DATENBANKSYSTEM FÜR RNA SIGNAL-STRUKTUREN IN ABHÄNGIGKEIT VON IHRER BIOLOGISCHEN HERKUNFT

DISSERTATION

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

Dem Fachbereich Biologie/Chemie vorgelegt von:
Dipl.-Biol. Werner Sälter

Bremen 2003

1. Gutachter: Prof. Dr. D. Blohm

2. Gutachter: Prof. Dr. R. Giegerich

Tag des Promotionskolloqiums:

Danksagung

An allererster Stelle möchte ich mich bei meinem Doktorvater, Herrn Prof. Dr. Dietmar Blohm, für die Betreuung dieser Arbeit, seine Hilfs- und Diskussionsbereitschaft und für seine Förderung und Unterstützung dieser Arbeit bedanken.

Des Weiteren möchte ich mich bei Herrn Prof. Dr. Manfred B. Wischnewsky, Arbeitsgruppenleiter der Abteilung Algebra im Fachbereich Mathematik und Prof. Dr. Otthein Herzog, Arbeitsgruppenleiter der Abteilung Künstliche Intelligenz im Fachbereich Informatik, für die interdisziplinäre Zusammenarbeit bedanken. In diesen Dank mit eingeschlossen sind die Arbeitsgruppenmitglieder Dr. Uta Bohnebeck, Dr. Thomas Waschulzik, Gerald Volkmann und insbesondere Dr. Manfred Nölte, die mir wertvolle Hinweise und Ratschläge zu Hintergrundwissen aus der Informatik gaben.

Bedanken möchte ich mich auch bei Herrn Karl-Heinz Glatting vom Deutschen Krebsforschungszentrum DKFZ. Als Systemadministrator des GENIUSnet war er mir oft eine wertvolle Hilfe bei technischen Problemen mit HUSAR.

Für die freundliche Arbeitsatmosphäre möchte ich mich bei allen Arbeitsgruppenmitgliedern der Arbeitsgruppe Biotechnologie und Molekulare Genetik bedanken. Besonders gilt der Dank meinem Kollegen Dr. Rainer Söller. Die oft langen Diskussionen über molekularbiologische Problemstellungen waren mir sehr wertvoll.

Ganz besonders danke ich meinen Eltern für Ihre Unterstützung während meiner Promotionszeit sowie meiner Freundin Bernarde für ihre Geduld und Aufmunterung.

Inhaltsverzeichnis

Inhaltsverzeichnis

1. Einleitung

1.1. Einführung

1.1.1. Bioinformatik und Genomforschung

Die Bioinformatik hat sich mittlerweile zu einem eigenständigen, neuen Forschungszweig entwickelt, ohne den die aktuellen Fortschritte in der Genomforschung, wie sie sich in der vollständigen Sequenzierung des humanen Genoms *(Macilwain, 2000)* und weiterer Genome von Modellorganismen *(The Arabidopsis Genome Initiative, 2000), (The C. elegans Sequencing Consortium, 1998), (Adams et al., 2000)* zeigen, nicht denkbar wären.

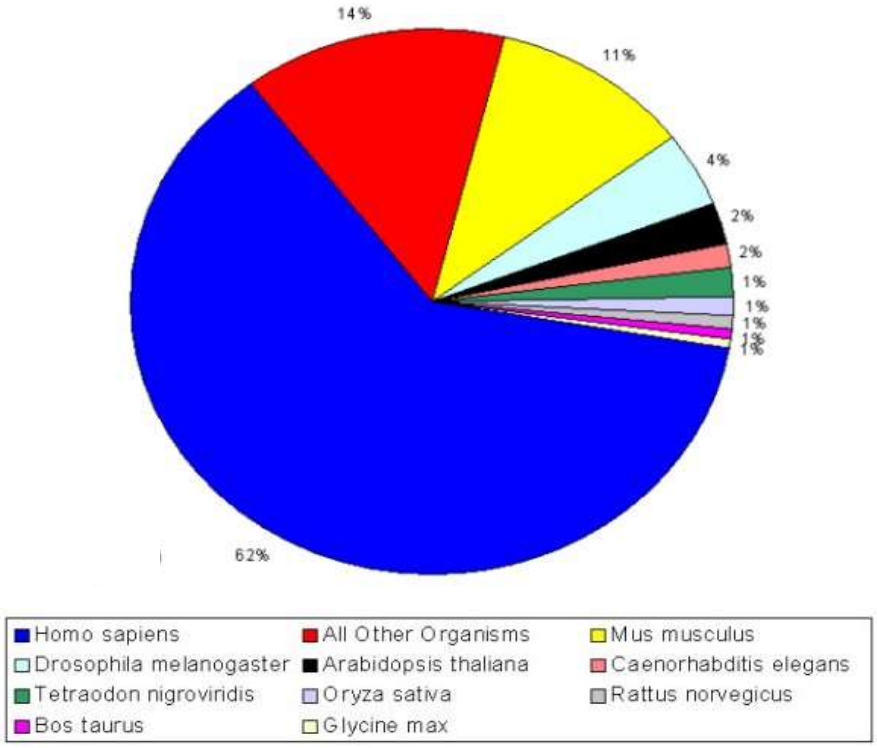


Abbildung 1. Anteil der analysierten Genomsequenzen von Modellorganismen an den Gesamtzahl der EMBL Datenbanksequenzen (EMBL Data Library Version 65)

Ein begrenztes Vorwissen über die biologischen Bausteine des Lebens, die Nukleinsäuren und Proteine vorausgesetzt, können unter Vorwegnahme von Experimenten im Labor mittels Algorithmen bereits mit hoher Sicherheit Vorhersagen über deren molekularbiologische Eigenschaften und Verhalten gemacht werden. Solche computergesteuerten „*in silico*“ Vorhersagen ersetzen teilweise Experimente im Labor und

werden in eigenen biologischen Zeitschriften publiziert *(Wingender et al., 1998)*. Das begrenzte Wissen, auf dem die Bioinformatik dabei aufbaut, liegt in den molekularen Daten internationaler Datenbanken wie EMBL Data Library *(Stoesser et al., 2001)* und GenBank *(Wheeler et al., 2001)* und weiteren etwa 500 molekularbiologischen Datenbanken *(Discala et al., 2000)* in Form strukturierter Beschreibungen von Sequenzen - den sogenannten Datenbank-Annotationen - vor. Durch die große Anzahl von bis zu 350 Genomprojekten *(Bernal et al., 2001)* entsteht aber auch eine große Menge an Sequenzrohdaten, den HTG Datenbankeinträgen („High Throughput Genomsequences“, siehe Abbildung 2.), die noch nicht oder unvollständig beschrieben bzw. annotiert sind, um sie einer sinnvollen Weiterverarbeitung zugänglich zu machen. Beispielsweise stellt die bioinformatische und experimentelle Zuordnung von Funktionen zu den HTG Einträgen des humanen Genoms nach dessen kompletter Sequenzierung auch die Bioinformatik vor massive Kapazitätsprobleme *(Claverie, 2000)*.

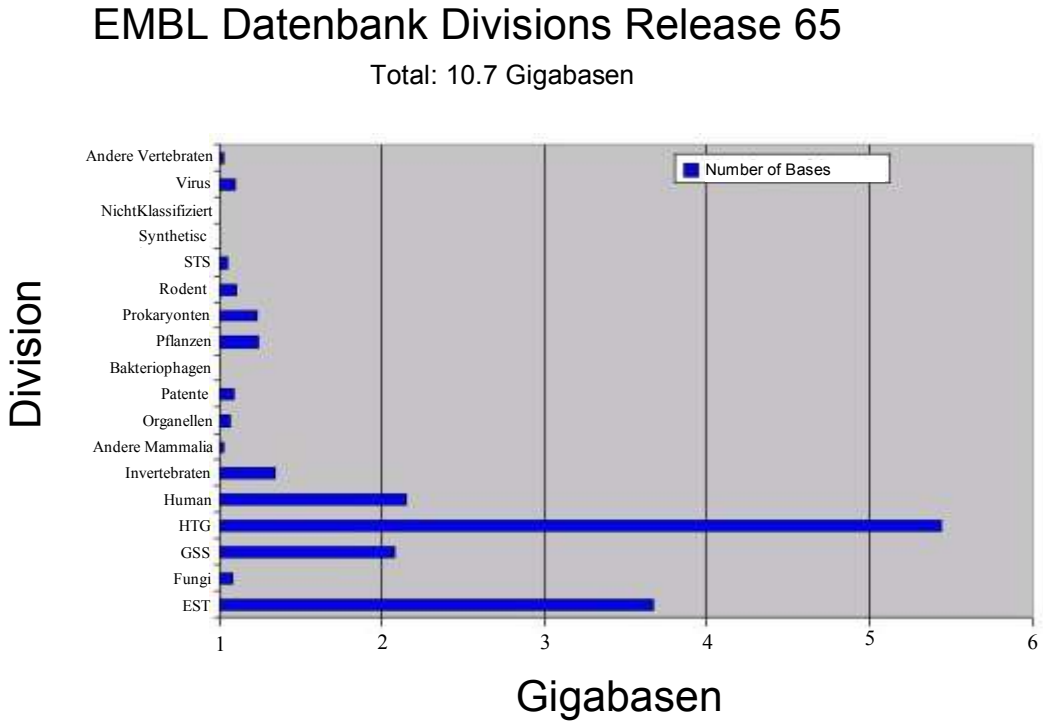


Abbildung 2. Verteilung der EMBL Gesamtbasenmenge. Erläuterung der Bezeichnungen: Synthetic = Artifiziell erstellte also synthetisierte Sequenzen; STS = „Sequence tagged sites“ sind Marker für die Kartierung einzelner Sequenzabschnitte; HTG = „High throughput genom sequences“ sind sogenannte Rohsequenzdaten von Sequenzen die noch unvollständig annotierte sind; GSS = „Genome Surveye Sequence“ sind Sequenzmarker zur Kartierung von Genomen; EST = „Expressed sequence tags“ sind cDNA Sequenzen von exprimierter mRNA.

So gibt es immer noch nach verschiedenen bioinformatischen Berechnungen keine eindeutige Aussage über die Gesamtzahl der in dem menschlichen Genom enthaltenen

Gene *(Smaglik, 2000)*. Sie schwankt zwischen 30.000 – 40.000 Genen *(Lander et al., 2001)* und bis zu 50.000[1] Genen bei *ab initio* Methoden, je nach den für die Vorhersage der Gene genutzten Algorithmen.

In dem Maße also, wie die Datenmengen über sequenzierte Gen- und analysierte Proteinsequenzen und deren Strukturen sowie komplette Genome in der Molekularbiologie anwachsen (siehe Abbildung 3), wachsen die Ansprüche an die Analysewerkzeuge der Bioinformatik. Die bioinformatische Software muss in der Lage sein, immer größere Datenmengen und spezialisiertere Datentypen in Datenbanken abrufbar zu speichern und immer detailliertere Vergleichsmöglichkeiten und Darstellungsformen der Daten bieten *(Benton et al., 1996)*.

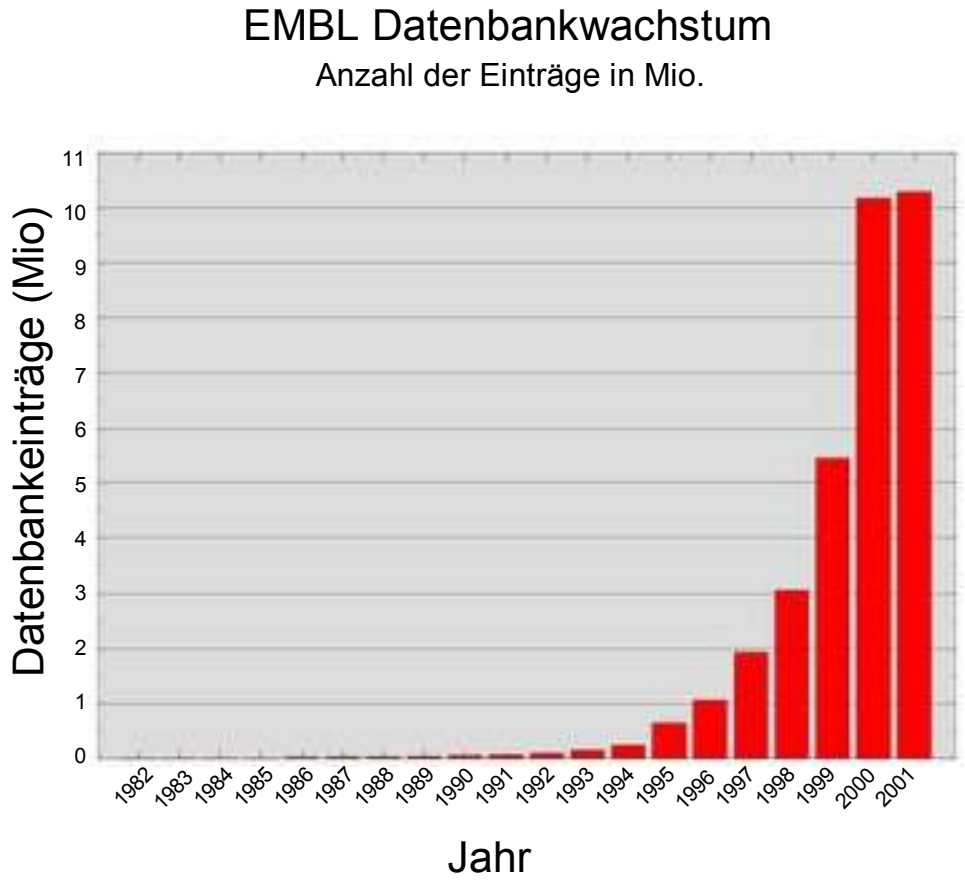


Abbildung 3. Wachstum der Datenbankeinträge in der EMBL Data-Library

1 Unveröffentlichte Genvorhersage mit FGenesh von SOFTBERRY Inc. zu finden unter: http://www.softberry.com/inf/humd_an.html.

Für diese Aufgabenstellungen sind in der Bioinformatik mittlerweile mehr als 500 “Software-Tools” entwickelt worden *(Rodriguez-Tome et al., 1998)*, die sich nach folgenden prinzipiellen Aufgabenstellungen gliedern lassen:

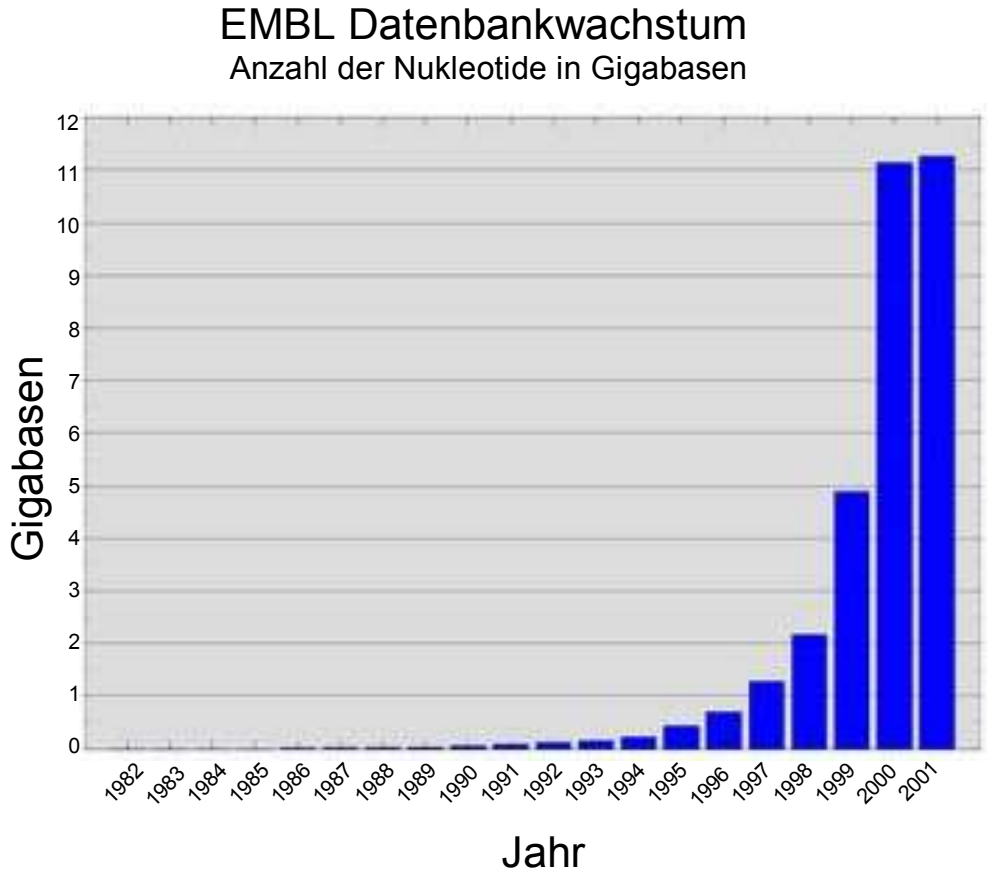


Abbildung 4. Wachstum der Nukleotidzahl in der EMBL Data-Library

- A. Die Planung und Auswertung insbesondere von „Large-Scale“ Sequenzprojekten, wie beispielsweise das „Contig-Assembly“ *(Zhang et al., 1999)* oder bei DNA-Microarray Experimenten *(Pan et al., 2002)*.
- B. Integration von Experiment-, Sequenz- und Strukturdaten in molekularbiologischen Datenbanken, sowie deren Vernetzung *(Karp et al., 1996, Frishman et al., 1998, Macauley et al., 1998)*.
- C. Auswertung gemeinsamer, homologer Merkmale von Sequenzen zur Identifizierung von Genen und regulatorischen Motiven *(Guigo et al., 2000; Brazma et al., 1998)*.
- D. Programme für die Vorhersage molekularbiologischer und physikochemischer Eigenschaften von Sequenzdaten wie beispielsweise die Ableitung der Proteinse-

quenz aus codierenden Bereichen einer Gensequenz oder die Modellierung der Protein *(Guex et al., 1999)* und RNA- *(Schuster et al., 1997)* Sekundär- und Tertiärstrukturen.

- E. Graphikprogramme für die Veranschaulichung der ausgewerteten Daten anhand von Sequenz- und Strukturdarstellungen und deren Kontext *(Kraemer et al., 1998 Dicks et al., 2000).*

Ausgehend von Nukleinsäure- oder Proteinsequenzen mit experimentell aufgeklärten Funktionen signifikante Sequenz- und Strukturübereinstimmungen oder -homologien aufzufinden, ist der Ausgangspunkt für die Konstruktion von metabolischen Netzwerken *(van Helden et al., 2000, Schuster et al., 1999)* Diese Netzwerke führen die Erkenntnisse aus den Forschungsbereichen, die sich mit den unterschiedlichen Eigenschaften von DNA, RNA oder Proteinen beschäftigen, den Genomics, Transcriptomics und Proteomics zusammen, um daraus den gesamten molekularen und zellulären Funktionszusammenhang nachzubilden *(Lengeler et al., 2000)*. Dieser Funktionszusammenhang wird in einem neuen Forschungszweig, der Systembiologie, zusammengefasst. Die Systembiologie generiert dabei eine neue Generation von Software, die die Interaktion der Zellkomponenten untereinander simuliert *(Kitano et al., 2002)*.

