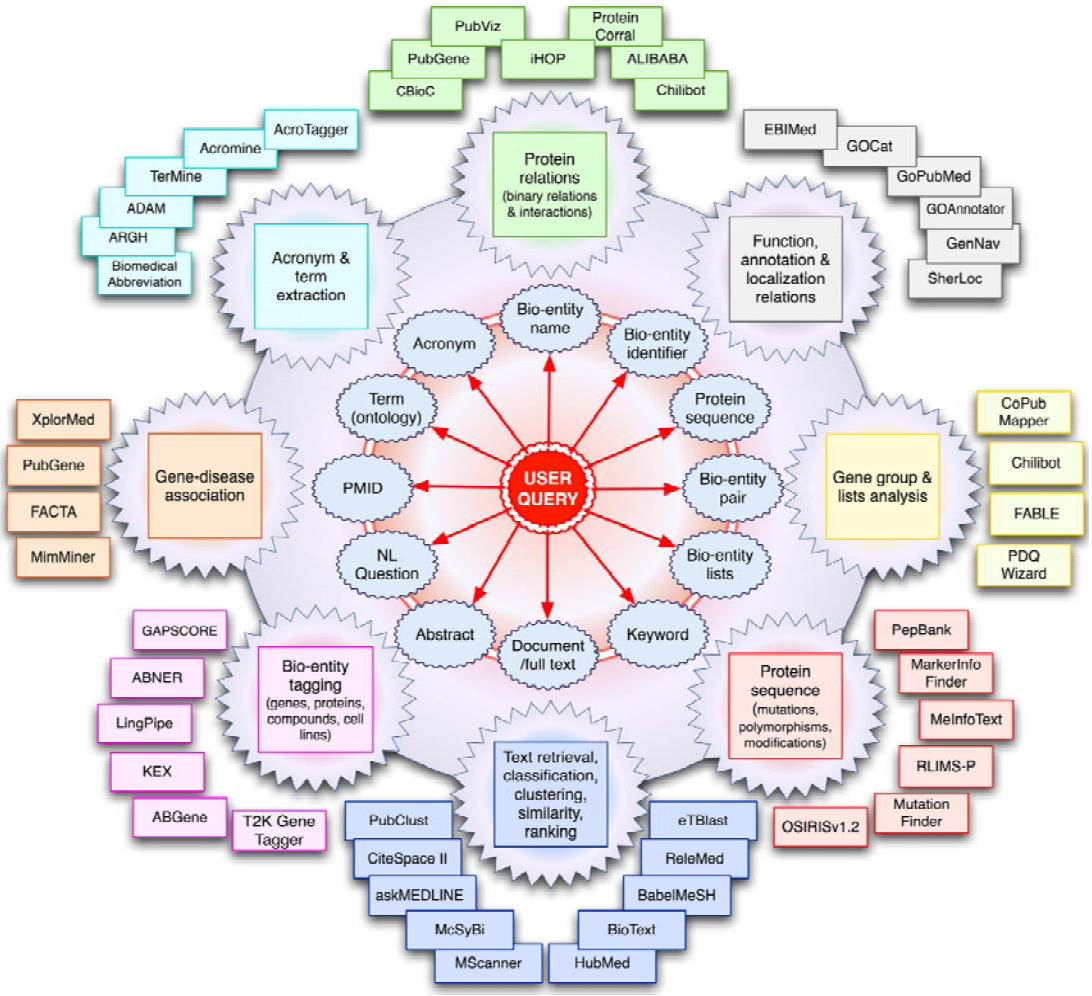


Abb. 5: Anwendungen des Text Mining in den biomedizinischen Wissenschaften. Aus (Krallinger *et al.*, 2008).

In dieser Arbeit werden nachfolgend die verwendeten Daten erläutert und generische Methoden zur Erstellung von Abbildungen dieser Transkript- und Proteinsequenzen auf genomische Referenzdatensätze dargestellt. Darüber hinaus wird die Vorgehensweise der Selektion geeigneter cDNA-Fragmente zur Generierung des ersten *Callithrix jacchus* cDNA-Chips erläutert. Anschließend werden zwei Software-Applikationen zur automatischen Funktionsannotation von ESTs und zur Verknüpfung verschiedenster Datenbankeinträge und Analyse mehrdeutiger biologischer Ausdrücke vorgestellt, die diesem generischen Ansatz folgen. Abschließend werden die erzeugten Abbildungen der Transkript- und Proteinsequenzen auf dem Genom graphisch sichtbar gemacht und mit den entwickelten Web-Applikationen zu einer Sequenz-Analyseplattform verknüpft.

2 Datengrundlage

Ziel dieser Arbeit ist die Abbildung von Transkript- und Proteinsequenzen auf Referenzdaten. Die Sequenzen, die für diese Arbeit benötigt werden sind in einer Vielzahl verschiedener Datenbanken gespeichert. Die einzelnen Datenbanken sind teils nur auf die Annotation eines Sequenztyps wie „expressed sequence tags“ (ESTs[1]), mRNAs, microRNAs[2] oder Proteinsequenzen spezialisiert. Organismus-übergreifende Datenbanken, die den Großteil aller zur Verfügung stehenden Sequenzen enthalten, sind die Referenzdatenbanken RefSeq (Pruitt *et al.*, 2007), die „UniProt Knowledgebase“ (The UniProt Consortium, 2008) und Ensembl (Flicek *et al.*, 2008). Weitere Datenbanken sind beispielsweise mirBase, deren Inhalt sich nur mit microRNAs und der Annotation dieser beschäftigt (Griffiths-Jones *et al.*, 2008). Andere sind darauf spezialisiert Zusammenhänge von menschlichen Genen und Krankheitsbildern darzulegen (Hamosh *et al.*, 2005).

Für einen effizienten Zugriff auf den Inhalt solcher Datenbanken ist es notwendig, die einzelnen Datensätze in lokalen Datenbanken abzuspeichern.

Im Folgenden wird eine Auswahl derjenigen Datenbanken vorgestellt, die die Grundlage für die weitere Arbeit bilden werden. Die Entity-Relationship-Modelle der einzelnen Datenbanken sind auf der beiliegenden CD-ROM enthalten (ERM; ERM.pdf). Die Software, die für den Import der Datenbanken entwickelt wurde, genauso wie die Softwarepakete für die nachfolgenden Projekte dieser Arbeit sind auf der CD-ROM enthalten. Die Inhalte der einzelnen Pakete sind in der Datei README.pdf aufgelistet.

[1] **ESTs** sind relativ kurze (bis 1000 Basen) transkribierte Nukleotidsequenzen, die durch Sequenzierung von cDNA-Bibliotheken erhalten werden. Die Qualität der Sequenzen ist relativ gering, da sie durch einmalige Sequenzierung entstehen (one-shot sequencing, single-pass cDNA).

[2] **microRNAs** (abgekürzt miRNA) sind eine Klasse der „nicht Protein-codierenden kurzen RNAs“, die zwischen 19 und 24 Nukleotide lang sind. miRNAs sind überall im Genom codiert. Die Mehrzahl der miRNA Gene (61%) ist in den Introns Protein-codierender Gene lokalisiert, sie können aber auch in Exons oder intergenisch vorkommen. miRNAs entstehen durch die Transkription der miRNA-Gene mit RNA-Polymerase II oder III (pol II, pol III). Nach der Transkription folgen weitere Prozessierungsschritte mit verschiedenen Enzymen, bis die funktionale miRNA vorliegt (siehe auch im Anhang). Ihre Funktion besteht in der Regulation von Translation und Degradierung von mRNA, indem sie an diese bindet. Diese mRNAs werden als microRNA targets bezeichnet.

2.1 Referenzdatenbanken

2.1.1 NCBI/Reference Sequence Database

Die “Reference Sequence (RefSeq) Database” des National Center for Biotechnology Information (NCBI) ist eine Sammlung nicht redundanter, annotierter DNA-, RNA- und Proteinsequenzen verschiedener Taxa (Pruitt *et al.*, 2007). Die Datenbank umfasst Sequenzen von Plasmiden, Organellen, Viren, Archaeae, Bakterien und Eukaryonten. Jeder Eintrag in der RefSeq-Datenbank repräsentiert ein einzelnes Molekül eines bestimmten Organismus. Das Ziel von RefSeq ist es einen umfassenden Standarddatensatz aller Sequenzen, die einem Organismus zugeordnet werden können, zu erstellen.

RefSeq-Sequenzen (DNA, RNA, Protein) stammen aus GenBank[1], unterscheiden sich im Wesentlichen aber dadurch, dass ein RefSeq-Eintrag eine Zusammenfassung aller für die Annotation einer biologischen Sequenz zur Verfügung stehenden Informationen aus GenBank darstellt. Auch sind diese nicht statisch, sondern werden, sobald weitere Daten erzeugt wurden, aktualisiert. Die Haupteigenschaften von RefSeq sind:

- Vermeidung von Redundanz
- Verknüpfung von Nukleotid- und Proteinsequenz
- Kontinuierliches Aktualisieren der Daten entsprechend dem aktuellen Wissensstand (Sequenz und Annotation)
- Validierung der Daten und Formatkonsistenz
- Beständige manuelle Re-Annotation der Daten durch NCBI und kollaborierende Gruppen

(NCBI, Oct. 2002). Weitere Unterschiede zu GenBank sind im Anhang erläutert.

[1] GenBank: Die GenBank Datenbank (Benson *et al.*, 2008) ist eine Sammlung aller öffentlich zur Verfügung stehenden Nukleotidsequenzen. Die Datenbank wird in Kollaboration von NCBI, EMBL (European Molecular Biology Laboratory), DDBJ (DNA Data Bank of Japan) und EBI (European Bioinformatics Institute) betrieben. GenBank verwaltet aktuell über 76 Millionen Sequenzen. Diese werden direkt an GenBank übermittelt, dann einer Qualitätskontrolle unterzogen. Die Annotation erfolgt nur durch den Übermittler der Sequenz.

RefSeq Einträge werden unter anderen mit folgenden Sequenz-basierten Informationen, folgend als Features bezeichnet, annotiert:

- Codierender Bereich des Transkriptes (CDS) und 5'- und 3'- nicht-codierende Bereiche (untranslated region: UTR)[1]
- Promotoren, Terminatoren[2]
- Repeat Regions[3]
- Variationen[4] etc.

Eine genaue Auflistung der verschiedenen Annotationsmerkmale findet sich auf der beigefügten CD-ROM.

2.1.1.1 RefSeq (Mammalia)

Für den Import von Säugetierdaten aus RefSeq (mRNA) in eine MySQL-Datenbank werden die von NCBI zum Download zur Verfügung gestellten Dateien im GenBank-Format verwendet. Auf Grund der Datenfülle, die für jeden RefSeq-Eintrag zur Verfügung steht, wurde eine Auswahl der Features getroffen, die die wichtigsten Daten enthalten und solche, die im Rahmen dieses Projektes benötigt werden. Diese beinhalten unter anderem Informationen über codierende Bereiche, Gen- und Proteinnamen, genomische Sequenzen und die Übersetzung in Aminosäuren sowie Cross-Referenzen.

Datenbank-Schema

Das Schema der MySQL-Datenbank ist so festgelegt, dass Daten, die bevorzugt gemeinsam abgerufen werden, in einer Tabelle zusammengefasst sind. Dies erspart zeitintensive Join-Operationen[5]. Die einzelnen Tabellen richten sich vor-

[1] Die **5'-UTR** (auch Leader-Sequenz) ist ein Teil der mRNA bzw. der DNA, welche für diese mRNA codiert. Sie beginnt am Transkriptionsstartpunkt und endet vor dem Translationsstartcodon der codierenden Region (**CDS**). Die **3'-UTR** schließt sich der CDS an. Sie beginnt hinter dem Translationsstoppcodon und reicht bis zum Polyadenylierungsstartpunkt.
[2] Als **Promotor** wird die DNA-Sequenz bezeichnet, die die Expression eines Gens reguliert. Die Promotorsequenz liegt meistens am 5'-Ende und somit vor dem RNA-codierenden Bereich. Als **Terminator** wird der Teil einer genetischen Sequenz bezeichnet, der das Ende eines Gens oder Operons markiert. Sie führt zur Termination der Transkription.
[3] Reapeat Regions sind Bereiche eines Chromosoms, in denen Widerholungen bestimmter Basenabfolgen (Repeats) variabler Länge gehäuft auftreten.
[4] In den Variationen enthalten sind Einzelnukleotidpolymorphismen (single nucleotide polymorphisms; SNPs), Insertionen und Deletionen. Als SNPs werden Punktmutationen auf dem Genom bezeichnet, die sich in mindestens 1% des Genpools einer Population etabliert haben (Mooney, 2005).
[5] Verknüpfungen der Einträge mehrerer Tabellen über gemeinsame Schlüsselwerte.

zugsweise nach den Features. Informationen, die außerhalb der Features des Genbank-Formats gespeichert sind, werden in eine eigene Tabelle geschrieben oder dem thematisch nächsten Feature zugeordnet.

Import von RefSeq in eine MySQL-Datenbank

Für den Import der Daten werden alle Dateien sequentiell gelesen. Prinzipiell können alle Einträge einzelner Organismen aus diesen gelesen werden. Empfohlen ist der Import aller gewünschten Organismen in einem Schritt, um den Leseprozess so zeitsparend wie möglich vornehmen zu können. Von inkrementellen Aktualisierungen dieser Datenbank ist abzusehen, da jeder Eintrag, der sich in der Datenbank befindet, zwischen den einzelnen Versionen geändert werden kann, und auf Veränderungen überprüft werden muss. Um diesem Umstand zu begegnen, wird eine neue Version in eine vom Aufbau identische Datenbank importiert. Wurde der Import erfolgreich abgeschlossen, können die so erzeugten Dateneinträge schnell in die eigentliche Datenbank übertragen werden. Bei der Übertragung werden Einträge, die bereits in der Datenbank enthalten sind, mit den aktuellen Daten überschrieben. Diese Methode hat den Vorteil, dass Einträge, die von RefSeq zur weiteren Annotation aus dem aktuellen Datensatz entfernt wurden, nicht verloren gehen. Damit endgültig aus RefSeq gelöschte Einträge nicht in der Datenbank verbleiben, werden alle nicht aktualisierten Einträge auf ihren Status überprüft. Lässt dieser auf die Endgültigkeit der Entfernung aus dem Datensatz schließen, werden diese gelöscht. Eine Auflistung möglicher Statuskommentare befindet sich auf der beigefügten CD-ROM.

RefSeq Datensätze beinhalten neben einer Vielzahl von Annotationen auch Referenzen zu entsprechender Literatur in PubMed[1] (NCBI, Oct. 2002).

Um bei Referenzen, die häufig auch Publikationen zu „Hochdurchsatz-Sequenzierungs“-Projekten enthalten, nur solche in die Datenbank zu importieren, die explizit Informationen zu dem betroffenen Eintrag enthalten, wird der Inhalt von PubMed einer Analyse unterzogen. Einträge in PubMed werden als unspezifische Literatur abgelehnt, wenn die Anzahl beschriebener Gene einen festgelegten Organismus-spezifischen Grenzwert überschreitet. Die Information über die in

[1] PubMed ist eine vom National Center for Biotechnology Information (NCBI) an der National Library of Medicine (NLM), entwickelte Datenbank für den Zugriff auf Literatur biomedizinischer Zeitschriften.

einer Literaturquelle beschriebenen Gene wird von Entrez Gene[1], einer weiteren Datenbank des NCBI (Maglott *et al.*, 2007) zur Verfügung gestellt. Aus diesen Daten wird eine Organismus-spezifische Statistik darüber erstellt, wie viele verschiedene Gene in einer bestimmten Literaturquelle vorkommen. Manuelle Evaluierung ergibt für jeden untersuchten Organismus einen spezifischen Grenzwert. Dieser gibt an, ab wie vielen Genen von „Hochdurchsatz-Sequenzierungs"-Projekten ausgegangen werden kann. Literaturreferenzen, die eine geringere Anzahl Gene erwähnen, werden importiert.

Implementierung der Import-Software

Die Implementierung erfolgt in Java. Die Software ist in Module aufgeteilt, so dass auf Formatänderungen seitens RefSeq und auf sich ändernde Anforderungen an den Inhalt der Datenbank schnell und ohne großen Aufwand reagiert werden kann.

Die einzelnen Module umfassen:

- Konfiguration des Datenbank-Schemas
- Das Einlesen der Daten
    - Erkennung der Features
    - Erkennung der Qualifier[2]
- Import ausgewählter Features und Qualifier
- Übertrag der Daten in die lokale RefSeq-Datenbank
- Statustest und Löschen ungültig gewordener Einträge

2.1.1.2 Integration der Positionierungen auf dem Genom

RefSeq positioniert alle in die Datenbank eingetragenen Sequenzen auf das entsprechende aktuelle Genom. Diese Positionierung erfolgt nur bei der Veröffentlichung eines neuen Genom Assemblies[3] und wird für diejenigen Einträge durchgeführt, die zu diesem Zeitpunkt fester Bestandteil des Datensatzes sind (nicht für Einträge, die gerade re-annotiert werden).

[1] Entrez Gene enthält Gen-basierte Informationen von Genomen, die schon vollständig sequenziert wurden. Sie werden manuell oder automatisch (aus RefSeq-Einträgen) oder aus anderen Modellorganismus-Datenbanken erzeugt.
[2] Qualifier: standardisierte Form der detaillierten Beschreibung eines Features.
[3] Ein Assembly beschreibt die Zusammenstellung der bei einem Genomprojekt erzeugten kurzen DNA-Sequenzen zu längeren Abschnitten. Diese stellen die Chromosomen dar, aus denen sie ursprünglich erzeugt wurden.

Integration der genomischen Positionierung in die lokale RefSeq-Datenbank

Die Koordinaten zu den jeweiligen RefSeq-Einträgen werden als Flat-File[1] zur Verfügung gestellt. In diesen Organismus-spezifischen Dateien sind alle Daten teilweise auch für alternative Genom Assemblies eines Organismus enthalten. Welches Assembly verwendet wird kann optional eingestellt werden. Mit der Standardeinstellung werden die Koordinaten des Referenz-Assemblies importiert.

2.1.2 UniProt Knowledgebase

Die „UniProt Knowledgebase“ (UniProtKB, (The UniProt Consortium, 2008)) stellt die zentrale Datenbank für Proteinsequenzen –UniProtKB- zur Verfügung. Bis in das Jahr 2002 koexistierten die EBI/SIB (European Institute of Bioinformatics / Swiss Institute of Bioinformatics) Swiss-Prot + TrEMBL Datenbanken und die PIR (Protein Information Resource) International Protein Sequence Database (PIR-PSD, (Barker *et al.*, 1999; Barker *et al.*, 2000; Barker *et al.*, 2001; Wu *et al.*, 2002)). Sie unterschieden sich sowohl in den enthaltenen Proteinsequenzen als auch in der Art der Annotation. Ab 2002 bildeten EBI, SIB und PIR das UniProt – Konsortium. Hauptziel dieses Zusammenschlusses ist es, eine Datenbank von hoher Qualität zur Verfügung stellen zu können, die alle Proteinsequenzen beinhaltet, welche vollständig klassifiziert, annotiert und mit Cross-Referenzen versehen frei zugreifbar ist (http://www.expasy.org/sprot/userman.html#what_is_uniprot).
UniProtKB besteht aus zwei Datenbanksektionen: Swiss-Prot (Boutet *et al.*, 2007) und TrEMBL. In der Swiss-Prot Sektion befinden sich manuell annotierte Proteinsequenzen. Die Annotation stammt dabei aus veröffentlichten Artikeln und aus ausgewerteten automatischen Analysen. TrEMBL dagegen enthält Sequenzen, die bislang nur automatischen Analysen unterzogen wurden. Sobald sie manuell annotiert wurden, werden auch diese Proteinsequenzen in Swiss-Prot integriert.

2.1.2.1 Swiss-Prot

Annotation

Jeder Swiss-Prot-Eintrag enthält Sequenzdaten, Referenzen auf das Protein beschreibende Literatur und die Taxonomie des Ursprungsorganismus. Darüber

[1] Ein Flat-File besteht aus Klartext und enthält einen Datensatz pro Zeile. Innerhalb eines Datensatzes werden einzelne Felder durch Begrenzungssymbole (Kommata, etc.) getrennt (http://en.wikipedia.org/wiki/Flat_File).

hinaus werden alle Einträge mit zusätzlichen Informationen zum Protein annotiert, wie zum Beispiel die Funktion eines Proteins, posttranslationale Modifikationen wie Phosphorylierung, Acetylierung oder Glycosylphosphatidylinositol-Anker (kurz GPI-Anker) Domänen und Bindungsstellen für Biomoleküle, Sekundär- und Quartärstruktur, Krankheiten, die mit Modifikationen in diesem Protein in Zusammenhang stehen, Konflikte, die die Sequenz betreffen, ebenso wie Spleißvarianten.
Swiss-Prot versucht möglichst umfassend zu annotieren. Um dies zu erreichen werden zusätzlich zu Literatur, die die neuen Sequenzen beschreibt, Reviews zur weiteren Annotation der schon bekannten Sequenzen verwendet. Diese Annotationen beinhalten vor allem Informationen über Proteinfamilien oder Gruppen von Proteinen. Zusätzlich werden Experten für die Annotation spezieller Proteingruppen herangezogen.

Minimale Redundanz

In der Swiss-Prot Sektion von UniProtKB wird großer Wert darauf gelegt, keine Redundanz in den Sequenzdaten auftreten zu lassen. Beschreiben mehrere Literaturquellen dasselbe Protein, werden diese Informationen so gut es möglich ist, in einem Datenbankeintrag vereinigt. Sollten Diskrepanzen in den zugehörigen Sequenzen auftreten, werden diese vermerkt. Normalerweise wird jedem Eintrag eine so genannte „Master-Sequenz“ zugewiesen. Sequenzvarianten, die durch alternatives Spleißen entstehen, können im Web angezeigt werden.
Für die Wahl zur „Master-Sequenz“ sollte eine Sequenz folgende Eigenschaften erfüllen:

- Sie ist die gebräuchlichste Sequenz.
- Sie ist die zu orthologen Sequenzen aus anderen Organismen ähnlichste Sequenz.
- Auf Grund der Länge oder Aminosäurezusammensetzung erlaubt sie die eindeutige Beschreibung von Domänen, Isoformen, Polymorphismen, posttranslationalen Modifikationen etc.
- Falls keine weiteren Informationen zu Sequenzen gegeben sind, wird die längste bevorzugt.

In besonderen Fällen werden dennoch mehrere Einträge für die Produkte eines Genes eingerichtet. Dies erfolgt dann, wenn alternatives Spleißen zu Varianten führt, die zwar vom selben Gen abstammen, aber nur wenige oder keine Exons gemein haben. In solch einem Fall ist der Unterschied der Proteine zu gravierend und die Varianten werden in verschiedenen Einträgen beschrieben (http://beta.uniprot.org/faq/30, Beispiel siehe Anhang).

Referenzen zu anderen Datenbanken

Swiss-Prot referenziert für einen Eintrag, wenn möglich, zu anderen Sequenz-basierten Datenbanken. Diese können Nukleotidsequenzen enthalten, beschreiben aber hauptsächlich Proteinsequenzen oder interpretieren die Tertiärstruktur eines Proteins. Zusätzlich wird auf Organismus-spezifische Datenbanken verwiesen. Momentan referenziert Swiss-Prot auf mehr als 50 verschiedene Datenbanken.

2.1.2.2 Translated EMBL (TrEMBL)

TrEMBL ist die nur automatisch annotierte Sektion der UniProtKB. Sie enthält Translationen aller codierenden Regionen der in GenBank vorliegenden Nukleotidsequenzen, Proteinsequenzen aus Publikationen oder direkt an UniProtKB übermittelte Sequenzen, die noch nicht in Swiss-Prot integriert wurden. TrEMBL erlaubt den schnellen Zugriff auf diese Sequenzen ohne den Annotationsstandard von Swiss-Prot zu verringern.

Automatische Annotation

Einträge in TrEMBL werden automatisch annotiert, indem Annotation aus gut beschriebenen Swiss-Prot-Einträgen auf diese übertragen wird. Die Daten werden nur dann übernommen, wenn sie derselben Proteingruppe (definiert laut InterPro, einer Datenbank für Proteinfamilien, -domänen, und -bindungsstellen) angehören. Diese Vorgehensweise nähert den Annotationsstandard von TrEMBL an Swiss-Prot an, und verbessert so die Qualität der verfügbaren Daten.

Minimale Redundanz

Sequenzen desselben Organismus werden in einem Eintrag zusammengefasst, wenn sie gleich lang und in 100% ihrer Aminosäuresequenz übereinstimmen, um die Zahl der Redundanzen zu verringern.

Datenbank-Schema und Import der UniProtKB

Die Sektionen Swiss-Prot und TrEMBL werden getrennt voneinander importiert und auch in verschiedenen Tabellen mit dem jeweiligen Präfix gespeichert. Dies beschleunigt die Anfragen für den in Kapitel 5 vorgestellten Service zur automatischen Annotation von „expressed sequence tags“ (ESTs).
Auf Grund der verschiedenen Anfragetypen, die in nachfolgenden Projekten (Kapitel 4, 5 und 6) gestellt werden, werden die zu importierenden Daten in zwei Untergruppen aufgeteilt und in verschiedenen Tabellen abgelegt. Dadurch können Informationen, die häufig gemeinsam abgefragt werden, ohne Speicher-intensive Join-Operationen abgerufen werden. Erstere enthält Gen- und Proteinnamen, deren Synonyme und so genannte Keywords, die die Haupteigenschaften des jeweiligen Proteins beschreiben. In die zweite werden ausgewählte Cross-Referenzen und die Aminosäuresequenz importiert.
Die Daten werden aus Flat-Files gelesen und wegen der in 2.1.1.1 aufgeführten Gründe nicht inkrementell aktualisiert. Ebenso können für den Import einzelne Organismen ausgewählt werden. Diese werden dann in Organismus-spezifische Tabellen geladen.

2.1.3 Ensembl

Ensembl (Flicek *et al.*, 2008) ist ein gemeinsames Projekt des European Bioinformatics Institute (EBI) und des Wellcome Trust Sanger Institute (WTSI). Ziel ist es ein Softwaresystem zu entwickeln, das ausgewählte eukaryontische Genome vollautomatisch annotiert und bei Bedarf weitere Informationen einpflegt. In nächster Zukunft wird sich Ensembl auf die Analyse und Annotation von Vertebraten konzentrieren. Ziel des Ensembl-Projektes ist:

- Genomische Daten automatisch und möglichst genau zu analysieren
- Aktuelle Analysen und Annotationen für bekannte Daten zur Verfügung zu stellen
- Ergebnisse solcher Analysen möglichst schnell im Web zu präsentieren
- Verteilung der Analysen an verschiedene bioinformatische Einrichtungen

Ensembl wurde auch deshalb in die Liste der Grundlagendatenbanken aufgenommen, weil zu jedem Transkript die Identifier der entsprechenden

Affymetrix Expression Analysis Arrays referenziert werden genauso wie Agilent Microarrays[1].

Datenbank-Schema und Import von Ensembl

Für alle in Ensembl gespeicherten Daten werden MySQL-Dumps[2] angeboten, das Datenbank-Schema wird unter http://www.ensembl.org/info/using/api/core/-schema/schema_description.html erläutert. Der Benutzer-definierte Import erfolgt vollautomatisch. Die Importsoftware enthält zahlreiche Konfigurationsmöglichkeiten. So können die Daten einzelner Organismen, sowie eine Auswahl von Organismen importiert werden. Die Importzeit hängt hier direkt von der Datenmenge der einzelnen Organismen ab, da die Dateien Organismus-spezifisch heruntergeladen werden. Konfiguriert werden kann auch, welche Organismus-spezifischen Teildatenbanken importiert werden sollen (http://www.ensembl.org/info/docs/api/index.html). Zur Auswahl stehen:

- **Core** Database: enthält den Hauptteil der Annotation wie zum Beispiel:
    - das mit der automatischen Genomanalyse und Annotationspipeline annotierte Set der Ensembl Gen-, Transkript- und Proteinmodelle
    - Sequence tagged sites (STS)[3] Marker Information
    - Microarray-Oligonukleotid Identifier
    - Sequenzen und weitere Referenzen zu anderen Datenbanken, etc.
- **cDNA** Database: enthält Alignments von cDNA und Proteinen mit dem Genom
- **Funcgen** Database: enthält Daten aus Funktionsanalysen
    - Daten, die mit Chromatin Immunoprezipitation erzeugt wurden
    - Daten zu epigenetischen Modifikationen und DNA-bindende Proteine
    - Externe Daten mit regulatorischem Hintergrund: Sequenzmotive, miRNA targets, etc.

[1] DNA-Microarrays sind unter anderem Werkzeuge für die Genexpressionsanalyse. DNA-Microarrays bestehen aus an Membran oder Glas angeordneten DNA-Molekülen (Oligonukleotide). Im Versuch werden aus biologischen Systemen extrahierte und markierte (z.B. fluoreszierend) Moleküle auf das Array aufgetragen. Die Oligonukleotide des Arrays binden diese spezifisch und erlauben so die Analyse der Art und Konzentration der gebundenen Moleküle (Ramsay, 1998).

[2] MySQL-Dump: komplettes Backup der in der Datenbank vorhandenen Einträge, das ein einfaches Importieren des Datenbank-Schemas und der Daten ermöglicht.

[3] Ein STS ist ein kurzer DNA-Abschnitt (200 bis 500 Basenpaare lang), der nur einmal in einem Genom vorkommt und dessen Sequenz und genomische Positionierung bekannt sind.

- **OtherFeatures** Database: enthält ein nicht redundantes Ensembl EST-Set und die jeweilige Position auf dem Genom
- **Variation** Database: enthält Einzelnukleotidpolymorphismen (SNPs: werden unterschieden in ‚synonymous' und ‚non-synonymous')[1], Genotypen, Allele, etc.
- **Vega**: (Vertebrate Genome Annotation) Einträge der Vega-Datenbank werden manuell annotiert und häufig aktualisiert. Die Datenbank enthält Einträge von bereits fertig gestellten Genomen von Vertebraten (Wilming *et al.*, 2008).

2.2 Andere Datenbanken

2.2.1 dbEST

dbEST (Boguski *et al.*, 1993) ist eine Teildatenbank von GenBank (Benson *et al.*, 2008), die die Sequenzinformationen für "single-pass" cDNA, oder ESTs für eine Vielzahl Organismen enthält. Momentan sind in dbEST circa 60 Millionen Sequenzen integriert, davon entfallen mehr als 8 Millionen allein auf humane Sequenzen (Stand Februar 2009). Genaue Angaben über die Anzahl der in dbEST enthaltenen Sequenzen können unter http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html eingesehen werden. dbEST wird in Flat-Files zum Download bereitgestellt. Die Daten werden auf Grund ihrer Menge nicht in eine Datenbank übertragen, sondern in Organismus-spezifische flat-Files umgeschrieben.

2.2.2 mirBase

In mirBase (Griffiths-Jones *et al.*, 2008) werden alle bisher bekannten microRNAs (Zhang *et al.*, 2008) gespeichert und mit Annotation versehen. mirBase besteht aus drei Teilbereichen:

- mirBase Sequence Database: Datenbank aller publizierten miRNA Sequenzen und deren Annotation
- mirBase Registry: Zuständig für die Zuweisung von Namen für neue miRNA Gene noch vor Veröffentlichung

[1] SNPs werden als ‚synonymous' bezeichnet, wenn durch die Veränderung der genetischen Sequenz keine Änderung der Proteinsequenz erfolgt, als ‚non-synonymous' falls die Proteinsequenz verändert wird.

- mirBase Targets: Datenbank für Vorhersagen zu miRNA Targets in Tieren und Menschen

Jeder Eintrag in der mirBase Sequence Database (Griffiths-Jones *et al.*, 2006) stellt eine vorhergesagte Haarnadelstruktur[1] eines miRNA-Transkriptes dar. Informationen zu Lokalisation und Sequenz der reifen miRNA (mature miRNA) sind hinzugefügt (http://microrna.sanger.ac.uk/).

Datenbank-Schema und Import

Der Download der mirBase Sequence Database als MySQL-Dump erfolgt vollautomatisch (ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/), die Daten werden unverändert als Datenbank „mirBase“ geladen.

2.2.3 Online Mendelian Inheritance in Man (OMIM™)

Die „Online Mendelian Inheritance in Man“-Datenbank (OMIM™, (Hamosh *et al.*, 2005)) ist eine umfassende Ressource über die Zusammenhänge von menschlichen Genen und Krankheitsbildern. Erstellt wird diese Datenbank von Wissenschaftlern und Ärzten mit Hilfe biomedizinischer Literatur. Jeder Eintrag besitzt eine Zusammenfassung über einen genetisch bestimmten Phänotyp und/oder ein Gen. Einträge werden, sofern möglich, mit anderen Datenbanken wie PubMed (NCBI, Oct. 2002), HGNC[2] (Bruford *et al.*, 2008) oder Sequenz-basierten Datenbanken verknüpft. OMIM™ wird in zwei Bereiche aufgeteilt: GeneMap und MorbidMap. GeneMap enthält alle Gene, deren zytogenetische Lokalisation bekannt ist, sowie weitere Informationen wie die Methode der Lokalisation, Genstatus (siehe Anhang der beigefügten CD-ROM) oder das korrespondierende Maus-Gen. MorbidMap listet alle in OMIM™ beschriebenen Krankheiten auf. Diese werden unterschieden in Einträge, die:

- einen Phänotypen beschreiben, dem kein eindeutiger Locus zugeordnet werden kann
- einen Phänotypen und ein Gen mit bekannter Sequenz beschreiben

[1] Haarnadelstrukturen können entstehen, wenn zwei Regionen des gleichen Moleküls mit einer palindromischen Nukleotidsequenz eine Doppelhelix bilden, die am Ende durch eine ungepaarte Schleife abgeschlossen wird.

[2] Ziel Des HUGO Gene Nomenclature Committee (HGNC) ist die Zuweisung eindeutiger und aussagekräftiger Namen und Kurznamen zu jeden menschlichen Gen. Die Datenbank enthält 2008 24.000 Einträge anerkannter Gennamen und damit verbundenen Informationen.