1.1.2. Thema und Gliederung

Eine zentrale Rolle spielt die Bioinformatik bei der Aufklärung von Genen. Dies gilt vor allem für deren Extraktion aus den in genomischen Rohdaten – den sogenannten Contigsequenzen - vorliegenden Informationen als auch für deren Charakterisierung in der Auswertung von DNA-Microarray Experimenten. Dazu gehört die Aufklärung ihrer Funktion sowie ihrer Regulation während der Expression in der Zelle. Diese Analyseziele gelten als vergleichsweise kennzeichnend für den Beginn der postgenomischen Ära.
Die geringe Anzahl an Genen (s.o.), die nach der Entschlüsselung des humanen Genoms gefunden worden sind, im Gegensatz zu den vorher prognostizierten 80 -100.000 Genen zeigt, dass ein wesentlicher Anteil der phänotypischen Vielfalt nicht auf der Vielfältigkeit unterschiedlicher Gene sondern der mit ihrer Expression verknüpften Regulationsprozesse beruht.
Das Thema der vorliegenden Arbeit ordnet sich in diesen Kontext ein, insofern es sich vor allem auf die Analyse des regulatorischen posttranskriptionalen Funktionszusammenhangs zwischen Transkription und Translation konzentriert. In diesem regulatorischen Netzwerk spielt die Gesamtheit der RNA, die als das Transkriptom bezeichnet wird, eine wichtige Rolle. Die Arbeit behandelt das Thema der Genexpressionskontrolle durch RNA-Signalstrukturen. Dazu wird die Funktion und der molekulare Aufbau solcher regulatorischen RNA Motive in Abschnitt 1.2 erläutert, sowie der bioinformatische Stand der Analysetechnik solcher Motive beschrieben. Die sich daraus ableitende Fragestellung wird in Abschnitt 1.4 erörtert.

Aus bioinformatorischer Sicht ist die vorliegende Arbeit in erster Linie eine Annotationsarbeit. Es sollen an der Regulierung der Genexpression beteiligte RNA-Signalstrukturen annotiert werden, so dass sie für ein Screeningverfahren nutzbar sind. Die dazu benötigten Datenquellen und Informatiktechnologie werden in Kapitel 2 vorgestellt. Die Zusammenfassung der erzielten Ergebnisse und deren Evaluierung folgt in Kapitel 3 und 4.

1.2. Posttranskriptionale regulatorische mRNA Motive

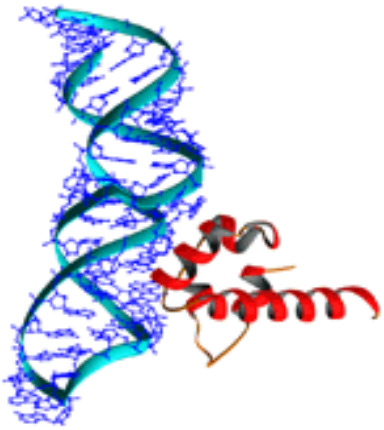


Abbildung 5. Dreidimensionale Darstellung einer RNA-Protein Bindung (RNA: Blau, Protein: Rot).

Seit der Aufklärung des Operonmodells durch Jacob und Monod *(Monod et al., 1961)* sind Nukleinsäuresequenz- und Strukturmotive als Teil des molekularen Regulationsapparates der Genexpression in allen lebenden Zellen bekannt. Sie bilden separat von der genetischen, codierenden Information für die Bildung von Proteinen als Promotor-, Enhancer- oder Attenuator-Strukturen eigene, steuernde Informationseinheiten in Nukleinsäuresequenzen. In jüngerer Zeit wurde zum ersten Mal unter anderem durch Shaw und Kamen *(Shaw et al., 1986)* gezeigt, dass solche Signalstrukturen auch in den nichttranslatierten Regionen am 5' und 3' Ende, den 5' und 3' UTRs („untranslated region"), sowie den eukaryotischen Introns der mRNA vorkommen. Sie regulieren durch *in trans* RNA-Protein (prototypische 3D Darstellung der RNA-Protein Interaktion ist in Abbildung 5. dargestellt) und *in trans / in cis* RNA-RNA Interaktion die Genexpression auf posttranskriptionaler Ebene *(McCarthy et al., 1995).* Im Folgenden werden diese Motivsequenzen und deren Sekundärstrukturen als RNA-Signalstrukturen bezeichnet und die Sequenzregionen als posttranskriptionale regulatorische Region, abgekürzt POSTREG.

1.2.1. Posttranskriptionale Funktion regulatorischer mRNA

1.2.1.1. mRNAs steuern ihr eigenes „Processing"

Die mRNA unterliegt nach der Transkription einer nachträglichen, als „Processing" bezeichneten Veränderung ihrer genetischen Information durch Umordnung ihrer molekularen Struktur. Dazu gehört das als „Splicen" bezeichnete Entfernen der nicht-codierenden Introns in eukaryotischen Genen *(Smith et al., 2000),* das Ersetzen einzelner Adenosinbasen durch Uracil beim „RNA-Editing" *(Kable et al., 1996)* und die Ergänzung der 5' und 3'Enden um die 5'Cap-Struktur und den Poly-A Schwanz. Bei dieser Reifung der Pre-mRNA sind Signalstrukturen in erheblichem Maße beteiligt. Aber auch bei anderen, mit dem Reifeprozess zusammenhängenden Ereignissen wie

zellulärem Transport der mRNA, Bestimmung der Halbwertszeit - also der Stabilität der mRNA - sowie der Translationsinitiation und Turn-over Rate als auch als Transkriptionsfaktor, spielen mRNA Motive eine wichtige Rolle. Folgende Tabelle gibt einen Überblick über Beispiele von Funktionen, die durch Postregs gesteuert werden.

Signalstruktur- oder Genname	*Funktionalität*	*Erstveröffentlichung*
MyoD	Transkriptionsfaktor	*(Rastinejad et al., 1993)*
Bicoid	Transport im Cytoplasma	*(Macdonald et al., 1988)*
AU-reiches Element	Transport zwischen Nucleus und Cytoplasma	*(Katz et al., 1994)*
CPE	Poly-Adenylierung	*(Fox et al., 1989)*
Fem-3	Poly-Deadenylierung	*(Ahringer et al., 1991)*
Mst(3)	Unterdrückung der Translation	*(Schafer et al., 1990)*
IRES Element	Initiation der Translation	*(Pelletier et al., 1988)*
IRE (5'UTR)	Stabilisierung	*(Aziz et al., 1987)*
AU-reiches Element	Destabilisierung	s.o.
Autokatalytische Introns Gruppe I	Splicen	*(Kruger et al., 1982)*

Tabelle 1. Posttranskriptionale Kontrollfunktionen der mRNA Signalstrukturen

Die genannten Kontrollfunktionen können sich, wie man anhand von Tabelle 1. sehen kann, überschneiden. So bestimmt der Grad der Adenylierung auch die Stabilität der mRNA, so dass eine Stabilitätsregulation sowohl direkt über destabilisierende Postreg-Sequenzen als auch über die Regulierung des Adenylierungsgrades erfolgen kann. Somit ist der Poly-A Strang und auch die 5' Cap-Struktur im eigentlichen Sinne eine Signalstruktur *(Preiss et al., 1999).* Diese Elemente werden aber häufig nicht als separate Signalstrukturen aufgeführt, weil sie konstitutionell zu einer gereiften mRNA gehören.

1.2.1.2. Induktion der Signalstrukturwirkung

Die Bindung von Trans- oder Cisfaktoren an eine Postreg-Sequenz ist der molekularbiologische Auslöser, der zur Entfaltung der Steuerwirkung von Signalstrukturen führt. Trans-aktivierende Faktoren sind überwiegend Proteine *(Draper et al., 1999),* aber auch RNA-Sequenzen sind Trans- oder Cisfaktoren. Trans-Faktoren sind separate RNA-Sequenzen wie die gRNA („guide RNA"), die an die mRNA andockt, um beispielsweise die Position der Insertion der Uracil-Base festzulegen *(Kable et al., 1996)*. Ein weiteres Beispiel für die Bindung zweier RNA-Stränge *in trans* ist das „Trans-Splicing" *(Agabian et al., 1990).* Für die posttranskriptionale Regulation der

Genexpression bedeutsame RNA *in cis* Bindungen ist die Interaktion der 5' Cap-Struktur mit dem Poly-A Schwanz *(Preiss et al., 1999)* oder das Selbstsplicen durch autokatalytische Introns *(Kruger et al., 1982)*.

Bei den Auslösern der Aktivität von Signalstrukturen ist zu unterscheiden zwischen der Bindung selbst und dem zellulären Ereignis, das der RNA-RNA oder RNA-Protein Interaktion unmittelbar vorausgeht und die Interaktion bewirkt. Üblicherweise unterscheidet man Gene, deren mRNA einerseits instabil, andererseits nach der Transkription schnell translatiert werden, von den Genen, die von der Zelle bevorratet werden, um auf wechselnde Zellzustände, die eine sofortige Reaktion der Zelle erfordern, antworten zu können. In der folgenden Tabelle sind beispielhaft solche Zustände zusammengefasst, die zu einer Bindung von RNA-Signalstrukturen über Trans- oder Cisfaktoren führen und damit die Signalstrukturwirkung induzieren oder inhibieren.

Auslösender Faktor	*Cis-Trans-Faktor*	*Translation*	*Literatur*
Hypoxia / Glycerol	?	+/-	*(Vassella et al., 2000)*
Stickstoffoxid/Wasserstoffperoxid	IRE-bindendes Protein	+/+	*(Hentze et al., 1996)*
Licht	mRNP-Komplex	+	*(Danon et al., 1994)*
Temperatur	?	+	*(Aly et al., 1994)*
Mechanischer Stress	Integrin	+	*(Chicurel et al., 1998)*
Hormone	mRNP-Komplex	+	*(Nielsen et al., 1990)*

Tabelle 2. Auslösende Stimuli für die Bindung von Signalstrukturen

1.2.1.3. Signalstrukturen steuern komplexe Entwicklungsprozesse

Die Bedeutung der Signalwirkung von Postreg-Sequenzen erschließt sich jeweils anhand der posttranskriptional regulierten Gene. Dies sind häufig Gene für zentrale Schaltprozesse in der Genexpression, die die Differenzierung der Zelle und damit den Entwicklungsprozess des Organismus festgelegen, wie die Beispiele in Tabelle 3. zeigen.

Gene	*Funktionalität*	*Reviews*
Tra-Komplex (Tra-1, Tra-2) D. Melanogaster, C. Elegans	Festlegung des Geschlechts	*(Cline et al., 1996)*
Wachstumsfaktoren und Protoonkogene	Zellwachstum und -teilung	*(Willis et al., 1999)*
Maternale, transkriptional arretierte mRNAs	Festlegung der Tochterzelltypen nach der Zellteilung durch zelluläre Verteilungsmuster	*(Stebbins-Boaz et al., 1997)*
Aconitase, Ferritin, Transferrinrezeptor, Succinate-Dehydrogenase	Regulation des Eisen- und Sauerstoffmetabolismus	*(Theil et al., 2000)*

Tabelle 3. Genabhängige, posttranskriptional regulierte Funktionen

Das letzte Beispiel zeigt, dass umfangreiche metabolische Zusammenhänge, wie der zelluläre Eisenmetabolismus, von einer RNA-Signalstruktur, in diesem Fall dem IRE (Iron Responsive Element), das später noch näher erläutert wird, als ihrer zentralen Schaltstelle gesteuert werden. Ist dieses mutiert, führt es zu einem pathogenen Zustand der Zelle, dass als Hyperferritinemia (OMIM Eintrag #600886) bezeichnet wird. Der metabolische Gesamtzusammenhang der zellulären Eisenregulation ist in der unteren Abbildung 6. gezeigt.

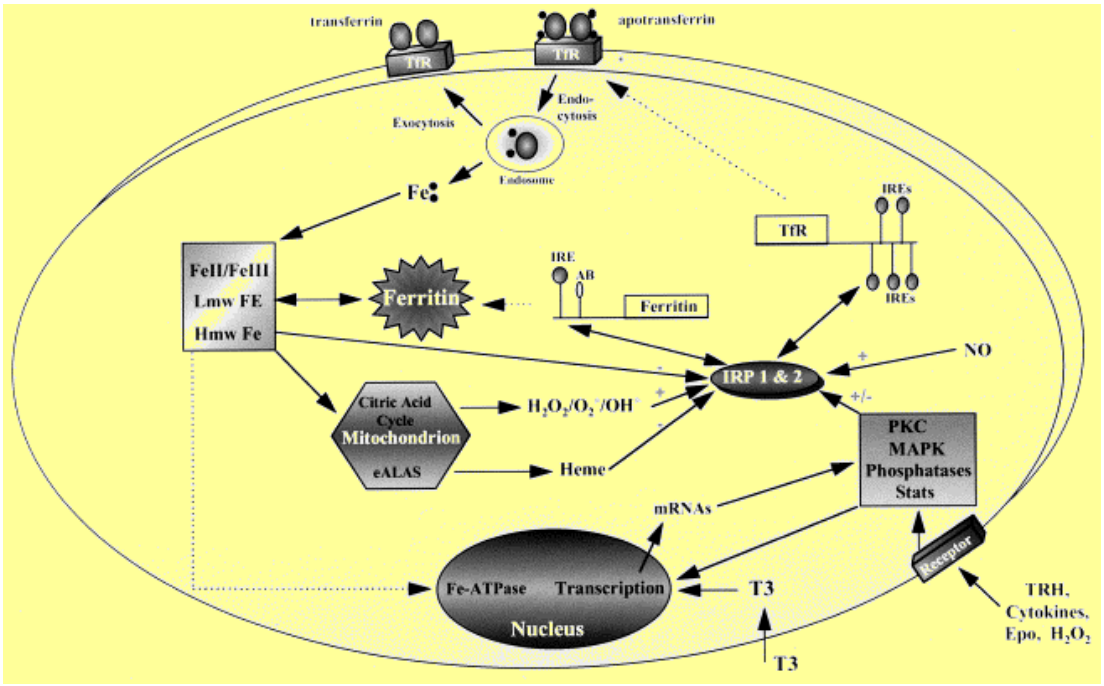


Abbildung 6. Funktionszusammenhang des durch das IRE Element gesteuerten Eisenmetabolismus. Das IRE ist jeweils in den Transkripten des Eisenspeichermoleküls Ferritin, des Eisenrezeptors Tranferrrin und der erythroid-spezifische D-Aminolevulinsäuresynthase (eALAS) enthalten .Aus (Thomson et al., 1999)

Oft sind die posttranskriptionalen Regulationsereignisse, die von Signalstrukturen gesteuert werden, positive Rückkopplungen der auf Transkriptionsebene getroffenen Entscheidungen und dienen als deren Verstärker *(Wickens et al., 1993)*. Bei einigen Beispielen, wie insbesondere in befruchteten Oocyten, ist das posttranskriptionale Regulationsereignis aber selbst Ausgangspunkt beispielsweise für die Festlegung des segmentalen Aufbaus des entstehenden Organismus *(Stebbins-Boaz et al., 1997)*.

1.2.2. Molekulare Komposition regulatorischer mRNA

1.2.2.1. Sequenzmerkmale

Signalstrukturen bestehen aus Sequenz- und Strukturelementen. Sequenzelemente sind definiert als eine eingeschränkt variierende Basenabfolge zwischen verschiedenen mRNA-Sequenzen, die dadurch in diesen Bereichen einem einheitlichen Muster entsprechen, d.h. Consensusmuster oder -sequenzen grenzen sich durch eine eingeschränkte Variabilität der Basen innerhalb des Consensusbereichs vom Rest der Sequenz ab. Innerhalb der Consensussequenzen lassen sich aufgrund der Variabilität der Basen zusätzliche Sequenzbereiche von der restlichen Consensussequenz abgrenzen, die sogenannten Kernregionen oder „core regions“. Sie zeichnen sich durch eine besonders beschränkte Variabilität der Basen an bestimmten Positionen aus und werden durch die vier Basen C, A, U, G gekennzeichnet, während die variableren Regionen durch den IUB-IUPAC Code angegeben werden.

Code	*Base*	*Bedeutung*
A	A	Adenin
C	C	Cytosin
G	G	Guanin
T/U	T	Thymin bzw. Uracil bei RNA
R	A und G	PuRin
Y	T und C	PYrimidin
M	A und C	AMino-Gruppe
K	G und T	Keto-Gruppe
S	G und C	Starke (Strong) Interaktion
W	A und T	schwache (Weak) Interaktion
B	C und G und T	nicht-A, B folgt A im Alphabet
D	A und G und T	nicht-C, D folgt C im Alphabet
H	A und C und T	nicht-G, H folgt G im Alphabet
V	A und C und G	nicht-T / nicht-U, V folgt U im Alphabet
N	A und C und G und T	beliebig (aNy)

Tabelle 4. IUB-IUPAC Basen-Code

Die Steuer- und Kontrollwirkung der Signalstrukturen vor allem in den UTR-Sequenzen wird durch definierte Basenabfolgen dieser Sequenzen vermittelt, die damit die Sekundärstruktur der Sequenz, die auch als Faltung bezeichnet wird, vorgeben. Eine definierte Basenabfolge der mRNA ergibt aber nicht nur eine Sekundärstruktur, sondern kann sich dynamisch in verschiedene Sekundärstrukturen falten, und umgekehrt können verschiedene Sequenzen sich in die gleiche Sekundärstruktur falten.

1.2.2.2. Sekundärstrukturmerkmale

Die Sekundärstrukturmerkmale der Signalstrukturen sind aus immer wiederkehrenden, grundlegenden Strukturelementen aufgebaut *(Mattaj et al., 1993, Nagai, 1992)*:

A. „Loop“ oder „Stemloop“

B. "Bulges"

C. „Interior Loops“ oder "Bubble"

D. “Multibranch loops” oder "Junctions"

E. "Pseudoknots"

F. „Stacking regions“ oder „stems“

G. „Dangeling ends"

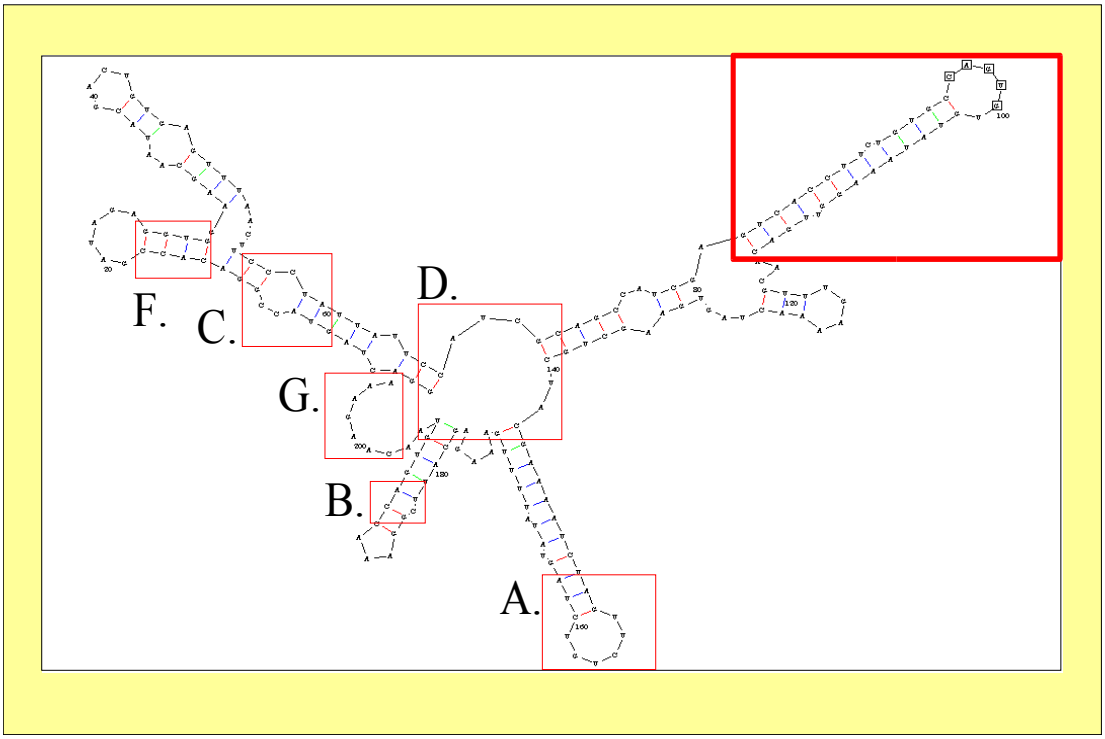


Abbildung 7. Strukturelemente und Beispiel einer Signalstruktur mit Consensussequenz (oben rechts)

Die thermodynamisch kleinste Sekundärstruktur, die gebildet werden kann, ist eine Faltung bestehend aus einem „Loop" mit minimal 3 Basen und einem „Stem" aus minimal 2 Basen *(Freier et al., 1986).* Solche minimalen Sekundärstrukturen kleinster Sequenzabschnitte, aus einer Schleife, dem „Loop", und einer Helix, dem „Stem", zusammengesetzt bezeichnet man als „Stemloops[2]" oder auch als „Hairpins". Sie sind die "Building Blocks" *(Noller et al., 1984)* der globalen RNA-Sekundärstruktur und damit der minimale Bestandteil der Sekundärstruktur einer Postregsequenz. Sie werden hier auch als Teilstrukturen bezeichnet. Viele Signalstrukturen bestehen aus einer Kombination dieser Strukturelemente. Sie kommen als einzelne Elemente in einer Teilstruktur verschieden häufig vor, so dass schon ein einziger Stemloop sehr komplex aufgebaut sein kann. Beispielsweise weist das Iron Responsive Element (IRE, siehe Abbildung 8.) ein spezifisches „consensus pattern" auf, das aus einem "bulge" mit einem Cytosin Rest, einem „hairpin loop" von 6 bp Länge mit dem Consensusmuster CAG(U/A)G und einer minimalen Länge des „stem" von 8 bp besteht. Über dieses „consensus pattern" hinaus kann von den bisher charakterisierten 15 verschiedenen IRE-Typen ein individuelles IRE bis zu 7 zusätzliche „Bulges" mit einer Sequenzlänge von 1-3 bp und einen „stem" von bis zu 37 bp Länge aufweisen

2 „Stemloop" wird synonym als Bezeichnung für den endständigen „Loop" als auch für die gesamte Hairpinstruktur gebraucht.

(Theil et al., 1998). Charakteristisch bei den strukturellen, homologen Merkmalen wie “Bulge” oder “Loop” ist also auch, dass sie eine eng variierende Sequenzlänge haben, sonst wird die Teilstruktur nicht als Bindungsstruktur von dem jeweiligen spezifischen Protein erkannt.

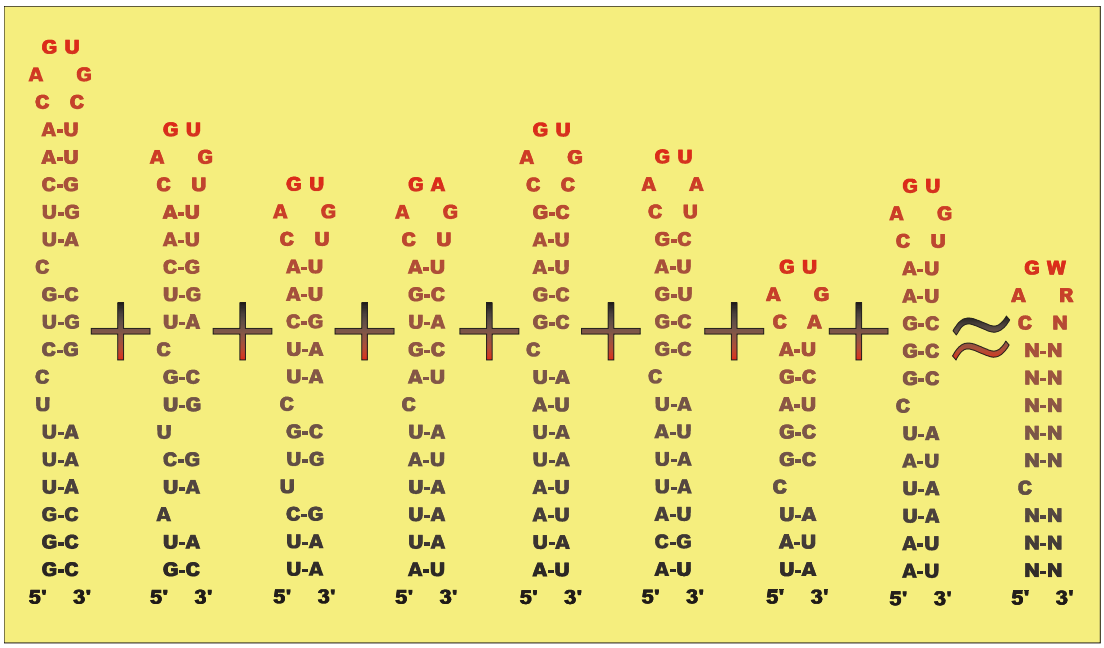


Abbildung 8. Consensussequenz und Struktur aus Beispielen einer Klasse (Beispiel Iron Responsive Element IRE)

Die Besonderheit von UTR-Signalstrukturen im Gegensatz zu regulatorischen Motiven in der DNA, ist, dass die Strukturelemente in einzelsträngigen RNAs häufiger eine zentrale Rolle bei Protein-Nukleinsäureinteraktionen spielen und damit ihre Mustermerkmale komplexer werden.

1.2.2.3. Signalstrukturwirkung basiert auf RNA Thermodynamik und Kinetik

RNA Sekundärstrukturen enthalten den größten Anteil der freien Energie der 3D Gesamtstruktur der RNA *(Flamm et al., 2000)*. Deshalb ist die Berechnung der Sekundärstruktur eine nützliche Näherung an die 3D Gesamtstruktur einer RNA-Sequenz. Die Bildung einer Sekundärstruktur ist ein hierarchischer Prozess *(Brion et al., 1997, Tinoco et al., 1999)*. Die Hairpins, aus denen sich die Gesamtstruktur zusammensetzt, bilden sich zuerst im Bereich von Millisekunden aus. Deshalb können sehr stabile Teilstrukturen, wie sie für Signalstrukturen oft charakteristisch sind *(McCarthy et al 1995),* die Gesamtstruktur der mRNA sehr stark beeinflussen. RNA-Sequenzen können auch alternative Konformationen einnehmen, die dann ebenfalls thermodynamische Energiewerte, vergleichbar den Werten der nativen Konformation, haben

(Emerick et al., 1993, Fresco et al., 1966, Hawkins et al. 1977). Native Konformationen grenzen sich von nicht nativen Konformationen durch hohe Energiebarrieren ab. Alternative Konformationen die in der frühen Phase des Faltungsprozesses eingenommen werden, führen zu einer zeitlich langwierigen Rückfaltung in die native Konformation *(Pan et al., 1999, Pan et al., 1997)*. Dieses kinetische Verhalten kann auch dazu führen, dass zwei unterschiedliche Konformationen (siehe Abbildung 9.) mit zwei verschiedenen Funktionen korrelieren *(Baumstark et al., 1997, Perrotta et al., 1998)*. Dies ist wahrscheinlich bei vielen RNA-Molekülen der Fall, nachgewiesen ist es aber beispielsweise bei dem Virus SV11 *(Biebricher et al., 1982, Biebricher et al. 1992, Zamora et al., 1995)*. Die Kinetik von unterschiedlichen Gesamtstrukturen, die voneinander durch relativ hohe Energiebarrieren getrennt werden und dadurch auch funktionale Schaltungen vermitteln, nennt man „molecular switches“. Sie treten bei den verschiedenen wichtigen Regulationsmechanismen auf, wie die folgende Tabelle zeigt.

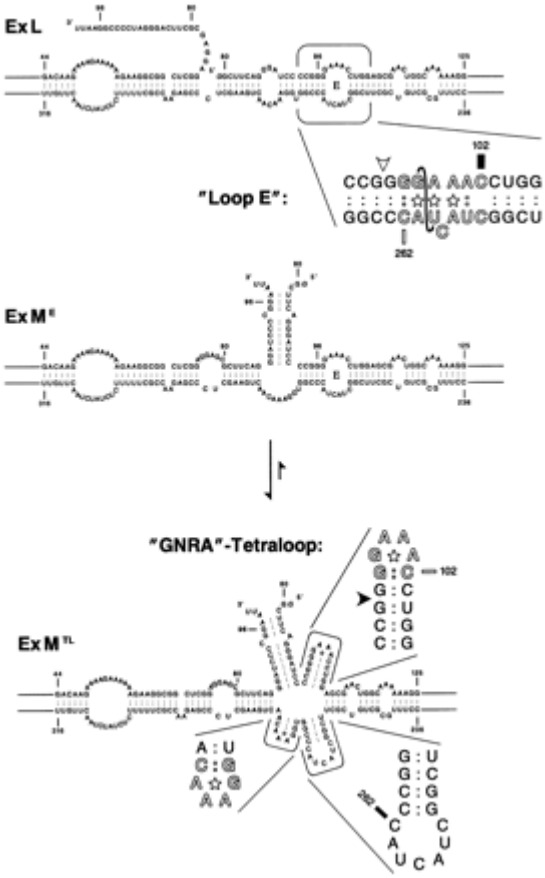


Abbildung 9. Darstellung des Umschaltens in der viroiden Konformation des Kartoffel spindle tuber Viroids. Dargestellt ist das Umklappen des GNRA Tetraloops in die E-Loop Konformation, die die aktive Konformation für die Ligierung ist. (Aus Baumstark et al. 1997)

Organismus / Molekulares System	*Funktion*	*Literaturangaben*
Viren	Viroide Replikation	*(Gultyaev et al., 1998, Hecker et al., 1988, Loss et al., 1991)*
E.coli, B.subtilis	Schaltung zwischen Terminator- und Antiterminator-Hairpins bei der Regulation der Genexpression	*(Babitzke et al., 1993, Putzer et al., 1992, Fayat et al., 1983)*
Hok/soc System R1	Plasmid Replikation	*(Nagel et al., 1999)*
28S rRNA	Elongation (EF-Tu, EF-G) der Proteinsynthese	*(Wool et al., 1992)*
tRNA	Initiation der tRNA Erkennung bei der Proteinsynthese	*(Lodmell et al., 1997, von Ahsen et al., 1998)*
Artifiziell	Siehe mol. Mech. der tRNA	*(Soukup et al., 1999)*
Artifiziell	Schaltung selbstschneidendes Ribozym vs. Ligasefunktion	*(Schultes et al., 2000)*

Tabelle 5. Beispiele für "Molecular switches"

1.3. Bioinformatik der RNA Analyse

1.3.1. Algorithmische Verfahren der Mustersuche

Von der Vielzahl der in der Informatik etablierten Verfahren sind nur einige für bioinformatische oder molekularbiologische Problemstellungen geeignet. Diese Verfahren, die bereits seit den Anfängen der Bioinformatik vielfach eingesetzt wurden, sollen im folgendem Kapitel kurz vorgestellt werden. Sie sind auch in der RNA-Informatik und insbesondere bei Sequenz- und Strukturvergleichen in der Bioinformatik gebräuchliche, algorithmische Methoden.

1.3.1.1. Linguistische Verfahren

Von Beginn an Eingang in das Spektrum bioinformatischer Verfahren haben die von Chomsky begründeten formalen Sprachen gefunden, die jeweils über analysierende oder generative Grammatiken verfügen. Übersichtsarbeiten sind dazu von den Arbeitsgruppen Haussler und Searls *(Searls et al., 1997 Searls et al., 2002 Sakakibara et al., 1994)* erschienen. Eine Grammatik ist ein Quadrupel von vier Mengen:

G = (V,A,R,s)

wobei

- V das Vokabular bestehend aus einer endlichen Menge von Symbolen ist
- A das Alphabet mit einer endlichen Menge an terminalen Zeichen, wobei $A \subset V$ ist.
- R die Menge aller Regeln mit: $R \subseteq (V - A) \times V$
- die Satzsymbole s $\in$ V - A

Anhand der Produktionsregeln lassen sich die verwendeten formalen Sprachen in vier Kategorien einteilen, die reguläre-, kontextfreie-, kontextsensitive- und unbeschränkte Grammatik. Sie beziehen sich jeweils hierarchisch als Teilmenge aufeinander:

$$regulär \subset kontextfrei \subset kontextsensitiv \subset unbeschränkt \quad .$$

Bei der kontextfreien Grammatik ist nur ein nicht-terminales Symbol erlaubt, während bei kontextsensitiven mehrere nicht-terminale Symbole erlaubt sind, solange die Länge der linken Regelseite der rechten Regelseite entspricht. Gibt es keine Beschränkung der Symbole auf der linken und rechten Regelseite, nennt man die Grammatik unbeschränkt. Der Regelaufbau erlaubt einen unterschiedlichen Grad der Detaillierung bei der Modellierung der Abhängigkeiten zwischen den Zeichen eines Satzes einer Sprache. Verschachtelte Abhängigkeiten der Zeichen können mit einer kontextfreien Grammatik repräsentiert werden. Für überkreuzende Abhängigkeiten benötigt man eine kontextfreie Grammatik. Die Basenpaare innerhalb einer RNA Sekundärstruktur sind verschachtelt und können daher mit einer kontextfreien Grammatik dargestellt werden. Allerdings trifft das nicht auf alle Sekundärstrukturen zu. Pseudoknots sind Sekundärstrukturen, die zusätzliche Sekundärkontakte zwischen Stemloops ausbilden und deshalb nur mit einer kontextfreien Grammatik beschrieben werden können. Für diese Problemstellung der vollständigen Beschreibung und Darstellung von RNA Sekundärstrukturen sind erste Lösungsvorschläge erarbeitet worden *(Tabaska et al., 1996)*. Auf die Darstellung eines praktischen Beispiels wird hier verzichtet, da in Kapitel 3.5.1. eine Anwendung einer formalen Sprache im Zusammenhang mit dieser Arbeit vorgestellt wird.

1.3.1.2. Hidden Markov Model HMM

Hidden Markov Modelle für Sequenzvergleiche basieren auf Sequenzprofilen. Ein Sequenzprofil ist nicht wie in der oberen Matrix ein Vergleich einer Base mit einer anderen, der dann mit einem Scorewert versehen wird, sondern eine Matrix, die die Wahrscheinlichkeit angibt, mit der eine Base durch eine andere ersetzt werden kann.

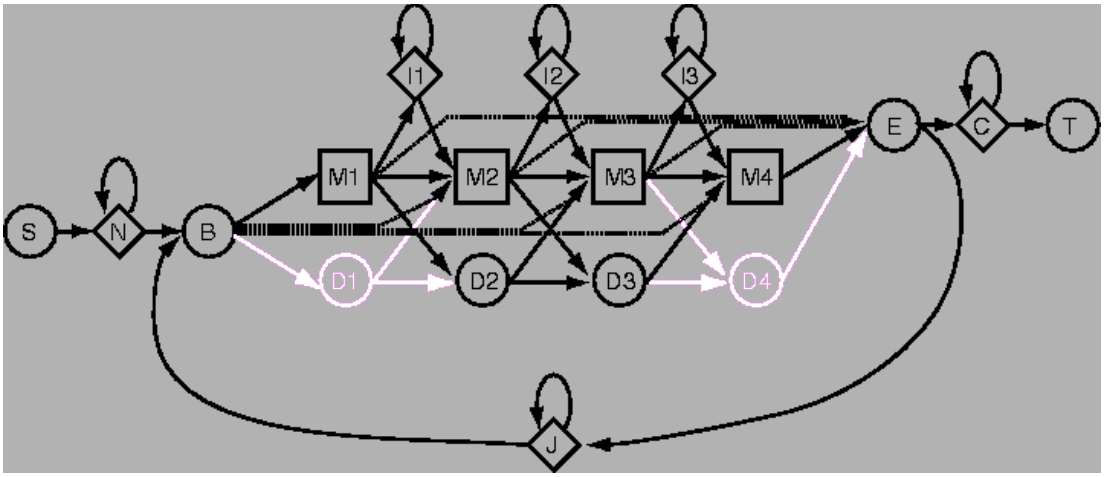


Abbildung 10. Beispiel für ein Hidden Markov Modell. Die Erläuterungen der Abkürzungen der repräsentierten Zustände sind wie folgt: [Mx] Matching Status x. Sender hat K Sendewahrscheinlichkeit [Dx] Löschungszustand x. Nichtsender. [Ix] Einfügezustand x. Sender hat K Sendewahrscheinlichkeit. [S] Start Zustand. Nichtsender. [N] N-terminal nichtalignierter Sequenzzustand. Sender für Zustandsänderung. Sender hat K Wahrscheinlichkeit. [B] Anfangszustand (für Eintreten in das Hauptmodell). Nichtsender. [E] Endezustand (für Verlassen des Hauptmodels). Nichtsender. [C] C-terminal unalignierter Sequenzzustand. Sendet bei Zustandsänderung mit K Wahrscheinlichkeit. [J] Zusammengeführte segmental unalignierter Sequenzzustand. Sendet bei Änderung mit K Sendewahrscheinlichkeit.

Wahrscheinlichkeitsprofile

Gegeben sei als Beispiel ein kleiner Sequenzabschnitt aus einem *Multiplen Alignment* (siehe Kapitel Alignment), das so geordnet ist, dass die große Ähnlichkeit zwischen verschiedenen Sequenzen repräsentiert wird.

CGGSLLNAN--TVLTAAHC

CGGSLIDNK-GWILTAAHC

CGGSLIRQG--WVMTAAHC

CGGSLIREDSSFVLTAAHC

Figure 3. Primäre Struktur von vier verwandten Proteinen.

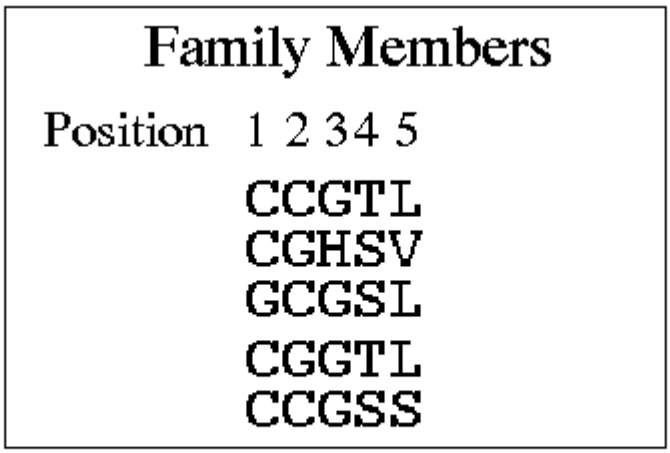
Die Sequenzen können betrachtet werden als ein kleiner Ausschnitt der Geschichte der Evolution dieser Sequenzfamilie. Ein gemeinsamer Vorfahre der Sequenzfamilie könnte so ausgesehen haben:

CGSLIREDWVLTAAHC

Figure 4. Der mögliche gemeinsame Vorfahre.