- einen mehrfach bestätigten Phänotypen beschreiben, zu dessen molekularem Hintergrund keine Informationen zur Verfügung stehen
- Phänotypen, bei denen Mendel'sche Vererbung vermutet wird, aber noch nicht begründet werden konnte

(http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html).

Datenbank-Schema und Import von OMIM™

Das Datenbank-Schema ist so konzipiert, dass es sowohl GeneMap und MorbidMap gemeinsam in einer Tabelle aufnehmen kann. Dabei werden korrespondierende Einträge aus beiden Teilbereichen von OMIM™ miteinander gekoppelt. GeneMap und MorbidMap stehen als Flat-Files zur Verfügung. Der Import wird mit GeneMap begonnen. MorbidMap wird danach unter Berücksichtigung der bestehenden Relation zwischen GeneMap und MorbidMap importiert. Dabei ist es möglich mehrere MorbidMap-IDs einer GeneMap-ID zuzuordnen.

2.2.4 Protein Data Bank (PDB)

Die „Protein Data Bank" (PDB, (Burkhardt *et al.*, 2006; Henrick *et al.*, 2008)) ist die Standarddatenbank für Informationen über die dreidimensionale Struktur biologischer Moleküle wie einzelne Proteine, Proteinkomplexe oder Nukleinsäuren. Diese Moleküle wurden mit Röntgenkritallographie, NMR und Elektronenmikroskopie bestimmt, sind also experimentell bestätigt (http://journals.iucr.org/a/issues/2008/01/00/sc5004/index.html).

Datenbank-Schema und Import der PDB

Die lokal erzeugte PDB-Datenbank enthält nur die Daten, die für dieses Projekt benötigt werden – nach Organismus aufgetrennt den PDB-Identifier, Bezeichnung der Proteinkette(n) und deren Sequenz(en). Sind weitere Informationen erforderlich, können auch hier durch die Modularität der Implementierung Anpassungen vorgenommen werden. Die Daten werden via FTP als einzelne Dateien geladen (ftp://ftp.wwpdb.org/pub/data/structures/all/pdb/). Die Dateien werden nach der Aminosäuresequenz (ASS) analysiert, d.h. assoziierte DNA- bzw. RNA-Fragmente werden nicht in Betracht gezogen. Die ASS wird nur in die Datenbank importiert, wenn sie eine Länge von mindestens 50 Aminosäuren auf-

weist. Die Einschränkung ist notwendig um insignifikante Positionierungen der Proteine auf das Genom zu vermeiden.
Damit die Sequenzen später auf einem Genom positioniert werden können, müssen die einzelnen Aminosäuren vom Dreibuchstabencode in den Einbuchstabencode übersetzt werden. 19 der 23 proteinogenen Aminosäuren können modifiziert in Proteinen vorkommen und sind mit speziellen Dreibuchstabencodes versehen. Diese werden in ihre nicht modifizierten Ausgangsaminosäuren zurückübersetzt. So wird beispielsweise aus 2-Amino-3-Cyclohexyl-Propionsäure (ALC) wieder ein Alanin oder aus Acetamidomethylcystein (CY1) ein Cystein (eine genaue Auflistung der Kürzel der modifizierten Aminosäuren und ihren korrespondierenden Ursprungsaminosäuren befindet sich auf der beigefügten CD-ROM). Sind keine Informationen über die ursprüngliche Aminosäure in den Dateien festgehalten, wird diese weggelassen.

2.2.5 Funktionsannotation

2.2.5.1 Gene Ontology (GO)

Das Gene Ontology (GO) Projekt stellt ein strukturiertes und kontrolliertes Vokabular für die Annotation der Funktion von Genen, Genprodukten und Sequenzen zur Verfügung. Die verwendeten Ontologien wurden zwischenzeitlich erweitert und verfeinert (The Gene Ontology Consortium, 2008).

GO deckt die folgenden Bereiche ab:

- Zellkomponente / Cellular component
- Biologischer Prozess / Biological process
- Molekulare Funktion / Molecular function

(http://www.geneontology.org/GO.doc.shtml).

Integration von Gene Ontology in die RefSeq-Datenbank

Gene Ontology wird über ein von Entrez Gene (Maglott *et al.*, 2007) angebotenes Gene/GO-Mapping in die RefSeq-Datenbank eingebunden. Die Informationen werden vollautomatisch heruntergeladen, eingelesen, analysiert und in die Organismus-spezifischen, bereits angelegten GO-Tabellen importiert. Dieser Vorgang kann für einzelne Organismen genauso wie für mehrere Organismen gleich-

zeitig durchgeführt werden. Auf Grund der Datenstruktur ist es zeitsparender die GO-Terme für alle gewünschten Organismen auf einmal zu importieren. Damit die Integritätsbedingungen der Datenbank nicht verletzt werden und die GO-Terme einer bestimmten RefSeq-ID zugewiesen werden können, ist eine Übertragung der GeneID auf RefSeq-ID nötig; diese wird als Schlüsselwert in die Tabellen importiert.

2.2.5.2 Functional Catalogue (FunCat)

Der FunCat (Ruepp *et al.*, 2004) ist ein hierarchisch strukturiertes, Organismus-unabhängiges Klassifizierungssystem zur Annotation von Funktions-Eigenschaften eines Proteins. Es besteht aus 28 Hauptkategorien wie Metabolismus, zellulärer Transport und Signaltransduktion. Unterkategorien bieten auf bis zu sechs Hierarchieebenen spezifischere Annotationen. In Version 2.1 werden insgesamt 1362 funktionelle Kategorien unterschieden. Auf den FunCat kann direkt zugegriffen werden, da die Entwicklung des Systems und die Weiterentwicklung der Datenbank am Institut für Bioinformatik und Systembiologie des Helmholtz Zentrum München erfolgt.

2.2.6 MIPS Mouse Functional Genome Database (MfunGD)

MfunGD (Ruepp *et al.*, 2006) ist eine Datenbank für annotierte Proteine der Maus. MfunGD integriert die Annotation der Funktion von Proteinen, Informationen über das Vorkommen von Proteinen in Proteinkomplexen (Ruepp *et al.*, 2008) und Protein-Protein-Interaktionen in einer Datenbank. Zusätzlich werden die Literaturreferenzen, mit deren Hilfe die Annotationen vorgenommen wurden, und Verknüpfungen zu weiteren Datenbanken - wie SIMAP (Rattei *et al.*, 2008), UniProtKB oder MGI - angegeben.

Import der MfunGD-Identifier in die RefSeq-Datenbank

Die Identifier der Mouse Functional Genome Database werden in die Cross-Referenzen-Tabelle für die Maus der RefSeq-Datenbank integriert. Dazu wird das RefSeq/MfunGD-Mapping verwendet, das bei jedem neuen Genom-Assembly berechnet wird. Die Integration der MfunGD-Identifier in RefSeq wird aus Effizienzgründen durchgeführt und ermöglicht leichteren Zugriff auf die Identifier von MfunGD. Dies wird für die in Kapitel 6 vorgestellte Software von Bedeutung sein.

3 Abbildung von Transkript- und Proteinsequenzen auf genomische Referenzdaten

Ein Ziel dieser Arbeit ist die generische Erzeugung von Relationen zwischen einzelnen Sequenzen verschiedener Typen wie mRNAs, microRNAs oder Proteinen. Die Relationen können Organismus-spezifisch oder Organismus-übergreifend berechnet werden. Diese Aufgabe lässt sich durch die Feststellung der absoluten Koordinaten der verschiedenen Sequenzen auf einem bestimmten Genom lösen. Zur Bestimmung der Positionen sind im Folgenden die Vorverarbeitung der Sequenzen, die Entwicklung von geeigneter Kontroll-Software und deren Optimierung für verschiedene Fragestellungen beschrieben. Zur Positionsberechnung wird Blat (Kent, 2002) verwendet, da die Modellierung von Intron/Exon Strukturen berücksichtigt wird. Ein weiterer Vorteil ist die erheblich schnellere Berechnungszeit im Vergleich zu Blast (Altschul *et al.*, 1997).
Im Folgenden wird die Vorgehensweise zur Positionierung der verschiedenen Sequenztypen erläutert, ebenso wie die angewendeten Optimierungsschritte zur Beschleunigung dieser Berechnungen. Weiterhin werden Analysen zur Erkennung von Pseudogenen und die Zusammenfassung der Positionierungen verschiedener Sequenztypen in einem so genannten Konfidenz-Track beschrieben. Abschließend erfolgt die Visualisierung der Positionierungen mit GBrowse (Stein *et al.*, 2002; Donlin, 2007).

Die Positionierung ganzer Datensätze eines bestimmten Organismus auf dessen Referenzgenom erfolgt immer dann, wenn entweder ein neues Assembly erstellt wurde, oder Sequenzen existieren, die noch nicht integriert wurden. Bei einer Aktualisierung werden die neuen Sequenzen für die Positionierung wie ein Datensatz behandelt, die Ergebnisse aber additiv in die bereits bestehende Datenbank integriert. Das zu diesem Zweck entwickelte Softwarepaket gliedert sich in drei Hauptbereiche auf: die Vorverarbeitung der zu positionierenden Sequenzen, den eigentlichen Positionierungsprozess und die Überprüfung und Integration der Ergebnisse in eine Datenbank.

Die Vorverarbeitung und die verwendeten Parameter (Blat) zum Positionieren der Sequenzen sind abhängig vom Sequenztyp. Sie sind für jeden Typ bereits voreingestellt, können aber individuell angepasst werden.

Durch die Menge der zu positionierenden Daten entsteht ein hoher Rechenaufwand, der durch geeignete Implementierungen verringert werden kann. Zur Optimierung stehen zwei Ansätze zur Verfügung: die Verminderung des Rechenaufwandes und die Parallelisierung von Rechenprozessen.

Optimierung durch Verminderung des Rechenaufwandes

Der wiederholte Rechenaufwand bei Veränderungen eines Assemblies oder Genprodukten kann mit Hilfe eines iterativen Verfahrens auf einen Bruchteil des „ab initio“-Ansatzes[1] verringert werden. Erste Analysen werden durchgeführt, um Veränderungen in den Chromosomensequenzen zu identifizieren. Unveränderte Genprodukte, die auf unveränderten Chromosomen aligniert sind, werden nicht neu positioniert und verringern so den Datensatz der zu positionierenden Sequenzen. Bei neuen Assemblies[2], an denen noch häufig Veränderungen und Korrekturen durch Re-Sequenzierungen vorgenommen werden, kann dieser Schritt ausgelassen werden. Eine weitere Reduzierung der Laufzeit wird erreicht, indem die verbleibenden Sequenzen jeweils nur auf das Chromosom positioniert werden, auf dem sie sich bereits im vorherigen Assembly befanden, anstatt sie auf dem gesamten Genom zu positionieren. Ändern sich die Koordinaten zu einer Sequenz nicht, bleibt sie dort bestehen. Da diese Berechnung bei Veränderung der Chromosomensequenz[3] ausgeführt wird, ist eine Abweichung der Position möglich. Die Abweichung, die die neue Positionierung auf dem Chromosom im Vergleich zur vorherigen aufweisen darf, kann individuell angepasst werden. Ist die Abweichung kleiner als die maximal mögliche, bleibt auch diese Positionierung erhalten. Alle verbleibenden Sequenzen werden abschließend auf das gesamte Genom aligniert.

[1] Der „ab initio“ (lat.: von vorn, von Anfang an) Ansatz ist die Berechnung der Positionierung aller Sequenzen auf das gesamte Genom.
[2] Erstes vollständiges Assembly eines Organismus.
[3] Veränderungen in der Sequenz entstehen durch Einfügen oder Löschen von nicht korrekten Chromosomenabschnitten in Folge verbesserter Sequenzierung und Neuberechnung des Assemblies.

Optimierung durch Parallelisierung

Die oben genannten Optimierungen können grundsätzlich vorgenommen werden, da sie den sequentiellen Ablauf der Rechenschritte beschleunigen. Die folgend beschriebenen Optimierungen sind von der zur Verfügung stehenden Hardware abhängig. Anstatt alle Berechnungen, die mit einem Assembly-Update anfallen sequentiell abzuarbeiten, bietet die Parallelisierung von Rechenprozessen weitere Möglichkeiten die Laufzeit zu verringern. Parallel berechnet werden können alle Positionierungen auf das Genom, indem dieser Prozess in Teilprozesse aufgeteilt wird. Bei der Alignierung der Sequenzen auf jeweils nur ein Chromosom können die Prozesse getrennt voneinander ausgeführt werden, die Resultate werden nach Beendigung der Berechnungen wieder zusammengeführt. Beim Alignieren auf das Genom spaltet man das Assembly in die einzelnen Chromosomen auf und verfährt wie zuvor. Sind ausreichende Rechenkapazitäten vorhanden, können auch die Sequenz-Datensätze weiter aufgespalten werden. Das heißt, pro Chromosom wird nicht nur ein Prozess, sondern es werden mehrere Prozesse (Anzahl der Teil-Sequenz-Datensätze) erzeugt.

Erfolgt bei der sequentiellen Berechnung der Import der Ergebnisse erst, wenn alle Prozesse erfolgreich abgeschlossen wurden, kann bei der parallelen Berechnung der Importprozess mit den Rechenprozessen gestartet werden, so dass Teilergebnisse von bereits beendeten Prozessen sofort importiert werden können. Sind Koordinaten für Datensätze verschiedener Sequenztypen zu berechnen, können diese gemeinsam gestartet werden. Während des Imports und der Analysen der Ergebnisse eines Sequenztyps können bereits die Rechenprozesse eines anderen Typs gestartet werden.

3.1 Positionierung transkribierter und prozessierter genetischer Sequenzen

Mit der Positionierung von transkribierten und prozessierten genetischen Sequenzen (mRNA) wird die Struktur und die Lage von Genen auf dem Genom bestimmt. Dazu werden die mit Hilfe der Polymerase-Kettenreaktion in „komplementäre DNA“ (cDNA) übersetzten mRNAs auf dem Genom angeordnet. Da Transkripte durch Spleißen der Intronsequenzen oder Polyadenylierung keine Kopie der genomischen Sequenz darstellen, sind Alignments mit 100% Sequenzi-

dentität selten; Ausnahmen sind Ein-Exon-Gene. Um die Sequenzidentität als Maß für die Qualität eines Alignments nicht mit zahlreichen nicht übereinstimmenden Basenpaarungen auf Grund der Polyadenylierung negativ zu beeinflussen, wird der Poly-(A)-Schwanz vor der Positionierung entfernt (siehe unten). Zur Berechnung der Lage von Genen auf dem Genom werden die Sequenzen der mRNAs aus RefSeq verwendet. Nicht experimentell verifizierte Einträge, die aus Genvorhersagen abgeleitet wurden, werden ausgeschlossen.

Entfernen der Polyadenylierung der cDNA-Sequenzen

Die Erkennung von Poly-(A)-Ketten, die durch Polyadenylierung entstanden sind, wird in zwei Schritten durchgeführt. Zuerst wird nach Sequenzen gesucht, die am 3'-Ende mindestens 10 Adenosine besitzen. Diese werden dann auf die 10 häufigsten Polyadenylierungssignale (Tian *et al.*, 2005) überprüft. Wird ein solches Signal in einem höchstens 50 Nukleotide messenden Abstand vom Beginn der Adenosine gefunden, werden diese von der cDNA abgetrennt.

3.2 Positionierung von Expressed Sequence Tags

"Expressed sequence tags" (ESTs) sind kurze Teilsequenzen transkribierter und prozessierter (Spleißen, Capping etc.) Nukleotidsequenzen (Protein-codierend oder nicht). Die Sequenzen werden durch „one-shot"-Sequenzierung einer cDNA einer Klon-Bibliothek erzeugt. Diese sind häufig von relativ geringer Qualität und mit heutiger Technologie auf 500 – 1000 Nukleotide begrenzt. Im September 2008 enthielt dbEST mehr als 50 Millionen Einträge, von denen etwa 8 Millionen auf den Menschen entfallen.

Zur Positionierung der in dbEST enthaltenen Sequenzen sind auf Grund der großen Datenmenge Optimierungen der Rechenprozesse zur Laufzeitverminderung unerlässlich. Weiterhin werden Sequenzen die weniger als 200 Basen lang sind aus dem Datensatz entfernt.

Um die Rechenkapazitäten optimal ausnutzen zu können, werden die zu positionierenden Sequenzen in mehrere Teilmengen aufgeteilt. Die Anzahl der Sequenzen in diesen Teilmengen kann vom Benutzer eingestellt werden, sollte allerdings nicht zu klein gewählt werden, da sonst die Vorbereitungsphase auf die Positionierung im Verhältnis zum eigentlichen Positionierungsvorgang zu zeitintensiv wird. Die Sequenzen einer Teilmenge werden jeweils als eigenständiger

Rechenprozess auf jedem Chromosom positioniert. Der Rechenaufwand lässt sich erheblich verringern, wenn bei solchen Änderungen bereits positionierte EST nur mit jenem Chromosom aligniert werden, auf dem sie sich vor der Veränderung des Assemblies befanden. Können diese Sequenzen mit hoher Signifikanz wieder auf dasselbe Chromosom an derselben Stelle positioniert werden, ist keine weitere Berechnung erforderlich. Damit Insertionen und Deletionen innerhalb des aktualisierten Assemblies nicht dazu führen, dass ESTs nicht mehr positioniert werden können, ist es möglich, die erlaubte Differenz zwischen alter und neuer Position anzugeben (je neuer das Assembly, desto größer). Alle ESTs, denen dennoch keine Position zugewiesen werden konnte, werden zusammen mit den neu zum Datensatz hinzugekommenen Sequenzen auf dem gesamten Genom positioniert. Der Rechenaufwand ist in diesem Fall nur für die erste Positionierung aller ESTs auf dem gesamten Genom hoch.

Eine anderer Ansatz, der weitere Zeit- und Rechenersparnis bedeutet, ist die Alignierung aller ESTs mit allen zur Verfügung stehenden mRNAs desselben Organismus, die schon positioniert wurden (Kapitel 3.1), und deren Koordinaten bekannt sind. Kann eine EST-Sequenz signifikant mit einer Referenzsequenz aligniert werden, übernimmt man die Start- und End-Koordinaten der EST-Sequenz aus dem Alignment und rechnet diese mit Hilfe der Koordinaten der Referenzsequenz auf das Chromosom zurück. Dies ist möglich, da mRNAs und ESTs beide keine Introns mehr enthalten. Um im Falle von Spleißvarianten die Intron/Exon Struktur der ESTs nicht zu verlieren, ist eine hohe Sequenzidentität und lückenlose Alignments zwischen EST und Referenzsequenz erforderlich. ESTs, die auf diese Weise keine Koordinaten zugeordnet bekommen, werden auf das Genom positioniert. Dieser Ansatz wurde schon in einer Kooperation und einem Webserver zur Annotation der Funktion von ESTs erfolgreich angewendet (Datson *et al.*, 2007; Waegele *et al.*, 2008).

Welche Methode eine kürzere Laufzeit verspricht hängt von der Qualität des Referenzdatensatzes ab. Enthält dieser nicht genügend Sequenzen, dann müssen mehr ESTs weiterhin auf das gesamte Genom positioniert werden. Optimale Laufzeiten und Resultate liefert eine Kombination aus beiden oben beschriebenen Ansätzen, welche weniger von der Qualität des Referenzdatensatzes abhängt. Zuerst werden die zu positionierenden ESTs mit dem Referenzdatensatz abge-

glichen. Alle ESTs, die keiner Referenzsequenz zugeordnet werden konnten, (hohe Sequenzidentität empfohlen) werden dann nur auf jeweils ein Chromosom positioniert (s. o.). Alle ESTs, die bis dahin noch keine Positionierung erhalten haben, werden zusammen mit neuen ESTs auf dem Genom positioniert.

3.3 Positionierung von Proteinsequenzen

Proteine sind Makromoleküle, die aus Aminosäureketten bestehen. Im Gegensatz zu mRNA bzw. cDNA werden mit dem Positionieren der Proteinsequenzen auf dem Genom keine UTRs (5' und 3') abgedeckt. Bei der Positionierung von neuen mRNAs lassen sich so bei Überlappung mit einem korrespondierenden Protein direkt die UTRs ablesen.

Für die Positionierung der Proteine eines Organismus stehen UniProtKB und die PDB zur Verfügung (siehe Kapitel 2.1.2 und 2.2.4). Jeder Eintrag in Swiss-Prot oder TrEMBL steht für eine einzelne Aminosäurekette. Bei Proteinen, die erst durch posttranslationale Modifikation aktiviert werden, wird die Vorläufersequenz verwendet. In PDB entspricht ein Eintrag einem Protein(-Komplex). Deshalb werden für jeden Eintrag der PDB-Datenbank die Ketten voneinander getrennt und einzeln positioniert. Für beide Datenbanken ist es möglich eine Mindestlänge für die zu positionierenden Sequenzen anzugeben. Voreingestellt ist eine Länge von mindestens 100 Aminosäuren.

Die PDB zeichnet sich als Datenbank dadurch aus, dass die Struktur enthaltener Proteine mit Röntgenkristallographie, NMR oder Elektronenmikroskopie bestimmt werden. Die so gewonnenen Daten liefern Informationen über den dreidimensionalen Aufbau und das Zusammenwirken der einzelnen Proteinketten in Proteinkomplexen. UniProtKB dagegen stellt die Aminosäuresequenzen der Proteine zur Verfügung. Swiss-Prot enthält manuell annotierte Proteine, TrEMBL automatisch generierte Aminosäuresequenzen basierend auf Transkripten oder Teiltranskripten[1]. Die Swiss-Prot-Annotationspipeline beinhaltet die Annotation von TrEMBL-Einträgen und deren Integration in Swiss-Prot[2].

[1] Diese sind durch Translation der partiellen CDS der in GenBank enthaltenen Nukleotidsequenzen entstanden.
[2] http://education.expasy.org/cours/Document/UniProtKB_Quickguid.pdf Stand September 2008

3.4 Positionierung von pre-microRNAs (mirBase) auf dem Genom

microRNAs sind kurze einzelsträngige RNAs mit etwa 22 Nukleotiden Länge. Die in mirBase enthaltenen Einträge beschreiben die pre-miRNA, eine durch Prozessierung der pri-miRNA (primäres Transkript des entsprechenden Genes) verkürzte Sequenz von ungefähr 70 Nukleotiden (pre-miRNA). Diese Sequenzen, die nicht der gesamten Sequenz eines Genes entsprechen, werden auf dem Genom positioniert. Dazu werden alle experimentell verifizierten microRNAs aus der mirBase Datenbank (Kapitel 2.2.2) ausgewählt und für den Positionierungsprozess in Flat-Files abgespeichert. Da die pre-miRNA Uracil enthält, werden diese in Thymin zurückübersetzt, so dass die RNA einer cDNA entspricht. Die Positionierung auf ein bestimmtes Genom wird mit allen experimentell verifizierten pre-miRNA aus mirBase durchgeführt. Mit diesem Ansatz können potentielle microRNA-Gene in einem Organismus bestimmt werden, die bereits in einem anderen Organismus experimentell bestätigt wurden.
In dieser Spezies-übergreifenden Positionierung kommt es vor allem darauf an, möglichst viele pre-miRNA auf entsprechende „Gene“ zu alignieren. Dazu werden die Sequenzen in drei Leserahmen in Aminosäuresequenzen übersetzt, das Genom in sechs Leserahmen. microRNA-Gene sind nicht Protein-codierend, dennoch erlaubt diese Art der Alignmentbildung eine höhere Sensitivität, da die Bildung der “initialen seeds“[1] für ein Alignment weniger strikt ist. Ausgleichend kann, falls striktere Alignmentbildung gewünscht ist, die minimale Sequenzidentität, die zwischen den alignierten Sequenzen bestehen soll, nach Bedarf eingestellt werden.
Da die Positionierung kurzer Sequenzen nicht sehr laufzeitintensiv ist, werden die Berechnungen auf alle Chromosomen eines Organismus sequentiell durchgeführt. Um die gewünschten Positionierungen aller microRNAs auf alle ausgewählten Genome zu erhalten, wird der Prozess für jedes Genom wiederholt. Zur optimalen Ausnutzung der zur Verfügung stehenden Rechenkapazität werden diese Prozesse parallel ausgeführt. Die wählbaren Genome können beliebig erweitert werden.

[1] Bei der Alignmentbildung mit Programmen wie FASTA (Pearson *et al.*, 1988; Pearson, 1990), Blast (Altschul *et al.*, 1990) oder Blat (Kent, 2002), werden zuerst lokale Alignments mit Teilsequenzen gebildet. Die besten (initial seeds) werden dann mit weiteren Teilalignments in beide Richtungen zu einem lokalen Alignment erweitert.

3.5 Positionierung bekannter und Identifizierung neuer potentieller Pseudogene

Pseudogene werden als Sequenzen genomischer DNA definiert, deren ursprüngliche Sequenz einem funktionsfähigen Gen zugeordnet werden kann, die aber Degenerationsmerkmale wie Frameshifts[1] und vorzeitige Stop-Codons aufweisen, die eine korrekte Expression verhindern (Balakirev *et al.*, 2003).

Es gibt - je nach Genese - drei verschiedene Pseudogen-Typen.

Prozessierte Pseudogene entstehen durch Retrotransposition, wenn mRNA revers transkribiert und die daraus entstandene cDNA wieder in das Genom eingebaut wird. Diese Art der Pseudogene entstammt nur aus den Exonbereichen eines transkribierten Gens. **Nicht-prozessierte** (duplizierte) **Pseudogene** entstehen durch Duplikation von Chromosomenabschnitten[2]. Der Verlust der Funktionalität eines Genes auf diesen Abschnitten kann durch fehlerhafte Duplikation oder Ansammlung von Mutationen[3], die die Sequenz verändern, eintreten.

Eine weitere Art von Pseudogenen sind Gene, die im Laufe der Zeit Mutationen angesammelt haben, so dass diese nicht mehr transkribiert oder translatiert werden können und damit keine Proteine mehr codieren können.

Neuere Erkenntnisse über Pseudogene zeigen, dass in einigen Fällen der Selektionsdruck, der auf die betroffene Sequenz einwirkt, genauso hoch ist wie bei codierenden Chromosomenabschnitten (Protein-codierenden Gene, tRNA und rRNA Gene, etc.). Dies bedeutet, dass selbst Pseudogene Funktionen haben können. Beispielsweise werden Pseudogene transkribiert und beeinflussen über ihre mRNA die Transkription oder Translation Protein-codierender Gene. Es wird angenommen, dass sich diese regulierenden Eigenschaften erst im Laufe der Evolution entwickelt haben (Gerstein *et al.*, 2006).

Für die Positionierung von Pseudogenen auf aktuellen Assemblies verschiedener Organismen gibt es zwei Varianten: die **Neupositionierung** der in

[1] Frameshifts sind Verschiebungen des ursprünglichen Leserahmens um ein oder zwei Basen. Bei der Translation entsteht nach einem Frameshift eine stark veränderte Proteinsequenz.

[2] Genduplikationen entstehen durch homologe Rekombination („crossover“ während der Meiose), Retrotransposition oder Duplikation ganzer Chromosomen.

[3] Durch Mutation oder Verlust einzelner Basen können Stop-Codons in einem offenen Leserahmen entstehen, so dass die Transkription eines Genes frühzeitig abbricht. Frameshifts führen zu funktionell eingeschränkten oder funktionslosen Proteinen. Mutationen im Promotor eines Genes können die Transkription eines Genes stilllegen.

Pseudogene.org beschriebenen Einträge (Karro *et al.*, 2007) oder die **Neuerkennung** von Pseudogenen (prozessierte und duplizierte Pseudogene) unter Verwendung der in Kapitel 3.1 erzeugten Positionierung der mRNAs.

3.5.1 De novo Positionierung der in Pseudogene.org enthaltenen Sequenzen

Die Datenbank Pseudogene.org enthält Pseudogene für verschiedenste Organismen. Für unser Projekt relevante Organismen sind die Mammalia. Zur Verfügung stehen *H. sapiens*, *P. troglodytes*, *M. musculus*, *R. norvegicus*, *C. familiaris* (Mensch, Schimpanse, Maus, Ratte, Hund). Da die Koordinaten der Pseudogene der Datenbank Pseudogene.org nicht für jedes Assembly berechnet werden, müssen die alten Koordinaten aktualisiert werden. Dazu geht man wie folgt vor:

Download der benötigten Daten: Die in Pseudogene.org enthaltenen Koordinaten zu einer bestimmten Version eines Assemblies stehen als Flat-File zum Download zur Verfügung (http://tables.pseudogene.org/flatfiles). Die dazu gehörenden Assemblies werden aus den Archiven von UCSC (Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway) (ftp://ftp.ncbi.nih.gov/genomes/<*Organismus*>/ARCHIVE/) bezogen.

Positionierung: Für die Positionierung der Pseudogene werden anhand der gegebenen Koordinaten die entsprechenden Sequenzen aus dem zugehörigen Assembly extrahiert. Sind die Koordinaten für den negativen Strang eines Chromosoms angegeben, werden diese Sequenzen in die revers komplementären Sequenzen übersetzt. Danach werden die Pseudogensequenzen auf das aktuelle Assembly positioniert. Jede Sequenz wird dabei nur auf dasjenige Chromosom positioniert, auf welchem diese auf dem alten Assembly lokalisiert war. Sequenzen, die selbst bei geringen Änderungen der Chromosomen-Sequenzen des neuen Assemblies nicht mehr signifikant positioniert werden können, werden aus dem Datensatz entfernt[1].

Überprüfung der neuen Positionierungen: Pseudogensequenzen, die positioniert werden können, werden auf die Qualität und Plausibilität der neuen Positionen

[1] Seit der Veröffentlichung von Pseudogene.org im Jahr 2007 (Karro *et al.*, 2007) wurde beispielsweise die Anzahl der humanen Pseudogene von 31768 auf 22645 und die der Maus von 15320 auf 15064 reduziert (Stand September 2008).

überprüft. Wichtig ist eine hohe Sequenzähnlichkeit – je weniger Veränderungen zwischen Assemblies auftreten, desto höher – und Alignments, die nur wenige und/oder kurze Lücken aufweisen (aus Pseudogene.org extrahierte Pseudogensequenzen sind ein genaues Abbild des Chromosomenabschnitts des veralteten Assemblies). Alignments, die nicht mindestens 90% der Eingabesequenz enthalten oder deren Länge größer als das 1,5-fachen der Eingabesequenz beträgt, werden gelöscht, genauso wie Alignments deren Sequenzidentitäten unter einem definierten Grenzwert liegen. Gibt es zu einer Sequenz mehrere mögliche Positionen, wird automatisch diejenige ausgewählt, die die höchste Sequenzidentität aufweist.

3.5.2 Identifizierung neuer potentieller Pseudogene

Mit Hilfe der in Kapitel 3.1 und 3.3 erzeugten Positionierungen der Genmodelle und Proteine können unabhängig von gegebenen Pseudogensequenzen Analysen zur Erkennung von Pseudogenen durchgeführt werden.

Prozessierte Pseudogene enthalten im Gegensatz zu einem großen Anteil der Protein-codierenden Gene keine Intronsequenzen mehr. Ist ein berechnetes Alignment nicht mindestens um 50% länger als das Transkript (mRNA nach Spleißen; Abb. 6) und unterschreitet die Sequenzidentität nicht einen festgelegten Grenzwert, wird das erzeugte Gen-Modell als potentielles Pseudogen zur Überprüfung in eine weitere Tabelle kopiert. Im Weiteren muss festgestellt werden, ob es sich bei den so gefilterten Modellen um falsch-positive Einträge handelt. Solche Einträge entstehen bevorzugt im Zusammenhang mit Ein-Exon-Genen, deren Alignment die gleiche Länge wie die mRNA besitzt. Gibt es zu einem potentiellen Pseudogen keine weiteren Genmodelle oder nur solche, die aus einem Exon bestehen, muss von einem Ein-Exon-Gen ausgegangen werden. Existiert dagegen mindestens ein Modell, das aus mehreren Exons besteht, wird weiterhin von einem potentiellen Pseudogen gesprochen.

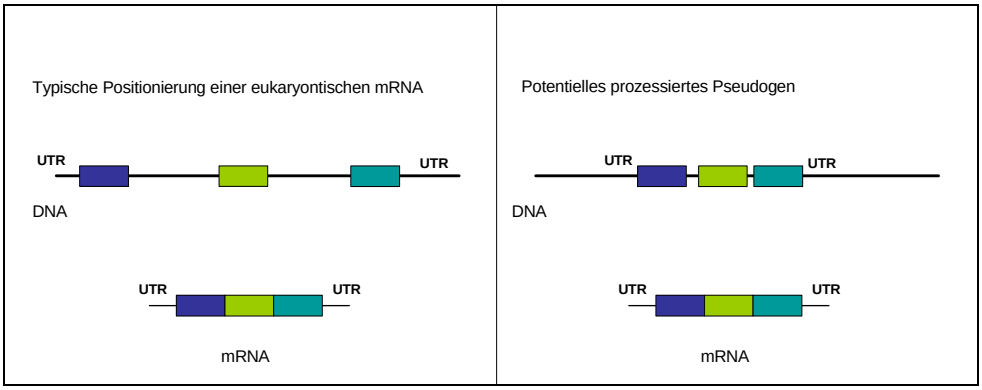


Abb. 6: Struktur eines typischen eukaryontischen Gens mit mehreren Exons; daneben die eines potentiellen Pseudogens. Der Längenvergleich verdeutlicht die Unterschiede der Längen, die eine Positionierung von einem Gen oder Pseudogen auf dem Genom einnimmt.

Identifizierung potentieller duplizierter Pseudogene

Die Erkennung von potentiellen Pseudogenen, die nach Duplikation von Genen oder ganzen Chromosomenabschnitten entstanden sind, erfolgt nach der Eliminierung der Alignments der prozessierten Pseudogene aus dem Genmodelldatensatz in zwei Analyseschritten. Es ist wichtig, dass die Positionierung der potentiellen Pseudogene nicht mit einer Positionierung eines Protein-codierenden Genes überlagert wird. Als Überlagerung gilt jede Überschneidung der beiden Positionierungen auf demselben Strang. So wird verhindert, dass ein potentielles Pseudogen auf einem genetisch „aktiven“ Chromosomenabschnitt zu liegen kommt.