Während des Reproduktionsprozesses einer Zelle, der Mitose, wird eine identische Kopie der Sequenz erzeugt. Über sehr lange Zeiträume kommt es jedoch zu Fehlern im Kopiervorgang. Diese Fehler können zusammengefasst werden als Deletionen, Insertionen und Substitutionen. Ein Resultat dieser Fehler sind die Abweichungen von dem Vorfahren mit konservierten Anteilen. Die konservierten Positionen erlauben es,

ein statistisches Modell der Sequenzfamilie zu erzeugen. Die folgende Abbildung 12. ist ein vereinfachtes statistisches Profil *(Gribskov et al., 1987)*. Das Modell zeigt für jede Position eine Wahrscheinlichkeitsverteilung. Entsprechend diesem Profil ist das Auftreten von C an der Position 1 0,8, die Wahrscheinlichkeit für G an der Position 2 ist 0,4 und so weiter. Die Wahrscheinlichkeiten sind entsprechend dem beobachteten Auftreten der einzelnen Reste in der Sequenzfamilie kalkuliert.



Position	1	2	3	4	5
Prob(C)	0.8	0.6	–	–	–
Prob(G)	0.2	0.4	0.8	–	–
Prob(H)	–	–	0.2	–	–
Prob(S)	–	–	–	0.6	0.2
Prob(T)	–	–	–	0.4	–
Prob(L)	–	–	–	–	0.6
Prob(V)	–	–	–	–	0.2

Abbildung 11. Statistisches Profil der Sequenzfamilie

Ein Profil gegeben, ist die Wahrscheinlichkeit einer Sequenz das Produkt der einzelnen Wahrscheinlichkeiten der Reste an den einzelnen Positionen. Beispielsweise ist die

Wahrscheinlichkeit der Sequenz CGGSV anhand des Profils:
0.8 * 0.4 * 0.8 * 0.6 * 0.2 = .031.
Ein statistisches Modell vorausgesetzt, kann die Wahrscheinlichkeit einer jeden Basenposition genutzt werden, um einen Score für die Sequenz zu berechnen. Um rechenintensive und computertechnisch gesehen teure Operationen zu vermeiden, werden die Werte dafür logarithmiert. Der Score wird errechnet aus der Addition aller log-Werte der Wahrscheinlichkeiten jeder Position. Die Anwendung der Methode basierend auf dem Logarithmus **e** ergibt den Score der Sequenz CGGSV:

$\log_e(0.8)+\log_e(0.4)+\log_e(0.8)+\log_e(0.6)+\log_e(0.2) = -3.48$

Üblicherweise werden bei der Scorebildung noch weitere Faktoren berücksichtigt. Beispielsweise haben Sequenzen variierende Längen, so dass Insertionen oder Deletionen mit Kosten gewichtet werden können. Die Scores für einzelne Reste im Sequenzstrang sind teilweise positionsabhängig. Diese Positionsabhängigkeit kann ebenfalls durch individuelle Gewichtung berücksichtigt werden, beispielsweise um zu verhindern, dass einzelne Reste die globale Struktur der Sequenz zerstören.

Hidden Markov Models

Hidden Markov Models (HMMs) bietet einen systematischen Ansatz, um Modellparameter zu evaluieren. Das HMM ist vom Typ eines dynamischen statistischen Profils. Wie ein normales Profil wird es durch die Analyse der Verteilung von Aminosäuren in einem Trainingsset verwandter Proteinsequenzen gebildet. Allerdings hat ein HMM eine komplexere Topologie als ein übliches Profil. Es kann als endliche Zustandsmaschine - *„finite state machine“* - betrachtet werden.

Typischerweise bewegen sich endliche Maschinen durch eine Serie von Zuständen und produzieren dabei eine Art von Ausgabe, wenn ein bestimmter Zustand erreicht wird oder wenn sie sich von einem Zustand zum anderen bewegen. Das HMM generiert eine Proteinsequenz durch Senden von Aminosäuren, wenn es durch verschiedene Zustände fortschreitet. Jeder Zustand hat eine Tabelle von Aminosäure-Ausgabewahrscheinlichkeiten, ähnlich den in den Profilmodellen beschriebenen. Es gibt zusätzlich Wahrscheinlichkeiten, um von einem Zustand zum anderen zu gelangen.

Diese strukturellen Ähnlichkeiten machen es möglich ein statistisches Modell einer Proteinfamilie zu erstellen. Das Modell in Abbildung 10 ist eine vereinfachtes statistisches Profil, ein Modell, das die Aminosäure Wahrscheinlichkeitsverteilung für jede Position in der Familie darstellt. Entsprechend diesem Profil ist die Wahrscheinlichkeit von C an der ersten Position 0.8, die Wahrscheinlichkeit von G an der Position 2 ist 0.4 und so weiter. Die Wahrscheinlichkeiten werden entsprechend der zu beobachtenden Frequenz der auftretenden Aminosäuren in der Proteinsequenzfamilie ermittelt.

1.3.1.3. Dynamische Programmierung

Unter der Bezeichnung „Dynamische Programmierung“ versteht man ein algorithmisches Optimierungsverfahren, bei dem ein gegebenes Problem in Teilprobleme

zerlegt wird und dann die Lösung rekursiv berechnet wird. Die Lösungssuche unterstellt, dass jede optimale Lösung des Teilproblems Teil der optimalen Gesamtlösung ist *(Giegerich et al., 2000)*. Das folgende Beispiel behandelt das so genannte globale Sequenzalignment anhand des Needleman – Wunsch Algorithmus *(Needleman et al., 1970)*, bei dem zwei Sequenzen verglichen werden. Das Ziel von Sequenzalignments ist es, die Ähnlichkeit von Sequenzen zu analysieren, indem die Basen von zwei oder mehreren Sequenzen positionsabhängig verglichen werden. Die verschiedenen Typen von Sequenzalignments werden in Kapitel 1.3.4 näher erläutert. Für das Beispiel seien zwei Sequenzen gegeben:

G A A T T C A G T T A (Sequenz #1)
G G A T C G A (Sequenz#2)

Sei M = 11 und N = 7 jeweils für die Länge der Sequenz #1 und Sequenz #2

Ein simples Schema für die Errechnung der Trefferquote ist:

- $S_{i,j}$ = 1 wenn der Basenrest an Position i von Sequenz #1 dem Basenrest an Position j von Sequenz #2 entspricht (match score); oder
- $S_{i,j}$ = 0 bei keiner Entsprechung der Basenreste (mismatch score)
- w = 0 (gap penalty)

Die Dynamische Programmierung beinhaltet drei Schritte:

1. Initialisierung
2. Matrix füllen (scoring)
3. Traceback (alignment)

Initialisierung der Matrix

Im ersten Schritt in einem Global Alignment mit dynamischer Programmierung wird eine Matrix erstellt mit der Spaltenzahl M + 1 und der Zeilenzahl N + 1, wobei M und N mit der Größe der Sequenzen korrespondieren, die aligniert werden sollen.
Da in diesem Beispiel angenommen wird, dass das Alignment keine anfänglichen Lücken, auch als Gaps bezeichnet, enthält (siehe Kapitel Alignment) und damit die Kosten für Gaps nicht als Anfangswerte in das Aligment eingehen, wird die Matrix zunächst mit 0 initialisiert.

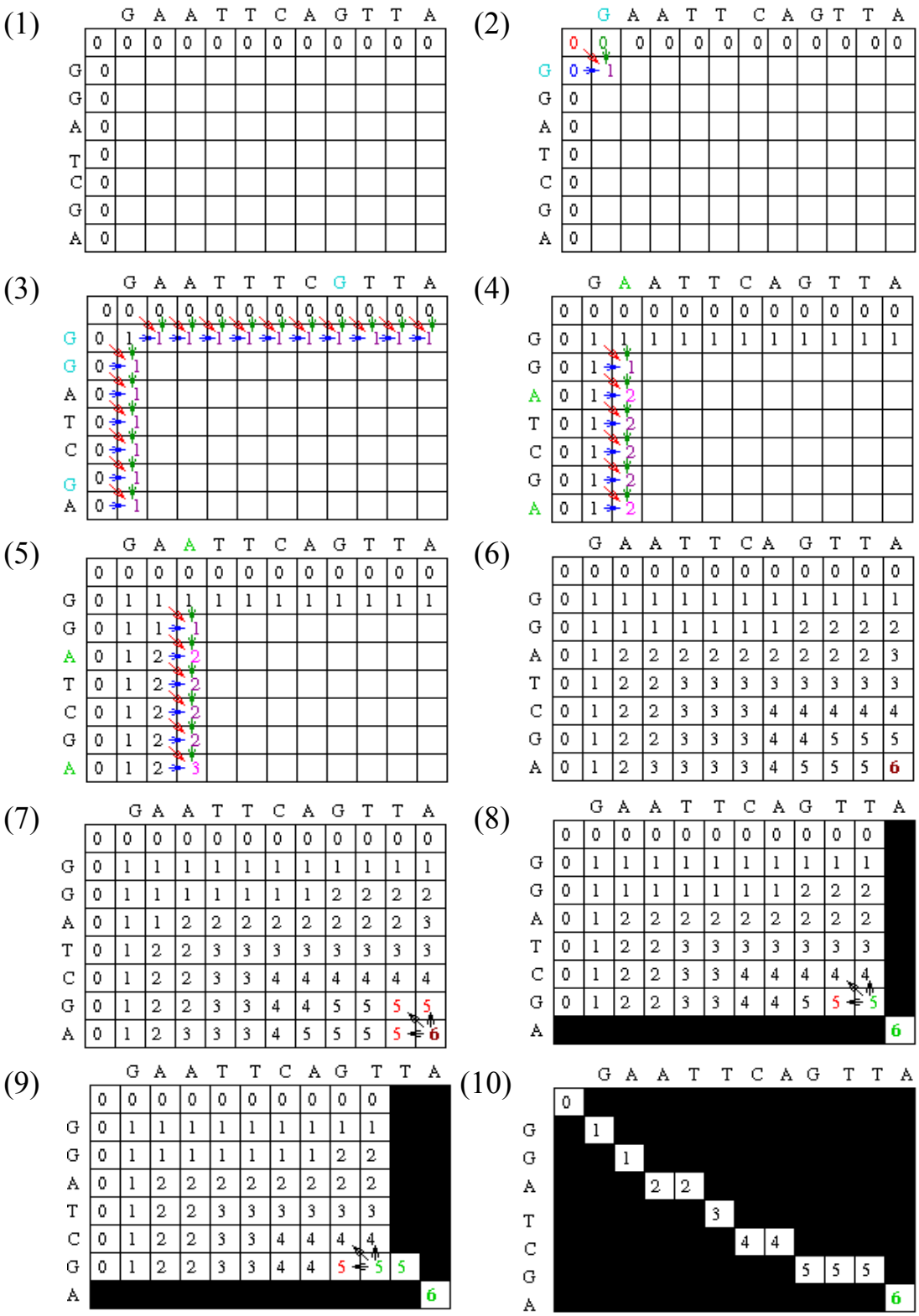


Abbildung 12. Schrittfolge der dynamischen Programmierung bei Sequqenzalignment

Aufbau der Matrix

Ein praktischer Ansatz, wie durch das Füllen der Matrix der maximale globale Alignment Score ermittelt werden kann, ist, in der linken oberen Ecke der Matrix anzufangen und den maximalen Score für jede Position in der Matrix zu $M_{i,j}$ zu finden. Um $M_{i,j}$ für jedes i,j zu finden muss der Score der Matrixpositionen links über und diagonal zu i,j , also $M_{i-1,j}$, $M_{i,j-1}$ and $M_{i-1,\ j-1}$ bekannt sein. Für jede Position $M_{i,j}$ ist der Maximumscore an der Position i,j definiert als:

Mi,j = MAXIMUM**[**

Mi-1, j-1 + Si,j (match/mismatch in der Diagonalen),
Mi,j-1 + w (Gap in der Sequenz #1),
Mi-1,j + w (Gap in der Sequenz #2)**]**

In der Abbildung zu dem Beispiel auf der nächsten Seite ist $M_{i-1,j-1}$ rot, $M_{i,j-1}$ grün und $M_{i-1,j}$ blau markiert.

Benutzt man diese Information, dann kann der Score an Position 1,1 in der Matrix kalkuliert werden. Weil der erste Basenrest in beiden Sequenzen ein G ist, ist $S_{1,1} = 1$. Unter der zu Beginn definierten Annahme ist w = 0. Daraus folgt, $M_{1,1} = MAX[M_{0,0} + 1, M_{1,\ 0} + 0, M_{0,1} + 0] = MAX\ [1, 0, 0] = 1$.

Der Wert 1 wird dann an Position 1,1 der Scoring-Matrix eingetragen (siehe Abbildung 10 (1)).

Da die Kosten für einen Gap (w) gleich 0 ist, kann der Rest von Reihe 1 und Spalte 1 mit dem Wert 1 gefüllt werden. Dies gilt für die gesamte Reihe 1. Beispielsweise ist der Wert in Spalte 2 das Maximum von 0 (für ein Mismatch), 0 (für ein vertikalen Gap) oder 1 (für horizontalen Gap). Der Rest von Reihe 1 (siehe Abbildung 10 (2)) kann auf die gleiche Weise gefüllt werden bis Spalte 8. An diesem Punkt ist ein G in beiden Sequenzen (blau). Daher ist der Wert das Maximum von 1 für einen Match, 0 für einen vertikalen Gap oder 1 für einen horizontalen Gap, also wieder 1. In dieser Weise kann der Rest von Reihe 1 und Spalte 1 ausgefüllt werden.

Die Zellen von Reihe 2 bekommen den Wert, der das Maximum von 1 (Mismatch), 1 (horizontaler Gap) oder 1 (vertikaler Gap) bildet, also 1. An der Position Spalte 2 Reihe 3 ist ein A in beiden Sequenzen so dass der hier einzutragende Wert das Maximum von 2 (Match), 1 (horizontaler Gap), 1(vertikaler Gap), so dass der Wert 2 ist. In der nächsten Position in Spalte 2 und Reihe 4 (siehe Abbildung 10 (5)) ist der Wert das Maximum von 1 (Mismatch) , 1 (horizontaler Gap), 2(vertikaler Gap), also 2. Für alle weiteren Zellen bis auf die Letzte in Spalte 2 entsprechen die Werte exakt dem in Reihe 4 (siehe Abbildung 10 (5)). Die letzte Reihe enthält den Wert 2 , da das Maximum aus 2 (Match), 1 (horizontaler Gap) und 2 (vertikaler Gap) gebildet wird. Mittels dieser Technik wird die Spalte 3 und alle weiteren Spalten mit Werten gefüllt (siehe Abbildung 10 (6)).

Rekursives „Traceback“

Nachdem die Matrix gefüllt ist, ist der maximale Alignment Score für die beiden Sequenzen 6. Der „Traceback“-Schritt bestimmt den tatsächlichen Maximumscore des gesamten Alignments. Mit dem hier genutzten simplen Score-Algorithmus sind mehrere maximale Alignments möglich. Der „Traceback“-Schritt beginnt in der M,J Position der Matrix, also der Zelle, die den maximalen Scorewert enthält (siehe Abbildung 10 (7)). Das „Traceback“ Verfahren betrachtet die Werte der benachbarten Zellen links (Gap in Sequenz #2), diagonal (Match/Mismatch) und darüber (Gap in Sequenz #1) (in der Abbildung rot markiert). Die Werte sind hier alle 5 (siehe Abbildung 10 (8)). Da die aktuelle Zelle den Wert 6 hat und der Score 1 ist für ein Match und 0 für jeden anderen Fall, so ist der zu wählende Nachbarwert der, der für den nächsten Match/Mismatch steht, also der diagonal liegende Wert. Das ergibt das folgende aktuelle Alignment:

```
(Seq #1)        A
                |
(Seq #2)        A
```

Der höchste Wert der benachbarten Zellen ist der Wert 5 (rot). Daraus folgt, dass in Sequenz #2 ein Gap eingefügt wird (siehe Abbildung 10 (9)). Das aktuelle Alignment sieht wie folgt aus:

```
(Seq #1)    T A
                |
(Seq #2)    _ A
```

Der folgende Schritt ist identisch, und es wird ein weiterer Gap in Sequenz #2 eingefügt:

```
(Seq #1)    T T A
                  |
(Seq #2)    _ _ A
```

Das „Backtracing“ wird fortgeführt bis Spalte 0 Reihe 0. Das daraus resultierende Alignment sieht dann so aus:

```
(Seq #1) G A A T T C A G T T A
           |    |    | |    |       |
(Seq #2) G G A _  T C _ G _  _ A
```

Das Beispiel ist damit abgeschlossen (siehe Abbildung 10 (10)). Für das „Backtracing“ können mehrere Möglichkeiten gegeben sein. In dem gegebenen Beispiel ist das beispielsweise das folgende Alignment:

```
(Seq #1) G _ A A T T C A G T T A
           |      |    | |    |       |
(Seq #2) G G _ A  _ T C _ G _  _ A
```

Die Berücksichtigung alternativer Lösungen stellt aber ein exponentielles Problem dar. Daher wird bei der dynamischen Programmierung in der Regel häufig nur eine Lösung ausgegeben.

1.3.2. Anwendung bioinformatischer Alignmentalgorithmen

Die Erkennung der beschriebenen, regulatorischen Motive setzt die Wahl einer geeigneten Suchmethodik für die Suche nach homologen Mustern in Nukleinsäuresequenzen voraus. Die Suchproblematik lässt sich in die Teilprobleme gliedern:

- Sequenzmustersuche
- Kombination Sequenz- und Strukturmustersuche

Bei dem Suchverfahren kann unterschieden werden zwischen der Suche nach neuen Beispielen einer schon bekannten Klasse von Signalstrukturen, die sich durch ein gemeinsames Consensusmuster auszeichnen und der Suche nach einer neuen Klasse, also einer komplett neuen Musterhomologie bzw. einem neuen Consensusmuster.

Die Sequenzmustersuche ist eins der ersten Problemstellungen, zu der Algorithmen entwickelt wurden, die heute zu den Standards der Bioinformatik gehören *(Needleman et al., 1970, Smith et al., 1981).* Sie begründen eine ganze Kategorie von Sequenzsuchprogrammen, die als Alignmentprogramme bezeichnet werden. Man spricht von „Pairwise-Alignment“, wenn zwei Sequenzen miteinander verglichen werden, und dem heute gebräuchlicheren, weil effektiveren „Multiple Alignment“, bei dem mehrere Sequenzen auf das Vorliegen von Sequenzhomologien untersucht werden können.

Für den Prozess der Identifizierung neuartiger, verständlicher und potenziell nützlicher Muster in Daten, das sogenannte Data Mining oder auch das KDD (Knowledge Discovery in Databases), stehen im Wesentlichen zwei Grundtypen von in der Genomforschung gebräuchlichen Analyseverfahren zur Verfügung: Ist ein Sequenzmotiv vorgegeben, dann kann man es über ein *Local Alignment* in Datenbanksequenzen recherchieren. Dieses Verfahren wird daher auch zum Datenbankretrieval verwendet. Ist kein Sequenzmotiv bekannt und wird nach einem neuen gesucht, wird das *Global Alignment* angewendet. Dabei werden Sequenzen ihrer ganzen Länge nach miteinander verglichen. Ein neueres Verfahren ist das sogenannte *Substring-Alignment*. Hier werden Sequenzen global miteinander verglichen, um verteilte lokale Ähnlichkeiten zwischen den Sequenzen heraus zu filtern.

1.3.2.1. Suche ohne Ausgangsbeispiel: Global Alignment

Das Verfahren, neue Sequenzmotive aus Datenbanksequenzen sichtbar zu machen, nennt man Multiples Global Alignment. Es funktioniert nach dem gleichen Prinzip der Distanzwertberechnung wie Local Alignments, allerdings ohne dass eine Ausgangssequenz benötigt wird. Die Distanzwertberechnung erfolgt aufgrund der Edit-Distanzwerte, die sich aus den drei verschiedenen Edit-Operationen ergeben (siehe Tabelle 6).

Editoperationen	*Substitute*	*Delete*	*Insert*
Typ	Basentausch	Basenlöschung	Baseneinfügung
Beispiel: G T	G T G -	- C G C	
Editdistanz 1	0	1 1	0

Tabelle 6. Distanzoperationen als Kosten in Sequenzalignments

Für die Auswertung der günstigsten Distanzwertsumme aller Basenpositionen nutzen die Alignmentprogramme Algorithmen der Dynamischen Programmierung, wie oben skizziert. Es sind für den Sequenzvergleich eine Vielzahl unterschiedlicher Algorithmen *(Gotoh et al., 1999)* entwickelt worden, die mittlerweile auch in der Lage sind, ganze Genomsequenzen untereinander zu vergleichen oder aus Contigsequenzen wieder ganze Genome in Form von Karten zu rekonstruieren *(Delcher et al., 1999, Florea et al., 2000)*. Unterschiede in den algorithmischen Verfahren liegen darin, dass z.B. die Berechnungen von Sequenzhomologien entweder auf zu minimierenden Sequenzunterschieden beruht *(MULTALIGN: Gupta et al., 1995)* oder auf Sequenzähnlichkeiten, die entsprechend maximiert werden *(CLUSTAL: Higgins et al., 1988, Higgins et al., 1992, Jeanmougin et al., 1998, Thompson et al., 1997),* oder aber das Verfahren ist für verteilte Motive in großen Sequenzen adaptiert *(DIALIGN: Morgenstern et al., 1996, Morgenstern et al., 1998, Morgenstern et al., 1999)*. Die Komplexität der Alignmentverfahren für die Suche nach Consensusmotiven macht es notwendig, deren Ergebnisse zu verifizieren *(Frech et al., 1997)*.Dies gilt vor allem für kleine Motive in großen Sequenzen. Ein Abgleich der Alignmentergebnisse ist daher notwendig, um den Grad der von dem Alignmentergebnis berechneten Ähnlichkeit der Sequenzen festzustellen. Beispielsweise hat ein optimales Alignment nicht immer die höchste Trefferzahl (siehe Abb. 13).

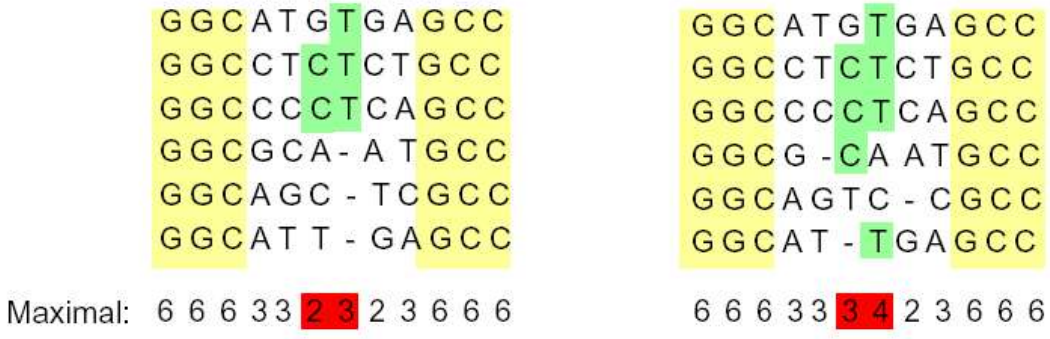


Abbildung 13. Beispiel für ein optimales Alignment mit Trefferzahl nach Alignmentprozess (links) und weiterer Optimierung (rechts) (Mit Clustal (s.o.) modifiziert aus einem Vortrag von T.Werner GCB 1996)

Durch Auswertung der alignierten Sequenzen mit Matrixprogrammen können beispielsweise optimierte Consensusmuster erstellt werden, wenn mehrere Beispiele eines Motivs aus einem Alignment zusammengefasst werden sollen *(Chen et al., 1995)*.

1.3.2.2. Suche mit einem oder mehreren Ausgangsbeispielen: Local Alignment- und Matrixanwendungen

Zwei Local Alignment-Programme, BLAST *(Altschul et al., 1990)* und FASTA *(Pearson et al., 2000)*, gehören zu den Standardwerkzeugen der Genomforschung. Von diesen zentralen Algorithmen sind diverse Programmvarianten für spezielle Anwendungen entstanden, wie etwa Blitz, 3D-Blast *(Brenner et al., 1995)* und Gapped-BLAST *(Altschul et al., 1997)*. Kommerzielle Softwarepakete für das Problem der Motivsuche (z.B. BLUEGENE, Magic Works GmbH), die mit neuronalen Netzen arbeiten, erlauben eine schnellere Analyse, um mit verbesserter Sensitivität eine Ausgangssequenz in Datenbanksequenzen zu finden. Solche Programme vergleichen schrittweise die extrahierte Motivsequenz mit gleich großen Subsequenzen einer Datenbanksequenz, um Sequenzübereinstimmungen festzustellen und arbeiten in der Regel auf Basis von Heuristiken.

Neuere Entwicklungen auf dem Gebiet der Motivsuche in Datenbanksequenzen erlauben nicht nur die Erkennung von möglichen Consensusmotiven in Datenbanksequenzen durch Sequenzübereinstimmung (MatInspector, *Quandt et al., 1995)*, sondern auch die nachträgliche Evaluation der gefundenen Sequenzähnlichkeiten durch Berücksichtigung zusätzlicher Parameter, wie beispielsweise der Distanz zum Startcodon (ConInspector, *Frech et al., 1997).*

1.3.3. Suche nach neuen Klassenbeispielen aus Datenbanken

Um “consensus pattern” von Signalstrukturen aus der Vielzahl variierender Sequenz- und Strukturelemente einer mRNA herauszufiltern, bedarf es bei der Suche nach neuen Beispielen einer Klasse einer Datensammlung bereits bekannter Signalstrukturen, um solche Muster vorzuklassifizieren und daraus Ähnlichkeitsmerkmale zu generieren. Für Proteine als auch DNA Signalstrukturen und Signalstrukturen der codierenden Region der mRNA sind solche Datensammlungen in Form von etablierten Datenbanken über Sequenzelemente bereits vorhanden (Proteinmotivdatenbank PROSITE *(Hofmann et al., 1999)* TRANSFAC *(Wingender et al., 2000* über Promotor- und Enhancermotive) und TRANSTERM *(Dalphin et al., 1999)* über Kontextsequenzen zu Start- und Stopcodon). Diese Datenbanken beruhen auf Literatursammlungen, die die experimentellen Resultate der aufgeklärten regulatorischen Funktionen und deren Sequenzeigenschaften zusammenfassen, sowie den Extraktionen der Sequenzen aus genomischen Datenbanken, die durch zusätzliche Annotationen neu strukturiert worden sind. Ein anderer Weg, experimentell abgesicherte Motivsequenzen direkt als Ausgangspunkt für die Definition neuer Klassen von Signalstrukturen zu nutzen, sind selektive Anreicherungsexperimente oder SELEX-Experimente *(„Systematic Evolution of Ligands by EXponential Enrichment“ Tuerk et*

al., 1990). Sie liefern potentielle Bindungsstrukturen, die für ein bioinformatisches Screening eingesetzt werden können *(Davis et al., 1995, Dandekar et al., 1998)*.

1.3.3.1. Suche nach neuen Klassen: Sequenz- und Strukturhomologie

Die Suche nach Stukturmustern ist komplexer, weil hier die Kombinatorik der Basenpaarbildung und die Auswertung der oben skizzierten Strukturelemente berücksichtigt werden muss. Das Problem bei der Analyse von Signalstrukturen ist, dass: 1) - eine Sequenz dynamisch verschiedenste Sekundärstrukturen - sogenannte suboptimale Faltungen – mit jeweils unterschiedlichen Teilstrukturen ausformen kann *(Zuker 1989)* und 2) - umgekehrt verschiedenste Sequenzen entweder in den thermodynamisch suboptimalen oder der optimalen Faltung identische Sekundärstrukturen bilden können *(Fontana et al., 1993)*. Die Sequenz kann also nur bedingt und nur in statistischer Hinsicht Aufschluss über homologe Strukturen geben.

Schuster *(Schuster, 1995)* hat aufgrund von Computerexperimenten errechnet, dass im Durchschnitt 7,5 Mutationen (Basenpaaraustausch im Doppelstrang) oder 15 einzelne Basenaustausche (d.h. Basenaustausche im Einzelstrang) notwendig sind, um zu einer gegebenen Sequenz mit Sekundärstruktur eine Sequenz zu finden, die die identische Sekundärstruktur einnimmt, wenn die Länge der Sequenz 100 bp beträgt. Dieses Ergebnis wurde aus einem Pool von 4×10^{24} Sequenzen mit unterschiedlicher Basenabfolge gewonnen. Von Schuster wurde auch festgestellt, dass über 10% aller Sequenzen aus einem Pool von 10^9 Sequenzen der Länge 100 bp bestimmte identische Sekundärstrukturen bevorzugt ausbilden. Konkret heißt das, von 218.820 zufälligen Strukturen wird eine identische Sekundärstruktur durschnittlich von 4907 Sequenzen gebildet, wobei einzelne Teilstrukturen von bis zu 22.719 Sequenzen gebildet werden können. Die Entstehung von homologen Sekundärstrukturen aus verschiedenen Sequenzen ist also statistisch gesehen nicht selten. Es kommt also bei der Suche nach homologen Teilstrukturen auch darauf an, mehr Consensusmerkmale als die Übereinstimmung der Struktur zu finden. Beispielsweise zeigen konkrete Signalstrukturen, dass gerade die verschiedenen Varianten der Sekundärstrukturen mit Signalstrukturwirkung auch konservierte, fixe topologische Längenmerkmale, wie die Länge von Loop- oder Stembereichen, haben. Dies sind vor allem im endständigen Bereich der Hairpins der Loop und der erste Stembereich wie etwa bei Histone *(Dominski et al., 1999)*.

Andererseits reicht dies oft auch nicht aus, da viele Signalstrukturen auch variierende Stukturmerkmale aufweisen und trotzdem einer Klasse zuzuordnen sind, wie beispielsweise das bFGF-Bindungsmotiv (siehe Abbildung 13). Ein zweites wichtiges Merkmal von Signalstrukturen ist daher die Kombination von Sekundärstrukturen und Consensusmotiv. Das bedeutet, dass nicht jeder Sequenzbereich in einer Sekundärstruktur beliebig variieren darf. Solche Consensusmotive sind häufig im oberen Teilstrukturberreich zu finden.

Die Topologie sowie bestimmte Basenabfolgen innerhalb einer Struktur bedingen sich also bei der Entfaltung ihrer Signalwirkung gegenseitig. Das Problem bei der Erkennung solcher Signalstrukturen ist die Unterscheidung zwischen möglichen funktionellen Sekundärstrukturen oder Teilstrukturen wie den "Hairpins", die meistens zwischen verschiedenen mRNA Typen konserviert vorliegen, und solchen Sekundärstrukturen, die von der mRNA in der Zelle willkürlich je nach den thermodynamischen Verhältnissen in der Zelle ausgebildet werden.

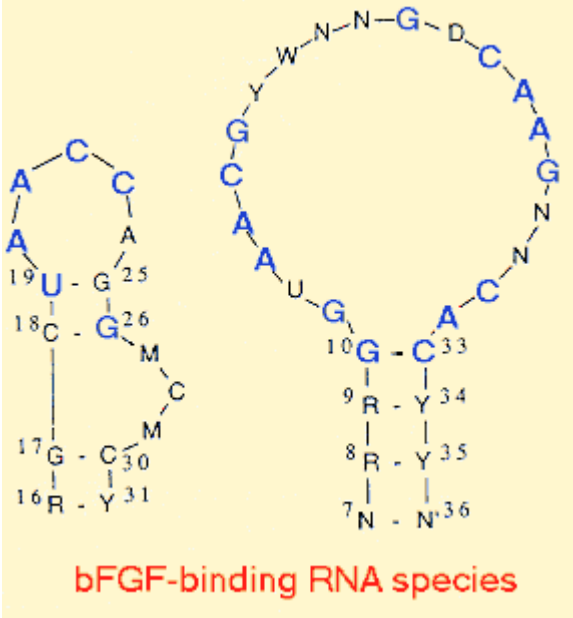


Abbildung 14. Beispiel für die Kombination von Sequenz und Strukturmerkmalen

Voraussetzung für die Suche nach Struktur- und Sequenzhomologien ist, dass entsprechende Algorithmen aus der Informatik für dieses Problem optimiert worden sind und dass die RNA-Sekundärstruktur anhand von thermodynamischen Parametern am Computer erstellt werden kann.

1.3.4. Berechnung der RNA Sekundärstruktur

Die bioinformatische Forschung hat mittlerweile diverse Programme entwickelt, die, neben der Vielzahl von Programmen zur Motiverkennung in Sequenzen, in der Lage sind, Sekundärstrukturen aus linearen RNA-Sequenzen mittels thermodynamischer Parameter *in-silico* zu konstruieren. Da, wie bereits beschrieben, die Struktur gerade bei mRNA Motiven wie katalytisch aktiven Regionen und Signalstrukturen ein deterministisches Element für die Wirkungsweise dieser Strukturen darstellt, entsteht darüber hinaus das Bedürfnis, die erstellten Sekundärstrukturen untereinander vergleichen zu können, um auch hier strukturelle Consensusmuster sichtbar zu machen.

1.3.4.1. Thermodynamische Parameter der computergesteuerten Strukturerstellung

Die RNA-Sekundärstruktur enthält den größten Anteil freien Energie einer Faltung im Gegensatz zum Protein *(Tinoco et al. 1999)*. Daher ist die Berechnung der 2D-Struktur eine sehr gute Näherung an die Gesamtstruktur eines RNA-Moleküls, obwohl es selbst nur ein Zwischenprodukt bei der Umfaltung der Gesamtstruktur ist. Das Prinzip der Sekundärstrukturberechnung basiert auf den von den Basen einer Sequenzabfolge gebildeten Basenpaarungen. Dabei gilt die Annahme, dass jede Base im Prinzip an einer Basenpaarung beteiligt sein kann und Basenpaarungen sich nicht überkreuzen, sondern sequenziell angeordnet sind. Diese Annahme trifft in der Regel auf die meisten RNA-Sekundärstrukturen zu, wenn man Pseudoknots unberücksichtigt lässt.

Pseudoknots sind RNA-Strukturen mit Überkreuzpaarungen und Basen mit Tertiärkontakten (Triplepaarungen). Von diesen relativ selten auftretenden Strukturen sind bisher folgende bekannt:

- „Group I“ Introns in P4-P6 Domäne
- „Hammerhead“-Ribozyme
- HDV Ribozyme
- Hefe tRNAphe
- L1 -Domäne von 23S rRNA *(Hermann et al., 1999).*

Durch die Konzentration auf sequenziell auftretende Basenpaarungen kann das mathematische Modell für die Berechnung von Sekundärstrukturen vereinfacht werden. Für zwei sich nicht überkreuzende Basenpaarungen der Basenpaare (i,j) und (k,l) gilt:

i < k < j < l

Dadurch können die möglichen, auftretenden Sekundärstrukturen wie folgt errechnet werden:

$$S_{kl} = S_{k+1,l} + \sum_{j=k+m}^{l} \prod kj \, S_{k+1,j-1} S_{j+1,l}$$

wobei $\prod kl = 1$ wenn die Positionen *k,l* eine Basenpaarung bilden (GC, CG, AU, UA, GU,UG) oder $\prod kl = 0$.

Das Vorgehen schildert das grundsätzliche Prinzip der Berechnung von möglichen Basenpaarungen. Diese Idee kann genutzt werden, um die Zahl der Basenpaarungen zu maximieren und damit die optimale Anzahl von Basenpaarungen zu berechnen:

$$M_{kl} = max\left\{ M_{k+1,l}; \left(M_{k+1,-1} + M_{j+1,l} + 1 \right) \forall \, j / \prod kj = 1 \right\} \quad .$$

Algorithmen für die Berechnung von RNA Faltungen basierend auf der dynamischen Programmierung funktionieren nach diesem Prinzip. Sie benutzen für die Berechnung von Sekundärstrukturen ein Loop-abhängiges Energiemodell. Bei diesem Standardenergiemodell wird die freie Energie als Summe *S* der Energien der Loops *l* berechnet:

$$E(S) = \sum_{l \in S} E(l)$$

Die Daten für die Berechnungen der freien Energie sind temperaturabhängig und stammen aus experimentellen Messungen von Schmelztemperaturen kleiner Oligosequenzen. Zwei Forschungsgruppen haben vor allem bei der Messung von RNA und DNA Parametern die Standards gesetzt: Douglas Turner und John SantaLucia *(SantaLucia et al., 1997) (SantaLucia et al., 1998)*. In den Labors werden die Schmelztemperaturen bei 1 molaren NaCl und 37°C gemessen. Für andere Temperaturen werden die Faltungen in der Regel durch Extrapolation berechnet.