Zuerst werden für jede mRNA, für die mehrere Genmodelle berechnet worden sind, die zum besten Alignment äquivalenten Genmodelle (geringe Abweichung in der Exonanzahl und Länge des Alignments) ausgewählt. Diese werden in einer weiteren Tabelle zwischengespeichert. Nur wenn bei der Überprüfung der zugehörigen Koordinaten keine Überschneidungen mit den Transkripten validierter Einträge von RefSeq bestehen, werden diese Modelle als potentielle Pseudogene abgespeichert.

3.6 Bewertung der Positionierungen

Nach dem Positionieren der einzelnen Sequenzen werden alle Positionierungen auf Plausibilität und Signifikanz getestet. Die Auswahl der besten Positionierungen ist abhängig vom jeweiligen Sequenztyp. Die Sequenzidentität, die für ein signifikantes Alignment gefordert wird, ist parametrisiert.

cDNA/mRNA: Im Falle der cDNA wird dasjenige Alignment behalten, das die höchste Sequenzidentität aufweist. Ist der Unterschied zwischen den besten und zweitbesten Alignments kleiner als ein vorgegebener Grenzwert, verbleiben diese Alignments in der Datenbank. Dieser Schritt kann nur dann korrekt durchgeführt werden, wenn die Pseudogenanalyse für die prozessierten Pseudogene durchgeführt wurde, damit keine potentiellen prozessierten Pseudogene an dieser Auswahl beteiligt werden.

Pseudogene: Bei der Auswahl der signifikanten Alignments wird unterschieden zwischen prozessierten und duplizierten Pseudogenen. Potentielle prozessierte Pseudogene verbleiben alle in der Datenbank. Duplizierte dagegen werden nur weiterverwendet, wenn ihr Alignment nicht mit dem eines Protein-codierenden Genes überlappt.

ESTs: ESTs sind kurze Sequenzen, deren Basenabfolgen mehrmals im Genom auftreten können. Deshalb werden zu jedem EST, wenn nötig mehrere Positionierungen beibehalten. Die Positionierung mit der höchsten Sequenzidentität (mindestens so hoch wie die geringste Sequenzidentität, die durch die Positionierungen aus Kapitel 3.2 festgelegt ist) wird in jedem Fall behalten, weitere Alignments werden abhängig von der jeweiligen erzielten Sequenzidentität und der Abweichung vom besten Treffer beibehalten oder verworfen (Abb. 7).

Proteinsequenzen: Proteinsequenzen, egal ob aus UniProtKB oder PDB, werden behandelt wie im Abschnitt cDNA/mRNA beschrieben. Auch hier muss vorher die Analyse auf prozessierte Pseudogene abgeschlossen worden sein.

microRNAs: Bei den microRNAs werden für jede Sequenz die besten drei Positionierungen beibehalten. Die Alignments dieser Positionierungen dürfen keine Lücken aufweisen, da diese Sequenzen keinem Spleißvorgang unterzogen werden. Alignments mit Lücken sind nicht korrekt. Die zweit- und drittbeste Positionierung darf des Weiteren nicht weniger als 90% der Sequenzidentität des besten Treffers aufweisen. Bei Organismus-übergreifenden Positionierungen von microRNAs wird nur der beste Hit erhalten.

1.Fall

hit_id	Name	Konfidenz
1	A123456	70
2	A123456	68
3	A123456	65
4	A123456	60
5	A123456	55
6	A123456	55
7	A123456	55
8	A123456	50
9	A123456	50
10	A123456	50

2.Fall

hit_id	Name	Konfidenz
1	A123456	100
2	A123456	100
3	A123456	99
4	A123456	99
5	A123456	98
6	A123456	97
7	A123456	95
8	A123456	95
9	A123456	90
10	A123456	90

Abb. 7: Beispiel für die Arbeitsweise des Entscheidungsalgorithmus zur Identifizierung der besten Positionierungen pro EST: Positionierung mit minimaler Sequenzidentität (SI) von 50%, erlaubte Abweichung zum besten Hit: 5% der SI des besten Hits. Minimale erlaubte Sequenzidentität für zweit-, drittbesten Hit etc.: 92%. Farbig unterlegt sind diejenigen Hits, die die Kriterien erfüllen.

3.7 Visualisierung der erzeugten Positionierungen

Zur Visualisierung der in den Kapiteln 3.1 bis 3.5 berechneten Positionierungen der ESTs, Proteine und mRNA (Gene und Pseudogene) wird der Generic Genome Browser (GBrowse) verwendet (Stein *et al.*, 2002; Donlin, 2007). GBrowse ist eine generische Web-basierte Applikation, die es dem Benutzer erlaubt, Regionen auf dem Genom zu betrachten oder Suchanfragen mit Gen- oder Proteinnamen oder Basenabfolgen zu stellen. Der Vorteil von GBrowse ist der schnelle Zugriff auf weitere zu den entsprechenden Sequenzen gehörenden Informationen. Diese können Gen- und Proteinnamen, die Sequenzidentität zur genomischen DNA, aber auch Verlinkungen auf die Ursprungsdatenbanken enthalten. Ein weiterer Vorteil ist die einfache Ankopplung weiterer Analyseprogramme wie Blast, Blat, SIMAP (siehe Kapitel 3.10) oder CRONOS (siehe Kapitel 6). Zusätzlich können weitere Datensätze - neben den oben beschriebenen - hochgeladen und angezeigt werden. Voraussetzungen zur Verwendung von GBrowse sind ein ApacheTomcat Server und eine MySQL-Datenbank.

Der UCSC Genome Browser (Zweig *et al.*, 2008) könnte ebenfalls zur Visualisierung verwendet werden. Es gibt aber in Bezug auf die speziellen Anforderungen Nachteile. Im Besonderen ist bei der Verwendung des UCSC Genome Browsers die integrierte Anzeige zur Charakterisierung der Sequenz wie der Sequenzidentität und assoziierten Gen- und Proteinnamen nicht möglich. Aus

den oben genannten Gründen wird der UCSC Genome Browser nicht zur Anzeige der Positionierungen verwendet. Nachfolgend wird deshalb nur die Erzeugung der Tracks für GBrowse erläutert.

Umformatierung der Koordinaten in das GFF3-Format

GBrowse stellt zum effizienten Datenimport der zu visualisierenden Sequenzdaten in die MySQL-Datenbank ein Uploadscript bereit. Dieses Script liest Dateien im "Generic Feature Format" (GFF – Format Version 3)[1] ein.
Da die oben berechneten Koordinaten der verschiedenen Sequenztypen in der Datenbank in Anlehnung an das PSL-Format (Blat) abgespeichert sind, werden diese nun in das benötigte GFF3-Format umgeschrieben. GBrowse unterstützt verschiedene Arten der Sequenzdarstellung genauso wie die Möglichkeit einer Sequenz weitere Informationen zuzuweisen. Jeder Sequenz werden so die Sequenzidentität und ihre Identifier aus den Ursprungdatenbanken mitgegeben. Später können diese dann mit Hilfe der GBrowse-Konfigurationsdatei zu den entsprechenden Datenbankeinträgen verknüpft werden. Nach Möglichkeit werden jeder Sequenz Gennamen und Proteinnamen sowie deren Synonyme zugewiesen. Mit Hilfe der Konfiguration lassen sich die einzelnen Sequenztypen mit Farbcodierungen versehen. Die zusätzlichen Informationen zu einer bestimmten Sequenz werden über GBrowse-Balloons[2] abrufbar gemacht.

mRNA (RefSeq): GBrowse bietet die Option UTRs sichtbar zu machen. Diese Eigenschaft wird für die Einträge aus RefSeq genutzt. Die Lage der CDS auf der mRNA und damit auch die Länge der 5'- und 3'-UTR sind in RefSeq (siehe Kapitel 2.1.1.1) enthalten, das heißt, diese muss auf die neu berechneten Koordinaten (inklusive Introns) des aktuellen Assemblys übertragen werden (Abb. 8). Würden die Start- und Stopp-Positionen der CDS eines Eintrags aus RefSeq nicht mehr zur Verfügung stehen, könnten diese mit Hilfe der Proteinpositionierungen berechnet werden.

Pseudogene: Insgesamt gibt es fünf Pseudogen-Tracks. Prozessierte und duplizierte Pseudogene werden jeweils für RefSeq und Swiss-Prot Einträge

[1] **G**eneric **F**eature **F**ormat. GFF ist ein zeilenbasiertes, Tab-separiertes Dateiformat zum Speichern von Sequenzeigenschaften und weiteren Erläuterungen zur Sequenz.
[2] Wird auf den Track einer Sequenz geklickt, öffnet sich ein Fenster mit weiteren Informationen.

berechnet und visualisiert. Ein weiterer Track ergibt sich aus den Alignments der von Gerstein et al. publizierten Pseudogene.

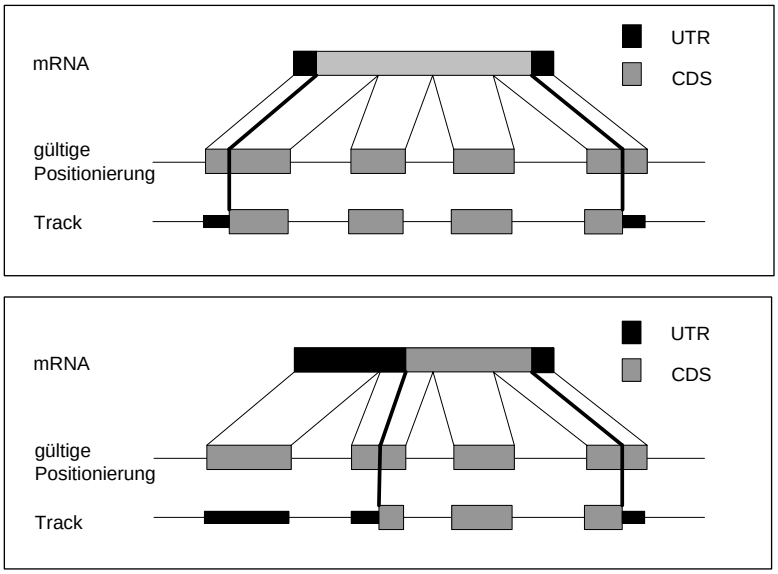


Abb. 8: Einfügen der UTR in die bereits berechneten Positionierungen auf dem Genom.

oben:
1. Fall: die UTRs befinden sich jeweils auf genau einem Exon.

unten:
2. Fall: eine der UTRs befindet sich auf mehreren Exons.
3. Fall: beide UTRs befinden sich jeweils auf mehreren Exons (ohne Abbildung).

ESTs: Auf Grund der Menge der EST-Daten werden zwei getrennte Tracks angelegt. Einer für Positionierungen auf dem Plus- und einer für den Minus-Strang. Außerdem werden mit einem eigenen Auswahlalgorithmus höchstens die besten zehn Positionierungen bestimmt, die dann in den entsprechenden Tracks zu sehen sind (Abb. 9).

hit_id	Konfidenz	
1	100	i=1; i<5
2	99	
3	99	i=3; i<5
4	95	
5	95	
6	95	
7	95	i=7; i>5
8	93	
9	93	
10	92	

hit_id	Konfidenz	
1	100	i=1; i<5
2	99	
3	99	
4	99	
5	99	
6	99	
7	99	
8	99	i=8; i>5
9	93	
10	92	

hit_id	Konfidenz	
1	100	
2	100	
3	100	
4	100	
5	100	
6	100	
7	100	
8	100	i=8; i>5
9	93	
10	92	

Abb. 9: Beispiel für die Auswahl der besten Positionierungen für den EST-Track: Es werden höchstens fünf Positionierungen ausgewählt. Positionierungen mit gleicher Konfidenz werden als Gruppe zur Auswahl hinzugefügt, solange der Grenzwert von fünf Positionierungen nicht überschritten wird (links und mitte). Ausnahme: Überschreitet die Gruppe der Positionierungen mit der besten Konfidenz den Grenzwert, werden diese dennoch zur Auswahl hinzugefügt (rechts).

Proteinsequenzen: Für die Positionierungen der Proteinsequenzen werden drei Tracks erzeugt. UniProtKB wird in zwei Tracks aufgeteilt – Swiss-Prot und TrEMBL. Der dritte Track enthält die Informationen zu den PDB-Einträgen. Mit Hilfe dieser Tracks können auch bei Fehlen der CDS – Informationen aus RefSeq die 5'- und 3'-UTRs erkannt werden.

microRNA: Für die microRNAs wird ein Track berechnet. Dabei kann bestimmt werden, ob alle zur Verfügung stehenden microRNAs einbezogen werden oder nur die bekannten microRNAs eines einzelnen Organismus.

3.8 Erzeugen der Konfidenz-Tracks

Konfidenz-Tracks sollen dazu dienen, schnell einen Überblick über den Erforschungsgrad und eine Abschätzung über die statistische Sicherheit exprimierter Chromosomenabschnitte eines Genoms zu geben. Es werden zwei verschiedene Arten von Konfidenz-Tracks erzeugt. Einer, der die Konfidenz aus höherwertigen Sequenzdaten (UniProtKB, RefSeq, PDB) extrahiert, der andere enthält die Konfidenz basierend auf den EST-Tracks. Dieser zweite Track kann auch dazu verwendet werden, um die Speicher- und Laufzeit-intensiven EST-Tracks zu substituieren. Konfidenz-Tracks werden jeweils für Plus- und Minusstrang separat erzeugt.

mRNA und Protein-Track: Dieser Track soll helfen, „aktive" Chromosomenabschnitte (Exons) nach Vorkommen in verschiedenen Datenbanken, Erforschungsgrad und Konfidenz einzuschätzen. Voraussetzung zur Erzeugung der Tracks sind die Positionierungen der entsprechenden Sequenzen aus den Kapiteln 3.1 und 3.3. Das Genom wird dazu abschnittsweise gescannt und jeder Base, die von diesen Sequenzen überlagert wird, ein Konfidenz-Wert zugewiesen. Dieser Wert ist additiv, das heißt, er setzt sich aus mehreren Werten zusammen, wenn eine Base von mehreren Sequenzen überlagert wird. Höhere Konfidenzwerte bedeuten hoher Erforschungsgrad/mehrfaches Vorkommen (siehe Abb. 10). Jedem der einzelnen Sequenztypen wurden empirisch bestätigte Werte zugewiesen, die die Konfidenz in die entsprechenden Typen widerspiegeln. In Tabelle 1 sind die einzelnen Sequenztypen mit der jeweiligen Gewichtung und Begründung zusammengestellt. Diesen Tracks sind die Farben von Hellblau über Dunkelblau bis Schwarz zugewiesen. Je dunkler die Farbe desto mehr

transkribierte und prozessierte Sequenzen befinden sich auf dem entsprechenden Chromosomenabschnitt.

EST-Track: Die beiden EST-Tracks dienen zur schnellen und übersichtlichen Anzeige der Bereiche des Genoms, auf denen ESTs positioniert werden konnten. Als zusätzlicher Parameter können Mindestlänge und maximale Länge der ESTs, die in die Tracks aufgenommen werden sollen, angegeben werden. Mit dieser Variante können die EST-Tracks auch in mehrere Tracks aufgeteilt werden, die jeweils ESTs selbst bestimmter Längenintervalle enthalten. Die Bewertung erfolgt wie zuvor beschrieben. Die Konfidenz-Werte entsprechen der absoluten Anzahl der EST-Sequenzen, die an einem Chromosomenabschnitt auftreten. Mit dem Konfidenz-Track für ESTs lässt sich somit die Gesamtheit der EST-Positionierungen in komprimierter Form darstellen. Werden die Einzelpositionierungen nicht benötigt, können die Berechnungen für die Tracks aus Kapitel 3.7 und der Import in die MySQL-Datenbank auf einen Bruchteil der Laufzeit verringert werden. Dies liegt vor allem daran, dass die Berechnung der Tracks für die ESTs entfällt und dadurch die Datenmenge für den Import auf 10% der ursprünglichen Datenmenge verringert wird.

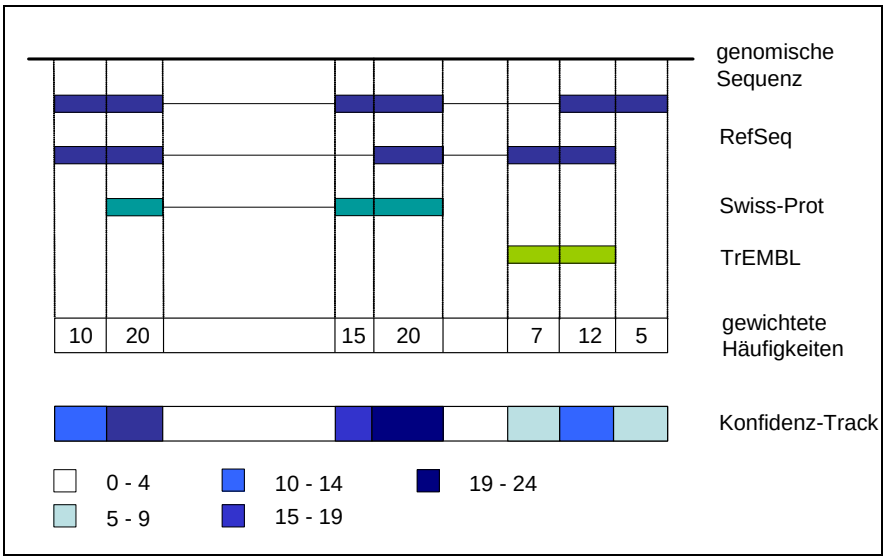


Abb. 10: Schematische Darstellung zur Erzeugung des mRNA/Protein Konfidenz-Track.

Sequenztyp	Gewichtung	Begründung
PDB	15	experimentell verifizierte Proteine
Swiss-Prot	10	Proteinsequenzen, die teilweise aus Genbanksequenzen translatiert wurden, kaum Validierung über kristallisiertes Protein
TrEMBL	2	vorläufige Proteinsequenzen, nicht validiert
RefSeq NM_	5	validierte Transkripte; der im Gegensatz zu Swiss-Prot niedrigere Wert berücksichtigt die Überrepresentation durch Splice Varianten
RefSeq XM_	2	Modelle aus Genvorhersagen

Tabelle 1: Gewichtungen der einzelnen Sequenztypen für die Erzeugung der Konfidenz-Tracks.

Optimierung zur Erzeugung der Tracks

Die Optimierung, die für eine effiziente Erzeugung der EST-Tracks nötig ist, findet auf zwei Ebenen statt. Zuerst werden die Tabellen der betroffenen Datenbanken mit Indizes versehen, die optimalen Zugriff auf die Daten erlauben. Die Software wurde in zweierlei Hinsicht optimiert: Laufzeit und Arbeitsspeicherverbrauch. Der Arbeitsspeicherbedarf lässt sich vermindern, indem man Chromosomen in Intervalle einer bestimmten Länge aufteilt. Diese werden dann nacheinander abgearbeitet. Nur diejenigen Daten, die sich im aktuellen Intervall befinden, werden in den Arbeitsspeicher geladen. Nach der Berechnung des Konfidenz-Tracks dieses Intervalls werden die Ergebnisse in eine Datei ausgelagert und der Speicher wieder freigegeben (siehe Abb. 11). Die Laufzeit bei der Erzeugung der Tracks ist größtenteils von der Geschwindigkeit der Datenbankanfragen abhängig. Diese lassen sich durch Anfrageoptimierung einerseits und Aufteilen der der Anfragen in mehrere Teilanfragen beschleunigen. Die Anfrageoptimierung hängt dabei vom SQL-Interpreter des Datenbankmanagementsystems ab. In diesem Fall war eine Abtrennung von Bedingungen, die die Ergebnismenge einschränken, in eine Unteranfrage die effizienteste Lösung. Weiterhin wird die Anfrage in vier Teilanfragen bearbeitet (entspricht Fällen 1-4 in Abb. 11) und die Intervalllänge auf 100 Mega-Basen festgelegt, was die Laufzeit, wie experimentell verifiziert, in Kombination mit den erzeugten Indizes nochmals beschleunigt.

3.9 Berechnung von Spleißvarianten

Aus Genen höherer Eukaryonten können durch den Prozess des alternativen Spleißens mehrere Proteinvarianten erzeugt werden. Bei diesem Vorgang werden verschiedene Exonkombinationen desselben Genes während der RNA-

Prozessierung zu einer reifen mRNA zusammengestellt. Alternatives Spleißen kann zu Unterschieden in den UTRs führen, genauso wie zu verschiedenen codierenden Exonsequenzen (Transkriptvarianten). Es bestimmt die intrazelluläre Lokalisation, enzymatische Aktivität, Proteinstabilität und posttranslationale Modifikationen vieler Proteine.

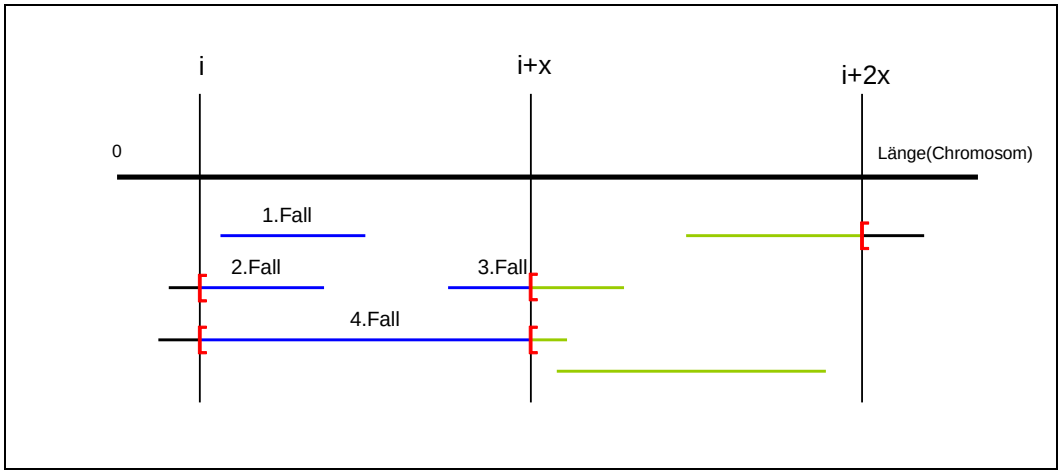


Abb. 11: „Intervall–Technik“ zur Minimierung des benötigten Arbeitsspeichers.
In einem Schritt werden alle Sequenzen bearbeitet, die ein Intervall der Länge x überschneiden. Berechnet werden die Tracks dann nur für diejenigen Bereiche der entsprechenden Sequenzen, die innerhalb des Intervalls [x, x+1[liegen. Diese Tracks werden in eine Datei geschrieben und der Arbeitsspeicher wieder freigegeben, bevor das nächste Intervall bearbeitet wird.

Diese unterscheiden sich in ihrer Funktionsweise durch Austausch, Verlust oder Gewinn von Proteindomänen (Stamm *et al.*, 2005). Es gibt verschiedene Arten des alternativen Spleißens (Abb. 12):

- Constitutive Splicing: Alle Introns eines Genes werden aus der pre-mRNA entfernt, alle Exons bleiben erhalten und bilden die reife mRNA.
- Exon Skipping: Beim Spleißen werden Exons mit den Introns entfernt.
- Intron retention: Introns verbleiben in der mRNA.
- Mutually exclusive Exons: Es werden Paare von Exons wechselnd aus der mRNA entfernt, so dass jeweils nur eins der beiden in der mRNA verbleibt.
- Alternative 5’/3’ Splice Sites: Beim Spleißen wird durch die Veränderung des 5’- oder 3’-Endes eines Exons, dieses verkürzt.
- Multiple Promoters: Durch das Spleißen entstehen mRNAs mit verschiedenen 5’-Enden.
- Multiple poly(A) sites: Das Spleißen variiert das Exon des 3’-Endes der pre-mRNA.

Jede dieser Spleißarten kann für sich, aber auch in Kombination mit anderen Arten auftreten. Dennoch haben alle Spleißvarianten mindestens ein gemeinsames Exon. In den Fällen von alternativen 5'- oder 3'-Splice Sites, besitzen Spleißvarianten jeweils mindestens ein Exon, in dem sie zu einem gewissen Anteil der Exonsequenz übereinstimmen.

Diese Beobachtung verwendet man nun zur automatischen Berechnung von Spleißvarianten. Voraussetzung ist ein Datensatz, bei dem die Positionierungen der einzelnen Sequenzen im PSL-Format (Blat) vorliegen (Kapitel 3.1), oder in einem anderen Format, das die Einzelpositionierung der Exons berücksichtigt. Die Berechnung erfolgt in drei Schritten:

- Einlesen des gesamten Datensatzes
- Bildung von Sequenzclustern, die in mindestens einem Exon überlappen. Hier kann der Grad der Überlappung (prozentualer Anteil der Gesamtlänge des Exons), die zur erfolgreichen Identifizierung von Varianten Voraussetzung sein soll, eingestellt werden.
- Verschmelzen von Clustern, die mindestens eine Sequenz miteinander gemein haben, zu einem Cluster.

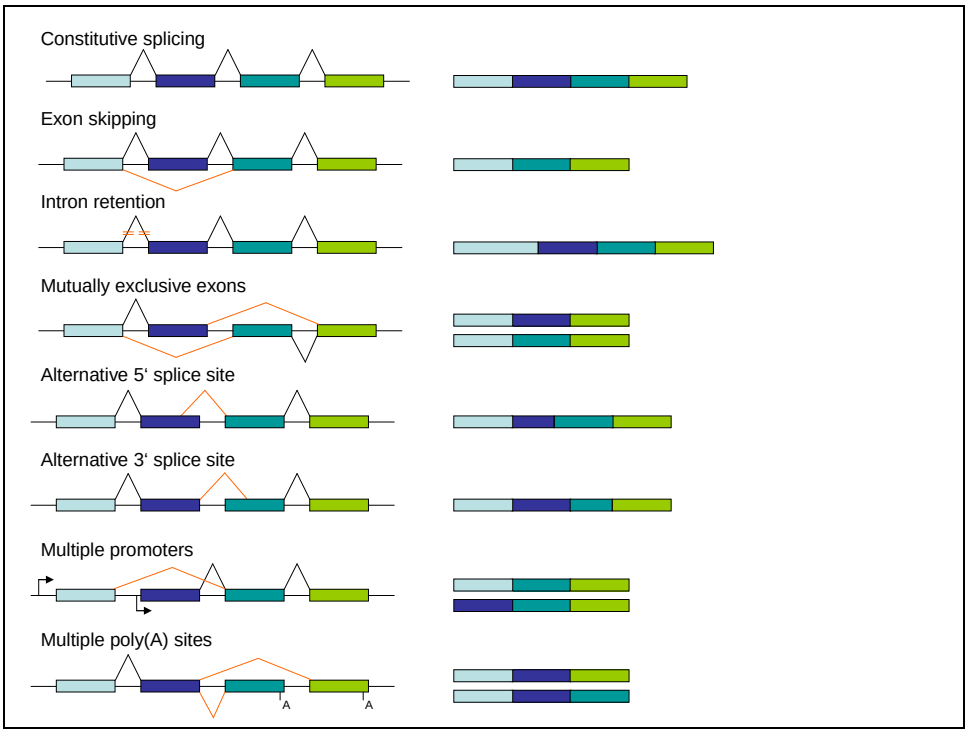


Abb. 12: Verschiedene Formen des alternativen Spleißens (rot) im Vergleich zum konstitutiven Spleißen (schwarz) (Matlin *et al.*, 2005)

Diese Methodik wurde bereits für die Berechnung der Spleißvarianten der “Mouse functional Genome Database“ MfunGD (Ruepp *et al.*, 2006) verwendet. Die berechneten Varianten wurden manuell überprüft und zu 100% in die Datenbank importiert.

3.10 SIMAP auf Nukleotidbasis

Einer der Hauptaspekte der Sequenzanalyse – sowohl auf genomischer als auch auf proteomischer Ebene – ist die Identifizierung ähnlicher Sequenzen. SIMAP (Similarity Matrix of Proteins) (Arnold *et al.*, 2005; Rattei *et al.*, 2006; Rattei *et al.*, 2008) wurde entwickelt, um den Rechenaufwand, der für wiederholte Berechnungen von Sequenzidentitäten und Sequenzalignments für Proteine entsteht, zu minimieren. SIMAP enthält für alle Proteine, die in die Datenbank integriert wurden, die vorberechnete Sequenzähnlichkeit (Vorberechnung erfolgt mit FASTA (Pearson, 2000), gefundene Hits werden mit Smith-Waterman-Algorithmus erneut berechnet) zu jedem anderen Protein in SIMAP. SIMAP lässt sich inkrementell erweitern, was den Rechenaufwand bei Aufnahme neuer Sequenzen in den Datensatz erheblich verringert.

Für den effizienten Zugriff auf ähnliche genomische Sequenzen wie homologe Gene beziehungsweise Pseudogene, wurde die Software, die für SIMAP entwickelt wurde, für Verwendung mit genomischen Sequenzen angepasst (im Folgenden Similarity Matrix of Genes – **SIMAG** - genannt). Anstelle der Aminosäuresequenzen der Proteine werden Sequenzen aus RefSeq verwendet, die Transkripten und Pseudogenen zugeordnet werden. Zur Berechnung der Sequenzähnlichkeit wird Blat anstelle von FASTA eingesetzt.

Im Folgenden werden evolutionäre Konzepte erläutert, die mit dem Einsatz von SIMAG untersucht werden können.

Homologe Gene sind Gene, die in ihrer Nukleotidsequenz eine hohe Übereinstimmung aufweisen und aus einer Ursequenz (oder Ur-Exons) hervorgegangen sind. Diese war, wie gefolgert wird, in einem Vorfahren vorhanden, von dem die verschiedenen Träger homologer Gene abstammen. Homologe Gene werden in orthologe und paraloge Gene unterteilt.

Orthologe Gene sind Gene, die in verschiedenen Organismen vorkommen. Die Gene stammen von einem gemeinsamen Urgen des letzten gemeinsamen Vorfahren ab (Last Common Ancestor, LCA).

Paraloge Gene sind eine Gruppe von Genen (Gen-Familie), welche in ein und demselben Organismus vorkommen. Sie werden als evolutionär entstandene Vervielfältigungen (Duplikationen[1]) eines einzelnen Ursprungsgens betrachtet, die sich anschließend getrennt voneinander weiterentwickeln. Klassische Beispiele für paraloge Gene sind olfaktorische Rezeptoren oder im Zellkern lokalisierte Hormonrezeptoren. Durch die voneinander unabhängige Weiterentwicklung der betroffenen Chromosomenabschnitte, gilt die Genduplikation als ein wesentlicher Mechanismus der Evolution. Durch Mutationen der Gen-Kopie können Gene mit neuen Funktionen aber auch **Pseudogene** entstehen[2].

Zur generischen Analyse verschiedener Sequenztypen wird neben SIMAG auch SIMAP an die in Kapitel 3.7 visualisierten Positionierungen von Proteinen und Transkripten mit Hilfe von GBrowse gekoppelt. Anfragen an SIMAP oder SIMAG können von dort direkt mit der verlinkten Sequenz aufgerufen werden. Diese beiden Analysevarianten erlauben einen umfassenden Einblick in die Relationen zwischen Sequenzen eines Organismus oder Organismus-übergreifend.

Für SIMAP und SIMAG finden sich verschiedene Anwendungen. Sollen auf der Ebene des Proteoms Analysen durchgeführt werden, wie zum Beispiel nach homologen Proteinen in verschiedenen Organismen im Allgemeinen oder nach so genannten “Clusters of Orthologous Groups of proteins“ – COGs - (Tatusov *et al.*, 2003), eignet sich SIMAP. Für die Analyse von Gen-Familien eines Organismus oder die Analyse von konservierten genomischen (Gen-) Sequenzen findet SIMAG Verwendung. In einigen Ausnahmefällen können nur entweder mit SIMAP oder SIMAG aussagekräftige Ergebnisse erzielt werden (Abb. 13). Wird beispielsweise nach einer Genduplikation in der Kopie der Leserahmen verändert und dadurch ein verändertes Protein erzeugt, kann SIMAP nur bedingt Ergebnisse liefern, SIMAG wird dennoch diese beiden Gene zusammenführen. Im Gegensatz dazu kann SIMAP ähnliche Proteine finden, auch wenn auf Ebene der

[1] **Duplikationen** bezeichnen in der Genetik allgemein Verdopplungen eines bestimmten Chromosomenabschnittes. (http://de.wikipedia.org/wiki/Genduplikation)
[2] http://www.zoolvet.unizh.ch/glossar/index.php?letter=*,
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.section.23

Nukleotidsequenzen erhebliche Unterschiede auftreten (synonyme Mutationen, die die Aminosäuresequenz nicht verändern). Orthologe Gene lassen sich mit SIMAG auch dann noch finden, wenn durch eine herbeigeführte Verschiebung des Leserahmens die Aminosäuresequenz verändert wird (Entstehung von Genen/Proteinen mit veränderter Funktion).

Mit der Integration der bereits bekannten Pseudogensequenzen zusätzlich zu den Protein-codierenden Transkripten in die Datenbank, können evolutionär bedingte Genverluste[1] erkannt werden.