1.3.4.2. Methoden der RNA-Sekundärstrukturerstellung

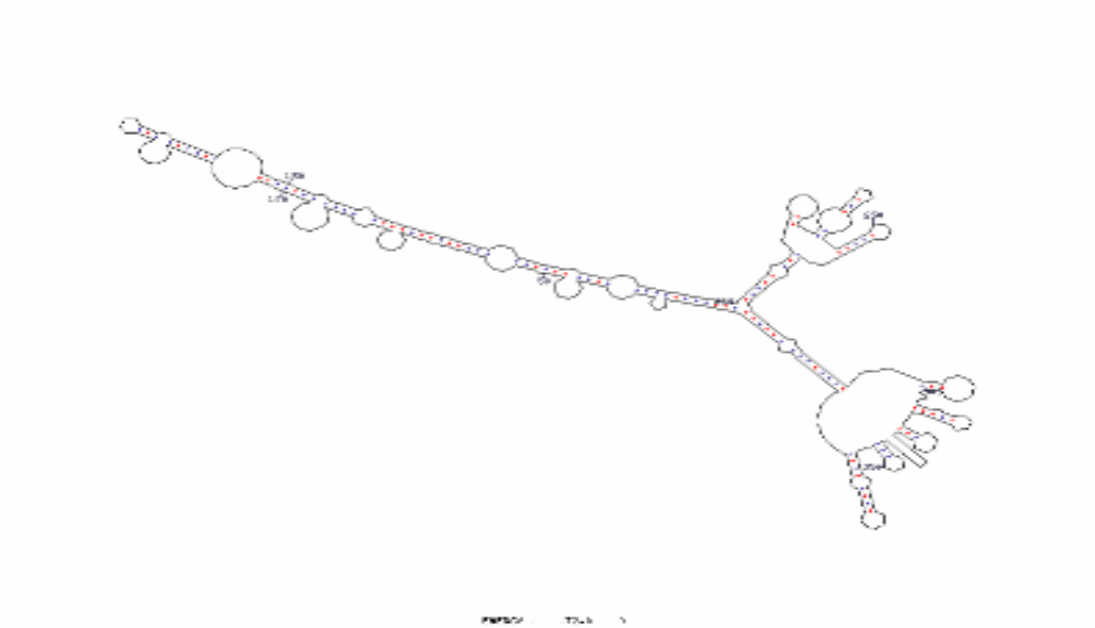


Abbildung 15. Beispiel einer grafischen Darstellung einer RNA-MFOLD Faltung (Transferringen) (Rot = C-G Paarung; Blau = G-U und A-U Paarung)

Die Modellierung von RNA Sekundärstrukturen ist seit der Entwicklung des MFOLD Algorithmus durch Zuker *(1989),* basierend auf experimentell ermittelten themodynamischen Parametern *(Freier et al. 1986, Mathews et al., 1999)* ein Standard in der RNA-Forschung.

Mittlerweile sind weitere Algorithmen für Faltungsprogramme entwickelt worden, die wie FOLDRNA auch die suboptimalen Sekundärstrukturen modellieren können, wie es das MFOLD Programm kann. Die wichtigsten werden in der Tabelle unten beispielhaft kurz vorgestellt.

Faltungsprogramm	*Algorithmus für die Berechnung der minimalen freien Energie*	*Literaturangaben*
MFOLD (UNIX) RNASTRUCTURE (Windows)	Dynamischer (rekursiver) Algorithmus, Bioinformatischer Standard mit Berechnung fast aller möglichen Sekundärstrukturen mit eingeschränkter Tertiärstrukturberechnung	*(Zuker et al., 1981, Mathews et al., 1999, Zuker et al., 1989)*
RNAFOLD	Teilweise schneller als MFOLD, Bestandteil des VIENNARNA-Paketes, Berechnung aller suboptimalen Sekundärstrukturen	*(Wuchty et al., 1999)*
STRUCTURELAB	Strukturvorhersage für Supercomputer	*(Shapiro et al., 1997)*
Kein Name	Monte-Carlo Simulation	*(Le et al., 1993)*
RNAFOLD	Heuristischer Algorithmus basierend auf Helix-orientierte Sekundärstrukturerstellung	*(Martinez et al., 1984, Martinez et al., 1990)*
Kein Name	Basierend auf der Modellierung des kinetischen Faltungsprozesses	*(Gultyaev et al., 1991)*
Kein Name	Hidden Markov Model	*(Mironov et al., 1993)*
tRNAscan	Hidden Markov Model	*(Lowe et al., 1997)*
Abgeleitet aus tRNAscan	Dynamischer Algorithmus für Tertiärstrukturen (Zeit- und Speicherintensiv)	*(Rivas et al., 1999)*
Kein Name	Monte-Carlo Simulation (Berechnung der wahrscheinlichsten Sekundärstruktur)	*(WuJu et al., 1998)*
ESSA	Interaktive Erstellung von Sekundästrukturen langer RNA-Sequenzen	*(Chetouani et al., 1997)*

Tabelle 7. RNA Faltungsalgorithmen

Die Berechnung der Sekundärstruktur mit den Algorithmen MFOLD und RNAFOLD hat eine Genauigkeit im Verhältnis zu den natürlich ausgebildeten Sekundärstrukturen von 70% korrekt gefalteter Sequenzen, die nicht größer sind als 700nt. Bei länger werdenden Sequenzen sinkt die Genauigkeit langsam auf bis zu 40% ab. Für die Auswertung der bei der Faltung mit MFOLD anfallenden Faltungsbeschreibungen ist zusätzliche Auswertungssoftware erstellt worden (*Schmitz et al., 1992),* unter anderem auch, um die Sekundärstrukturen zu visualisieren *(Matzura et al., 1996).* Kinetische Analysen zum Faltungsverhalten lassen sich mit speziell dazu entwickelten Programmen erstellen *(Evers et al., 1999, Gultyaev et al., 1995).* Diese kinetischen Analysen *in silico* von RNA Sequenzen lassen mittlerweile auch Vorhersagen zu den bereits erwähnten „molecular switches“ (siehe 1.2.2.3) zu. So gibt es das Programm

paRNAss, das solche Schalter in RNA-Sequenzen vorhersagt und modelliert *(Giegerich et al., 1999)*.

```
21  ENERGY = -9.3  APHISH1+
    1 A       0    2    0    1
    2 A       1    3    0    2
    3 A       2    4    0    3
    4 G       3    5   19    4
    5 G       4    6   18    5
    6 C       5    7   17    6
    7 U       6    8   16    7
    8 C       7    9   15    8
    9 U       8   10   14    9
   10 U       9   11    0   10
   11 U      10   12    0   11
   12 U      11   13    0   12
   13 A      12   14    0   13
   14 A      13   15    9   14
   15 G      14   16    8   15
   16 A      15   17    7   16
   17 G      16   18    6   17
   18 C      17   19    5   18
   19 C      18   20    4   19
   20 A      19   21    0   20
   21 C      20   22    0   21
```

Abbildung 16. Ausgabeformat von MFOLD und RNASTRUCURE (Beispiel: Histone 3' Stemloop)

Die Modellierung der RNA Sekundärstruktur gibt bereits ausreichenden Aufschluss über strukturelle und damit funktionelle Eigenschaften einer RNA-Sequenz *(Schuster 1995)*. Das aktuelle Forschungsinteresse spiegelt Arbeiten zur Verbesserung der Strukturvorhersagemethode von MFOLD wider, um die Tertiärstruktur einer RNA Sequenz zu berechnen *(Bevilacqua et al., 1998, Tabaska et al., 1998)*.

1.3.5. Suchmethoden für Strukturhomologie

1.3.5.1. RNA-Strukturbeschreibung

Um eine erstellte RNA-Sekundärstruktur einer Weiterverarbeitung beispielsweise in einem Strukturalignment zugänglich zu machen, müssen die Eigenschaften der Struktur zunächst geparst werden. Mit Parsing bezeichnet man in der Informatik ein Verfahren, das Wortproblem für ein festes terminales Alphabet und eine feste Grammatik für beliebige vorgegebene Wörter der freien Sprache zu lösen, und zwar in der Regel durch zeichenweise Analyse. Parsing-Algorithmen gibt es in der Bio-

informatik beispielsweise für die Gen-Expressionsanalyse *(Sutcliffe et al., 2000)*, RNA-Ligand *(Leclerc et al., 1998)* und DNA-Ligand Interaktionen *(Haq et al., 2000)* und die Strukturvorhersage, wie sie das Programm MFOLD ausgibt *(Lefebvre et al., 1995)*. Die Parser beziehen sich dabei auf unterschiedliche Repräsentationen und Grammatiken, in die die jeweiligen molekularen Konformationsdaten umgesetzt werden. In der Bioinformatik sind verschiedene Repräsentationsformalismen speziell für die Darstellung von RNA-Sekundärstrukturen *(Grate et al., 1994, Grate et al., 1995)* inklusive Pseudoknots *(Rivas et al., 2000, Brown et al., 1996)* erstellt worden, in die die Strukturinformation umgesetzt werden kann

1.3.5.2. RNA-Strukturalignment

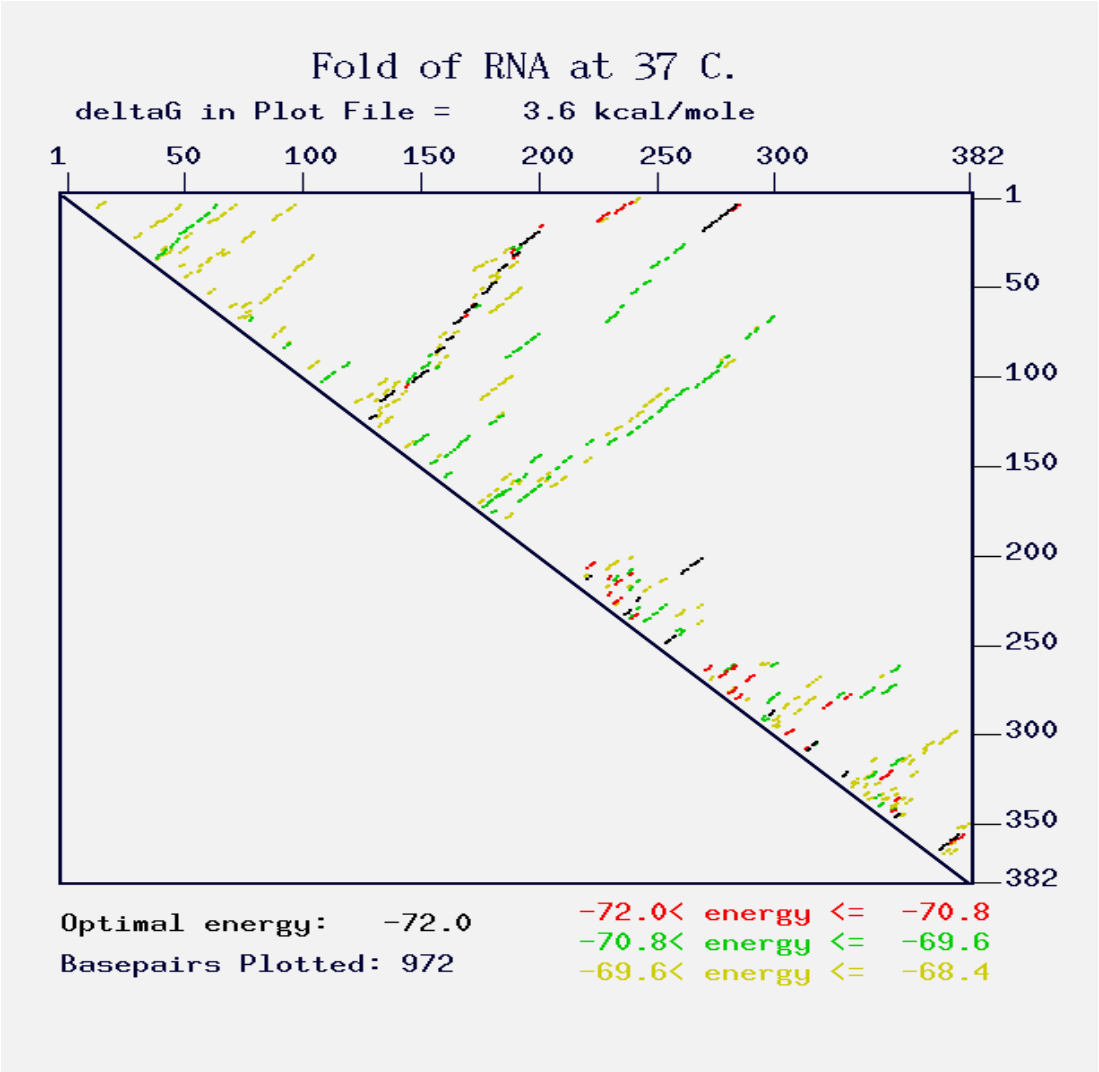


Abbildung 17. Dotplot Matrix einer RNA Sekundärstruktur. Die farbigen Linien stellen jeweils eine Hairpin-Struktur dar.

Ein erster konservativer Ansatz für das Auffinden phylogenetisch konservierter Sequenzregionen mit Sekundärstruktur war der Einsatz von Multiplen Alignments *(Corpet et al., 1994)*. Dies erforderte jedoch die manuelle Optimierung des Alignments, da die Struktureigenschaften nicht selbst unmittelbar im Alignmentverfahren berücksichtigt werden können, sondern als Know-How des Benutzers vorliegen müssen.

Auf der graphischen Darstellung der gefalteten Sekundärstrukturen von MFOLD der Basenpaar Punktmatrix (siehe Abb. 17) bzw. „DOT PLOT Matrices“ *(Jacobson et al., 1993)* aufsetzend, sind Arbeiten erschienen, die konservierte Sekundärstrukturen aus einer großen Anzahl von gefalteten Sequenzen sichtbar machen können *(Davis et al., 1995, Davis et al., 1996)* Die Analysemethode besteht darin, durch ein Alignment Sequenzübereinstimmungen zu detektieren, um dann diese Region mit MFOLD zu falten und anschließend Alignment und Faltung über die Basenpaar Punktmatrix zu vergleichen. Dabei werden alle Faltungen und Faltungsvarianten entsprechend den Alignmentpositionen der Sequenz übereinander gelegt. Dadurch lassen sich konservierte Sekundärstrukturen sichtbar machen. Die Einschränkung in der Nutzbarkeit der von Davis entwickelten Methoden, konservierte Teilstrukturen aus Sequenzen zu extrahieren, ist, dass durch das Verfahren selbst nicht festgestellt werden kann, ob bestimmte Strukturelemente in der Sekundärstruktur ein mögliches Consensusmotiv aufweisen, wie es bei den meisten Signalstrukturen der Fall ist. Durch manuelle Optimierung muss der höchste Grad der Konservierung zwischen Sequenz und Struktur gefunden werden, was durch abgestufte Farbgebung der Matrixpositionen unterstützt wird.
Ein pures Multiple Alignment ist zu insensitiv für einzelne Basenaustausche in Consensusmotiven, wenn diese sehr kurz und teilweise über mehrere Strukturelemente verteilt sind. Dies gilt auch für einen ähnlichen Ansatz *(Luck et al., 1996),* der ebenfalls auf der Basenpaar-Punktmatrix aufbaut und die Optimierung des Vergleichs durch eine Maximierung der Übereinstimmung zwischen Struktur und Sequenz, also der globalen Consensusstruktur, erzielt. Auch hier besteht das Problem, verteilte Motive, die mit Strukturelementen korrelieren, heraus zu filtern. Daher wurden in den genannten Arbeiten entweder nur kurze Sequenzen (200nt) oder Sequenzen aus der codierenden Region als Programmtestdaten verwendet, die eine erheblich geringere Mutationsrate und Basenvariabilität aufweisen als z.b. UTR Sequenzbereiche.

Ein Verfahren, komplette Strukturbeschreibungen direkt in das Alignmentverfahren einzubinden, ist von Gautheret *(Gautheret et al., 1990)* vorgestellt worden. Das Alignmentverfahren ist in dem Programm RNAMOT *(Laferriere et al., 1994)* implementiert. Es nutzt die oben skizzierten Alignmentverfahren, um nach RNA-Sekundärstrukturen zu suchen. Dabei besteht das Problem, Strukturinformation wie Position des Stems, Positon des Loops usw. für die Recherche in lineare Sequenzinformation umzusetzen, ohne dass die Strukturinformation verloren geht. Durch Unterteilung der Teilsequenz, aus der die Sekundärstruktur gebildet wird, in die jeweiligen helikalen und einzelsträngigen Abschnitte (Strukturelemente) und deren Übertra-

gung in eine spezielle Grammatik (siehe 2.3.2. genutzte Datenformate) werden diese Sequenz- und Strukturinformation miteinander kombiniert.

Descriptor für IRE
H1 s1 H2 s2 H3 s3 H3 s4 H2 H1
H1 6:18 9 SHS:SDS
H2 0:1 1 G:C
H3 5:5 1 KKYRM:KTRDA
s1 0:1 T
s2 1:3 TKC
s3 6:6 CAGTGH
s4 0:1 M
R H1 H2 H3
M 8

Tabelle 8. Beispiel einer Descriptorbeschreibung anhand des "Iron Responsive Elements"

Die Bezeichner „H“ und „s“ repräsentieren in dem obigen Beispiel doppelsträngige, helicale und einzelsträngige Regionen. Sie werden in der ersten Zeile in der Reihenfolge ihres Vorkommens in der Gesamtstruktur annotiert. Die Reihenfolge wird durch die Angabe von Zahlen festgehalten. Darunter folgt die Annotation jedes einzelnen Strukturelements. Sie besteht aus dem Bezeichner, der Angabe der minimalen und maximalen Länge sowie zusätzlich der Anzahl der maximal erlaubten Mismatches bei den helicalen Regionen. Am Ende jeder Zeile kann optional eine Basensequenz im IUB-IUPAC Code angegeben werden, die in dem Strukturelement enthalten sein muss. Dabei wird unterschieden zwischen Groß- und Kleinbuchstaben. Basen die als Großbuchstaben angegeben werden, müssen in der komplementären Struktur enthalten sein, während das Vorkommen von Basen, die als klein geschriebene Buchstaben angegeben sind, den Scorewert verbessern.
Die letzten beiden Zeilen dienen dazu, einerseits die Priorität in der Reihenfolge der zu suchenden Strukturelemente zu verändern (R) und andererseits die Anzahl der insgesamt zulässigen Mismatches anzugeben (M).

Der große Vorteil dieser Methodik, Struktur- und Sequenzangaben in einer Beschreibung zu vereinen, besteht in der Möglichkeit, komplexe, umfangreiche Sekundärstrukturen zu beschreiben, die auch tertiäre Kontakte haben können wie Pseu-

doknots. Andererseits ist aber auch der Umfang der Strukturangaben der zeitkritische Faktor im Laufzeitverhalten des Programms. Sind die Sequenzangaben rar und die Strukturangaben umfangreich so steigt die Laufzeit stark an. Es ist also für einen performanten Einsatz des Programms sinnvoll, die Reihenfolgeoption zu nutzen um die Laufzeit zu reduzieren.

Durch die Kombination von Struktur-, Längen-, und Sequenzmerkmalen sind für ein Strukturelement mehrere Lösungen möglich. Um die optimale Lösung zu finden ist eine Evaluierung der Ergebnisse als Scoring-Schema in RNAMOT implementiert. Das Scoring-Schema sieht eine Reduzierung der gewichteten Übereinstimmung vor, wenn:

- die Sequenzabfolge mit der Sequenzabfolge des Strukturelements überein stimmt (bei Basen, die mit kleinen Buchstaben angegeben sind)
- die Summe der Längen aller gefundenen, komplementären Helices dem Maximum entspricht
- die Summe der freien Energie aller Helices dem Minimum entspricht.

Im Endergebnis sind also die Sekundärstrukturen mit den niedrigsten Scorewerten enthalten. An den letzten beiden Evaluierungsschritten kann man erkennen, dass das Scoring-Schema darauf ausgerichtet ist, das jeweils energetisch stabilste Motiv als die optimale Lösung zu behandeln. Dies trifft sicherlich für die meisten RNA-Signalstrukturen auch zu, kann aber in einigen Fällen zu Problemen führen, bei denen nicht das Merkmal der Stabilität, sondern Sequenz- oder Strukturmerkmale die Funktionsweise der Signalstruktur dominieren.