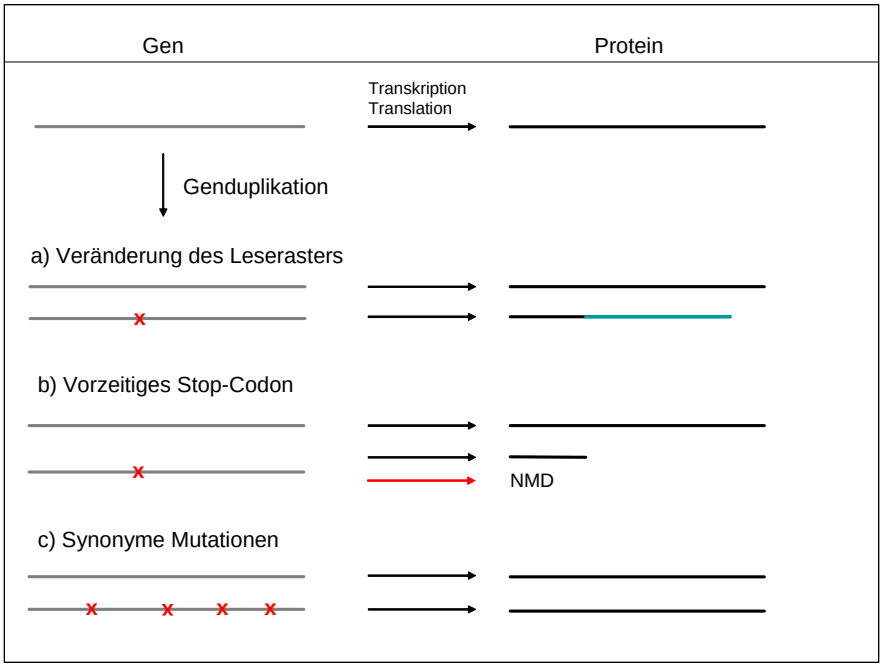


Abb. 13: spezifische Anwendungen von SIMAP und SIMAG: a und b sind Fälle in denen SIMAG im Gegensatz zu SIMAP umfassendere Ergebnisse liefert (NMD=nonsense-mediated decay (Maquat, 2002)); Bei Fall c verhält es sich umgekehrt.

Automatisierung des Aufbaus der SIMAG-Datenbank

Verwendete Daten

Für SIMAG werden Nukleotidsequenzen aus RefSeq (Kapitel 2.1.1) verwendet. In der ersten Version wird die Menge der Sequenzen auf solche eingeschränkt, die experimentell verifiziert worden sind. Diese umfassen Gene genauso wie Pseudogene. Momentan enthält die Datenbank Sequenzen der Organismen *H. sapiens*, *M. musculus, R. norvegicus, C. familiaris* und *B. taurus*. Weitere Organismen wie die Primaten werden bereits integriert. Zur Verfügung stehende zusätzliche Organismen aus der Domäne der Eukaryonten (zum Beispiel Schaf, Opossum und Hase; *O. aries*, *M. domestica*, *O. cuniculus*) werden für den Import

[1] Ein Genverlust tritt ein, wenn Mutationen funktionale Gene inaktivieren. Dieser Prozess wird auch als Pseudogenisierung bezeichnet (Gross; 2006).

vorbereitet. Geplant sind bereits die Integration von Daten ausgewählter Insekten und Pflanzen. Zum aktuellen Zeitpunkt enthält SIMAG bereits mehr als 75.000 verschiedene Sequenzen.

Aufbau der SIMAG-Datenbank

Zum Aufbau der Datenbank wird das zur Verfügung gestellte Java Archive (simapadmin.jar) verwendet. Um die Sequenzen aus RefSeq in SIMAG importieren zu können, werden diese in einem Vorverarbeitungsschritt als multi-FASTA Dateien abgespeichert. Folgende Schritte werden durchgeführt:

1. Anlegen der Datenbank
   - add_database/update_database: Mit diesem Modul werden die Sequenzen aus den multi-FASTA-Dateien in die Datenbank importiert. Werden Sequenzen eines neuen Organismus importiert, wird die “add_database“-Methode aufgerufen. Werden zu einem bereits in der Datenbank bestehenden Organismus neue Sequenzen importiert, wird die “update_database“-Funktion.
2. Berechnung der Sequenzähnlichkeiten
   - export_wokunits: Dieses Modul erzeugt für die neu importierten Sequenzen Dateien, mit denen dann die Berechnung der Sequenzidentitäten durchgeführt wird. Berechnungen werden “all-against-all“ für die neuen Sequenzen, und jeweils mit jeder der bereits in SIMAP enthaltenen Sequenzen angestellt.
   - Berechnung der Sequenzidentitäten (SI): Die SI werden mit Blat (Kent, 2002) berechnet. Die Größe des initialen Seeds wird auf acht Nukleotide eingestellt.
   - import_results: Die Ergebnisse aus dem vorangegangenen Schritt werden dann in die Datenbank importiert.
3. Abschließen des Imports und Exportieren der Daten in Flat-Files
   - finish_import: Abschließen des aktuellen Berechnungsvorganges und Vorbereitung der importierten Daten auf mögliche Updates.
   - export_release: Erstellen der Flat-Files und Indizes für einen schnellen Zugriff auf alle integrierten Datensätze.

Automatisierung und Optimierung

Zur leichteren Anwendung wurden die Schritte des vorherigen Abschnittes automatisiert. Die Entscheidung, welche Methode zum Datenbankaufbau verwendet wird, wird dem Nutzer von der Kontroll-Software abgenommen. Die Optimierung erfolgt durch Parallelisierung der drei Schritte aus Abschnitt zwei (betrifft den Export der Workunits, die Berechnung der Sequenzidentitäten und den Import der Ergebnisse in die MySQL-Datenbank). In Abbildung 14 ist die Vorgehensweise zum Erstellen oder Aktualisieren der SIMAG-Datenbank visualisiert. Horizontale Kästen werden sequentiell bearbeitet (gelb, orange, blau und grün). Jeder Folgeprozess hängt vollständig von seinem Vorgänger-Prozess ab, weshalb diese Schritte nicht parallelisiert werden können. Die Prozesse des dreigeteilten orangenen Kastens dagegen können (bedingt) parallelisiert werden (vertikale schwarze Balken). Die einzelnen Prozesse werden in kurzem zeitlichem Abstand gestartet. Der Export aller Workunits ist bei großen Datenmengen zeitintensiv. Deshalb wird noch während des Exports, der zu berechnenden Dateipaare (enthalten die Sequenzen, die miteinander aligniert werden), die Berechnung der Alignments für schon exportierte Dateipaare begonnen. Sobald die Ergebnisse vorliegen, werden sie ohne Verzögerung in die Datenbank importiert. Der Rechenaufwand der bei der Berechnung der Alignments entsteht, wird mit Hilfe der SGE (Sun Grid Engine; www.sun.com) auf mehrere Rechner verteilt, und dadurch die Laufzeit aller Rechenprozesse direkt proportional zur Anzahl der Rechner verringert.

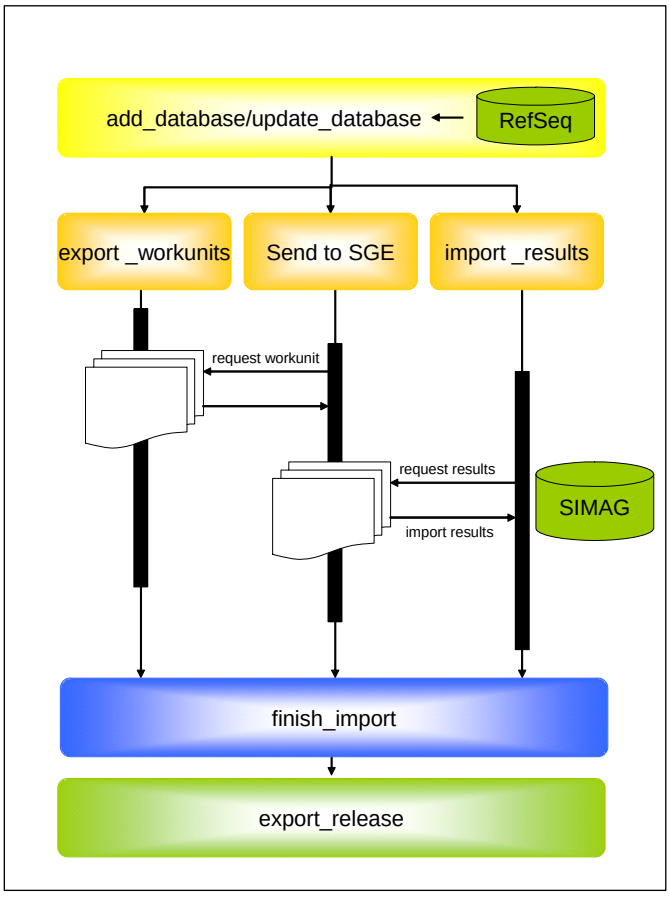


Abb. 14: Schematischer Ablauf des Aufbaus der SIMAG-Datenbank. Prozesse der gelben, orangen, blauen und grünen Kästen laufen sequentiell ab. Die Prozesse des dreigeteilten orangen Kastens werden parallel ausgeführt.

4 Erstellung des ersten *Callithrix jacchus* DNA-Microarrays

Nach dem erfolgreichen Abschluss der Entwicklung des im vorangegangenen Kapitel beschriebenen Softwarepaketes und dem daraus gewonnenen Know-how, ergab sich eine erfolgreiche Kooperation mit der “Division of Medical Pharmacology, Leiden/Amsterdam Center for Drug Research and Leiden University Medical Center, The Netherlands”. Zielsetzung dieser Kooperation war die Erstellung des ersten DNA-Microarrays mit Oligonucleotiden des Weißbüschelaffen (*Callithrix jacchus*) (EUropean MArmoset MicroArray, EUMAMA). Mir kam die Aufgabe zu, die bereits sequenzierten EST-Sequenzen mit Gennamen und offenen Leserahmen (open reading frames; ORFs) zu annotieren. Dazu wurde die bestehende Software aus Kapitel 3 verwendet und um projektspezifische Anforderungen erweitert.

4.1 Hintergrundinformation

Callithrix jacchus (Weißbüschelaffe) ist ein kleiner in Brasilien beheimateter Neuweltaffe, der immer öfter als Modellorganismus für die biomedizinische Forschung und die Entwicklung von Medikamenten zum Einsatz kommt. Nicht nur seine nahe genetische, physiologische und metabolische Verwandtschaft zum Menschen, sondern auch mehrere physiologische Unterschiede zu Altweltaffen erklären die Verwendung dieses Primaten als Modellorganismus. Im Gegensatz zu dem immer größer werdenden Interesse an *Callithrix jacchus* stehen Werkzeuge für genetische beziehungsweise Expressionsdatenanalysen kaum zur Verfügung.
Obwohl schon im Jahr 2004 ein Genomprojekt initiiert wurde, sind erst wenige mRNA Sequenzen für die Öffentlichkeit zugänglich. Um Genexpressionsanalysen und Analysen des Metabolismus durchführen zu können, sollte ein DNA-Microarray hergestellt werden. Dazu wurde mit “large-scale EST sequencing“ begonnen, um die benötigte Sequenzinformation zu generieren.

4.2 Problemstellung

Nachdem unser Kooperationspartner die Extraktion und Sequenzierung der “expressed sequence tags (ESTs) aus dem Hippocampus erfolgreich abgeschlossen hatte, bestand unsere Aufgabe darin, die so gewonnenen Sequenzen mit Gennamen und gegebenenfalls mit offenen Leserahmen (ORFs) zu annotieren.

4.3 Qualität der Sequenzen

Die ESTs werden mit der Didesoxymethode nach Sanger mit einer Leseweite (reads) von 800-850 Basen sequenziert. Für die Annotation werden 3441 qualitativ hochwertige Sequenzen ausgewählt, die mindestens den PhredScore Q15 über eine Länge von mehr als 600 Basen aufweisen können. Ein Q-Wert von 15 bedeutet dabei, dass die Wahrscheinlichkeit einer falsch sequenzierten Base bei 3% liegt. Bei 87,5% der 3441 Sequenzen liegt diese sogar nur bei 1% (PhredScore: Q20; der PhredScore wird nach der Formel $Q = -10\log_{10}(p)$ berechnet. P ist die geschätzte Fehlerwahrscheinlichkeit für eine Base (Ewing *et al.*, 1998)).

4.4 Positionierung der Sequenzen

Zur Annotation der *Callithrix jacchus* – Sequenzen bedienen wir uns der in Kapitel 3 beschriebenen Software zur Ähnlichkeitsuntersuchung zwischen Sequenzen. Bevor jedoch mit der Annotation der ESTs begonnen werden kann, müssen die Sequenzen vorbereitet (preprocessing) werden.

4.4.1 Vorbereiten der Sequenzen

4.4.1.1 Eliminieren identischer Sequenzen

Da bei Sequenzierungs-Projekten nicht vermieden werden kann, dass mehrere Sequenzen eines Datensatzes identisch sind, oder Bruchstücke einer längeren Sequenz neben dem Original koexistieren, werden redundante Sequenzen aus dem zu annotierenden Datensatz eliminiert. Zur Erkennung identischer DNA-Sequenzen wird die Blat-Software verwendet (Parameterwahl entsprechend der Web-Version der UCSC). Sequenzen, die in ihren überlappenden Bereichen in mindestens 95 Prozent ihrer Basen übereinstimmen, werden geclustert und nur die jeweils längste Sequenz verbleibt im Datensatz.

4.4.1.2 Trimmen und Orientierung

Sequenzen, die mit der Sanger-Methode erzeugt werden, weisen am 5'- und 3'-Ende erhöhte Fehlerraten auf. Daher werden vom 5'-Ende 20 Basen entfernt (Abb. 15). Um der mit der Länge kontinuierlich abnehmenden Sequenzqualität zu begegnen werden die Sequenzen, die eine Länge von 700 Basen überschreiten am 3'-Ende um 15% gekürzt, in jedem Fall aber auf maximal 700 Basen (Abb. 16).

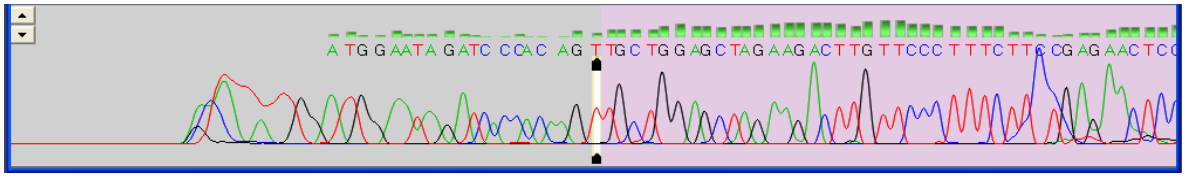


Abb. 15: Chromatogramm einer *Callithrix Jacchus* EST-Sequenz;
rechts von der Markierung Sequenzabschnitte mit höchster Sequenziergenauigkeit; Zu sehen ist das 5'-Ende der Sequenz.

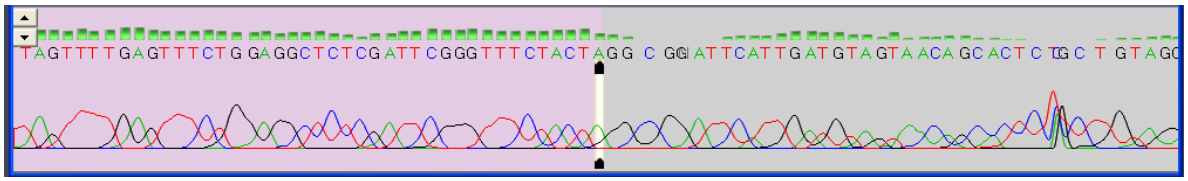


Abb. 16: Chromatogramm einer *Callithrix jacchus* EST-Sequenz;
links von der Markierung Sequenzabschnitte mit höchster Sequenziergenauigkeit; Zu sehen ist das 3'-Ende der Sequenz.

Nach diesem Trimmen haben die ESTs des Datensatzes eine durchschnittliche Länge von 678 Basen. Da ESTs auf dem komplementären Strang vom Gen-Ende aus sequenziert wurden (Abb. 17), müssen die sich daraus ergebenden Sequenzen in die revers komplementären Sequenzen zurückübersetzt werden (a→t, c→g, g→c, t→a).

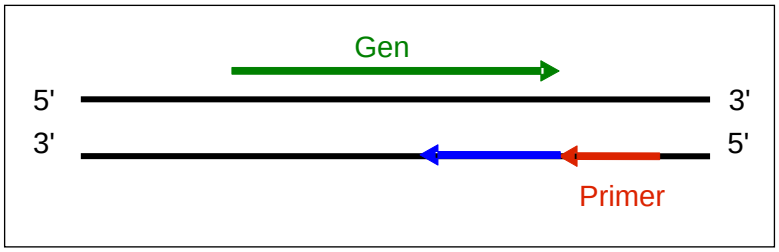


Abb. 17: Schematische Darstellung der EST-Sequenzierung

4.4.2 Positionierung der ESTs

Die Positionierung der EST-Sequenzen erfolgt mit der in Kapitel 3 beschriebenen Software. Verwendet wird der Abschnitt für die Positionierung von ESTs. In einem ersten Schritt werden alle ESTs auf einen Referenzdatensatz aus mRNAs positioniert. Im zweiten Schritt erfolgt die Positionierung derjenigen ESTs, die keiner Referenzsequenz zugeordnet werden konnten, auf das Genom. Die zusätzliche Positionierung der EST-Sequenzen auf die komplette genomische Sequenz ermöglicht durch die weniger stringenten Parameter eine höhere Sensitivität bei der Annotation der EST-Sequenzen.

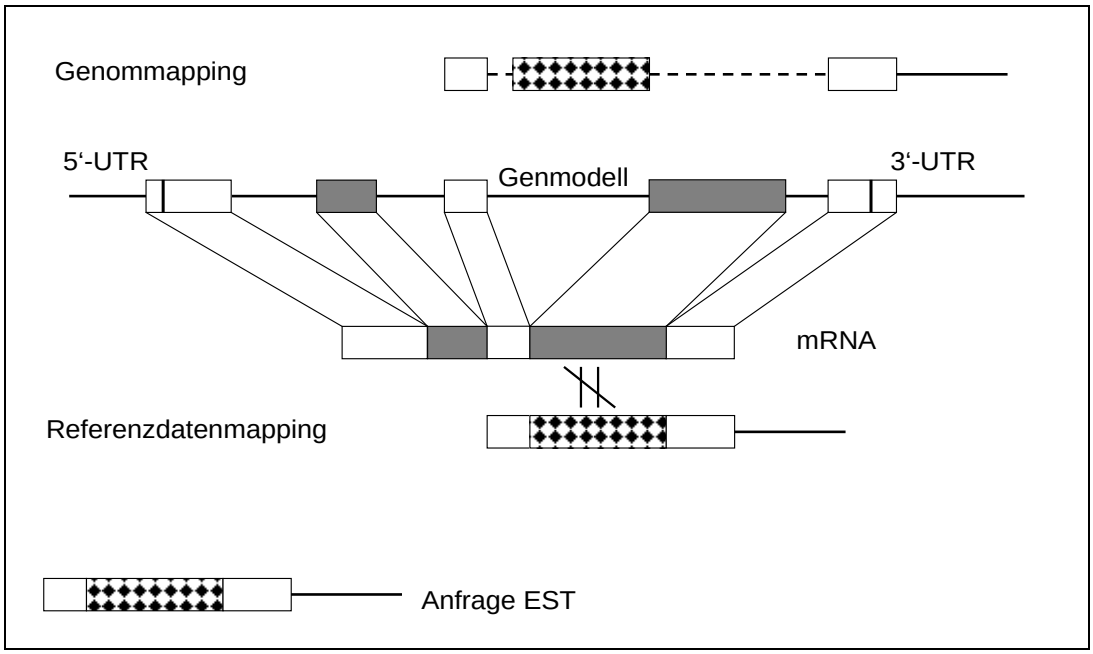


Abb. 18: Verdeutlichung des Nutzen der zusätzlichen Positionierung auf das Genom. Während beim Positionieren auf den Referenzdatensatz die Sequenzidentität durch ein vertauschtes Exon nicht erreicht werden kann, ermöglicht die Analyse der Exon-Überlappung mit Hilfe der Genompositionierung eine erfolgreiche Annotation.

Auch kann dadurch ein nicht vollständiger Referenzdatensatzes in Hinsicht auf Spleißvarianten ausgeglichen werden (Abb. 18). Als Annotationsgrundlage wird der RefSeq-Datensatz (Transkripte) und das humane Genom verwendet, weil die Referenzdatensätze von RefSeq am vollständigsten sind und kein anderer naher Verwandter des *Callithrix jacchus* so gut annotiert ist wie der Mensch.
Die Positionierung wird in zwei Phasen unterteilt: in eine Testphase zur empirischen Festlegung geeigneter Parameter und die eigentliche Positionierung der ESTs.

4.4.2.1 Testphase der EST-Positionierung

In einem ersten Test werden alle Sequenzen mit ausreichender Sequenzqualität auf Sequenzähnlichkeit mit Transkripten aus einem Referenzdatensatz untersucht. Dieser besteht aus allen aktuell in RefSeq aufgeführten humanen mRNA-Sequenzen. Manuelle Überprüfung der erfolgten Zuordung von EST zu Transkript ergibt optimale Ergebnisse bei einer Sequenzidentität von mindestens 50 Prozent (im Verhältnis zur untersuchten EST-Sequenz).
ESTs, die keine Ähnlichkeit mit einer Referenz-mRNA aufweisen, werden mit einem weniger stringenten Identitätswert von 30 Prozent auf die aktuelle Version des humanen Genoms positioniert. Trotz des geringen Grenzwertes konnten 60

ESTs nicht positioniert werden. Diese Sequenzen wurden nochmals mit Hilfe der Chromatogrammdaten manuell auf ihre Qualität hin inspiziert. Von den 60 ESTs müssen 19 ESTs auf Grund schlechter Qualität gelöscht (Abb. 19), 26 ESTs um 100 Basen gekürzt (Abb. 20) werden. 15 Sequenzen bleiben unverändert erhalten.

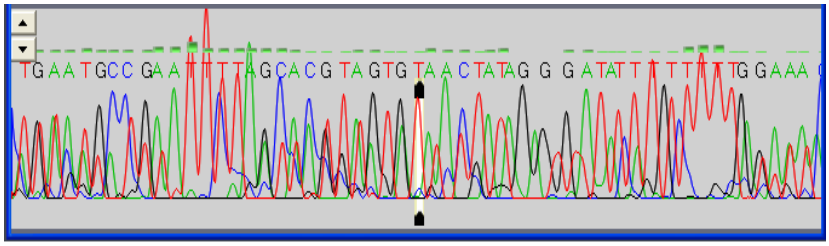


Abb. 19: Chromatogramm eines auf Grund der schlechten Qualität gelöschten ESTs. Deutlich zu erkennen sind die sich überlagernden Signale der fluoreszierenden Basen. Gezeigt werden die Basen der Positionen 126 bis 178.

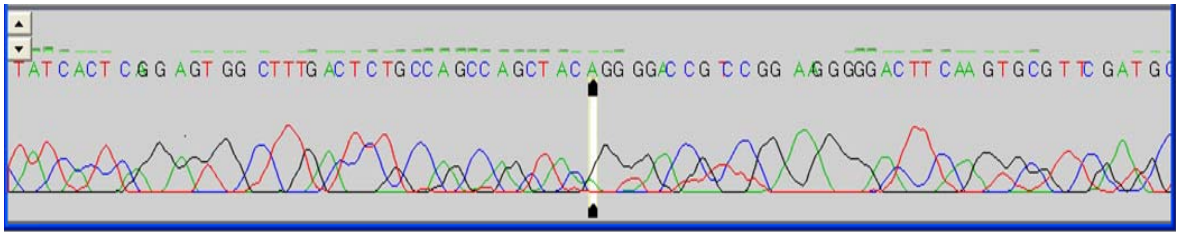


Abb. 20: Beispiel für die schlechte Auftrennung längerer Sequenzen; Positionen 781 bis 859.

4.4.2.2 Positionierung der *Callithrix jacchus* EST-Sequenzen

Anschließend wird mit der Analyse der verbleibenden 3215 EST-Sequenzen begonnen. Erster Schritt ist die Ähnlichkeitsanalyse zwischen den EST-Sequenzen und dem Referenzdatensatz (RefSeq; *Homo sapiens* Transkripte inklusive Gen-Modellen; Stand Dezember 2006). Die voreingestellten Parameter der Blat-Installation bleiben dabei unverändert.

Diejenigen Sequenzen, die nicht einem Transkript mit signifikanter Sequenzidentität zugewiesen werden können, werden in einem zweiten und dritten Schritt auf dem humanen Genom (RefSeq; Assembly Build 36.1; März 2006) positioniert. Auch hier wird eine minimale Sequenzidentität von 50 Prozent verwendet. Die Parameter werden im zweiten Schritt so gewählt, dass die erzielten Ergebnisse mit der Web-Version des UCSC-Genome Browsers vergleichbar sind. ESTs ohne signifikanten Treffer werden nochmals mit veränderten Parametern auf dem Genom positioniert. Diesmal wird die Option gewählt, bei der die zu analysierende EST-Sequenz und die genomische Sequenz mit allen sechs Leserahmen in Aminosäuresequenz (ASS) übersetzt werden. Da der genetische Code degeneriert ist, wird bewirkt, dass auf DNA-Ebene Sequenzen mit geringer Identität, die in ihrer ASS große Ähnlichkeit aufwiesen, dennoch gefunden werden.

Diese Art der Positionierung wird nur angewendet, falls in den ersten beiden Schritten keine Ergebnisse erzeugt werden konnten, da sie sehr zeitintensiv ist.

4.5 Annotation

Im Anschluss an die Ähnlichkeitsanalyse werden die Sequenzen mit Gennamen annotiert. Ist der Bereich der EST-Sequenz, der die Proteinsequenz codiert (EST-Sequenz abzüglich der 3'-UTR) lang genug, erfolgt zusätzlich die Annotation mit offenen Leserahmen (ORFs) beziehungsweise partiellen codierenden Sequenzen (CDS).

4.5.1 Annotation der Gennamen

Die Annotation der Gennamen erfolgt mit unterschiedlichen Methoden je nachdem, ob eine Zuordung zu einem Transkript erfolgt ist oder eine Positionierung auf dem Genom existiert. Existieren zu einer EST-Sequenz Referenz-Transkripte, werden maximal die besten fünf für eine Annotation in Betracht gezogen. Dazu werden aus der in Kapitel 2.1.1 zur Verfügung gestellten Datenbank die Gennamen und Synonyme zu jedem Referenzdatensatz extrahiert und analysiert. Bestehen zwischen den Annotationen der einzelnen Transkripte signifikante Unterschiede, werden diese manuell überprüft und gegebenenfalls korrigiert.
Existiert kein Referenz-Transkript, wird versucht anhand der erzeugten Koordinaten und der durch Blat vorhergesagten Exonstrukturen Überschneidungen zu humanen Genen zu finden. Dabei werden Referenz-Transkripte nur dann in Betracht gezogen, wenn sie in mindestens einem Exon mit den für das EST neu berechneten Exonbereichen auf demselben DNA-Strang überlappen. Für jedes EST wird eine Liste der in Frage kommenden Transkripte erzeugt und die jeweiligen Gennamen analysiert. Stimmen diese nicht überein beschreiben sie verschiedene Gene und werden nochmals manuell überprüft.

4.5.2 Annotation der offenen Leserahmen (open reading frames; ORFs)

Damit ESTs nur mit ORFs hoher Qualität annotiert werden, wird eine weitere Sequenzanalyse nötig. Die Analyse erfolgt mit Swiss-Prot als Referenzdatensatz. Über die Analyse der Sequenzähnlichkeit auf Proteinebene wird sichergestellt, dass in den ESTs vorhandene Verschiebungen im Leserahmen (so genannte Frameshifts) erkannt werden, die die Aminosäuresequenz (ASS) eines Proteins

korrumpieren. Dadurch werden ORFs nur dann annotiert, wenn auch die korrespondierende ASS zum EST eine hohe Sequenzähnlichkeit zu einem Protein aus dem Referenzdatensatz aufweist.

4.5.2.1 Vorbereitung der Sequenzen

Damit die ESTs mit den Proteinsequenzen aus Swiss-Prot verglichen werden können, werden sie zuerst revers komplementär übersetzt und dann mit Hilfe des Standard-Codes in die korrespondierende Aminosäuresequenz überführt. Da nicht bestimmt werden kann, ab welcher Base der ESTs der richtige Leserahmen beginnt (die ESTs beschreiben das 3'-Ende eines Transkriptes), werden die durch den Tripletcode vorgegebenen drei möglichen Leserahmen (frame1-3; Start bei Position 1-3 der EST-Sequenz; Abb. 21) jeweils bis zu einem Stopp-Codon übersetzt und bei einer Mindestlänge von 100 Aminosäuren in die zu analysierenden Proteinsequenzen aufgenommen.

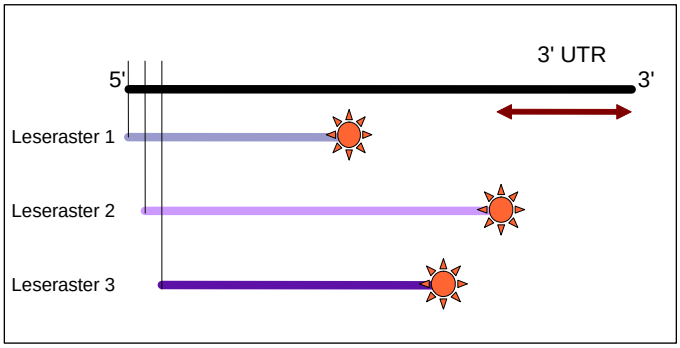


Abb. 21: Vorbereitung der EST-Sequenzen auf die Positionierung auf dem Proteinreferenzdatensatz;
Translation der verschiedenen Leserahmen in ASS;
orange dargestellt: Stop-Codons

4.5.2.2 Positionierung und Analyse der ORFs

Die Zuordung der in ASS übersetzten ESTs zu Proteinen erfolgt wiederum mit der Blat-Software. Es werden die Standardparameter für die Positionierung auf dem Proteinreferenzdatensatz gegen den Proteinreferenzdatensatz Swiss-Prot (UniProtKB: Swiss-Prot (*Homo sapiens*) Stand Dezember 2006) verwendet. Als signifikante Treffer werden nur Zuordnungen gezählt, deren Alignment bei Position 1 der ASS beginnt und deren Alignment mindestens 100 identische AS enthält. Danach wird für diejenigen ESTs, die bei der ersten Ähnlichkeitsanalyse gegen den RefSeq-Referenzdatensatz einen Treffer aufweisen, die Qualität der übersetzten ASS überprüft. Dazu wird das Verhältnis zwischen der erwarteten Proteinlänge und der tatsächlich beobachteten Proteinlänge (übersetzte ASS der ESTs) untersucht. Die erwartete Proteinlänge ergibt sich dabei aus der Länge der

codierenden Sequenz (CDS) des korrespondierenden Referenzset-Hits (RefSeq) abzüglich der Länge, der CDS, die nicht durch die EST-Sequenz abgedeckt wird (Abb. 22).

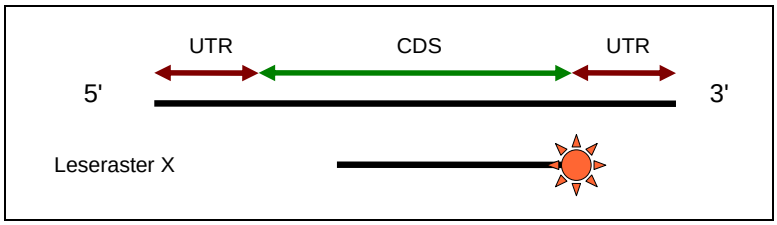


Abb. 22: Analyse der Qualität der partiellen CDS.

Als hochwertige ORFs zählen Fragmente von Protein-codierenden Sequenzen, bei denen die Länge der übersetzten ASS (einer der frames 1-3) mindestens 80 Prozent der erwarteten CDS einnehmen.

4.6 Auswertung der Positionierungen und der erfolgten Annotation

4.6.1 Analyse der Sequenzähnlichkeit

Die Analyse zur Sequenzähnlichkeit wird mit 3215 ESTs ausreichend hoher Sequenzqualität durchgeführt. Davon können 2250 Sequenzen Transkripten des Referenz-Sets von RefSeq zugeordnet werden. Von den verbleibenden 965 ESTs werden insgesamt 934 auf dem humanen Genom positioniert. Mit der ersten Parameterwahl (analog der Web-Version von Blat) können 598 Sequenzen mit der zweiten (Übersetzung der ESTs in Proteinsequenzen; sechs Leserahmen) weitere 335 Sequenzen positioniert werden (siehe Abb. 23). Nur 31 EST konnten weder einem Transkript zugeordnet noch auf dem Genom positioniert werden und wurden nicht weiter analysiert. Wie die Statistik verdeutlicht, kann der Hauptanteil der ESTs bereits bei der ersten Ähnlichkeitsanalyse einem Referenz-Transkript zugeordnet werden.

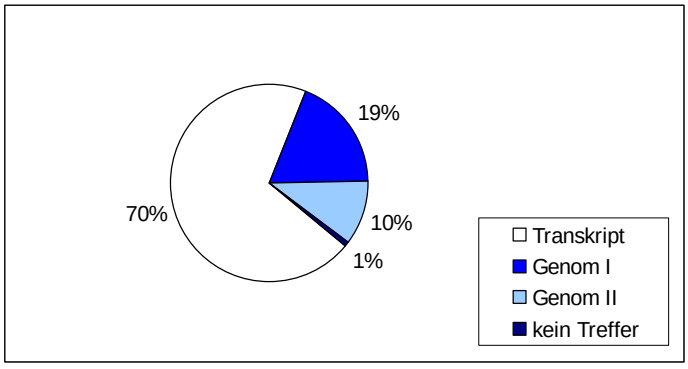


Abb. 23: Statistik zur Positionierung der ESTs

Bei der nachfolgenden Analyse der Sequenzähnlichkeit in Bezug auf das Genom werden nochmals knapp ein Drittel der Sequenzen positioniert. Dies bewirkt eine höhere Sensitivität bei der Annotation.

4.6.2 Annotation der Gennamen

Die Annotation erfolgt in zwei Schritten: Im ersten Schritt können 2250 ESTs mit Referenz-Transkript direkt Gennamen und Synonyme aus den Referenz-datensätzen zugewiesen werden. Im zweiten Schritt können mit Hilfe der genauen genomischen Positionsangaben bekannter humaner Gene aus den 933 (598 + 335) nur auf das Genom positionierbaren Sequenzen weitere 279 ESTs eindeutig mit einem Gen beziehungsweise einem Gen-Modell in Verbindung gesetzt werden (RefSeq ValidCoordinates; Stand März 2006). 16 ESTs überlappten mehrere Gene mit verschiedenen Gennamen (Abb. 24). Diese wurden manuell analysiert und 13 davon konnten eindeutig einem einzelnen Gen zugeordnet werden. Damit konnten über die Genompositionierung weitere 292 ESTs mit einem Gennamen annotiert werden, was einer Steigerung um 13 Prozent entspricht.

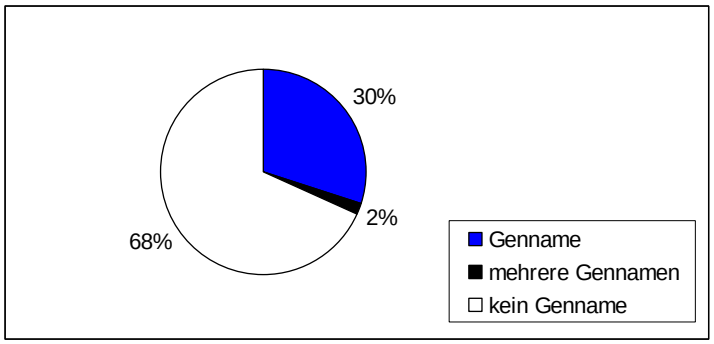


Abb. 24: Statistik zur Annotation mit Gennamen mit Hilfe der zusätzlichen Positionierung auf dem Genom.

In einem Fall (Sequenz H06-53.4_D03_013.ab1 in Abb. 25), in dem Überlappungen mit mehreren Genen (Magmas (NM_016069) und CORO7 (NM_024535), Abb. 25) bestehen, liegt die Vermutung nahe, dass es sich bei diesem EST um ein Fragment eines Fusionsgens[1] handelt, zumal zwei ähnliche humane Sequenzen in GenBank (BC032732, AL833954) gefunden werden.

[1] Ein Fusionsgen ist ein Hybrid-Gen, das aus zwei vormals voneinander getrennten Genen gebildet wurde.

4.6.3 Annotation der offenen Leserahmen (ORF-Annotation)

Mit der Positionierung auf dem Proteinreferenzdatensatz konnten 856 der in Aminosäuresequenz (ASS) übersetzten ESTs einem Protein zugeordnet werden. Davon haben 701 ESTs ein Alignment, das mit dem N-Terminus der entsprechenden ASS beginnt. Von diesen wiederum besitzen 641 EST bereits eine Positionierung auf einem Referenz-Transkript und können auf die Qualität des ORFs überprüft werden (siehe Abb. 26).

Von den 641 ESTs, deren Position auf einem Protein am N-Terminus beginnt und für die ein Referenz-Transkript zur Verfügung steht, ergibt die Analyse der CDS 31 ESTs, deren zu erwartete CDS (die des Referenz-Transkripts: von der Position der ersten Basenübereinstimmung des ESTs mit der mRNA bis zum Ende der CDS; Abb. 22) viel größer ist, als die CDS des Referenz-Transkripts. 610 ESTs, das sind 95 Prozent, erfüllen die gestellte Anforderung einer CDS-Länge von mindestens 80 Prozent der erwarteten CDS-Länge.

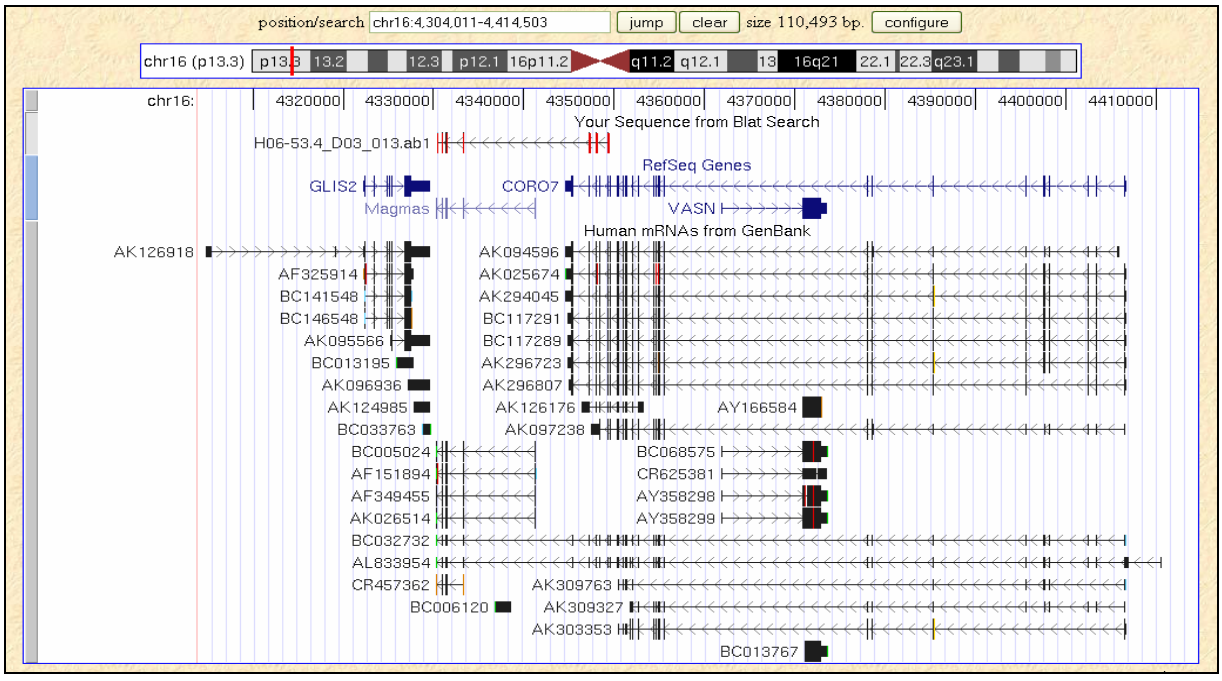


Abb. 25: Ansicht der Positionierung der Sequenz H06-53.4_D03_013.ab1 und der beiden Gene Magmas und CORO7 mit UCSC Blat Genome Browser. Angezeigt werden humane mRNAs aus RefSeq und GenBank (Juni 2008).

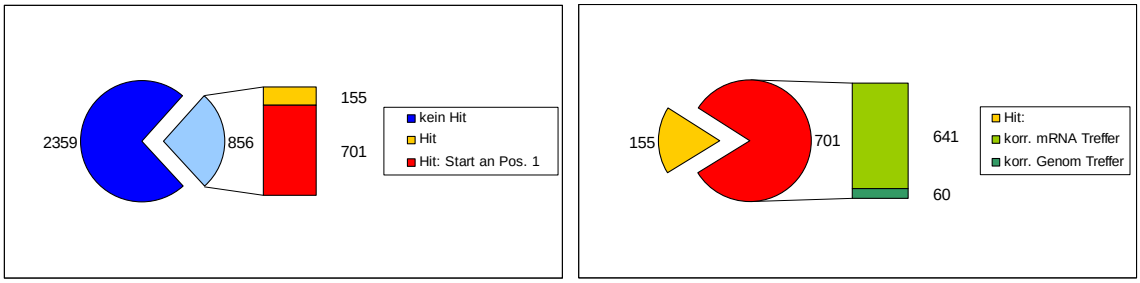


Abb. 26: Statistik zur Annotation der partiellen CDS
Links: Unterscheidung der ESTs in ESTs mit Hit und ohne Hit auf ein Referenzprotein; weiter wird zwischen Hits mit Alignmentstart an erster Position und Hits ohne Alignmentstart an erster Position unterschieden.
Rechts: Weitere Differenzierung der ESTs mit Hits auf Proteinsequenzen, die an erster Position starten, in Hits mit einem Hit auf dem Referenzdatensatz (Transkript) und ohne einem Hit auf ein korrespondierendes Transkript.

4.7 Veröffentlichung der annotierten Sequenzen

Insgesamt wurden 3215 EST-Sequenzen an GenBank und dbEST übermittelt:

610 Sequenzen, die über eine partielle CDS verfügen an GenBank, die restlichen 2605 an dbEST (GenBankIDs: EF214838-EF215447 und EH380242-EH382846). Diese beinhalten auch die 31 ESTs, die nicht auf das humane Genom positioniert werden konnten. Sie sind vermutlich Organismus-spezifische Gene.

Die Ergebnisse dieses Projektes wurden bereits veröffentlicht (Datson *et al.*, 2007).

5 Automatische Funktionsannotation von ESTs (OREST)

Nach dem erfolgreichen Abschluss der Annotation der *Callithrix Jacchus* EST-Sequenzen (Datson *et al.*, 2007) erfolgte die Umsetzung in eine vollautomatische Annotationsplattform (bereits veröffentlicht; (Waegele *et al.*, 2008)). Acht Säugetiere (*H. sapiens*, *M. musculus*, *R. norvegicus*, *B. taurus*, *C. familiaris*, *M. mulatta, P. troglodytes*, *M. domestica*) und die Bäckerhefe (*S. cerevisiae*) wurden in das System integriert. Eine Erweiterung um zusätzliche Organismen ist möglich. Parallel zur Annotation mit einem genomischen Referenzdatensatz wird ein Referenzdatensatz mit Proteinsequenzen zur Verfügung gestellt. Zusätzlich zu Gennamen und Synonymen wird die Funktion des Genproduktes annotiert. Hier kann zwischen dem “MIPS Functional Catalogue“ kurz FunCat (Ruepp *et al.*, 2004) oder den Annotationstermen von Gene Ontology (GO (The Gene Ontology Consortium, 2008)) gewählt werden. Diese beiden Annotationsschemata ermöglichen umfangreiche Analysen ganzer Datensätze. OREST selbst bietet mit der Einbindung der PROMPT-Software (Schmidt *et al.*, 2006) eine solche Analyse an: die statistische Auswertung der erzeugten Daten in Hinblick auf die Unter- beziehungsweise Überrepräsentation bestimmter Annotationsterme (FunCat oder GO).

Bei Verwendung des humanen Referenzdatensatzes bietet OREST, falls gewünscht, die Annotation mit krankheitsrelevanten Genen (OMIM™ – MorbidMap) (Hamosh *et al.*, 2005) an. Diese können, wie die Annotation der Funktion, mit PROMPT analysiert werden.

Um diese Analyse-Pipeline auch für Datensätze aus Hochdurchsatz-*in silico* - Analysen von über 30.000 Sequenzen anbieten zu können, sind Modifikationen und Optimierungen erfolgt, mit denen die Analysedauer erheblich verkürzt werden konnte, ohne Beeinträchtigungen der Annotationsqualität hinnehmen zu müssen.

5.1 Aufbau der benötigten Datenressourcen

Für die Erzeugung der Referenzdatensätze der Mammalia werden die bereits in Kapitel 2.1 beschriebenen Datenbanken RefSeq[1] und Swiss-Prot[2] verwendet. Im

Falle von *S. cerevisiae* werden die benötigten Daten aus CYGD[3] (Guldener *et al.*, 2005) verwendet. Um ein einheitliches Datenformat zwischen Mammalia und *S. cerevisiae* zu erhalten, werden alle Einträge aus CYGD in die bestehenden lokalen Datenbanken RefSeq und Swiss-Prot integriert. Transkripte werden in das von der RefSeq-Datenbank vorgegebene Format überführt, Proteine in das der Swiss-Prot-Datenbank. Zusätzlich wird eine Tabelle erzeugt, die in Format und Inhalt der Koordinaten-Tabelle der RefSeq-Datenbank entspricht. Diese Vereinheitlichung des Formats resultiert in einer wesentlich kompakteren Software und geringerem Wartungsaufwand. Zusätzlich wird im Fall der Hefe der Datenzugriff effizienter, da die benötigten Daten für die Annotation bereits im gewünschten Format vorliegen.

Um nicht bei jedem zu annotierenden Datensatz ein neues Referenzset zu erzeugen, werden diese in Dateien zwischengespeichert (so wie sie für die späteren Berechnungen benötigt werden) und nur bei Aktualisierung der zu Grunde liegenden Datenbanken neu erzeugt. Dies bedeutet eine erhebliche Zeitersparnis, da die Zugriffe auf den Datenbankserver erheblich sinken. Somit können die schon vorhandenen Datenbanken ohne Leistungseinbußen auch für dieses Projekt mitverwendet werden.

Um geringe Zugriffszeiten zu gewährleisten, wurde eine weitere Datenbank angelegt, in die Daten, die für die Annotation der ESTs benötigt werden in geeigneter Form importiert werden. In dieser Datenbank werden auch bei der Analyse des Datensatzes anfallende Zwischenergebnisse und Ergebnisse gespeichert.

Die Annotation der Funktion wird in vielen Datenbanken, so auch Entrez Gene (die Referenzen zu den Einträgen sind in RefSeq integriert worden; siehe Kapitel 2.2.5) und Swiss-Prot mit Annotationstermen von Gene Ontology (GO) durchgeführt. Damit auch die hierarchisch aufgebaute Klassifizierung der Funktion (Functional Catalogue, FunCat) annotiert werden kann, wird eine Tabelle importiert, in der jeder GO-ID genau eine FunCat-ID zugewiesen ist. Diese Zuordnung erfolgte manuell und wurde im Laufe dieses Projektes erweitert.

[1] Das genomische Referenzset wird aus den manuell validierten Transkripten der RefSeq-Datenbank zusammengestellt. Die genomischen Koordinaten jedes Genes des Referenz-Transkripts werden aus den für das aktuelle Assembly berechneten und zur Verfügung gestellten Dateien herausgefiltert.

[2] Für das proteomische Referenzset werden alle Proteinsequenzen, die in Swiss-Prot gelistet sind, verwendet.

[3] CYGD fungiert als Datenquelle für beide Referenzdatensätze: Transkripte und Proteine. Ebenso werden die aktuellen Koordinaten eines jeden Gens aus CYGD extrahiert.

Für die Annotation krankheitsrelevanter Gene/Genprodukte werden alle Einträge der Referenzdatenbanken RefSeq und Swiss-Prot mit OMIM Morbid-IDs verknüpft. Zwei Einträge stehen dabei in Relation zueinander, wenn ihre jeweils annotierten Gennamen und Synonyme in mindestens einem dieser Terme übereinstimmen. Um Relationen mit mehrdeutigen Gennamen zu vermeiden, wird eine Mindestlänge von drei Buchstaben pro Gennamen festgelegt. Von mehrdeutigen Gennamen spricht man, wenn identische Namen für verschiedene Gene verwendet werden. Zum Beispiel steht der Genname „TF“ für ‚Transferrin’ (NM_001063) und ‚Coagulation factor III’ (‚thromboplastin’, ‚tissue factor’, NM_001993). Die resultierenden Wertepaare werden aus Effizienzgründen in zwei Tabellen importiert; eine Tabelle für Relationen zwischen RefSeq-IDs und ihren korrespondierenden Morbid-IDs und eine weitere für die Relationen zwischen Swiss-Prot-IDs und deren Morbid-IDs.

5.2 Arbeitsweise des fertigen Servers

Die Hauptfunktion des Servers besteht darin, zeitsparend Datensätze von bis zu 50.000 EST-Sequenzen mit Funktionsannotation zu versehen und kann in vier konsekutive Schritte aufgeteilt werden: Auswahl der Parameter, mit der die Berechnungen durchgeführt werden, Vorverarbeitung und Positionierung der eingegebenen Sequenzdaten, Annotation der Funktion und statistische Analyse der erzielten Resultate. Eine schematische Darstellung des Annotations-Workflows ist in der nachfolgenden Abbildung (Abb. 27) vermerkt.

5.2.1 Parameterauswahl und Validierung der Eingabe-Sequenzen

Der erste Schritt der EST-Analyse in OREST ist die Auswahl geeigneter Parameter. Parametrisiert wurden die Wahl des Referenzorganismus, die Art des Referenzsets, die Art der Funktionsannotation und die Annotation krankheitsrelevanter Gene/Genprodukte. Als Referenzorganismen stehen verschiedene Säugetiere (Mensch, Maus, Ratte und weitere fünf Säugetiere) und die Bäckerhefe *Saccharomyces cerevisiae* für die Analyse von Pilzdatensätzen zur Verfügung. Abhängig von der phylogenetischen Verwandtschaft der Eingabesequenzen zum gewählten Referenzset kann der Benutzer die minimale Sequenzidentität auf einen Wert zwischen 50% und 90% festlegen. Die Art des Referenzsets bestimmt, ob die zu analysierenden Daten auf einem Datensatz positioniert werden, der

entweder aus genomischen Sequenzen bzw. Transkripten oder – nach Übersetzung in alle sechs Leserahmen – Proteinsequenzen besteht.

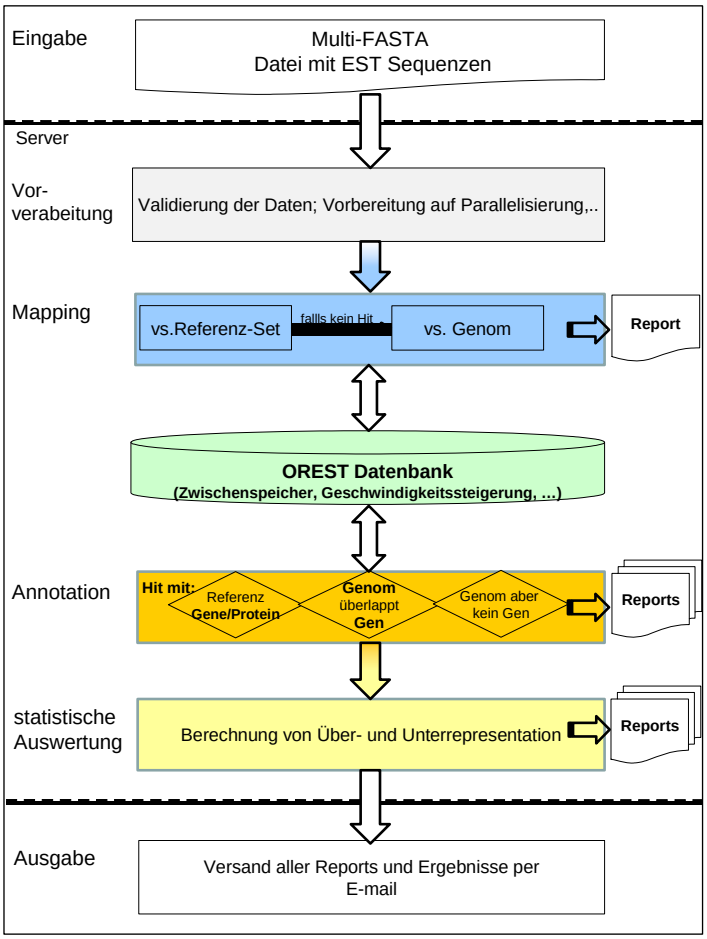


Abb. 27: OREST Workflow. OREST verarbeitet EST-Datensätze in Form von Dateien im multi-FASTA-Format. Die Analyse erfolgt in vier Schritten: Vorbereitung der Daten, Positionierung der ESTs auf Referenzdatensatz und Genom, Annotation und statistische Analyse.

Im Bereich der Funktionsannotation kann zwischen GO- und FunCat-Annotationstermen gewählt werden. Falls als Referenzorganismus der Mensch selektiert wurde, kann die Annotation um krankheitsrelevante Informationen ergänzt werden. Bevor die EST-Daten endgültig zur Analyse an den Server übermittelt werden, wird ihr Format überprüft.

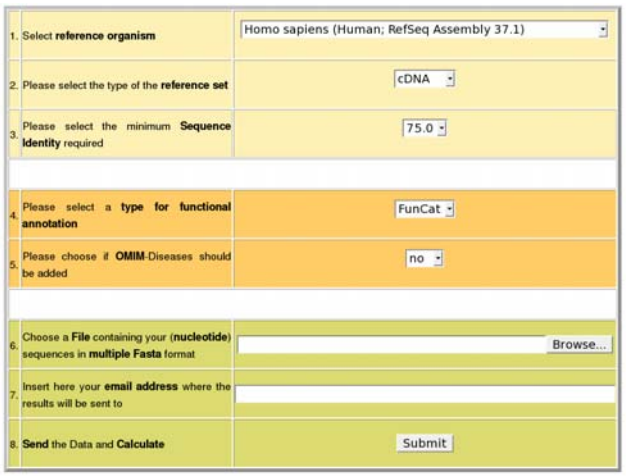


Abb. 28: Startseite mit Eingabemaske von OREST.

5.2.2 Vorprozessierung und Positionierung der EST-Sequenzen

In die Berechnungen werden nur Sequenzen einbezogen die eine Mindestlänge von 100 Basen aufweisen. Wurde eine Annotation mit Hilfe des Proteinreferenzdatensatzes gewählt, werden die ESTs für jeden der sechs Leserahmen in Aminosäuresequenzen übersetzt (ASS). Die Analyse wird nicht mit der gesamten ASS durchgeführt, sondern mit dem putativen N- oder C-Terminus der Sequenz, welcher mindestens 20 Aminosäuren umfassen muss.

Für die Positionierung der ESTs auf die Sequenzen des gewählten Referenzdatensatzes wird Blat (Kent, 2002) verwendet. Die Prozedur erfolgt in zwei aufeinander folgenden Schritten: Im ersten Schritt werden die vorprozessierten EST-Sequenzen auf Sequenzähnlichkeit mit den Sequenzen aus den Referenzdatensätzen überprüft. Falls keine signifikanten Treffer mit den Referenz-Transkripten oder -Proteinen erzielt wurden, wird diese Analyse mit den revers-komplementären Sequenzen wiederholt. In diesem ersten Schritt kann es aus verschiedenen Gründen zu insignifikanten Treffern kommen. Ein Grund ist schlechte Sequenzqualität, ein anderer die Unvollständigkeit der Referenzdatensätze. Sequenzen, die aus letzterem Grund nicht annotiert werden können, werden im zweiten Schritt auf dem Genom positioniert. ESTs, die auf dem Genom positioniert werden können und die eine Überschneidung mit Exonbereichen eines Genes[1] aufweisen, verbleiben in dem zu annotierenden Datensatz. Nicht positionierbare ESTs werden mit „kein Treffer“ gekennzeichnet. Die Berechnung der Sequenzähnlichkeit mit Hilfe der Referenzsets wird verwendet, um die Anzahl der ESTs, die auf das Genom positioniert werden zu minimieren. Wie oben erwähnt, kann die minimal geforderte Sequenzidentität für Transkript- und Genomreferenzset zwischen 50% und 90% eingestellt werden. Bei der Analyse der Sequenzähnlichkeit zum Proteinreferenzset ist die minimale Sequenzidentität auf 80% festgelegt um falschpositive Treffer zu vermeiden.

5.2.3 Annotation der Funktion

Um die täglich in Experimenten zahlreich erzeugten EST-Daten nutzen zu können, wird eine systematische Annotation jeder einzelnen Sequenz benötigt. Zur Anno-

[1] Für die Positionierung der Sequenzen auf das Genom werden bei der Berechnung der Überlappungen mit Exonbereichen die aktuell gültigen Koordinaten (valid coordinates) von NCBI beziehungsweise CYGD verwendet. Die in diesem Datensatz abgelegten Einträge entsprechen denen des Referenzdatensatzes für Transkripte.

tation von eukaryontischen Genprodukten existieren zwei häufig verwendete Annotationsschemata: “MIPS Functional Catalogue“ (FunCat) und Gene Ontology (GO). Für Datensätze mit Säugetiersequenzen kann die Annotation mit beiden Schemata durchgeführt werden, für die Hefe ist bei OREST nur die Annotation mit dem FunCat-Schema möglich. Die GO-Terme werden pro EST aus den korrespondierenden RefSeq- beziehungsweise Swiss-Prot-Einträgen extrahiert. Falls der Benutzer für weiterführende Analysen ein hierarchisches Annotationsschema bevorzugt, werden die GO-Terme in FunCat-IDs übersetzt. Ist die Hefe der gewählte Modellorganismus, wird der FunCat direkt aus CYGD übernommen.

Für die Funktionsannotation werden die ESTs in zwei Teilmengen aufgeteilt: ESTs mit Ergebnissen aus der Sequenzähnlichkeitsanalyse und ESTs mit einer Positionierung auf dem Genom. ESTs, die Treffer auf Referenz-Transkripten oder –Proteinen aufweisen, werden mit den FunCats/GO-IDs und deren Beschreibung aus den korrespondierenden Referenzeinträgen annotiert. GO-IDs mit dem Evidence-Code „IEA“ (inferred by electronic annotation) werden wegen der hohen Fehleranfälligkeit ausgeschlossen (weitere Evidence-Codes und deren Beschreibung sind auf der beigefügten CD-ROM enthalten).

Pro EST-Sequenz werden nur die besten drei Referenz-Transkripte/-Proteine zu den Ergebnissen der Analyse hinzugefügt. Zu ESTs, die kein korrespondierendes Gen aufweisen, werden die berechneten Koordinaten auf dem Genom zur Verfügung gestellt. Wurde als Referenzorganismus der Mensch ausgewählt, ist die zusätzliche Annotation beziehungsweise die Vorhersage von Assoziationen zwischen Genen und Krankheiten (d.h. Krankheiten, deren Genotypen bekannt sind) möglich. Zu diesem Zweck werden die Swiss-Prot- und RefSeq-IDs mit Hilfe der bereits berechneten Übersetzungstabellen in Morbid-Map-IDs übersetzt.

5.2.4 Statistische Auswertung

Mit Hilfe statistischer Verfahren kann OREST auf Grund der Erzeugten Datengrundlage folgende Fragestellungen beantworten:

- Weisen aus Krebszellen isolierte ESTs eine nicht zufällige Anreicherung bestimmter Funktions-Klassen auf?
- Hängen gefundene ESTs mit einer bestimmten molekularen Funktion zusammen?
- Können EST Marker gefunden werden, die auf Krankheiten hindeuten?

Die statistische Auswertung der EST-Analyse erfolgt mit einer in das Software-Paket integrierten und auf die Anforderungen angepassten Version der PROMPT-Software. Mit ihr werden die annotierten FunCat-, GeneOntology- und OMIM-Informationen derjenigen Gene, denen ESTs zugeordnet wurden, daraufhin untersucht, ob sie über- beziehungsweise unterrepräsentiert sind. Für jeden zur Verfügung stehenden Annotationsterm der ESTs wird die Art der Repräsentation im Verhältnis zum Auftreten einer Annotation im gesamten Genom (die Menge aller bekannten Gene) berechnet. Die Bewertung, ob das erzielte Ergebnis signifikant ist, wird mit dem so genannten „e-Score“ vorgenommen. Der „e-Score“ beschreibt die Wahrscheinlichkeit, mit der die Differenz zwischen beobachteter und gegebener Verteilung (Annotation der ESTs aus Experiment, Annotation aller Gene des Referenzsets) zufällig auftritt. Der „e-Score“ wird mit der von Castillo-Davis (Castillo-Davis *et al.*, 2003) beschriebenen Methode unter Verwendung einer hypergeometrischen Verteilung mit konservativer Bonferronikorrektur berechnet. Als Ergebnis liefert OREST also nicht nur die Annotation der ESTs mit Funktionsklassen und Krankheiten sondern auch eine statistische Signifikanzanalyse zu deren Vorkommen.

5.2.5 Ausgabe der Ergebnisse

Die Ergebnisse der gesamten Analyse werden per E-Mail an den Benutzer versandt und bestehen aus zwei verschiedenen Berichtarten: den Annotationen und den auf den Annotationen basierenden statistischen Auswertungen. Die Annotationsergebnisse werden entsprechend ihrer Qualität (Positionierung auf Eintrag des Referenzsets oder Genom) weiter unterteilt. Entsprechend werden die statistischen Analysen für die annotierten Gene ausgegeben. Falls die Annotation mit krankheitsrelevanten Genen gewünscht wurde, werden zwei weitere Berichte erstellt. Die jeweiligen Dateiformate sind im Anhang genau erklärt.

5.2.6 Optimierung

Für die Verarbeitung von bis zu 50.000 ESTs ist eine optimale Ausnutzung der zur Verfügung stehenden Ressourcen essentiell. So werden die Referenzdatensätze nur bei einem Update der entsprechenden Datenbanken neu erzeugt, bis dahin werden sie arbeitsspeicherneutral zwischengespeichert. Bei der Positionierung der ESTs wird zeit- und arbeitsspeicherintensives Positionieren auf das

Genom nach Möglichkeit vermieden. Das erreicht man über das Positionieren der ESTs auf einen nur einen Bruchteil der Genomsequenz einnehmenden Referenzdatensatz. Nur diejenigen ESTs ohne einen Treffer auf einer Sequenz des Referenzdatensatzes, werden auf dem Genom positioniert. Die Vorteile dieses Ansatzes liegen in der kürzeren Berechnungsdauer der Positionierung eines ESTs und in der schnelleren Annotation, da Treffer auf ein Referenz-Transkript oder –Protein direkt annotiert werden können, wohingegen bei Positionierung auf das Genom erst Überlappungen mit auf dem Genom liegenden Genen berechnet werden müssen, um die Annotation durchführen zu können. Für die Positionierung wird Blat anstelle von Blast verwendet, da es, wie gezeigt wurde, um den Faktor 500 bzw. 50 (mRNA/EST und Proteinsequenz) (Kent, 2002) schneller ist. Zur weiteren Optimierung der Laufzeit werden die Berechnungen zur Positionierung nach Möglichkeit parallel durchgeführt. Das bedeutet, dass die EST-Sequenzen für die Positionierung auf Referenzdatensätzen in Teilmengen aufgetrennt und auf mehrere CPUs verteilt berechnet werden. Die Positionierung der restlichen ESTs auf das Genom erfolgt für jedes Chromosom einzeln. Die erzeugten Positionierungen werden erst in der Datenbank sequenzspezifisch zusammengeführt. Für die Annotation werden nach Möglichkeit alle für die Annotation immer wieder benötigten Ressourcen aus der MySQL-Datenbank in den um ein Vielfaches schnelleren Arbeitsspeicher geladen. Die speicherintensive Analyse von bis zu 50.000 ESTs macht es notwendig, die in der Datenbank zwischengespeicherten Informationen für die Annotation in Teilmengen zu laden und zu verarbeiten.

5.2.7 Implementierung

Die Software, die OREST zu Grunde liegt, ist in Java 1.5 unter Verwendung der JavaBean-Technologie implementiert. Als Server für den Webzugriff wird ein Sun Java System Application Server der Version 9.1 verwendet. Für die Positionierung der ESTs kommt die Blat-Software der Version 34 (64bit-Version) zum Einsatz. Die Verteilung der benötigten Rechenaufgaben zur parallelen Prozessierung erfolgt mit der Sun Grid Engine (www.sun.com). Zur Zwischenspeicherung der Ergebnisse und um effizienten Zugriff auf alle benötigten Daten zu gewährleisten, dient ein MySQL-Server Version 5.0.27-standard.

6 Cross-referencing (CRONOS)

Die redundante Speicherung von Transkript-Sequenzen in verschiedensten Datenbanken wie sie bei Genomprojekten entstehen, führt bei Projekten, in denen Daten aus vielfältigen Quellen integriert werden müssen, zu erheblichen Problemen. Die manuelle Auflösung dieser Redundanzen ist sehr zeitaufwändig. In der Vergangenheit wurden verschiedene Ansätze wie die Berechnung von Sequenzidentitäten oder die Verwendung von Gennamen entwickelt, um Identifier miteinander in Relation zu stellen. Gen- und Proteinnamen für Moleküle wurden in der Vergangenheit von Wissenschaftlern festgelegt, die an diesen Molekülen geforscht haben. Als Resultat dieser unkontrollierten Nomenklatur wurden verschiedene Proteinnamen für identische Proteine, und umgekehrt, ein Name für verschiedene Proteine verwendet. Dasselbe gilt auch für Gennamen wie beispielsweise IGEL, welcher nun als Synonym für PHF1 und MS4A2 weiter verwendet wird. Mit dem HUGO Gene Nomenclature Committee (Povey *et al.*, 2001) wurde eine Organisation gegründet, die sich der Problematik im Bereich der Gennamen annimmt. Trotzdem werden Proteinnamen von den entsprechenden Datenbanken wie UniProtKB (The UniProt Consortium, 2008) und RefSeq (Pruitt *et al.*, 2007) unterschiedlich vergeben.

Anwendungen wie MatchMiner (Bussey *et al.*, 2003), welche die Relationen mit Hilfe der Gen- und Proteinnamen erzeugen, sind von dem oben genannten Problem der mehrdeutigen Nomenklatur betroffen. Anwendungen, die Verknüpfungen zwischen zwei Datenbankeinträgen mit Hilfe der Sequenzidentität aufzubauen, sind dagegen von dieser Nomenklatur unabhängig. Dennoch ist der Treffer mit der höchsten Sequenzähnlichkeit nicht notwendigerweise dasselbe Gen und ein Grenzwert von 100% Sequenzidentität wie er in PICR (Cote *et al.*, 2007) verwendet wird, kann nicht alle korrekten Relationen wiedergeben. Dies folgt aus der Tatsache, dass in verschiedenen Datenbanken Genmodelle und daraus abgeleitete Proteinsequenzen oft nicht konsistent sind. Dies tritt vor allem am N-Terminus der Proteinsequenz auf (zum Beispiel das Gen IARS; RefSeq NM_002161 und UniProtKB P41252). Eine andere Problematik entsteht durch die Handhabung von Spleißvarianten in Swiss-Prot. Spleißvarianten werden in einem Datenbank-Eintrag von einer so genannten Master-Sequenz (Kapitel 2.1.2.1)

repräsentiert. Dies führt zwangsläufig zu erheblichen Unterschieden in der Aminosäuresequenz korrespondierender Einträge anderer Datenbanken.
Erschwerend kommt hinzu, dass jede der einzelnen Datenbanken jedem Eintrag eigene, Datenbank-spezifische Identifier zuweist, was die Kompatibilität der verschiedenen Datenbanken erheblich einschränkt. Weitere Komplikationen treten wegen instabiler Identifier auf, die beispielsweise auf Grund von verbesserten Genmodellen ersetzt oder gelöscht werden.
Diese Inkonsistenzen haben beträchtliche Auswirkungen auf spätere bioinformatische Anwendungen, die auf diesen Daten aufbauen. Mehrdeutige molekulare Nomenklatur betrifft beispielsweise die Analyse von Protein-Netzwerken, die mit Hilfe von Text-Mining erzeugt werden. Im Text-Mining bislang nicht berücksichtigte mehrdeutige Namen ermöglichen das Einbringen fehlerhafter Protein-Protein-Interaktionen.