1.4. Fragestellung

Es sollen mit Hilfe von Data-Mining Werkzeugen funktionale RNA-Signalstrukturen verlässlich in genomischen Sequenzdaten identifiziert werden. Dazu sollen die wiederkehrenden, fixen oder homologen Sequenz- und Strukturmerkmale der konkret bekannten Signalstrukturen aus der Literatur extrahiert und als Erkennungsmerkmal möglicher neuer Signalstrukturen genutzt werden. Diese Nutzung kann erfolgen, indem die regulatorische Sequenzregion als Sequenzmotiv (siehe 1.3.2.) oder als gefaltete Sekundärstruktur (siehe 1.3.3.) als Vorlage für Filterkriterien dienen. Sie können dazu angewendet werden, entweder direkt nach neuen Klassenbeispielen (siehe 1.3.1.1.) zu suchen oder um daraus typische Ähnlichkeitsmerkmale zu gewinnen, die dazu geeignet sind, neue Signalstrukturklassen (siehe 1.3.1.2.) zu finden.

Ziel der vorliegenden Arbeit ist es, für dieses skizzierte Vorgehen eine RNA-Signalstrukturdatenbank zu erstellen, die die auf der Auswertung der Literatur basierten Signalstrukturen sowie deren Klassenbeschreibung in Form von annotierten Sequenz- und Strukturrepräsentationen enthält. Des weiteren ist eine Software als Recherchewerkzeug bereitzustellen, um auf Grundlage der Klassenbeschreibungen in Sequenzeinträgen molekularbiologischer Datenbanken nach neuen Signalstrukturbeispielen einer gegebenen Klasse recherchieren zu können. Die Software

soll durch Homologievergleich bisher noch nicht charakterisierte Signalstrukturen detektieren. Für den dabei erforderlichen Strukturvergleich ist die Erstellung der Sekundärstrukturen aus den in den Datenbanken vorhandenen Sequenzeinträgen notwendig. In der Bioinformatik sind dafür entsprechende Programme zur Strukturberechnung, wie z.B. verschiedene Versionen des Programms MFOLD (UNIX) *(Zuker et al., 1994, Zuker et al., 1989)* und RNASTRUCTURE (WindowsNT) *(Mathews et al. 1999, Mathews et al., 2002)* und außerdem Strukturberechnungs- und Vergleichsprogramme vorhanden. Daran anknüpfend soll mit einem neuen Ansatz versucht werden, die Ausgabe der Strukturberechnung von RNASTRUCTURE für Sequenzen, in denen es gilt, noch unbekannte Sekundärstrukturen aufzufinden, mit einem Strukturvergleichsprogramm zu kombinieren. Mit dieser Programmkombination sollen in einem Anwendungsbeispiel vor allem die RNA-UTRs von RNA-Sequenzen, die in Zellen des Hirns exprimiert werden, nach dem Vorkommen von Signalstrukturen untersucht werden. Diesem Ziel entsprechend beinhaltet die Arbeit folgende Teilprojekte: Aufbau von Sequenzdatenbanken vor Ort (Datenbankprojekt) und die Entwicklung Bereitstellung und von Software für den Strukturvergleich (Softwareprojekt). Schließlich ist bei erfolgreichem Abschluss des Softwareprojektes eine Sekundärstrukturrecherche vor Ort durch zu führen (Rechercheprojekt).

Parallel zu diesem Vorgehen ist ein Kooperationsprojekt im Fachbereich Informatik der Universität Bremen angesiedelt, in dem meine Kollegin Uta Bohnebeck die vorhandenen Klassenbeschreibungen als Fallbasis nutzt, um ein Data-Mining Verfahren zu entwickeln, das aus den Struktur- und Sequenzangaben der Signalstrukturklassen ein Ähnlichkeitsmaß generiert. Dieses Ähnlichkeitsmaß soll als Grundlage für die Identifizierung von komplett neuen Signalstrukturklassen in genomischen Sequenzdaten dienen.

1.4.1. Erstellung einer Datenbank für posttranskriptionale regulatorische RNA

In einem Datenbankprojekt ist eine Datenbank für die RNA-Motivsequenzen anzulegen, mit der nach Signalstrukturbeispielen bekannter Klassen recherchiert werden kann. Die Notwendigkeit für diesen Schritt ergibt sich aus der Tatsache, dass die Klassenbeschreibungen der Signalstrukturen erst aus den Sequenzdaten der internationalen molekularbiologischen Datenbanken gewonnen werden müssen. Zudem sind die Postreg-Sequenzen nur zum geringen Teil in den Datenbanken annotiert, vor allem die bekannteren Signalstrukturen wie z.B. das „Iron Responsive Element“ IRE. Da also weder die Detailinformationen über Motivsequenzen noch deren zellspezifisches Vorkommen in der EMBL Data Library oder GenBank separat recherchierbar sind, müssen sie erst in einer neuen Datenbank zusammen gestellt werden. Ausgangspunkt der Recherchen sind die datenbanktechnisch zu erfassenden, in der Literatur bereits dokumentierten Signalstrukturen aus der mRNA nichttranslatierten Region.

1.4.2. Anwendungsbeispiel des bioinformatischen Verfahrens auf mRNA-Sequenzen aus Hirnzellen

Besondere Bedeutung hat das Expressionsverhalten sogenannter regulatorischer Gene in noch nicht ausdifferenzierten, pluripotenten Zellen. Sie sind für die gewebetypische Ausdifferenzierung einer Zelle essentiell. Ein Beispiel für die komplizierte Wirkungsweise dieser Gene ist das komplexe Gewebe des Hirns, in dem regulatorische Gene für die Ausdifferenzierung von bis zu ca. 10.000 verschiedenen Neuronen und unterschiedlichen Gliazellen verantwortlich sind. Die Wirkungsmechanismen der im Gehirn während seiner Entwicklung in der Ontogenese und im späteren Dauerstadium der Neurone exprimierten regulatorischen Gene sind noch wenig charakterisiert. Das besondere Interesse gilt den nichttranslatierten Regionen der mRNA neuronaler Gene. In diesen Sequenzabschnitten am 5'- und 3'-Ende eines transkribierten Gens sind Sequenzstrukturen enthalten, die die Expression eines Gens auf posttranskriptionaler Ebene im Cytoplasma steuern. Besonders in der nichttranslatierten Region am 3'Ende (3'UTR) sind solche Signalstrukturen bekannt, die das Expressionsverhalten der Gene durch regulatorische Mechanismen steuern und damit in der Differenzierung von Neuronen eine wichtige Rolle spielen können, wie an Muskelzellen bereits gezeigt worden ist *(Rastinejad et al., 1993)*.

Arbeitshypothese ist, dass in den nichttranslatierten Regionen der mRNA hirnspezifischer, also in Zellen des Gehirns exprimierter Gensequenzen weitere bisher unbekannte Signalstrukturen vorhanden sind *(Stewart et al. 1992)*. Die Etablierung eines Suchverfahrens zur Erkennung und Charakterisierung von Signalstrukturen in hirnspezifischen UTR-Sequenzen auf bioinformatischem Weg ist das Ziel der Dissertation. Voraussetzung für das Erreichen dieses Ziels ist das Auffinden von signifikanten Sequenz- und / oder Strukturübereinstimmungen (Homologien) in mRNA-Sequenzen, die im Hirn exprimiert werden.

Posttranskriptionelle Regulation von Genen, die in Hirnzellen exprimiert werden, sind bisher bereits in vielen Fällen experimentell auf molekularbiologischem Weg nachgewiesen worden, wie die untere Tabelle beispielhaft aufführt.

Posttranskriptional reguliertes Gen	*Zelltyp*	*Funktion*	*Literaturangaben*
Sekretorische Phospholipase A(sPA2), induzierbare Stickstoffoxid Synthase (iNOS)	Astrocyten	Ethanol abhängige Enzymregulierung	*(Wang et al., 2001)*
Glukose Transporter GLUT1 u. GLUT3	Gliazellen und Neuronen	Altersabhängige Regulation der Glukoseaufnahme	*(Khan et al., 1999)*
L-Thyronine Rezeptor beta 1	Oligodendrocyten	Thyroninaufnahme	*(Baas et al., 1998)*
Glutaminsäure Decarboxylase	Neuronen	Regulation der exitatorischen/inhibitorische Plastzität	*(Cao et al., 1996)*
Id-Transkriptionsfaktor	Unreife Neuronen	Regulationsprozesse in der Neurogenese	*(King et al., 1994)*
Gap-43	Unreife Neuronen	Regulationsprozesse in der Neurogenese via Hud-Elav Komplex	*(Chung et al., 1997, Steller et al., 1996)*

Tabelle 9. Beispiele für posttranskriptional regulierte Gene in Hirnzellen

1.4.3. Auswertung von Ähnlichkeitsmerkmalen zum Erkennen neuer Signalstrukturklassen

Ähnlichkeitsmerkmale oder homologe Merkmale von RNA Signalstrukturen sind die Merkmale, die zwischen den Beispielen *einer Signalstrukturklasse* auf Strukturebene und auf Sequenzebene an den spezifischen Andockstellen für ein Protein keine oder nur geringfügige Abweichungen zeigen. Sie werden bei den aus der Literatur bekannten Signalstrukturen als deren Identifikationsmerkmal herausgestellt (s.o.). Das Erkennungsproblem von funktionellen Teilstrukturen ist, ähnlich wie auf der Ebene der Consensusmotive, statistisch signifikante, innerhalb einer Standardabweichung nicht variierende Sequenzen der Strukturelemente und deren spezifische ebenfalls konservierte Längenangaben zu extrahieren. Für die Extraktion von homologen Merkmalen ist daher eine Charakterisierung von Signalstrukturen auf statistischer Basis notwendig, die die topologischen Consensus-Eigenschaften der einzelnen Strukturelemente einer Signalstrukturklasse beschreibt. Eine Consensusbeschreibung soll ebenfalls die in den Signalstrukturvarianten vorkommende Verteilung der Consensusmotive in den Strukturelementen berücksichtigen, die ein wichtiges Merkmal der meisten Signalstrukturen sind und für den nächsten Schritt, nämlich die Interpretation der gewonnen statistischen Daten genutzt werden soll.

1.4. Fragestellung

Die auszuwertenden Merkmale werden zusammenfassend als Ähnlichkeitsmaß bezeichnet. Die einzelnen Komponenten des Ähnlichkeitsmaßes (Aufschlüsselung der Strukturelementgrößen und Basensequenzen) werden in einem Merkmalsvektor zusammengefasst.

Ähnlichkeitsmerkmale:

- Kernregion
- Consensussequenz
- Typ der Strukturelemente der Sekundärstruktur
- Korrelation Strukturelement-Consensussequenz
- Abfolge der Strukturelemente
- Sequenzlänge der Strukturelemente

Um in einem ersten, bioinformatischen Ansatz die Komplexität der Fragestellung zu reduzieren, werden als Schwerpunkt dieser Arbeit zunächst nur solche Signalstrukturen betrachtet und bearbeitet, die aus einem Stemloop bestehen und hier auch als Teilstruktur der globalen Sekundärstruktur bezeichnet werden. Diese Betrachtungsweise schließt alle Strukturelemente aus, die nicht in einem Stemloop vorkommen können. Das sind Multibranchloops und Pseudoknots.

Die Menge an Teilstrukturen in einer optimalen oder suboptimalen Sekundärstrukturfaltung von beispielsweise MFold kann bei einer Sequenzlänge von 500bp etwa 20 - 30 Stemloops umfassen, und die Menge der gefalteten thermodynamisch möglichen Sekundärstrukturen pro Sequenz beträgt ca. 15-20 Faltungen. Für die daraus resultierende Menge an zu untersuchenden Teilstrukturen bei unterstellten genomischen Sequenzmengen soll ein neues Data Mining Verfahren erstellt werden, das die topologischen und sequenzspezifischen Merkmale extrahiert um die homologen Merkmale durch Vergleich der einzelnen bereits oben beschriebenen Strukturelemente zu ermitteln. Dabei ist durch entsprechende statistische Angaben die Korrelation von Consensusmotiv und bestimmten Strukturmerkmalen (z.B. „Loop“, „Bulge“, „Stem“ etc.) in den Signalstrukturen nachzuweisen.

Das Problem bei der Detektion von unbekannten, funktionellen Teilstrukturen in der globalen Sekundärstruktur einer mRNA Sequenz ist, dass es, wie oben skizziert, eine Vielzahl solcher Teilstrukturen oder Stemloops in einer RNA Sequenz gibt. Das IRE (Iron Responsive Element) ist z.B. eine Stemloopstruktur im 5'UTR der mRNA des Ferritingens, das eine wichtige Rolle bei der Regulation der Eisenaufnahme in eukaryotischen Zellen spielt. Die Sequenz des Ferritingens von verschiedenen Organismen hat aber noch ca. 15 weitere Stemloops. Weiterhin bildet sich der IRE-Stemloop dieser mRNA in nur einer von verschiedenen, thermodynamisch möglichen Zuständen. Man spricht dabei von suboptimalen Faltungen, im Unterschied zu dem thermodynamisch günstigsten, energieärmsten Zustand (siehe 1.3.3.1.), der als optimale Faltung der mRNA Sequenz bezeichnet wird. Die Schwierigkeit, den IRE-Stemloop von anderen Stemloop-Strukturen zu unterscheiden und zu identifizieren, nimmt zu,

sobald die Tatsache hinzu kommt, dass IRE-Strukturen von Ferritingenen aus den verschiedenen Organismen zwar identisch in der Funktion sind, aber in ihrer Struktur abweichende Varianten bilden können. Dies trifft auf alle bisher bekannten komplex oder einfach aufgebauten Signalstrukturen zu. Signalstrukturen sind also eine Kombination aus fixen und variablen Strukturmerkmalen. Die Herausbildung dieser Varianten in der Vielzahl möglicher Stemloopstrukturen von gleichen oder verwandten Genen verschiedener Organismen erschwert eine eindeutige Identifikation der Signalstrukturen. Die Aufgabenstellung, solche Signalstrukturen über die wiederkehrenden, fixen oder homologen Merkmale der Stemloopstrukturen als Erkennungsmerkmal möglicher neuer Signalstrukturen aus der großen Menge an Struktur- und Sequenzdaten zu identifizieren, ist typisch für Data-Mining Verfahren, wie sie vermehrt in der Bioinformatik Anwendung finden *(Bassett et al., 1999)*.

1.4.4. Zielvorstellung

Zusammenfassend soll das Ziel erreicht werden, eine Datenbank aufzubauen, in der RNA- Signalstrukturklassen annotiert vorliegen. Die Annotationen sollen sowohl die Motivsequenzen und die Literaturangaben als auch Analysemethodik und biologische Herkunft dokumentieren, auf denen die Klassenbeschreibungen basieren. Diese Annotationen sollen sowohl die Consenusangaben zu Sekundärstruktur- wie auch Sequenzinformation enthalten und einfach zu erweitern und zu warten sein. Dazu wird eine Software erstellt, die ein strukturiertes Annotationsverfahren beschleunigt und eine Arbeitsumgebung zur Verfügung stellt, die eine flexible Auswahl von analysierenden genomischen Datensätzen wie auch von Klassenbeschreibungen der Datenbank erlaubt. In diese Software integriert wird ein Screeningverfahren basierend auf Sequenz- und Strukturvergleich technisch realisiert, das die Suche nach neuen Klassenbeispielen in hirnspezifischen UTR-Sequenzen ermöglicht. Darüber hinaus sollen in einem ergänzenden zweiten Schritt die Klassenbeispiele als Fallbasis in einem Data-Mining Verfahren für die Extraktion eines Ähnlichkeitsmaßes eingesetzt werden, das dazu geeignet ist, komplett neue Signalstruktur-Klassen über eine genomische Sequenzdatenmenge zu definieren.

2. Material und Methoden

2.1. Hardwareausstattung

Die Arbeiten wurden auf einer Sun Sparc Station 10 von Sun Microsystems und einem Pentium 166 MHZ Rechner ausgeführt, der in einem heterogenen Client-Server Netzwerk eingebunden ist. Als Netzwerk- und Datenserver dient die Sparc-Station 10.

2.2. Softwareeinsatz

Die für die molekularbiologische Analyse erstellten Algorithmen wurden zunächst als Software für den lokalen Einsatz entwickelt. Seit der rasanten Entwicklung des Internet und der Weiterentwicklung des HTML-Formats als dessen Standard-Format für den Datenaustausch im World Wide Web (WWW) wird seit etwa Mitte der 90er Jahre bioinformatische Software auch über das Internet in Form von Diensten angeboten. Das bedeutet, dass die Software auf den Internetservern lokal ausgeführt wird, und anschließend das Ergebnis dem Anwender auf dem Server online zur Verfügung steht oder über elektronische Mail (Email) zugänglich gemacht wird. Damit bietet das Internet lokalen, molekularbiologisch arbeitenden Forschungsgruppen ohne komplette, bioinformatische Softwareausstattung die Möglichkeit am bioinformatischen Fortschritt teilzuhaben. Sowohl Datenbanken über molekularbiologische Literatur und Genom-Sequenzdaten sowie Algorithmen zur Suche und Analyse solcher Daten werden als Internet-basierte Dienste angeboten. Allerdings kommt es gerade wegen des aktuellen Internet-Booms hin und wieder zu Überlastungen der Netzverbindungen und damit zu zusätzlichem Zeitaufwand bei der Nutzung des Internet. Deshalb oder weil eine Software besonders häufig benötigt wird, ist es oft aus zeitökonomischen Gründen und weil nicht alle Programme im Internet als Dienst zugänglich sind, weiterhin notwendig, auf vor Ort installierte Programme zurückzugreifen. Solche Software ist entweder als „Scientific Freeware/Shareware“ oder kommerzielle Softwarepakete erhältlich.

2.2.1. Internet-basierte Dienste

Alle Dienste, die im Zusammenhang mit den hier dargestellten Forschungsarbeiten benutzt wurden, sind in der unteren Tabelle zusammengestellt.

2.2. Softwareeinsatz

Internetdienst	*Internetadresse*	*Anwendung*	*Literatur*
ENTREZ	http://www.ncbi.nlm.-nih.gov/Entrez	Sequenzrecherche und Analyse	*(Wheeler et al., 2000)*
SRS	http://srs.ebi.ac.uk/	Sequenzrecherche und Analyse	*(Etzold et al., 1993, Etzold et al., 1993 Schaftenaar et al., 1996)*
W2H, Inter-netfähiges HUSAR-Sys-tem	http://genius.emb-net.dkfz-heidelberg.de/menu/w2h/w2hdkfz/index.html	Sequenzrecherche und Analyse	*(Senger et al., 1998)*
MFOLD-Ser-ver	http://www.ibc.wustl.e-du/~zuker/rna/	Berechnung der RNA Sekundär-struktur über 1600bp	*(Zuker et al., 1989)*
BLAST 2	Http://www.ebi.ac.uk/blastall/	Suche nach lokaler Sequenzhomologie	(*Altschul et al., 1990, Altschul et al., 1997)*
FASTA 3	Http://www.ebi.ac.uk/fasta3/	Suche nach lokaler Sequenzhomologie	*(Pearson et al., 2000)*

Tabelle 10. Benutzte bioinformatische Softwaredienste und Ressourcen des Internet

Über das Suchsystem ENTREZ, das die Literaturdatenbank MEDLINE mit der DNA Sequenzdatenbank GENBANK und weiteren Datenbanken und Genom-orientierten Informationsdiensten verknüpft, ist der größte Anteil der Literaturrecherchen durchgeführt worden.