Im Folgenden werden korrespondierende Einträge der Datenbanken RefSeq, UniProtKB und Ensembl miteinander verknüpft. Es wird beschrieben, wie mehrdeutige Gen- oder Proteinnamen aus diesen Datenbanken extrahiert werden, und mit Hilfe der eindeutigen Terme Verknüpfungen erstellt werden können. Diese werden mit Hilfe von Sequenzalignments und der Sequenzidentität verifiziert. Die Verknüpfungen können mit der Web-Applikation CRONOS (Waegele *et al.*, 2008) aufgerufen werden.

6.1 Vorgehensweise

6.1.1 Erzeugung der Relationen zwischen Einträgen verschiedener Datenbanken

Die Berechnung der Relationen wird mit Daten von sechs Organismen (fünf Säugetiere: Mensch, Maus, Ratte, Rind, Hund und die Fruchtfliege) aus den Datenbanken UniProtKB, RefSeq und Ensembl durchgeführt. Mensch, Maus und Ratte sind momentan die am häufigsten beforschten Organismen basierend auf der Anzahl der in PubMed enthaltenen Einträge (10,3; 0,9 und 1,25 Millionen Eintragungen). Die anderen Organismen sind als weitere Modellorganismen von Bedeutung und konnten auf Grund derselben Datengrundlagen direkt in die Berechnungen einbezogen werden.

Die Relation zweier Genprodukte, welche die Grundlage der Umrechnungen von Identifiern ist, wird wie folgt definiert: Zwei Datenbankeinträge stehen in Relation zueinander, wenn sie mindestens einen Gen- oder Proteinnamen gemein haben.
Die Relationen werden für jeden Eintrag in konsekutiven Schritten berechnet, bis weitere Einträge mit identischen Eigenschaften gefunden werden.

- Zwei Einträge haben mindestens einen Gennamen oder ein Synonym zu einem Gennamen gemein.
- Zwei Einträge haben mindestens einen Gen- und Proteinnamen gemein. Dieser Schritt ist notwendig, da in manchen Datenbanken Gennamen als Proteinnamen verwendet werden und umgekehrt.
- Zwei Einträge haben mindestens einen Proteinnamen oder ein Synonym zu einem Proteinnamen gemein.

Dieser Vorgang wird zuerst mit Daten aus RefSeq und UniProtKB durchgeführt. Wenn zu einem RefSeq-Eintrag mehrere Relationen erzeugt werden können, werden solche Einträge aus UniProtKB bevorzugt, die aus Swiss-Prot stammen. Danach werden die Relationen zwischen RefSeq- und Ensembl-Einträgen berechnet. Falls keine Relation zwischen RefSeq und Ensembl hergestellt werden kann, aber eine Relation zwischen Ensembl und UniProtKB besteht, wird Ensembl über UniProtKB mit RefSeq in Relation gebracht.
In einem weiteren Schritt werden nun transitive Relationen aus den oben definierten Relationen berechnet. Ziel ist die Erzeugung nicht-redundanter Identifier-Triplets, die einander entsprechende Einträge der drei Primärressourcen enthalten. Diese Triplets werden direkt in CRONOS importiert. Der vierte Schritt beinhaltet die Integration aller Datenbank-Einträge, die bislang noch nicht importiert werden konnten. Dies betrifft den Import jener Einträge, die nur in jeweils einer der oben erzeugten Relationen auftreten, genauso wie solche, die in keiner Relation zu einem anderen Eintrag stehen.
Abschließend werden allen in CRONOS enthaltenen Einträgen (Triplets) vollständige nicht-redundante Listen der Gen- und Proteinnamen zugewiesen.
UniProtKB, RefSeq und Ensembl enthalten jeweils eine Vielzahl Referenzen auf weitere Datenbanken, die Informationen über Stoffwechselwege, Annotationen über die Funktion oder über Domänen eines Proteins enthalten. Diese zusätzlichen Informationen erlauben es in CRONOS Anfragen nach sich entsprechenden Einträgen mit bis zu 18 verschiedenen Identifiern – abhängig vom jeweiligen Orga-

nismus - zu stellen (siehe Anhang). Somit ist CRONOS die zentrale Drehscheibe bei der Umsetzung von Identifiern verschiedener Datenbanken.

6.1.2 Erstellen der Menge der mehrdeutigen Gen- und Proteinnamen

Wie oben beschrieben, ist die Erzeugung der Relationen von den einem Eintrag zugewiesenen Gen- und/oder Proteinnamen abhängig. Um zu verhindern, dass falsche Relationen auf Grund von zweideutigen Gen- oder Proteinnamen (ein Name beschreibt verschiedene Proteine) auftreten, werden für jeden Organismus separate Listen dieser Terme erzeugt. Organismus-spezifische Listen sind nötig, da Bezeichnungen, die in einem Organismus zweideutig, in einem anderen eindeutig sein können. Zum Beispiel ist ADORA2 ein zweideutiger Genname von *H. sapiens* aber nicht von *M. musculus*. Umgekehrtes gilt für die Bezeichnung GALT, die bei *M. musculus* zwei Gene beschreibt. Im ersten Schritt werden alle zweideutigen Terme innerhalb einer Datenbank extrahiert. Falls ein Name in mindestens zwei Einträgen vorkommt, die verschiedene Gene (Spleißvarianten werden als ein Gen betrachtet) oder Proteine beschreiben, wird dieser spezielle Name als zweideutig eingeordnet und nicht für die Erzeugung der Relationen verwendet. Dieser Vorgang wurde bis zur Erkennung von automatisch prüfbaren Eigenschaften manuell erledigt, danach automatisch. Da die einzelnen Einträge auch in Zukunft – gemäß dem aktuellen Wissensstand – weiter annotiert werden, erfolgt diese Analyse bei jeder Aktualisierung von CRONOS.

RefSeq

In dieser Analyse werden nur Datenbankeinträge verwendet, die auf Evidenzen wie cDNA basieren, also keine Modelle sind (Einträge mit Präfix NM). Für die Gennamen-Analyse werden alle Gennamen und deren Synonyme in eine nicht redundante Liste eingelesen und sequenziell auf ihr Vorkommen in verschiedenen Einträgen (NM) hin überprüft. Wird ein solcher Genname gefunden, muss geklärt werden, ob es sich um Spleißvarianten handelt, oder um verschiedene Gene. RefSeq speichert Spleißvarianten eines Genes in verschiedenen Einträgen. Dazu werden die zu den RefSeq-Einträgen gehörenden Gene-IDs[1] extrahiert. Werden zwei oder mehr verschiedene Identifier gefunden, so wird der Genname als zwei- oder mehrdeutig erkannt. Die Analyse der Proteinnamen erfolgt analog zu den

[1] Eindeutige ID, die jedem Eintrag in Entrez Gene zugewiesen wird.

Gennamen. Dazu werden alle Terme, die zu den Proteinnamen gehören extrahiert und analysiert.

UniProtKB

Die Analyse der UniProtKB-Einträge erfolgt nur mit der manuell annotierten Sektion – Swiss-Prot. Einträge der TrEMBL-Sektion werden nicht verwendet, da sie vorläufig sind und noch keiner weiteren Prüfung durch Experten unterzogen wurden. Bei der Analyse der Gen- und Proteinnamen macht man sich die Annotationsweise von UniProtKB zu Nutzen. Um Redundanzen zu vermeiden vereint UniProt alle von einem Gen codierten Proteinprodukte einer Spezies in einem Eintrag. Unterschiede in den zusammengefassten Sequenzen sind innerhalb dieses Eintrages vermerkt (‚Sequence annotation', ‚Alternative products', ‚General annotation'). Nur wenn die Sequenzen und die Struktur zweier Produkte zu verschieden sind, werden die einzelnen Varianten in getrennten Einträgen beschrieben. Für die Analyse bedeutet dies, dass alle Gennamen und Proteinamen, die in verschiedenen Swiss-Prot Einträgen mehrfach vorkommen, automatisch potentielle mehrdeutige Terme sind. Einträge, die diese Namen enthalten, müssen danach noch darauf hin untersucht werden, ob sie aus Spleißvarianten entstandene Proteine beschreiben[1]. Nur wenn dies nicht der Fall ist, werden sie den Listen der zweideutigen Gen- oder Protein-Namen zugeführt.

Ensembl

Jeder Datenbankeintrag in Ensembl ist in Hierarchie-Ebenen aufgegliedert (Abb. 29). Die Informationen, die für diese Analyse benötigt werden, befinden sich auf zwei Hierarchie-Ebenen. Auf der ersten Ebene befindet sich das zu beschreibende Gen. Jedem Gen sind auf einer zweiten Ebene ein oder mehrere Transkripte zugeordnet. Diesen Transkripten sind jeweils Gennamen zugewiesen, die innerhalb eines Eintrages nicht variieren. In seltenen Fällen beschreiben zwei verschiedene Gen-Einträge dasselbe Gen; diese bekommen – wie überprüft wurde – identische Gennamen und Synonyme zugewiesen (zum Beispiel SEEK1: ENSG00000137336 und ENSG00000204540). Mehrdeutige Gennamen sind folglich solche Namen, denen mehr als eine GeneStableID zugeordnet werden

[1] Momentan kann in Swiss-Prot anhand der Annotation aller Gennamen eines Proteins entschieden werden, ob zwei Einträge Spleißvarianten darstellen oder nicht. Nur wenn alle Gennamen übereinstimmen handelt es sich um Spleißvarianten.

können. Zusätzlich dürfen die zu den GeneStableIDs gehörenden Gennamen und Synonyme nicht übereinstimmen.

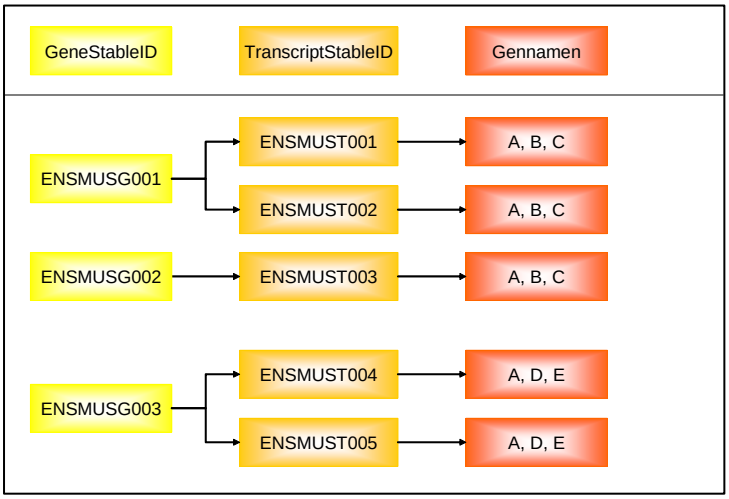


Abb. 29: Hierarchischer Aufbau der Ensembl-Einträge; Jeder GeneStableID sind ein oder mehrere TranscriptStableIDs zugeordnet. Diesen sind wiederum jeweils eine Liste der zugehörigen Gennamen und Synonyme zugewiesen.

In einem zweiten Schritt der Extraktion aller mehrdeutigen Gennamen, werden korrespondierende Gennamen aus den manuell annotierten Sektionen von RefSeq und UniProtKB analysiert.

Einträge beider Datenbanken, die denselben Gennamen enthalten und in einer “one-to-many“ oder “many-to-many“ Relation zueinander stehen (z.B.: ein Swiss-Prot-Eintrag wird mehreren RefSeq-Einträgen zugeordnet), werden auf irreführende Annotation hin untersucht. Dies wurde manuell durchgeführt, indem weitere Informationen wie die Funktion der Proteine oder die Sequenzidentitäten der betroffenen Einträge zu Hilfe genommen wurden. In den meisten Fällen können durch Ausschließen des zweideutigen Gennamens korrekte “one-to-one“ Relationen erzeugt werden. Ein Beispiel für eine “one-to-many“-Relation ist in Abb. 30 erläutert; Die korrekte “one-to-one“-Relation in Abb. 31.

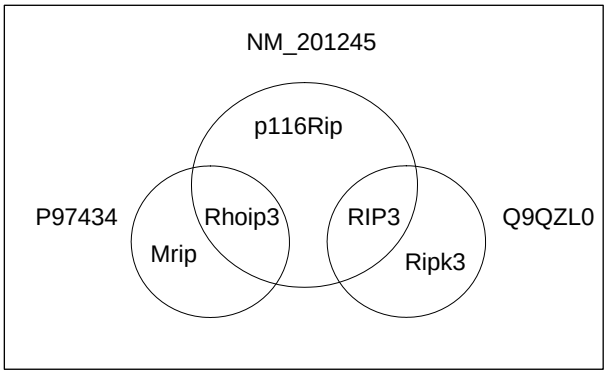


Abb. 30: Beispiel für eine inkorrekte “one-to-many“-Relation; Ein RefSeq-Eintrag (NM_201245) wird auf zwei Swiss-Prot-Einträge (P97434 und Q9QZL0) referenziert. Nur die Relation zwischen NM_201245 und P97434 ist korrekt. Durch Ausschließen des Gennamen RIP3 kann die korrekte Zuweisung erfolgen (Abb. 31).

Diese Beobachtung kann nicht nur bei Gennamen sondern auch bei den Proteinnamen gemacht werden (Abb. 32). Als Beispiel dienen die in der Abbildung ge-

zeigten Datenbankeinträge, die Gen- und Proteinnamen sind voneinander getrennt aufgelistet. Eingeklammerte Terme (siehe Kapitel 6.1.3 Informationsgehalt von Gen- und Proteinnamen) werden nicht für die Erzeugung der Relationen verwendet (siehe Kapitel 6.1.1).

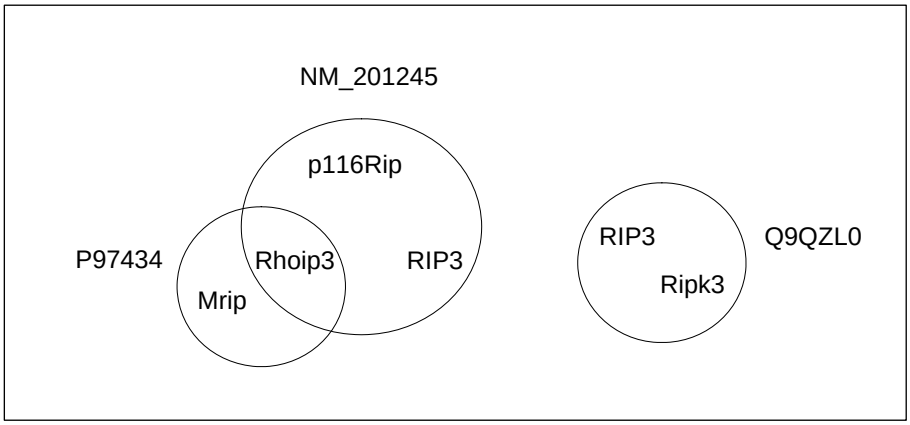


Abb. 31: Korrekte "one-to-one"-Relation nach Ausschließen des zweideutigen Gennamens.

Die Berechnung der Relationen mit Gennamen erzeugt keine Ergebnisse, die darauf folgende Berechnung mit Gen- und Proteinnamen verknüpft korrekt RefSeq (NM_000067) mit Swiss-Prot-Eintrag P00918 (grün). Mit P00915 kann NM_000067 über den Proteinnamen 'Carbonic anhydrase B' verknüpft werden (rot). Nach Prüfen der verschiedenen Proteinnamen und zusätzlicher Berechnung der Sequenzidentitäten wird 'Carbonic anhydrase B' als irreführende Annotation markiert und die Verknüpfung zwischen NM_000067 und P00915 gelöscht.

NM_000067	P00918	P00915
(CA2) CAII Car2 CA-II	(CA2)	(CA1)
Carbonic anhydrase II carbonate dehydratase II carbonic dehydratase carbonic anhydrase B carbonic anhydrase II, CA II	Carbonic anhydrase 2 Carbonate dehydratase II Carbonic anhydrase II (EC 4.2.1.1) CA-II Carbonic anhydrase C (CAC)	Carbonic anhydrase B (CAB) Carbonic anhydrase 1 (EC 4.2.1.1) Carbonic anhydrase I Carbonate dehydratase I CA-I

Abb. 32: Nicht korrekte "one-to-many"-Relation auf Grund irreführender Proteinnamen. Die Erzeugung einer Relation ist mit Gennamen nicht möglich, deshalb wird eine Relation mit Gen- und Proteinnamen (CA-II) und eine weitere mit Proteinnamen erzeugt. Bei der Überprüfung der "one-to-many"-Relationen wird deutlich, dass der Term "Carbonic anhydrase B" eine falsche Relation verursacht. Dieser wird als mehrdeutig markiert und die Verknüpfung aufgelöst.

Nicht alle "one-to-many"-Relationen enthalten zweideutige Terme. Diese Fälle setzen sich meist aus einem RefSeq- und zwei Swiss-Prot-Einträgen zusammen. Grund hierfür sind Spleißvarianten von Proteinen eines Gens, die in einzelne

Einträge aufgetrennt werden. Seltener tritt die umgekehrte Kombination auf. Grund hierfür ist das Clustern mehrerer Gene in einem Swiss-Prot-Eintrag, die in RefSeq getrennt voneinander gespeichert werden. Unregelmäßigkeiten in einzelnen Datenbankeinträgen (Spleißvarianten in verschiedenen Einträgen sind in den entscheidenden Features unterschiedlich annotiert), die mit automatischen Analysen nicht bearbeitet werden können, werden in diesem manuellen Schritt mitbehandelt.

6.1.3 Informationsgehalt von Gen- und Proteinnamen

Nicht alle Annotationen, die in Datenbanken als Gen- oder Proteinnamen zur Verfügung gestellt werden, eignen sich zur Erzeugung der in 6.1.1 beschriebenen Relationen. Vor allem solche Eintragungen, mit denen keine Relationen aufgebaut werden können, da sie keine Gen- oder Proteinnamen sind, genauso wie Teilmengen von Namen, die zu einem unverhältnismäßigen Anstieg der Fehlerquote führen, sollen im Folgenden untersucht werden.

Analyse der Länge eines Gennamens

Zur Analyse der Gennamenlänge werden jeweils alle Gennamen einer Datenbank (RefSeq[1], UniProtKB, Ensembl) entsprechend ihrer absoluten Länge (Anzahl alphanumerischer Zeichen inklusive Sonderzeichen wie “-“ oder “/“) in neun Klassen eingeteilt. Die Klassen eins bis acht enthalten jeweils die Gennamen der Längen eins bis acht, die neunte Klasse alle Gennamen, die aus neun und mehr Zeichen bestehen. Die Abbildungen (Abb. 33: links) zeigen für jede der drei Hauptdatenbanken die Organismus-spezifischen Klassenstärken für *H. sapiens* und *M. musculus*.

Die Statistik über die Gennamenlängen für Swiss-Prot enthält Gennamen am häufigsten in den Klassen der Namenlänge vier, fünf und sechs. Erstellt man eine polynomische Trendlinie so verhält sie sich näherungsweise normalverteilt um die Klasse der Länge fünf. In RefSeq und Ensembl ist der Trend zur Normalverteilung in den Klassen eins bis sieben erkennbar. Deutlich zu sehen ist der im Gegensatz zu Swiss-Prot erhebliche Anteil der „Gennamen“ in den Klassen der Länge acht

[1] Zu dieser Analyse werden nur Einträge verwendet, deren Existenz durch Evidenzen (full-length cDNA, ESTs etc.) gesichert ist. RefSeq-Einträge besitzen den Status ’REVIEWED’, ’VALIDATED’ oder ’PROVISIONAL’. In keinem Fall dürfen sie Ergebnis von Genvorhersagen sein, d.h. als ’PROVISIONAL’ gekennzeichnete Einträge werden ausgeschlossen. UniProt-Einträge stammen aus Swiss-Prot.

und „größer als acht“. Einzelanalysen der Klassen acht und „größer als acht“ der Datenbanken RefSeq und Ensembl zeigen den Grund für die Abweichung der Verteilung im Gegensatz zu Swiss-Prot. In diesen Datenbanken treten in den Gennamen-Feldern gehäuft Terme auf, die keine Gennamen darstellen. Es sind hauptsächlich Terme, die aus alphabetischem Präfix mit numerischem Suffix bestehen. Diese sind meist projektbezogene oder automatisch erzeugte Identifier (z.B. 5830435C13Rik oder LOC641199), können aber auch Annotationen wie die Masse eines Proteins sein. Diese Terme werden später noch diskutiert werden.

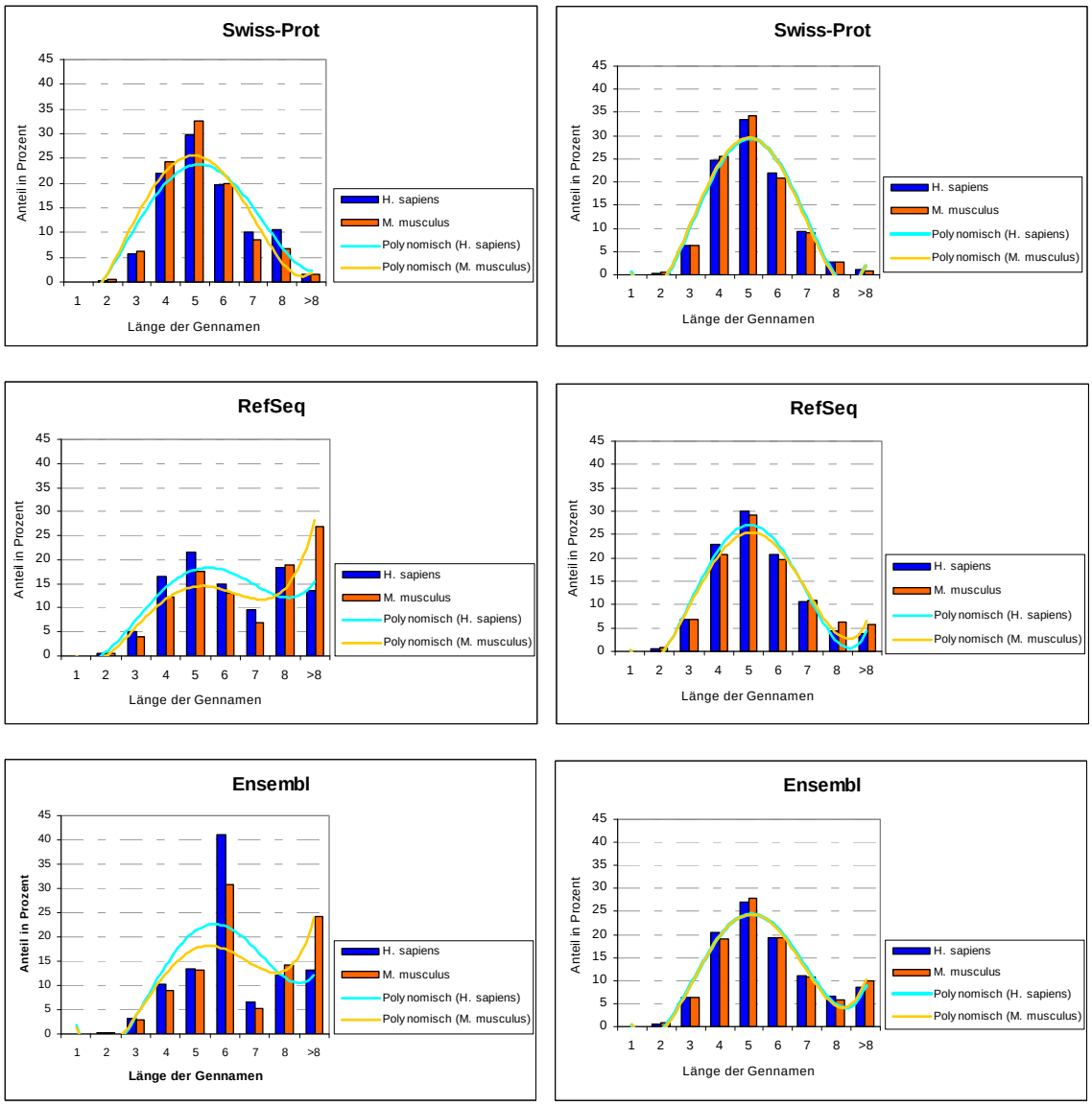


Abb. 33: Statistik zur Häufigkeit von Gennamenlängen in den Referenzdatenbanken. Links abgebildet sind die Häufigkeiten aller Gennamen, rechts die Häufigkeiten aller Terme ohne Identifier.

Abhängigkeit der mehrdeutigen Gennamen von ihrer Länge

Für die Statistik wird jede Längenklasse aus dem vorhergehenden Abschnitt für jede Datenbank einzeln analysiert. Die Menge der zweideutigen Gennamen wird ebenso in neun Klassen aufgeteilt. Danach wird der jeweilige prozentuale Anteil der zweideutigen Gennamen für jede Längenklasse berechnet. Als Referenzwert wird der Anteil aller zweideutigen Gennamen an der Gesamtheit aller Gennamen[1] verwendet (Abb. 34).

Deutlich zu erkennen sind die sehr hohen Anteile der zweideutigen Gennamen in den Klassen eins bis drei in jeder der einzelnen Datenbanken. In der Klasse vier findet jeweils eine Annäherung an den Referenzwert statt, der ab Klasse fünf unterschritten und danach nicht wieder überschritten wird.

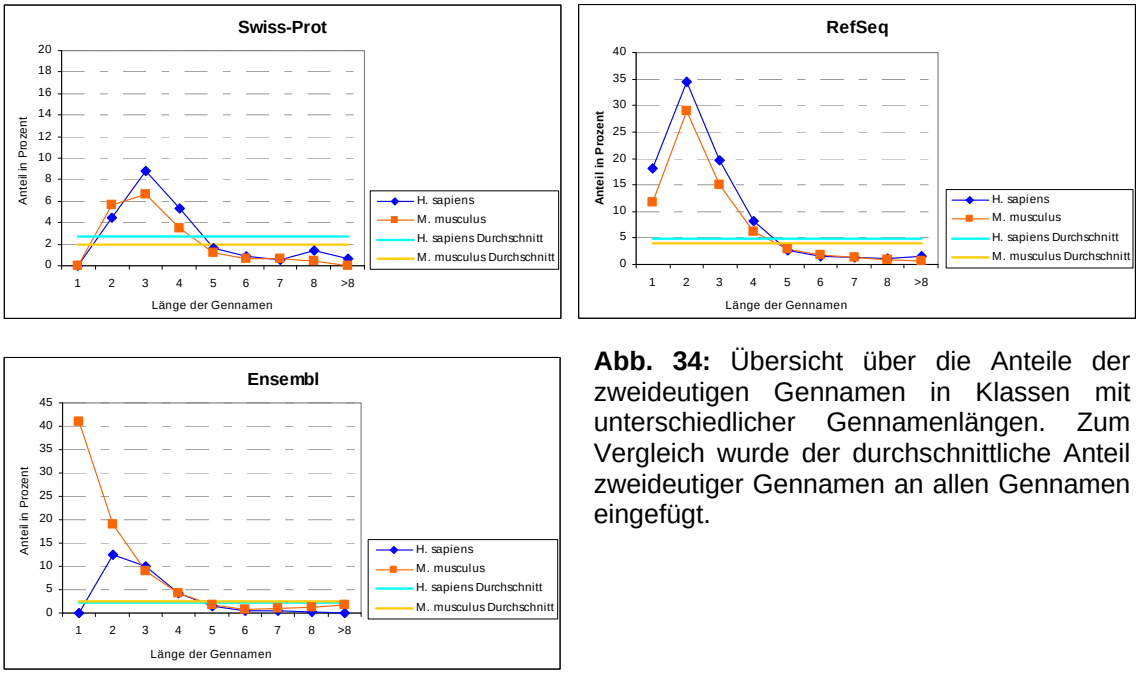


Abb. 34: Übersicht über die Anteile der zweideutigen Gennamen in Klassen mit unterschiedlicher Gennamenlängen. Zum Vergleich wurde der durchschnittliche Anteil zweideutiger Gennamen an allen Gennamen eingefügt.

Um die Fehleranfälligkeit bei der Erzeugung der Relationen und den Aufwand der manuellen Inspektionen möglichst gering zu halten, werden die Relationen nur noch mit Gennamen mit einer Mindestlänge von vier Buchstaben erzeugt. Gennamen der Länge vier werden trotz überdurchschnittlichem Vorkommen auf Grund ihres Anteils an der Gesamtmenge weiterverwendet. Mit dem Ausschluss der fehleranfälligsten Terme kann das durchschnittliche Auftreten eines mehrdeutigen Gennamen um bis zu 50% verringert werden. Abbildung 35 zeigt den Anteil der

[1] Die hierfür verwendeten Gennamenlisten enthalten bereits keine Terme mehr, die projektbezogene oder automatisch erzeugte Identifier darstellen. Ebenso gehen Masseangaben nicht in die Statistik ein.

zweideutigen Terme aufgeteilt in zwei Klassen: Gennamen mit einer Länge von bis zu drei Buchstaben und Namen mit vier und mehr Buchstaben.

Trotz der Nachteile von Gennamen mit weniger als vier Buchstaben, lassen sich diese in CRONOS suchen. Zum Beispiel liefert eine Suchanfrage mit dem Tumorsuppressor “p53“ die entsprechenden Einträge mit dem offiziellen Gennamen “TP53“.

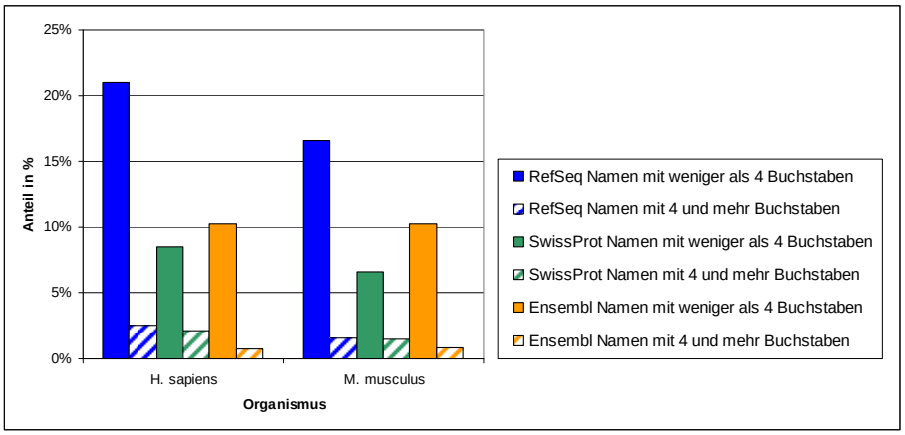


Abb. 35: Anteil der mehrdeutigen Gennamen innerhalb der Gruppe der nicht verwendeten und verwendeten Gennamen.

6.1.4 Irrtümlich als Gen- oder Proteinnamen gespeicherte Terme

Zur Identifizierung von Termen, die in den drei Referenzdatenbanken als Gen- oder Proteinnamen gespeichert sind aber keine anerkannten Gen- oder Proteinnamen darstellen, wird wie folgt vorgegangen: Zuerst werden die betroffenen Terme aller Datenbanken in einer nicht redundanten Liste gespeichert und sortiert. Auffallend häufig auftretende Kombinationen aus einem gemeinsamen alphabetischen Präfix und differierenden numerischen Suffixen wurden manuell auf ihren Informationsgehalt und Eignungsgrad hin überprüft. Für nicht geeignet erwiesen sich automatisch erzeugte und projektbezogene Identifier, Angaben über Eigenschaften eines Genprodukts, genauso wie die Genklasse der olfaktorischen Rezeptoren und zu allgemein formulierte Terme wie “CALCIUM BINDING PROTEIN“ oder “HISTONE FAMILY MEMBER“.

Automatisch erzeugte Identifier

Terme, die mit einem “LOC“ beginnen, haben ihren Ursprung in Entrez Gene (Maglott *et al.*, 2007). Sie werden neuen Sequenzen (Genmodellen) zugewiesen, wenn noch kein offizieller Genname vorhanden ist. Sie setzen sich aus dem Präfix

"LOC" und der GeneID[1] zusammen. Diese Terme werden für die Berechnung der Relationen nicht verwendet.

Pojektbezogene Identifier

Identifier dieser Art stammen aus verschiedenen cDNA-Projekten wie "RIKEN Mouse Gene Encyclopedia Project" (Kawai *et al.*, 2001), "National Institutes of Health (NIH) Full-Length cDNA Project" (Gerhard *et al.*, 2004), "Kazusa cDNA Project" (Nagase *et al.*, 2006) oder "Full-Length long Japan cDNA Project" (Okazaki *et al.*, 2004; Ota *et al.*, 2004). Weitere Informationen zu diesen Projekten finden sich im Anhang.

Andere Identifier

Unter diese Kategorie fallen Identifier, die aus zwei Buchstaben gefolgt von mindestens fünf Ziffern bestehen. Diese Regelung betrifft Clone-IDs (zwei Buchstaben, fünf Ziffern) und GenBank-IDs (zwei Buchstaben, sechs Ziffern). Zusätzlich sind Ziffernfolgen betroffen und alle Identifier der Referenzdatenbanken UniProtKB (Primary Accession, Secondary Accession und Entry-Name), RefSeq (Transkript- und Protein-Identifier) und Ensembl.

Eigenschaften von Genprodukten

Terme wie Massenangaben, Angaben über die Lokalisation eines Gens und EC-Nummern sind nicht eindeutig einem Gen oder Protein zuzuordnen. Sie werden als ungeeignet eingestuft.

Olfaktorische Rezeptoren

Olfaktorische Rezeptoren stellen die größte Gen-Klasse in Mensch und Maus (drei bis fünf Prozent aller Gene (Young *et al.*, 2002)). Charakteristisch für diese Klasse ist die Entstehung vieler Gene durch Genduplikationen und Verlust einzelner Gene durch die Entstehung von Pseudogenen. Diese Gene bilden so genannte Cluster auf dem Genom, die bis zu über 100 Gene besitzen. Da die Gene der olfaktorischen Rezeptoren sich evolutionär schnell weiterentwickeln, hybridisieren mehr als die Hälfte der Gene (*H. sapiens*) auf mehr als eine Position auf dem Genom. Ein Alignment aller olfaktorischen Rezeptoren in Mensch und Maus zeigt, dass Gene, die nah beieinander liegen oft sehr ähnliche Sequenzen haben, d.h.

[1] Eindeutige ID, die jedem Eintrag in Entrez Gene zugewiesen wird.

es gibt Gene in dieser Klasse, die eine Sequenzidentität von über 90% aufweisen. Diese können sich im gleichen Cluster befinden aber auch auf verschiedenen Chromosomen (Young *et al.*, 2001). Ältere Gennamen für Olfaktorische Rezeptoren, die mit dem Präfix “Olfr“ beginnen, wurden zur Bezeichnung von ganzen Gen-Clustern (Zhang *et al.*, 2002) verwendet und werden deshalb nicht für die Berechnung von Relationen zugelassen. Seither haben sich die Richtlinien des HUGO Gene Nomenclature Committees (HGNC) in Hinsicht auf die Namensgebung für Olfaktorische Rezeptoren geändert. Diese werden nun mit dem Präfix „OR“ und - basierend auf einem evolutionären Modell nach (Glusman *et al.*, 2001) - mit der Angabe über die Familie und Subfamilie des jeweiligen Rezeptors und eine laufende Nummer für jedes einzelne Gen gekennzeichnet (HORDE (Safran *et al.*, 2003) und ORDB (Crasto *et al.*, 2002)) Diese Gennamen werden weiterhin verwendet.

Einige der oben beschriebenen Identifier sind überdies nur in einer der Hauptdatenbanken vertreten, so dass mit ihnen keine Relationen erzeugt werden können. Ein Ausschluss dieser Terme bewirkt vielmehr eine Beschleunigung der Analyse auf mehrdeutige Gen- oder Proteinnamen und der anschließenden Berechnung der Relationen.

6.1.5 Validierung der erzeugten Relationen

Um ein Maß für die Verlässlichkeit der vorgestellten Methode zu erhalten, werden die Sequenzen aus UniProtKB und die translatierten codierenden Sequenzen der referenzierten RefSeq-Einträge miteinander aligniert. Das Alignment wird mit JAligner (http://jaligner.sourceforge.net) berechnet. Die Sequenzidentität wird berechnet als die Anzahl übereinstimmender Aminosäuren geteilt durch die Länge der längeren Sequenz. Dies ist notwendig, um auch bei erheblichen Unterschieden in den Sequenzlängen einen Identitätswert zu erhalten, der die Länge des zu Grunde liegenden Alignments widerspiegelt. Da Sequenzähnlichkeiten zwischen RefSeq- und TrEMBL-Einträgen auf Grund der Tatsache, dass TrEMBL-Einträge meist nur aus Proteinfragmenten bestehen, gewöhnlich sehr gering sind, kann das zu Grunde liegende Alignment angezeigt werden.

6.1.6 Implementierung und Optimierung

Implementierung

Die Software zur Erkennung der zweideutigen Gen- und Proteinnamen und für den Aufbau der Relationen ist in Java implementiert. Die Referenzdatenbanken (Kapitel 2.1) liegen auf MySQL- und PostgreSQL-Servern.

Die Software, die den Zugriff auf CRONOS gewährleistet, ist in Java 1.5 unter Verwendung der JavaBean- und WebService-Technologie implementiert. Als Server für die Webzugriffe wird ein Sun Java System Application Server der Version 9.1 verwendet. Zur Speicherung der Ergebnisse und um effizienten Zugriff auf alle benötigten Daten zu gewährleisten, dient ein MySQL-Server.

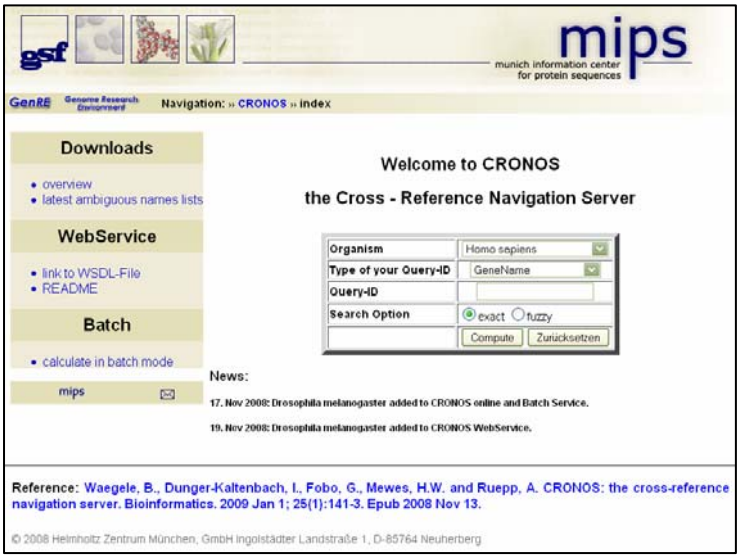


Abb. 36: Startseite zu CRONOS.

Zur Abwehr des als SQL-Injection bekannten Problems, bei dem versucht wird eine Sicherheitslücke auszunutzen um Kontrolle über den Server zu erhalten, werden alle Anfragen mit der Technik der gebundenen Parameter in “Prepared Statements“ ausgeführt.

Optimierung

Mit der Unterbringung der Hauptreferenzdatenbanken auf drei eigenständigen Servern, kann die Analyse der Gennamen auf Mehrdeutigkeit ohne Performanceeinbußen parallel ausgeführt werden. Bei der Berechnung der Relationen ist die Reihenfolge der Verwendung von Gen- und Proteinnamen entscheidend. Es wird mit der Kombination begonnen, die mit der höchsten Wahrscheinlichkeit eine Relation erzeugen kann (1.Genname:Genname; 2.Genname:Proteinname; 3.Protein-

name:Proteinname). Konnte eine Relation erzeugt werden, werden die nachfolgenden Kombinationen nicht mehr betrachtet. Nach demselben Prinzip werden die Relationen zu UniProtKB-Einträgen erzeugt. Zuerst werden Relationen mit Swiss-Prot-Einträgen aufgebaut, nur wenn kein Swiss-Prot-Eintrag gefunden wird, folgt eine Anfrage an TrEMBL-Einträge.

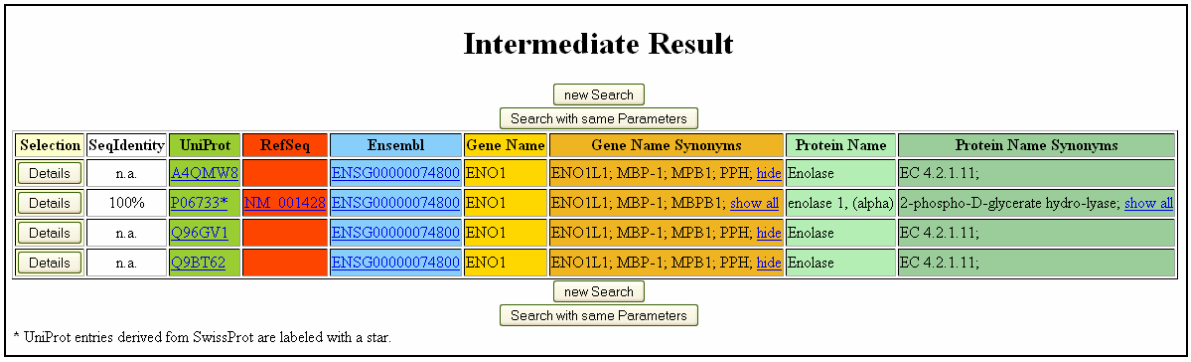
Aktualisierungen der Datenbank, die eine Neuberechnung aller Relationen einschließen, werden in regelmäßigen Abständen durchgeführt. Damit in CRONOS dennoch schnell neue Datenbankeinträge integriert werden können, werden inkrementell Zwischenversionen erzeugt. Dafür werden nur für neue Einträge Relationen mit bereits integrierten CRONOS-Einträgen berechnet und diese importiert. Diese Art des Updates kann nur für Zwischenversionen verwendet werden, da Veränderungen in den bereits in CRONOS enthaltenen Einträgen (Aktualisierung in den Hauptdatenbanken) auftreten können. Aus diesem Grund sollte in bestimmten Zeitintervallen CRONOS vollständig neu berechnet werden.

6.2 Auswertung der Ergebnisse

CRONOS ist eine schnelle und präzise Ressource für die Umrechnung von Identifiern verschiedener Referenzdatenbanken. Im Augenblick enthält CRONOS Einträge von fünf Säugetieren und der Fruchtfliege. Eine Erweiterung von CRONOS ist möglich. Mit den Einträgen von UniProtKB, RefSeq und Ensembl sind die Daten der drei am häufigsten benutzten Gen- und/oder Protein-Datenbanken in das System integriert. Sollte eine Suche nach Referenzen mit Gen- oder Proteinnamen durchgeführt werden, wird das Primärergebnis zusammen mit der Sequenzidentität (siehe 6.1.5) und allen zugeordneten Gen- und Proteinnamen angezeigt (Abb. 37).

Mehr als 17.500 (90%) der humanen Swiss-Prot-Einträge konnten mit 24.000 (95%) der entsprechenden RefSeq-Einträge in Beziehung zueinander gebracht werden. Die größere Anzahl RefSeq-Einträge erklärt sich aus der Tatsache, dass RefSeq Spleißvarianten von Genen in separaten Einträgen abspeichert, wohingegen Swiss-Prot üblicherweise einen repräsentativen Eintrag erstellt, der alle Produkte eines Genes enthält. Zur Validierung der erzeugten Relationen wurde die Identität zwischen den Proteinsequenzen verschiedener Datenbanken berechnet. Dazu wird JAligner verwendet und die Sequenzidentität in den

Primärergebnissen und den Endergebnissen (Abb. 39) angezeigt; das jeweilige Alignment kann zusätzlich angezeigt werden.



Intermediate Result

new Search | Search with same Parameters

Selection	SeqIdentity	UniProt	RefSeq	Ensembl	Gene Name	Gene Name Synonyms	Protein Name	Protein Name Synonyms
Details	n.a.	A4QMW8		ENSG00000074800	ENO1	ENO1L1; MBP-1; MPB1; PPH; hide	Enolase	EC 4.2.1.11;
Details	100%	P06733*	NM_001428	ENSG00000074800	ENO1	ENO1L1; MBP-1; MBPB1; show all	enolase 1, (alpha)	2-phospho-D-glycerate hydro-lyase; show all
Details	n.a.	Q96GV1		ENSG00000074800	ENO1	ENO1L1; MBP-1; MPB1; PPH; hide	Enolase	EC 4.2.1.11;
Details	n.a.	Q9BT62		ENSG00000074800	ENO1	ENO1L1; MBP-1; MPB1; PPH; hide	Enolase	EC 4.2.1.11;

new Search | Search with same Parameters

* UniProt entries derived fom SwissProt are labeled with a star.

Abb. 37: Ergebnis bei Suche mit Gen- und Proteinnamen mit dem Gennamen ENO1.

Die Auswertung des vollständigen CRONOS-Datensatzes für *Homo sapiens* verdeutlicht die hohe Qualität der Relationen. Zirka 85% aller RefSeq-Swiss-Prot-Paarungen weisen eine Sequenzidentität zwischen 90% und 100% auf. Geringere Werte sind häufig das Resultat von Spleißvarianten (RefSeq), die signifikant kürzer sind als die entsprechende Master-Sequenz aus Swiss-Prot (siehe Kapitel 2.1.2.1). In der folgenden Abbildung (Abb. 38) ist die Qualität der Relationen jeweils mit und ohne den Ausschluss der zweideutigen Gen- und Proteinnamen gezeigt. Auf der X-Achse werden die Sequenzidentitäten absteigend in 5%-Schritten angetragen. Auf der Y-Achse jeweils der Anteil der RefSeq-Swiss-Prot-Paarungen, die in diese Sequenzidentitäts-Klasse fallen. So sind knapp 80% der CRONOS-Einträge zu 95% identisch.

Auf Grund der Tatsache, dass 20% der “bona fide”-Relationen Sequenzidentitäten von 90%-95% aufweisen, stellt CRONOS umfangreichere Daten zur Verfügung als Anwendungen, die nur Relationen mit 100% Sequenzidentität zulassen.

Nicht alle Gennamen eignen sich zur Berechnung von Relationen. Gennamen mit weniger als vier Buchstaben sind wesentlich fehleranfälliger in der Verwendung; bis zu 20% dieser Gennamen wurden positiv auf Zweideutigkeit getestet (Abb. 35). Weitere Fehlerquellen sind Gennamen, die aus historischen Gründen verschiedenen Genen zugewiesen wurden, zum Beispiel MDR1, welches früher für einen ABC Transporter (jetzt ABCB1) und für ein Gen der TBC1-Domänen-Familie (TBC1D9) verwendet wurde. Die Konsequenzen dieser mehrfachen Zuordnungen resultiert in falsch berechneten Relationen zwischen Datenbankeinträgen. Um diese zu vermeiden, wurden Listen mit diesen zweideutigen Gennamen erstellt.

Diese wurden manuell geprüft und enthalten ca. 1900 Terme für *H. sapiens*. Weitere Analysen zeigten, dass in 18,1 Millionen PubMed-Kurzzusammenfassungen mehrdeutige Terme 1,25 Millionen mal auftreten. Diese Listen sind wichtig für bioinformatische Anwendungen wie Text-Mining. Die Erstellung von Protein-Protein-Interaktions-Netzwerken unter Verwendung von Gen- und/oder Proteinnamen führt zu einer beachtlichen Zahl falsch-positiver Interaktionen, wenn die mehrdeutigen Namen nicht entfernt werden. Die Umrechnung von Gen- und Proteinnamen in Identifier anderer Datenbanken ist die Hauptanwendung von CRONOS. UniProtKB, RefSeq und Ensembl enthalten Verknüpfungen zu einer Vielzahl zusätzlicher Ressourcen, wodurch mit weiteren Identifiern Umrechnungen gestartet werden können. So kann der Benutzer AgilentMicroarray-Identifier, die in Ensembl enthalten sind, leicht in UniProtKB Identifier umrechnen. Dies erlaubt die einfache Umrechnung aller überexprimierten cDNAs eines Microarray-Experimentes in Identifier anderer Datenbanken.

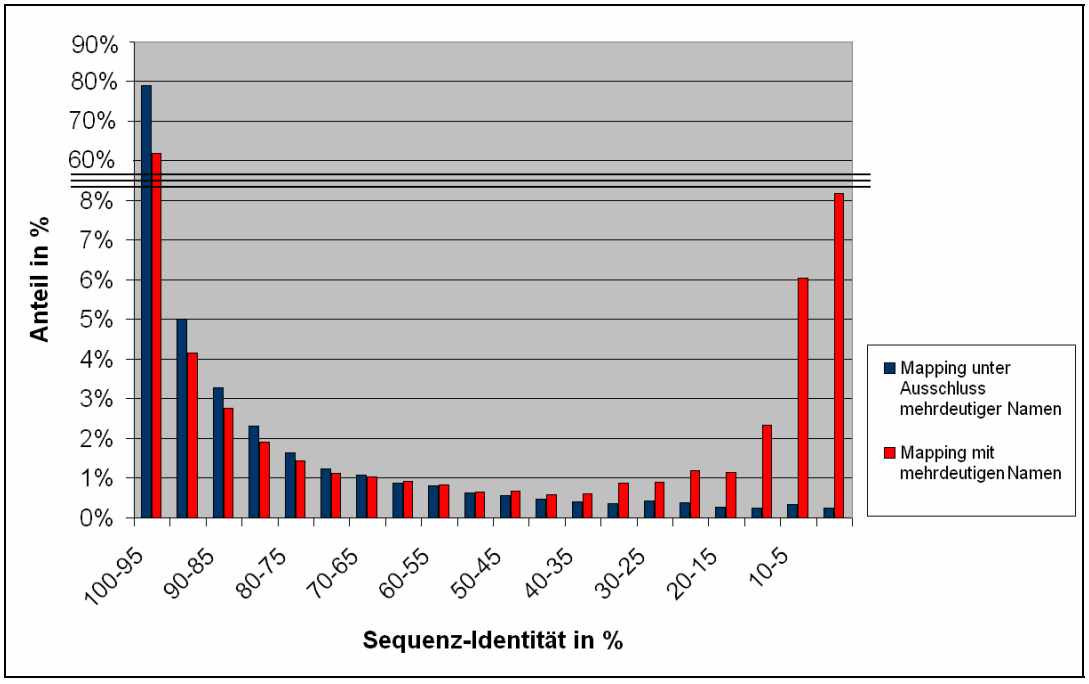


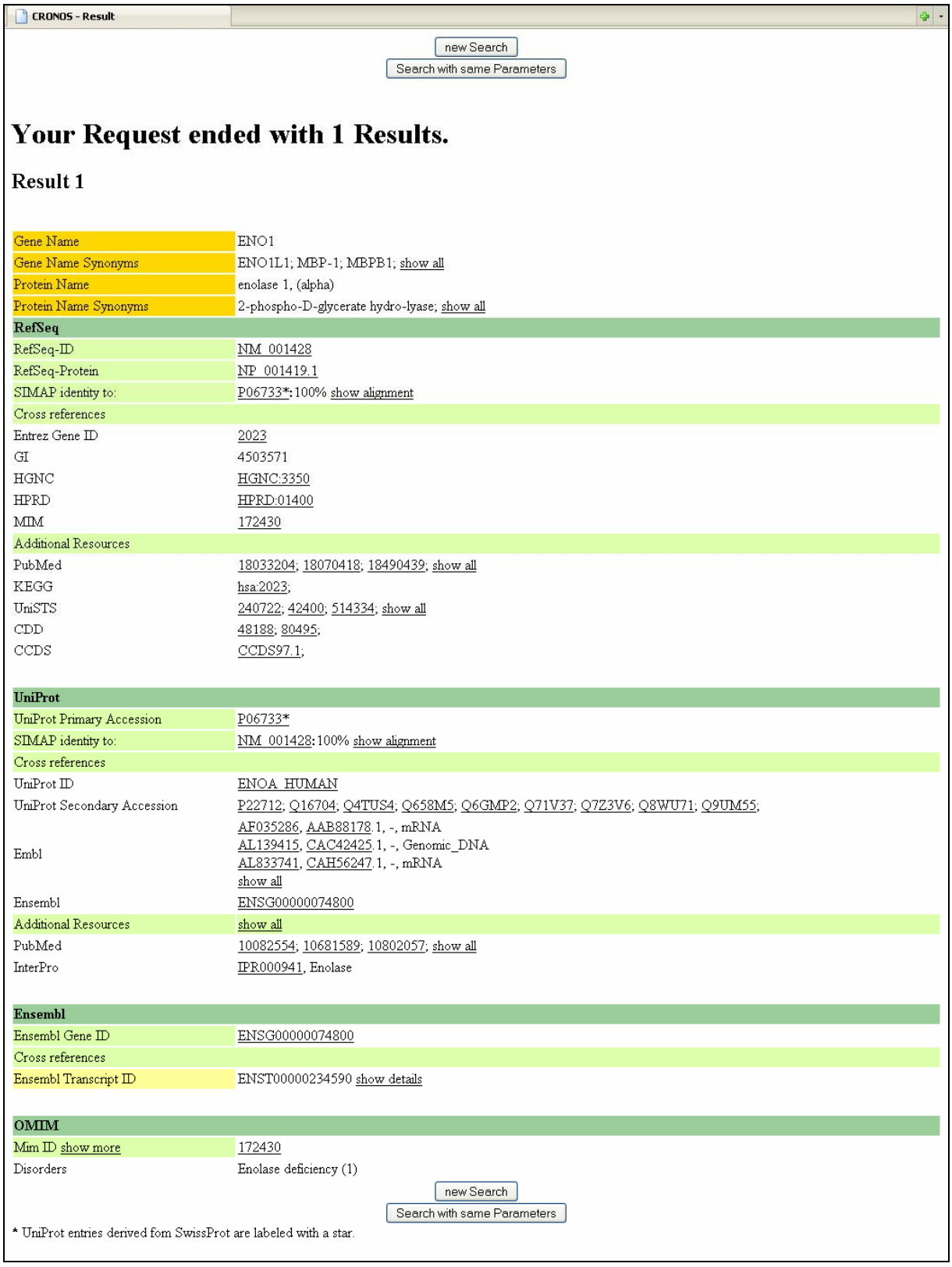
Abb. 38: Qualität der Relationen mit und ohne Ausschluss der zweideutigen Gen- und Proteinnamen.

CRONOS bietet neben Anfragen mit Gen- oder Proteinnamen, die Möglichkeit Suchanfragen mit bis zu 18 verschiedenen Identifier-Typen zu stellen. Konvertierungen ganzer Identifier-Listen können im Batch-Modus vorgenommen werden. Das Ergebnis wird als Datei (CSV-Format; kompatibel zu allen gängigen Tabellenkalkulationsprogrammen) per E-Mail versandt. Für einen vollständig

automatisierten Zugriff wird ein Webservice zur Verfügung gestellt (für weiterführende Informationen siehe Anhang).

Die Ergebnisseiten von CRONOS (Abb. 39) enthalten alle relevanten Informationen der UniProtKB-, RefSeq- und Ensembl-Einträge und falls möglich auch OMIM-Einträge[1] (McKusick, 2007). Zusätzliche Informationen über Stoffwechselwege, die Funktion oder die Domänen eines Proteins werden ebenfalls angezeigt. CRONOS ermöglicht somit einfachen Zugriff auf Cross-Referenzen der am häufigsten genutzten Datenbanken. Manuell geprüfte Listen zweideutiger Gen- und Proteinnamen werden online zur Verfügung gestellt.

[1] OMIM-Einträge werden mit Hilfe ihrer jeweiligen Gennamen CRONOS-Einträgen zugeordnet.



CRONOS - Result

new Search
Search with same Parameters

Your Request ended with 1 Results.

Result 1

Gene Name	ENO1
Gene Name Synonyms	ENO1L1; MBP-1; MBPB1; show all
Protein Name	enolase 1, (alpha)
Protein Name Synonyms	2-phospho-D-glycerate hydro-lyase; show all
RefSeq	
RefSeq-ID	NM_001428
RefSeq-Protein	NP_001419.1
SIMAP identity to:	P06733*:100% show alignment
Cross references	
Entrez Gene ID	2023
GI	4503571
HGNC	HGNC:3350
HPRD	HPRD:01400
MIM	172430
Additional Resources	
PubMed	18033204; 18070418; 18490439; show all
KEGG	hsa:2023;
UniSTS	240722; 42400; 514334; show all
CDD	48188; 80495;
CCDS	CCDS97.1;
UniProt	
UniProt Primary Accession	P06733*
SIMAP identity to:	NM_001428:100% show alignment
Cross references	
UniProt ID	ENOA_HUMAN
UniProt Secondary Accession	P22712; Q16704; Q4TUS4; Q658M5; Q6GMP2; Q71V37; Q7Z3V6; Q8WU71; Q9UM55;
Embl	AF035286, AAB88178.1, -, mRNA AL139415, CAC42425.1, -, Genomic_DNA AL833741, CAH56247.1, -, mRNA show all
Ensembl	ENSG00000074800
Additional Resources	show all
PubMed	10082554; 10681589; 10802057; show all
InterPro	IPR000941, Enolase
Ensembl	
Ensembl Gene ID	ENSG00000074800
Cross references	
Ensembl Transcript ID	ENST00000234590 show details
OMIM	
Mim ID show more	172430
Disorders	Enolase deficiency (1)

new Search
Search with same Parameters

* UniProt entries derived fom SwissProt are labeled with a star.

Abb. 39: Detaillierte Auflistung aller Referenzen, die mit dem Gennamen ENO1 in Relation stehen.

7 Zusammenfassung

Mit dieser Arbeit wurde ein generisches Software-System entwickelt, das heterogene Sequenzdaten automatisch verarbeiten und graphisch anzeigen kann. Transkript- und Proteinsequenzen werden dazu auf genomische Referenzdaten abgebildet. Zu dieser Thematik wurden für eine bessere Performance Optimierungen vorgenommen, die eine vollständige Neuberechnung der Abbildungen bei Assembly-Updates vermeidet und damit die Rechenzeit verringert. Die Erkennung von Spleißvarianten wird bereits erfolgreich zur automatischen Annotation von Einträgen der Mouse functional Genome Database (MfunGD) verwendet. Ebenso wurden Chromosomenabschnitte berechnet, die putative Pseudogene enthalten. Die Analyse erfolgte sowohl für prozessierte als auch für duplizierte Pseudogene. Durch die Modularität des Softwarepaketes ist es auch möglich Peptide entweder auf einem Genom zu positionieren oder aber mit Hilfe eines Protein-Referenzdatensatzes zeitsparend zu charakterisieren. Für die Organismus-übergreifende Analyse von Sequenzhomologien auf DNA-Ebene wurde die SIMAP-Technologie zur Verwendung mit Nukleotidsequenzen angepasst. In die Datenbank wurden bereits mehr als 75.000 verschiedene Säugetiersequenzen aus RefSeq importiert.

Mit Hilfe der oben entwickelten Technologie zur Erstellung von Abbildungen auf Referenzdaten wurde in einer internationalen Kooperation das erste DNA-Microarray des *Callithrix jacchus* erstellt (Datson *et al.*, 2007). Die Annotation der aus dem Hippocampus gewonnenen EST-Sequenzen mit Gennamen erfolgte über die automatische Positionierung der ESTs auf einen Referenzdatensatz. Aus den am besten übereinstimmenden Referenzsequenzen wurden die Gennamen auf die ESTs übertragen. In seltenen Ausnahmen erfolgte die Auswahl des Gennamens manuell. Aus den annotierten Sequenzen wurde eine Teilmenge von 1448 3'-ESTs vom EUPEAH-Konsortium (europäisches Projekt: „Glucocorticoid Hormone Programming in Early Life and its Impact on Adult Health“) ausgewählt, die zusammen mit 68 bereits in GenBank enthaltenen cDNAs auf dem Array repräsentiert sein sollten. Diese wurden mit dem GeneChip® Custom Express® Array Programm von Affymetrix zum Design und zur Fertigung des Arrays weiterverarbeitet. Der endgültige Chip enthält gesamt 1649 Proben, 97 davon als Kontrollproben und 1552 Proben zur Abbildung von 1541 *C. jacchus* Transkripten.

Ca. 95% der Proben wurden aus den neu annotierten Sequenzen ausgewählt. Der DNA-Chip wurde bereits erfolgreich zur Expressionsanalyse der Gene im Hippocampus und anderen Geweben verwendet (Datson *et al.*, 2007). Noch immer wird bei vielen anderen Transkriptom-Analysen trotz der Tatsache, dass Ungenauigkeiten in den Proben die Hybridisierungsintensität beeinflussen (Level der false-negatives oder zu niedrig exprimierter Gene steigt), mit humanen DNA-Chips gearbeitet. Solange es keine bessere Alternative gibt (zum Beispiel mit Hilfe des vollständig sequenzierten Genoms), ist unser DNA-Chip das einzige Hilfsmittel zur Analyse der Genexpression des *C. jacchus*.

Nach dem erfolgreichen Abschluss der Kooperation, erfolgte die Weiterentwicklung der bisherigen Annotations-Pipeline zu einer Web-Applikation (OREST) für die vollautomatische Funktionsannotation von ESTs (Waegele *et al.*, 2008). Im Unterschied zu anderen Annotations-Pipelines wie ESTExplorer (Nagaraj *et al.*, 2007), PartiGene (Parkinson *et al.*, 2004) oder EST2uni (Forment *et al.*, 2008) sind für die Verwendung von OREST keine Installationen von Software oder Datenbanken nötig. Es erfüllt Anforderungen wie einfache Bedienbarkeit, die Analysemöglichkeit von ESTs verschiedenster phylogenetischer Gruppen und ermöglicht eine erste Charakterisierung des Datensatzes.

Zudem können mit Hilfe der schrittweisen Positionierung (auf Referenzdatensatz und Genom) ESTs häufiger einer Referenzsequenz zugeordnet werden als mit anderen beschriebenen Anwendungen. Annotiert werden die Gennamen und die assoziierte Funktion genauso wie mit Genen in Zusammenhang stehende Krankheiten (OMIM, (Hamosh *et al.*, 2005)). Mit der zusätzlichen Integration einer statistischen Analyse können Aussagen über die Über- und Unterrepräsentation von Genen aus bestimmten Funktionsgruppen im Gegensatz zu definierten Referenzdatensätzen getroffen werden.

Mit den Technologien, die aus der Genomik hervorgegangen sind, wie die der Hochdurchsatz-Sequenzierung oder der vergleichenden Transkriptomik (cDNA-Chips) wurden unter anderem das HapMap Projekt (zur Bestimmung der Variabilität des menschlichen Genoms; (The International HapMap Consortium, 2003; The International HapMap Consortium, 2005)) und die SNP-Genotypisierung möglich. Diese Technologie wird in genetischen Studien zu Krankheiten und pharmakogenomischen Studien (Analyse der Reaktion eines Individuums auf medizinische Behandlung) eingesetzt. Mit den genannten Studien und geeigneten

Analysewerkzeugen wird es in naher Zukunft möglich sein, Medikamente zur Bekämpfung von Krankheiten auf den Patienten individuell anzupassen ("personalized medicine" (Ginsburg *et al.*, 2001; Gilad *et al.*, 2006; Kim *et al.*, 2007)).

Zur Auflösung der redundanten Speicherung von Transkript-Sequenzen in verschiedenen Datenbanken und der Möglichkeit Identifier einer Datenbank in die Identifier einer anderen Datenbank umzurechnen, wurde die Web-Applikation CRONOS (Waegele *et al.*, 2008) entwickelt. Damit können nun Daten aus vielfältigen Quellen in nicht redundante Datensätze überführt werden. In der Vergangenheit wurden ähnliche Applikationen wie MatchMiner (Bussey *et al.*, 2003) oder PICR (Cote *et al.*, 2007) entwickelt. Diese ordnen korrespondierende Sequenzen über den Vergleich von Gennamen oder Sequenzen zu. Nachteil bei den Gennamen-basierten Ansätzen sind mehrdeutige Terme, die verschiedene biologische Einheiten beschreiben. Diese werden in MatchMiner nicht berücksichtigt und resultieren in falschen Zuordnungen. PICR umgeht dieses Problem, indem Sequenzen, die eine Identität von 100% aufweisen, miteinander in Relation gestellt werden. Korrespondierende Sequenzen, die sich nur minimal unterscheiden werden mit diesem Ansatz nicht gefunden. Mit CRONOS können mehrdeutige Gen- und Proteinnamen identifiziert werden. Diese werden nicht für die Berechnung der korrespondierenden Sequenzen verwendet. Damit werden falsche Zuordnungen verhindert und Sequenzen, auch wenn sie nicht zu 100% identisch sind, einander zugeordnet. Die Organismus-spezifischen Listen mehrdeutiger Gen- und Proteinnamen können zur Verbesserung weiterer bioinformatischer Anwendungen wie Text Mining eingesetzt werden. Text Mining wird bereits in Bereichen der Funktionsannotation und zur Erkennung von Interaktionen wie Inhibition oder Aktivierung und die Analyse auf Protein-Protein-Interaktionen verwendet. Vollautomatischer Zugriff auf CRONOS (gerade in Hinblick auf die Verwendung in institutsfremden Text Mining Applikationen) wird über Web-Services gewährleistet. Für Einzelabfragen steht eine Online-Version zur Verfügung. Diese kann nicht nur direkt sondern auch über GBrowse verwendet werden.

GBrowse dient vorrangig der graphischen Darstellung der erzeugten Abbildungen auf dem Genom. Es ist eine generische Web-basierte Applikation, die es dem

Benutzer erlaubt, Regionen auf dem Genom zu betrachten oder Suchanfragen mit Gen- oder Proteinnamen oder Basenabfolgen zu stellen. Zu den erstellten “Tracks“ für die verschiedenen Sequenztypen wurden Konfidenz-Tracks erzeugt, die kompakt wiedergeben, wie häufig ein Chromosomenabschnitt von Transkript- und Proteinsequenzen überdeckt wird. Informationen wie Gennamen, Proteinnamen, deren Synonyme und die Sequenzidentität zur genomischen Sequenz sind den abgebildeten Sequenzen zugeordnet worden. Für jede Sequenz ist der Zugriff auf den entsprechenden Eintrag ihrer Ursprungsdatenbank direkt möglich. Zusätzlich können eigene Daten zu den bereits erstellten hochgeladen werden und im Kontext mit diesen betrachtet werden.
Aus dieser graphischen Oberfläche heraus lassen sich die beiden sequenzspezifischen Anwendungen SIMAP/SIMAG und CRONOS für den gewählten Eintrag direkt aufrufen und führen zu den gewünschten Ergebnisseiten (siehe Abb. 40).

In dieser Arbeit wurden bioinformatische Konzepte und Lösungen vorgestellt, die es ermöglichen diverse biologische Daten miteinander in Relationen zu stellen. Dazu gehörten neben der generischen Erzeugung der Abbildungen von Transkript- und Proteinsequenzen auf genomische Referenzdaten, auch die Anwendung der entwickelten Methoden in einer internationalen Kooperation zur Erstellung des ersten *C. jacchus* DNA-Microarrays und die Weiterentwicklung zu einer Pipeline, die die automatische Annotation der Funktion für ganze EST-Datensätze ermöglicht. Weiterhin wurden Analysen zu Spleißvarianten und Pseudogenen durchgeführt und die SIMAP-Technologie auf Nukleotidsequenzen erweitert. Abschließend wurden diese unabhängig voneinander verwendbaren Web-Applikationen zu einer umfassenden Sequenz-Analyse-Plattform verknüpft.

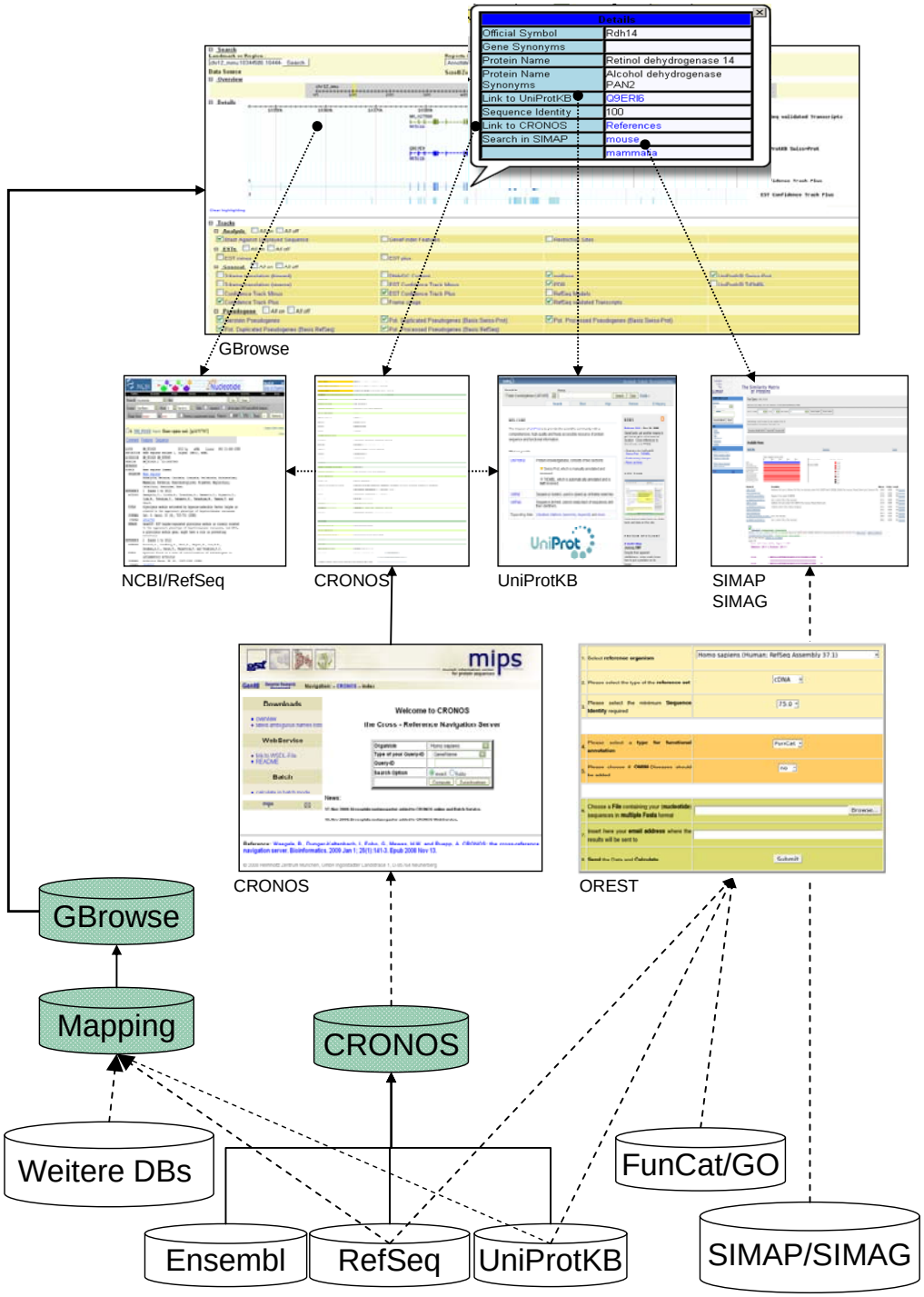


Abb. 40: Übersicht über die in den Kapiteln drei bis sechs entwickelten Applikationen. Darstellung der Mehrschichten-Architektur der gesamten Analyse-Plattform. Von unten nach oben: Grundlagendatenbanken und Datenbanken, die mit Daten aus diesen erzeugt werden (farbig hinterlegt); Applikationsschicht (CRONOS und OREST); Darstellungsschicht; Pfeile: durchgezogene Linie → „wird aus Daten aufgebaut“; gestrichelte Linien→“verwendet Daten aus“; gepunktete Linie→ Verknüpfung zu Web-Applikation oder Web-Seite.

A. Anhang

Weitere Bemerkungen zu Datenbanken

RefSeq

RefSeq	GenBank
von NCBI aus bereits existierenden Daten erzeugt	Autoren übermitteln Daten
NCBI bearbeitet Entries wenn neue Daten bekannt werden	nur Autoren können Veränderungen vornehmen
Je nur ein Entry pro Molekül eines best. Organismus	mehrere Einträge pro locus
	Einträge können sich widersprechen
nur Modellorganismen	alle Organismen
Datenbank exklusiv bei NCBI	Datenaustausch zwischen INSDC-Mitgliedern
verhält sich wie Review	verhält sich wie Primärliteratur
verlinkt identifizierte Proteine und Transkripte	verlinkt identifizierte Proteine
Zugriff über Nukleotid- und Proteindatenbanken	Zugriff über NCBI Nukleotid Datenbanken

Tabelle 2: Unterschiede RefSeq/GenBank (NCBI, Oct. 2002).

Swiss-Prot

Beispiel für das Erzeugen zweier Einträge für Spleißvarianten eines Genes ist das humane Gen ACE. In diesem Fall wurden zwei Einträge erzeugt: P12821 und P22966. Der erste beschreibt zwei Varianten, die in somatischen Geweben vorkommen, der andere eine gewebespezifische Form, die in Spermatocyten vorkommt. Der zweite Eintrag unterscheidet sich in 67 N-terminalen Aminosäuren (AS). Die Beschreibung eines Signalpeptides und zwei Polymorphismen innerhalb dieser 67 AS und eine Veränderung der dreidimensionalen Struktur in einem gesonderten Eintrag machen die Annotation klarer (http://beta.uniprot.org/faq/30).

mirBase

microRNAs sind eine Klasse der nicht Protein-codierenden small RNAs, die zwischen 19 und 24 Nukleotide lang sind. miRNAs sind überall im Genom codiert. Die Mehrzahl der miRNA Gene (61%) ist in den Introns Protein-codierender Gene lokalisiert, sie können aber auch in Exons oder intergenisch vorkommen. miRNAs entstehen durch die Transkription der miRNA-Gene mit RNA-Polymerase II oder III (pol II, pol III). Nach der Transkription folgen weitere Prozessierungen mit verschiedenen Enzymen, bis die reife miRNA vorliegt:

- Transkription des Genes zur primary miRNA (pri-miRNA) mit 5' Cap und 3' poly-(A) tail mit pol II oder III (jedes Gen einzeln oder polycistronisch).
- Pri-miRNA wird von einem Komplex, der aus RNAse III (Drosha) und DGCR8 besteht, zur pre-miRNA zurechtgeschnitten (ca. 70 Nukleotide lang zu stem-loop Struktur geformt).
- Transport der pre-miRNA ins Cytoplasma mit dem Transporter Exportin 5
- Prozessierung der pre-miRNA in 19-24 Nukleotide langen doppelsträngigen miRNA:miRNA*-Komplex mit Dicer (RNAse III) und TRBP.
- Die reife miRNA (mature miRNA) verbindet sich mit RISC (RNA-induced silence complex) und reguliert die Genexpression während miRNA* abgebaut werden (Mechanismus unbekannt) *(Zhang et al., 2008)*

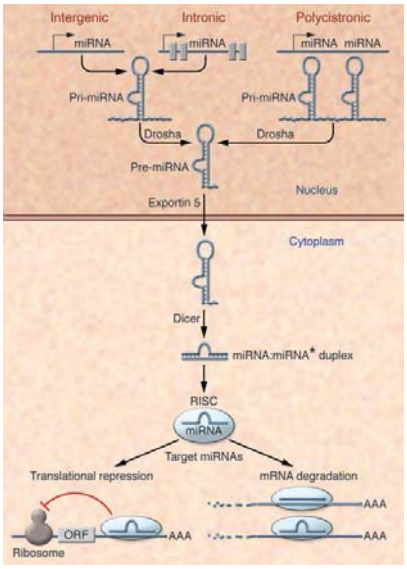


Abb. 41: miRNA Biogenese und Funktion. (van Rooij *et al.*, 2007).

OREST

Dateiformate

Alle Dateien werden als „comma separated value files“ (csv-files, kompatibel mit allen gängigen Tabellenkalkulationsprogrammen) gespeichert.

a) map_reference_gene_genename.txt und map_genome_genename.txt: enthalten die ESTs, die auf ein Referenz-Gen- oder –Protein positioniert werden konnten. Datei enthält zeilenweise:

- query_id (Input EST identifier)
- target_id
- Genname
- Synonyme
- Sequenzidentität des Alignments
- Funktion (GO/FunCat)
- optional: OMIM-Identifier (nur Identifier aus MorbidMap)

b) map_reference_gene_no_genename.txt:

wie oben. Es fehlen die Spalten für Gennamen und Synonyme.

- query_id (Input EST identifier)
- target_id
- Sequenzidentität des Alignments
- Funktion (GO/FunCat)

c) map_genome_no_genename.txt:

ESTs mit einem Hit auf dem Genom mit Überlappung eines Genes, aber ohne verfügbare Annotation:

- query_id
- target_id
- Sequenzidentität des Alignments

ESTs mit einem Hit auf dem Genom ohne Überlappung mit einem Gen. Enthält zusätzlich die Koordinaten des Hits:

- query_id
- Sequenzidentität des Alignments
- Positionierung: Chromosom mit Strang und Exon-Koordinaten

d) no_hit.txt:

enthält alle Identifier derjenigen Sequenzen, die weder auf einen Referenzdatensatz noch auf das Genom positioniert werden konnten.

e) all *.stat files:

enthalten Statistiken über die Über- und Unterrepräsentation annotierter Funktionen zu den Eingabesequenzen.

- annoTerm: FunCat, GO oder OMIM
- popFreq: relative Häufigkeiten der entsprechenden Annotation (pop steht für Population und enthält die Häufigkeiten der Annotationen aller Gene des Referenzdatensatzes eines Organismus)
- popFrac: wie popFreq; als Bruch geschrieben.
- studyFrac: relative Häufigkeiten der entsprechenden Annotation des Eingabedatensatzes als Bruch geschrieben.
- studyFreq: relative Häufigkeiten der entsprechenden Annotation des Eingabedatensatzes
- diffStudyMinusPop: Differenz zwischen den relativen Häufigkeiten (pop und study)
- raweScore
- eScore: eScore indiziert die statistische Signifikanz; z.B. die Wahrscheinlichkeit mit der die beobachtete Anreicherung einer Annotation zufällig auftritt.
- Beschreibung
- contributingGenes enthält alle Gennamen, die mit dieser Annotation verknüpft wurden.

CRONOS

Liste aller von CRONOS unterstützten Identifier-Typen

	H. sapiens	M. musculus	R. norvegicus	B. taurus	C. familiaris
Gene Name	x	x	x	x	x
Protein Name	x	x	x	x	x
RefSeq	x	x	x	x	x
UniProtKB	x	x	x	x	x
Ensembl Gene ID	x	x	x	x	x
Ensembl Transcript ID	x	x	x	x	x
Ensembl Protein ID	x	x	x	x	x
GI	x	x	x	x	x
GeneID	x	x	x	x	x
HGNC	x				
MfunGD		x			
MGI		x			
MIM ID	x				
Morbid ID	x				
EMBL	x	x	x	x	x
PIR	x	x	x	x	x
ORF Names	x	x	x	x	x
dbSNP	x	x	x	x	x
UniSTS	x	x	x	x	x
CDD	x	x	x	x	x
CCDS	x				
Affymetrix gene chips	15x	11x	8x	x	x
agilentcgh	x	x	x		
agilentprobe	x				

Tabelle 3: Organismus-spezifische Liste der möglichen Eingabe-Identifier.

Typ	Beschreibung	Web-Zugriff
RefSeq	Reference Sequence Project	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProtKB		http://www.uniprot.org/
Ensembl		http://www.ensembl.org/index.html
GI	GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
GeneID	Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
HGNC	Hugo Gene Nomenclature Commitee	http://www.genenames.org/
MfunGD	Mouse functional Genome Database	http://mips.gsf.de/genre/proj/mfungd/
MGI	Mouse Genome Informatics	http://www.informatics.jax.org/
MIM/Morbid ID	Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
EMBL	EMBL-EBI	http://www.ebi.ac.uk/
PIR	Protein Information Resources	http://pir.georgetown.edu/
dbSNP	Single Nucleotide Polymorphism	http://www.ncbi.nlm.nih.gov/projects/SNP/
UniSTS	sequence tagged sites	http://www.ncbi.nlm.nih.gov/sites/entrez?db=unists
CDD	Conserved Domain Database	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
CCDS	Consensus CDS Database	http://www.ncbi.nlm.nih.gov/CCDS/

Tabelle 4: Identifier-Typen mit Beschreibung und Web-Link.

Projektbezogene IDs

RIKEN Mouse Gene Encyclopaedia Project

In diesem Projekt wurde versucht, einen möglichst vollständigen Satz der transkribierten Sequenzen der Maus (Mus musculus) zu erhalten. In diesem Projekt erzeugte Klone mit signifikanter Ähnlichkeit zu bereits bekannten Genen wurden mit deren Gennamen oder Proteinnamen annotiert. Klone, die unbekannte Gene enthielten bekamen „Gennamen“ der Form ’<*Riken Clone Identifier*>Rik; Riken cDNA <*Riken Clone Identifier*> gene’ zugewiesen; z.B.: 2310022A10Rik; RIKEN cDNA 2310022A10 gene (Kawai *et al.*, 2001).

National Institutes of Health (NIH) Full-Length cDNA Project

Das Ziel dieses Projektes ist es, für jedes Gen ausgewählter Säugetiere (Mensch: Homo sapiens, Maus: Mus musculus, Ratte: Rattus norvegicus, Rind: Bos taurus) Klone zur Verfügung zu stellen, die jeweils einen vollständigen offenen Leserahmen (complete open reading frame; complete ORF) enthalten. Diese Sammlung genetischer Sequenzen ist auch als Mammalian Gene Collection (MGC) bekannt. Die Sequenzen sind in GenBank mit dem Präfix MGC: gekennzeichnet (MGC133001) (Gerhard *et al.*, 2004).

Kazusa cDNA project (KIAA und mKIAA)

Das Kazusa cDNA Projekt wurde ins Leben gerufen um humane cDNAs in ihrer Gesamtheit zu sequenzieren und damit die codierenden Sequenzen unbekannter Gene vorherzusagen. Insbesondere wurden lange mRNA-Sequenzen betrachtet. Den unbekannten Genen wurden systematisch IDs zugeteilt, die sich aus KIAA (“KI” steht für “Kazusa DNA Research Institute” und “AA” sind Referenz Buchstaben) und einer vierstelligen Nummer zusammensetzen. Da innerhalb dieses Projektes die Funktion der KIAA-Proteine „in vivo“ untersucht wird, wurden zu diesen die homologen cDNAs aus dem Modellorganismus Maus isoliert und sind als mKIAA Gene bekannt (KIAA0340, mKIAA0198) (Nagase *et al.*, 2003; Nagase *et al.*, 2006).

Full-Length long Japan Project

Mit diesem Projekt sollte eine Basis für das humane Transkriptom und die funktionelle Genomik geschaffen werden. Resultat ist die „full-length long Japan“ (FLJ) Kollektion sequenzierter humaner cDNAs. Identifiziert werden können Sequenzen dieses Projektes anhand des Kürzels „FLJ“ (bei Maus „mFLJ“) mit nachfolgender laufender Nummer (FLJ13171, mFLJ00237) (Okazaki *et al.*, 2004; Ota *et al.*, 2004).

Deutsches Krebsforschungszentrum

Das deutsche cDNA Konsortium wurde als Teil des German Genome Projects gegründet. Ziel war die Charakterisierung vollständiger Sequenzen (zum damaligen Stand) neuer humaner Transkripte auf cDNA-Ebene. Jede der 1500 erzeugten cDNAs dieses Projektes enthält die vollständige Protein-codierende Region. Sie sind mit dem Präfix DKFZ gekennzeichnet (DKFZp434F162). (Wiemann *et al.*, 2001).

CRONOS Webservice

Verfügbare Methoden

- public boolean **isinRedList**(String name,String organism_3_letter)
  gibt „true“ zurück, falls „name“ ein zweideutiger Gen- oder Proteinname ist, sonst „false“
- public String **cronosWS**(String input_id, String organism_3_letter, int query_int_id, int target_int_id)
  gibt den entsprechenden Eintrag für “input_id” im Ausgabetyp “target_int_id” zurück

Parameter

- name: Gen- oder Proteinname
- organism3Letter: 3-buchstabige Abkürzung für den Organismus

Organismus	organism3letter
Homo sapiens	hsa
Mus musculus	mmu
Rattus norvegicus	rno
Bos taurus	bta
Canis familiaris	cfa
Drosophila melanogaster	dme

Tabelle 5: CRONOS: Abkürzungen der zur Verfügung stehenden Organismen.

- inputId: Identifier einer best. Datenbank (NM_12345, ENSMUST00004567, etc.)
- queryIntId: Typ des Input-Identifiers als Integer (Tabellen zur Übersetzung der Input-Typen in Integer-Werte siehe unten) (z.B.: ist die inputId NM_12345, dann ist die queryIntId 3; RefSeq)
- targetIntId: Typ der Target-Datenbank als Integer (z.B. falls in Ensembl Gene IDs umgerechnet werden soll, ist die targetIntId 5)

Tabellen zur Umrechnung der Query- und Target Database IDs in Integer

1. Werte für Query_id und Target_id der am meisten gebrauchten Identifier-Typen:

Integer Wert	Query/Target-Typ		Integer Wert	Query/Target-Typ
1	Gene Name		14	HGNC
2	Protein Name			
3	RefSeq		17	MfunGD
4	UniProtKB		18	MGI
5	Ensembl Gene ID			
6	Ensembl Transcript ID			
7	Ensembl Protein ID			
8	GI			
9	GeneID			
10	EMBL			
11	PIR			
12	DBSNP			
13	UniSTS			

Tabelle 6: CRONOS: Umrechnungstabelle für die Query- und Target-Database IDs in Integer für die Referenzdatenbanken.

2. Werte für Query_id und Target_id: Expressions-Analyse, Mensch und Maus

Integer Wert	Query/Target-Typ		Integer Wert	Query/Target-Typ
200	affy_hc_g110		500	affy_mg_u74a
210	affy_hg_u133_plus_2		510	affy_mg_u74av2
220	affy_hg_u133a_2		520	affy_mg_u74b
230	affy_hg_u133a		530	affy_mg_u74bv2
240	affy_hg_u133b		540	affy_mg_u74c
250	affy_u133_x3p		550	affy_mg_u74cv2
260	affy_hg_u95a		560	affy_moe430a
270	affy_hg_u95av2		570	affy_moe430b
280	affy_hg_u95b		580	affy_mouse430_2
290	affy_hg_u95c		590	affy_mouse430a_2
300	affy_hg_u95d		600	affy_mu11ksuba
310	affy_hg_u95e		610	agilentprobe
320	affy_hg_focus			
330	affy_hugenefl			
340	affy_hg_u133_x3p			
350	agilentcgh			
360	agilentprobe			

Tabelle 7: CRONOS: Umrechnungstabelle für die Query- und Target-Database IDs in Integer für die Expressionsanalyse in Mensch und Maus.

3. Werte für Query_id und Target_id: Expressions-Analyse, Ratte, Rind und Hund

Integer Wert	Query/Target-Typ		Integer Wert	Query/Target-Typ
700	affy_rg_u34a		800	affy_bovine
710	affy_rg_u34b			
720	affy_rg_u34c		900	affy_canine
730	affy_rat230_2			
740	affy_rae230a			
750	affy_rae230b			
760	affy_rn_u34			
770	affy_rt_u34			
780	agilentprobe			

Tabelle 8: CRONOS: Umrechnungstabelle für die Query- und Target-Database IDs in Integer für die Expressionsanalyse in Ratte, Rind und Hund.

B. Abbildungsverzeichnis

C. Tabellenverzeichnis

D. Literaturverzeichnis

The Gene Ontology Consortium, 2008. The Gene Ontology project in 2008. Nucleic Acids Res **36**(Database issue): D440-4.

The International HapMap Consortium (2003).The International HapMap Project. Nature **426**(6968): 789-96.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature **437**(7063): 1299-320.

The UniProt Consortium, 2008. The universal protein resource (UniProt). Nucleic Acids Res **36**(Database issue): D190-5.

The yeast genome directory, 1997. Nature **387**(6632 Suppl): 5.

Afshari, C. A., Nuwaysir, E. F. and Barrett, J. C. (1999). Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. Cancer Res **59**(19): 4759-60.

Ahn, N. G. and Resing, K. A. (2001). Toward the phosphoproteome. Nat Biotechnol **19**(4): 317-8.

Altmaier, E., Ramsay, S. L., Graber, A., Mewes, H. W., Weinberger, K. M. and Suhre, K. (2008). Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication. Endocrinology **149**(7): 3478-89.

Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., et al. (2008). Text mining for biology--the way forward: opinions from leading scientists. Genome Biol **9 Suppl 2**: S7.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol **215**(3): 403-10.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**(17): 3389-402.

Arnold, R., Rattei, T., Tischler, P., Truong, M. D., Stumpflen, V. and Mewes, W. (2005). SIMAP--the similarity matrix of proteins. Bioinformatics **21 Suppl 2**: ii42-6.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature **455**(7209): 64-71.

Balakirev, E. S. and Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet **37**: 123-51.

Barker, W. C., Garavelli, J. S., Hou, Z., Huang, H., Ledley, R. S., McGarvey, P. B., Mewes, H. W., Orcutt, B. C., Pfeiffer, F., et al. (2001). Protein Information Resource: a community resource for expert annotation of protein data. Nucleic Acids Res **29**(1): 29-32.

Barker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B. C., Srinivasarao, G. Y., Xiao, C., Yeh, L. S., Ledley, R. S., et al. (2000). The protein information resource (PIR). Nucleic Acids Res **28**(1): 41-4.

Barker, W. C., Garavelli, J. S., McGarvey, P. B., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S., Ledley, R. S., Mewes, H. W., et al. (1999). The PIR-International Protein Sequence Database. Nucleic Acids Res **27**(1): 39-43.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2008). GenBank. Nucleic Acids Res **36**(Database issue): D25-30.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**(7146): 799-816.

Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993). dbEST--database for "expressed sequence tags". Nat Genet **4**(4): 332-3.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007). UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol Biol **406**: 89-112.

Bruford, E. A., Lush, M. J., Wright, M. W., Sneddon, T. P., Povey, S. and Birney, E. (2008). The HGNC Database in 2008: a resource for the human genome. Nucleic Acids Res **36**(Database issue): D445-8.

Burkhardt, K., Schneider, B. and Ory, J. (2006). A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. PLoS Comput Biol **2**(10): e99.

Bussey, K. J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay, W. and Weinstein, J. N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol **4**(4): R27.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., et al. (2005). The transcriptional landscape of the mammalian genome. Science **309**(5740): 1559-63.

Castillo-Davis, C. I. and Hartl, D. L. (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19**(7): 891-2.

Chittur, S. V. (2004). DNA microarrays: tools for the 21st Century. Comb Chem High Throughput Screen **7**(6): 531-7.

Costa, F. F. (2007). Non-coding RNAs: lost in translation? Gene **386**(1-2): 1-10.

Cote, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R. and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics **8**: 401.

Crasto, C., Marenco, L., Miller, P. and Shepherd, G. (2002). Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Res **30**(1): 354-60.

Cuzin, M. (2001). DNA chips: a new tool for genetic analysis and diagnostics. Transfus Clin Biol **8**(3): 291-6.

Datson, N. A., Morsink, M. C., Atanasova, S., Armstrong, V. W., Zischler, H., Schlumbohm, C., Dutilh, B. E., Huynen, M. A., Waegele, B., et al. (2007). Development of the first marmoset-specific DNA microarray (EUMAMA): a new genetic tool for large-scale expression profiling in a non-human primate. BMC Genomics **8**: 190.

Dhiman, N., Bonilla, R., O'Kane, D. J. and Poland, G. A. (2001). Gene expression microarrays: a 21st century tool for directed vaccine design. Vaccine **20**(1-2): 22-30.

Donlin, M. J. (2007). Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics **Chapter 9**: Unit 9 9.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8**(3): 186-94.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., et al. (2008). Ensembl 2008. Nucleic Acids Res **36**(Database issue): D707-14.

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F. and Blanca, J. M. (2008). EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics **9**: 5.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. and Mewes, H. W. (2001). Functional and structural genomics using PEDANT. Bioinformatics **17**(1): 44-57.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res **14**(10B): 2121-7.

Gerstein, M. and Zheng, D. (2006). The real life of pseudogenes. Sci Am **295**(2): 48-55.

Gilad, Y. and Borevitz, J. (2006). Using DNA microarrays to study natural variation. Curr Opin Genet Dev **16**(6): 553-8.

Ginsburg, G. S. and McCarthy, J. J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol **19**(12): 491-6.

Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001). The complete human olfactory subgenome. Genome Res **11**(5): 685-702.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res **34**(Database issue): D140-4.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. and Enright, A. J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res **36**(Database issue): D154-8.

Gross L (2006) When Less Is More: Losing Genes on the Path to Becoming Human. PLoS Biol 4(3): e76

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., Garcia-Martinez, J., et al. (2005). CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res **33**(Database issue): D364-8.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res **33**(Database issue): D514-7.

Henrick, K., Feng, Z., Bluhm, W. F., Dimitropoulos, D., Doreleijers, J. F., Dutta, S., Flippen-Anderson, J. L., Ionides, J., Kamada, C., et al. (2008). Remediation of the protein data bank archive. Nucleic Acids Res **36**(Database issue): D426-33.

Hieter, P. and Boguski, M. (1997). Functional genomics: it's all how you read it. Science **278**(5338): 601-2.

Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. and Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE **2005**(283): pe21.

Hollywood, K., Brison, D. R. and Goodacre, R. (2006). Metabolomics: current technologies and future trends. Proteomics **6**(17): 4716-23.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., et al. (2004). Integrative

annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol **2**(6): e162.

Johnson, S. M., Lin, S. Y. and Slack, F. J. (2003). The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. Dev Biol **259**(2): 364-79.

Johnston, R. J. and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature **426**(6968): 845-9.

Karolchik, D., Hinrichs, A. S. and Kent, W. J. (2007). The UCSC Genome Browser. Curr Protoc Bioinformatics **Chapter 1**: Unit 1 4.

Karro, J. E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P. and Gerstein, M. (2007). Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res **35**(Database issue): D55-60.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., et al. (2001). Functional annotation of a full-length mouse cDNA collection. Nature **409**(6821): 685-90.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res **12**(4): 656-64.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res **37**(Database issue): D767-72.

Kim, S. and Misra, A. (2007). SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng **9**: 289-320.

Krallinger, M., Erhardt, R. A. and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. Drug Discov Today **10**(6): 439-45.

Krallinger, M. and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. Genome Biol **6**(7): 224.

Krallinger, M., Valencia, A. and Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol **9 Suppl 2**: S8.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., et al. (2007). EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res **35**(Database issue): D16-20.

Kurian, K. M., Watson, C. J. and Wyllie, A. H. (1999). DNA chip technology. J Pathol **187**(3): 267-71.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**(6822): 860-921.

Lin, S. Y., Johnson, S. M., Abraham, M., Vella, M. C., Pasquinelli, A., Gamberi, C., Gottlieb, E. and Slack, F. J. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. Dev Cell **4**(5): 639-50.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res **35**(Database issue): D26-31.

Maquat, L. E. (2002). Nonsense-mediated mRNA decay. Curr Biol **12**(6): R196-7.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet **9**: 387-402.

Matlin, A. J., Clark, F. and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol **6**(5): 386-98.

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet **80**(4): 588-604.

Mercer, T. R., Dinger, M. E. and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. Nat Rev Genet.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. (1999). MIPS: a database for genomes and protein sequences. Nucleic Acids Res **27**(1): 44-8.

Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E. and Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics **85**(1): 1-15.

Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform **6**(1): 44-56.

Nagaraj, S. H., Deshpande, N., Gasser, R. B. and Ranganathan, S. (2007). ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res **35**(Web Server issue): W143-7.

Nagase, T., Kikuno, R. and Ohara, O. (2003). The Kazusa cDNA project for identification of unknown human transcripts. C R Biol **326**(10-11): 959-66.

Nagase, T., Koga, H. and Ohara, O. (2006). Kazusa mammalian cDNA resources: towards functional characterization of KIAA gene products. Brief Funct Genomic Proteomic **5**(1): 4-7.

NCBI (Oct. 2002). The NCBI handbook [Internet], Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; Chapter 1,2,18, GenBank: The Nucleotide Sequence Database, PubMed: The Bibliographic Database, The Reference Sequence (RefSeq) Project.

Nelson, P. T., Wang, W. X. and Rajeev, B. W. (2008). MicroRNAs (miRNAs) in neurodegenerative diseases. Brain Pathol **18**(1): 130-8.

Nita-Lazar, A., Saito-Benz, H. and White, F. M. (2008). Quantitative phosphoproteomics by mass spectrometry: past, present, and future. Proteomics **8**(21): 4433-43.

Okazaki, N., Kikuno, R., Ohara, R., Inamoto, S., Koseki, H., Hiraoka, S., Saga, Y., Kitamura, H., Nakagawa, T., et al. (2004). Prediction of the coding sequences of mouse homologues of FLJ genes: the complete nucleotide sequences of 110 mouse FLJ-homologous cDnas identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries. DNA Res **11**(2): 127-35.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet **36**(1): 40-5.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., et al. (2005). The MIPS mammalian protein-protein interaction database. Bioinformatics **21**(6): 832-4.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A. and Blaxter, M. (2004). PartiGene--constructing partial genomes. Bioinformatics **20**(9): 1398-404.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol **183**: 63-98.

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132**: 185-219.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85**(8): 2444-8.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M. and Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). Hum Genet **109**(6): 678-80.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res **35**(Database issue): D61-5.

Ramsay, G. (1998). DNA chips: state-of-the art. Nat Biotechnol **16**(1): 40-4.

Rattei, T., Arnold, R., Tischler, P., Lindner, D., Stumpflen, V. and Mewes, H. W. (2006). SIMAP: the similarity matrix of proteins. Nucleic Acids Res **34**(Database issue): D252-6.

Rattei, T., Tischler, P., Arnold, R., Hamberger, F., Krebs, J., Krumsiek, J., Wachinger, B., Stumpflen, V. and Mewes, W. (2008). SIMAP--structuring the network of protein similarities. Nucleic Acids Res **36**(Database issue): D289-92.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., et al. (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res **36**(Database issue): D646-50.

Ruepp, A., Doudieu, O. N., van den Oever, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Skornia, C., et al. (2006). The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context. Nucleic Acids Res **34**(Database issue): D568-71.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., et al. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res **32**(18): 5539-45.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., et al. (2003). Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res **31**(1): 142-6.

Sassen, S., Miska, E. A. and Caldas, C. (2008). MicroRNA: implications for cancer. Virchows Arch **452**(1): 1-10.

Schena, M. (1996). Genome analysis with gene expression microarrays. Bioessays **18**(5): 427-31.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270**(5235): 467-70.

Schmidt, T. and Frishman, D. (2006). PROMPT: a protein mapping and comparison tool. BMC Bioinformatics **7**: 331.

Shalon, D., Smith, S. J. and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res **6**(7): 639-45.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol **26**(10): 1135-45.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A. and Soreq, H. (2005). Function of alternative splicing. Gene **344**: 1-20.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., et al. (2002). The generic genome browser: a building block for a model organism system database. Genome Res **12**(10): 1599-610.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**: 41.

Tian, B., Hu, J., Zhang, H. and Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res **33**(1): 201-12.

van Rooij, E. and Olson, E. N. (2007). MicroRNAs: powerful new regulators of heart disease and provocative therapeutic targets. J Clin Invest **117**(9): 2369-76.

Venter, J. C., Levy, S., Stockwell, T., Remington, K. and Halpern, A. (2003). Massive parallelism, randomness and genomic advances. Nat Genet **33 Suppl**: 219-27.

Waegele, B., Dunger-Kaltenbach, I., Fobo, G., Montrone, C., Mewes, H. W. and Ruepp, A. (2008). CRONOS: the cross-reference navigation server. Bioinformatics.

Waegele, B., Schmidt, T., Mewes, H. W. and Ruepp, A. (2008). OREST: the online resource for EST analysis. Nucleic Acids Res.

Watson, S. J., Meng, F., Thompson, R. C. and Akil, H. (2000). The "chip" as a specific genetic tool. Biol Psychiatry **48**(12): 1147-56.

Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., et al. (2001). Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Res **11**(3): 422-35.

Wilming, L. G., Gilbert, J. G., Howe, K., Trevanion, S., Hubbard, T. and Harrow, J. L. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Res **36**(Database issue): D753-60.

Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z. Z., Ledley, R. S., Lewis, K. C., Mewes, H. W., et al. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res **30**(1): 35-7.

Young, J. M., Friedmann, C., Williams, E. M., Ross, J. A., Torres-Priddy, L. and Trask, B. J. (2001). Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum Mol Genet **11**(5): 535-546.

Young, J. M. and Trask, B. J. (2002). The sense of smell: genomics of vertebrate odorant receptors. Hum Mol Genet **11**(10): 1153-60.

Zhang, B. and Farwell, M. A. (2008). microRNAs: a new emerging class of players for disease diagnostics and gene therapy. J Cell Mol Med **12**(1): 3-21.

Zhang, X. and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. Nat Neurosci **5**(2): 124-33.

Zweig, A. S., Karolchik, D., Kuhn, R. M., Haussler, D. and Kent, W. J. (2008). UCSC genome browser tutorial. Genomics **92**(2): 75-84.