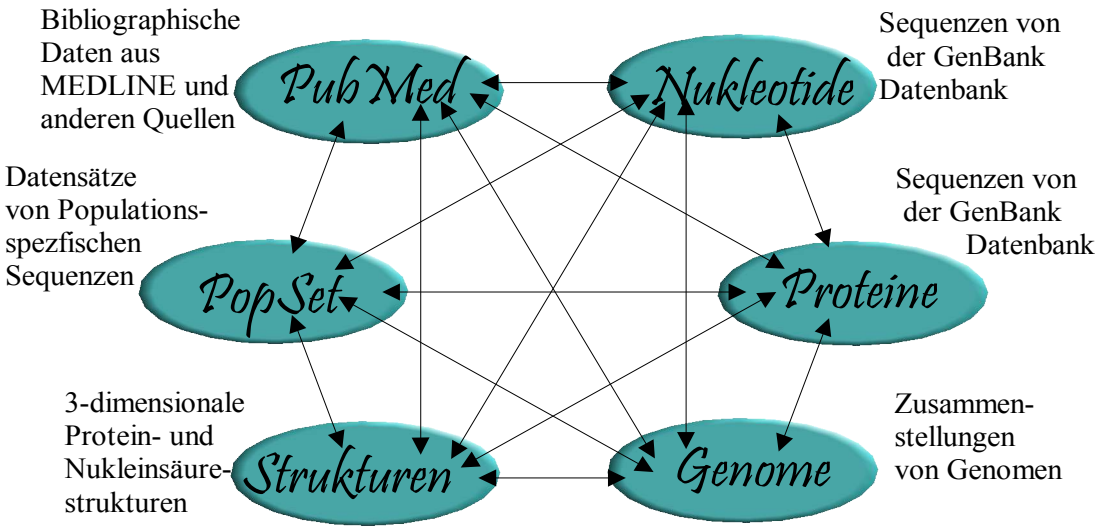


Abbildung 18. Vernetzte Dienste unter ENTREZ

Die für den Aufbau der Datenbank benötigten Sequenzeinträge aus der EMBL-Data Library wurden mit Hilfe des Systems SRS *(Etzold et al., 1996, Etzold et al., 1996)* extrahiert. Bevor diese Internet-basierten Dienste über das WWW zugänglich waren, wurden Datentransfers mittels des Terminaldienstes x-term und dem vt100 Protokollstandard unter UNIX über das Internet durchgeführt. Mit Hilfe von x-term konnte über ein Benutzerkonto auf dem Internetserver „GENIUSnet“ des Deutschen Krebsforschungszentrums DKFZ in Heidelberg, der auch EMBNET Knotenpunkt ist, Kommandos ausgeführt werden, um beispielsweise Sequenzen zu recherchieren und herunter zu laden oder zu analysieren. Dafür war als eine der noch wenigen so erreichbaren bioinformatischen Arbeitsumgebungen das HUSAR System auf GENIUSnet eingerichtet, das in wesentlichen Programmteilen auf dem GCG-Softwarepaket der „Genetics Computer Group“ der Universität Wisconsin beruht (*Senger et al. 1998)*. Auch dieses System ist mittlerweile nicht nur über W3Husar, sondern auch als eigenes WWW –basiertes System zugänglich *(Womble, 2000)*. Bevor Publikationen über das Internet recherchiert werden konnten, wurde die lokale MEDLINE Datenbankinstallation der Staats- und Universitätsbibliothek Bremen genutzt und die daraus resultierenden Daten wurden aus Hardcopy-Vorlagen in den Computer eingegeben.

2.2.2. Software vor Ort

Die Liste der lokal benutzten Software zeigt die Tabelle unten. Teilweise ist die Software auch als Clientsoftware von speziellen Internetservern explizit für die Auswertung der von ihnen gelieferten Ergebnisse entworfen worden, wie z.B. Visual BLAST und Visual FASTA *(Durand et al. 1997)* zur Visualisierung von Ergebnissen der diversen FASTA und BLAST-Server im Internet.

Software	*Betriebssystem*	*Anwendung*	*Literatur*
RNASTRUCTURE[3]	MSWindows 98/NT©	Sekundärstrukturberechnung	*(Mathews et al., 1999)*
RNAMOT	UNIX	Struktursuche in Datenbanken	*(Laferriere et al., 1994)*
RNABOB	UNIX	Struktursuche in Datenbanken	http://www.genetics.wustl.edu/eddy/
DNASTAR	MSWindows 98/NT©	Sequenz-Alignment	*(Burland et al., 2000)*
VisualBLAST VisualFASTA	MSWindows 98/NT©	Visualisierung Sequenzhomologie	*(Durand et al., 1997)*
Genedoc	MSWindows 98/NT©	Alignment-Editor	*(Nicholas et al., 1993)*
Matinspector	MSWindows 98/NT©	Matrixsuche	*(Quandt et al., 1995)*
MatInd	MSWindows 98/NT©	Matrixerstellung	*(Quandt et al., 1995)*
Patser	UNIX	Matrixsuche	*(Hertz et al., 1990)*
WConsensus	UNIX	Matrixerstellung	*(Hertz et al., 1999)*

Tabelle 11. Lokal installierte Software

Von den beiden Programmen „Patser“ und „WConsensus“ wurde auch der Quellcode von G. Hertz via FTP zur Verfügung gestellt.

2.2.3. Datenbanksysteme und Programmierwerkzeuge

Als Datenbanksystem wurde das relationale Datenbank Managementsystem PARADOX 7.0 für Personal Computer eingesetzt. PARADOX 7.0 wird über die Borland Database Engine (BDE) verwaltet, die direkt über die komponentenbasierte Programmiersoftware C++BUILDER angesprochen werden kann. Der C++BUILDER gehört zu den sogenannten RAD Werkzeugen (Rapid Application Development) und basiert auf der von INPRISE (ehemals Borland) entwickelten Pascal-Programmierumgebung DELPHI. Der C++BUILDER in der benutzten Version 3.0 integriert die

3 RNASTRUCTURE basiert auf dem MFOLD Algorithmus und stellt somit eine Portierung von MFOLD auf MS-Windows mit einigen zusätzlichen Ein- und Ausgabefiltern dar.

Compiler für die Programmiersprachen C und C++ als auch Pascal. Zur Integration von HTML-Links in die Software wurden außerdem Datenbankkomponenten von TSM Software in die Entwicklungsumgebung installiert und eingebaut.

2.3. Datenquellen

Bioinformatiker sind für ihre Arbeiten auf molekularbiologische Datenbestände angewiesen und ihnen steht mittlerweile eine Vielzahl von Datenquellen und etwa 500 molekularbiologische Datenbanken zur Verfügung. Eine Übersicht über Listen molekularbiologischer Datenbanken bietet die untere Tabelle.

Datenbank oder Liste	*Literaturangaben*
LiMB (Listing of Molecular Biological Databases, wird seit 1997 nicht mehr weiter gepflegt)	*(Lawton et al., 1989, Burks et al., 1988)*
MBDL (Molecular Biology Database List)	*(Burks et al., 1999)*
DBCat (Database Catalogue)	*(Discala et al. 2000)*
DATABANKS	*(Etzold et al., 1996)*

Tabelle 12. Metadatenbanken über molekularbiologische Datenbanken

Bei den Datenquellen hat es, seit die erste Sequenzdatenbank GenBank 1983 etabliert worden ist, eine rasante Entwicklung gegeben, die nicht nur die Zahl der Datenbanken dokumentiert, sondern auch den teilweise hohen Spezialisierungsgrad der sich in den Listen bzw. Metadatenbanken durch eine besondere Kategorisierung widerspiegelt wird. In der Aufzählung sind nicht die bibliographischen Datenbanken enthalten, da sie wegen des umfangreichen Anteils medizinischer Literatur im strengen Sinne keine molekularbiologischen Datenbanken sind. Als Literaturdatenquelle diente die medizinisch-biologische, bibliographische Datenbank MEDLINE. Es gibt zwar die alternative bibliographische Datenbank Current Contents LifeScience vom Institute of Scientific Information (ISI) bzw. in der kommerziellen Version von DIALOG als Literaturquelle, aber die Datenbank MEDLINE ist als erste Datenbank gebührenfrei im Internet erreichbar gewesen und hat den Vorteil, dass sie direkte Verweise zu Sequenzeinträgen der Sequenzdatenbanken enthält.

2.3.1. Molekularbiologische Datenbanken

Für den Aufbau der Datenbank wurden sowohl bibliographische Daten als auch DNA Sequenzdaten benötigt. Als Quelle für die Sequenzrecherche und die Sequenzeinträge in POSTREG hätten sowohl die Sequenzdatenbank GenBank (GB), die DNA Data Base of Japan (DDBJ), als auch EMBL Data-Library herangezogen werden können, weil sie aufgrund des täglichen Austauschs der Updates inhaltlich identisch sind. Genutzt wurde die EMBL Data-Library, weil sie im Internet über europäische Server

besser zu erreichen ist und weil die verwendete Software, wie z.B. MATINSPECTOR, auf das EMBL Format abgestimmt ist:

Datenbanken	*Institution*	*Literaturangaben*
EMBL Data-Library (Seit 1997 auch EMBL Nukleotide Sequence Database)	European Bioinformatics Institute (EBI), Hinxton, GB	*(Stoesser et al., 2002)*

Datenbanken	*Institution*	*Literaturangaben*
MEDLINE (Freie Internetversion)	National Library of Medcine (NLM), Fredericksburgh, USA	*(Klemencic et al., 1999)*
Protein Database of Brookhaven (PDB)	Chemical Database Service (CDS) Daresbury Laboratory, Daresbury, GB	*(Berman et al., 2000, Bhat et al., 2001)*

Tabelle 13. Datenbanken die als Literaturquellen benutzt wurden

Die Datenbank PDB wurde als Ressource für die dreidimensionale Darstellungen von RNA Posttranskriptionalen Regionen genutzt. Insbesondere enthält die Datenbank kristallographische Daten über die Interaktion von Proteindomänen mit RNA-Abschnitten.

2.3.2. Datenformate

Um Daten direkt aus molekularbiologischen Datenbanken als auch aus Software Anwendungen nutzen zu können, müssen deren Formate berücksichtigt und verarbeitet werden können. Da im Folgenden verschiedene Formate erwähnt werden, die für den Aufbau der PORD Datenbank verarbeitet wurden, sollen hier die verwendeten Formatbeschreibungen in zusammen gefasster Form beschrieben werden. Die EMBL Data Library besitzt das sogenannte Flat-File Format. Es enthält definierte Feldbezeichner im Zweibuchstabencode, die die jeweiligen Feldinhalte anzeigen und über die die Tabelle 13 auf der übernächsten Seite aufklärt. Weiterhin gibt es auch noch formale Merkmale der Formatbeschreibung, die Position und Abstände zwischen Feldbezeichnung und Inhalte festlegen. Diese Definitionen sind bei EMBL üblicherweise natürlich die Zeilenposition 1-2 für die Feldcodierung, Zeilenposition 6 für den Beginn der Feldinhalte und maximal 80 Zeichen pro Zeile. Diese globalen Einstellungen werden im weiteren Text als globale Formatmerkmale bezeichnet.

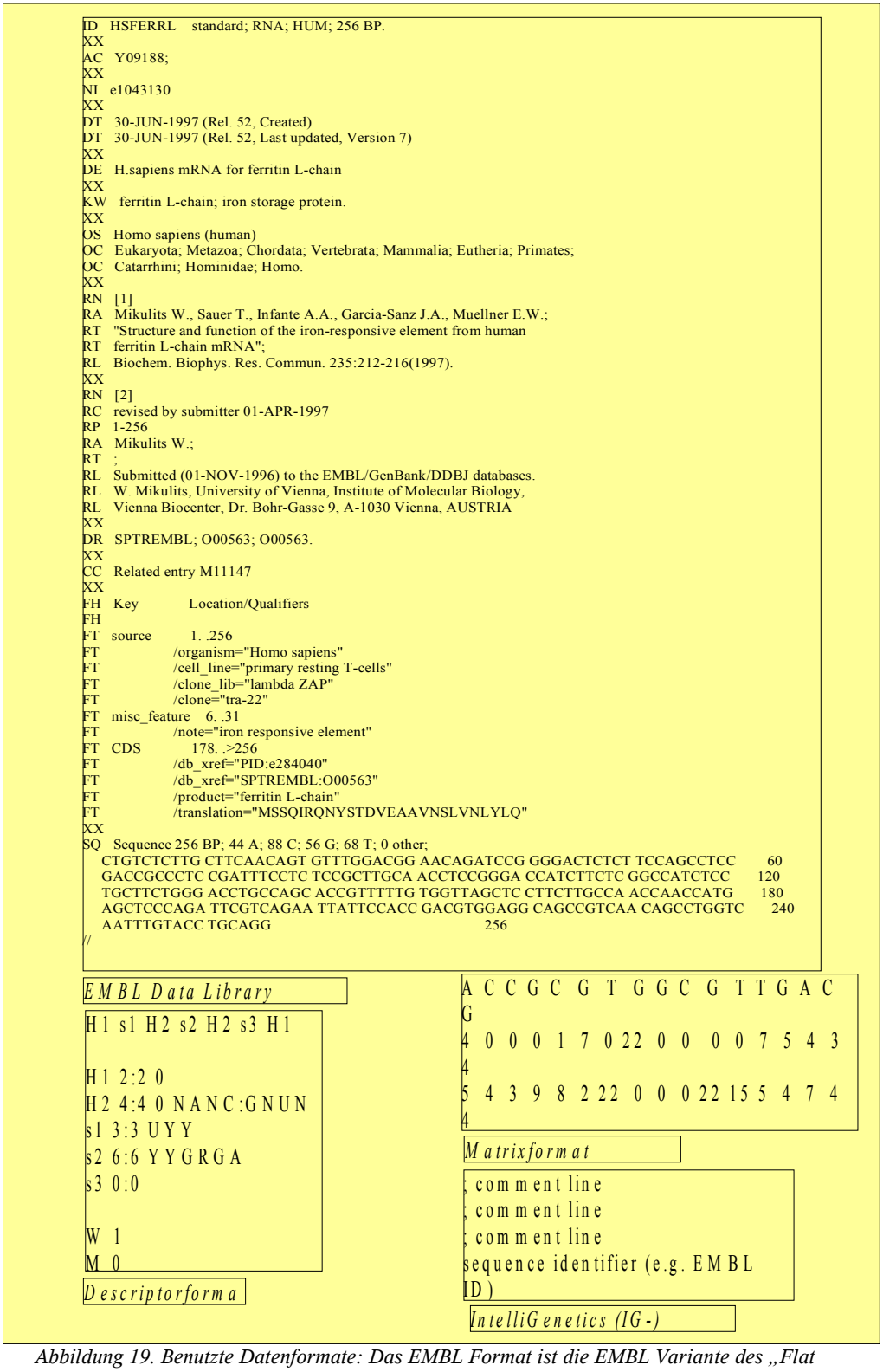


Abbildung 19. Benutzte Datenformate: Das EMBL Format ist die EMBL Variante des „Flat File“Formats. Die Descriptorbeschreibung stammt von RNAMOT. Das IntelliGenetics-Format

ist eine Variante des FASTA Formats. Beide Formate werden für die Annotierung von Sequenzen genutzt die keine Datenbankannotationen besitzen und durch eigene Annotationen beschrieben werden sollen. Das Matrixformat ist ein Beispiel das von dem Programm MatInspector genutzt wird.

AC - accession number (>=1 pro Eintrag)
SV - new sequence identifier (>=1 pro Eintrag)
DT - date (2 pro Eintrag)
DE - description (>=1 pro Eintrag)
KW - keyword (>=1 pro Eintrag)
OS - organism species (>=1 pro Eintrag)
OC - organism classification (>=1 pro Eintrag)
OG - organelle (0 or 1 pro Eintrag)
RN - reference number (>=1 pro Eintrag)
RC - reference comment (>=0 pro Eintrag)
RP - reference positions (>=1 pro Eintrag)
RX - reference cross-reference (>=0 pro Eintrag)
RA - reference author(s) (>=1 pro Eintrag)
RT - reference title (>=1 pro Eintrag)
RL - reference location (>=1 pro Eintrag)
DR - database cross-reference (>=0 pro Eintrag)
FH - feature table header (0 or 2 pro Eintrag)
FT - feature table data (>=0 pro Eintrag)
CC - comments or notes (>=0 pro Eintrag)
XX - spacer line (viele pro Eintrag)
SQ - sequence header (1 pro Eintrag)
bb - (blanks) sequence data (>=1 pro Eintrag)
// - termination line (beendet jeden Eintrag; 1 pro Eintrag)

Tabelle 14. Liste der EMBL Bezeichner

Ein weiteres wichtiges Datenformat dieser Arbeit ist das Descriptor Format, das dazu dient, Sekundärstrukturen von RNA Sequenzmotiven zu beschreiben und von dem RNAMOT Algorithmus genutzt wird. In der Formatbeschreibung kann zwischen helicalen Regionen (H) und Einzelstrangregionen (s) unterschieden werden und es erlaubt die Angabe der Anzahl von GU Paarungen, die man als Wobblepositionen bezeichnet, und Mismatches. Es lassen sich nicht nur Sekundärstrukturen beschreiben sondern mit dem RNAMOT-Algorithmus und einer Descriptor-Vorlage nach gleich-

artigen Sekundärstrukturen suchen. Für die Suche nach Sequenzen ohne Strukturinformation gibt es die bereits erwähnten Matrixanwendungen. Die Matrices werden ebenfalls in einem definierten Format angegeben, wobei zwischen den Formaten unterschieden wird nach horizontalen und vertikalen Matrixformaten. Das Programm Matinspector verlangt beispielsweise vertikale Matrices wie es in dem nebenstehenden Beispiel angegeben ist. Dabei erfolgt die Angabe der Basentypen von oben nach unten in der Reihenfolge A, C, G, T.

```
UI  - 0
AU  - Rundlof AK
AU  - Carlsten M
AU  - Giacobini MM
AU  - Arner ES
TI  - Prominent expression of the selenoprotein thioredoxin reductase in the
      medullary rays of the rat kidney and thioredoxin reductase mRNA variants
      differing at the 5' untranslated region.
LA  - ENG
PT  - JOURNAL ARTICLE
DA  - 20000418
DP  - 2000 May 1
IS  - 0264-6021
TA  - Biochem J
PG  - 661-668
IP  - Pt 3
VI  - 347
JC  - 9YO
AB  - The mammalian selenoprotein thioredoxin reductase is a central enzyme in
      protection against oxidative damage or the redox control of cell function.
```

Abbildung 20 Beispiel für MEDLINE Format für Literaturangaben

Für den Datenaustausch zwischen verschiedenen Anwendungen und Datenbanken werden häufig Intermediärformate verwendet, die die ursprünglichen Inhalt beispielsweise von Sequenzeinträgen aus Datenbanken auf kurze Sequenzabfolgen und ein Namenskürzel sowie einen Kurzkommentar zusammenkürzen und somit für die Weitergabe von Ausschnitten aus Sequenzen an verschiedene Algorithmen zur Bearbeitung geeignet sind. Zu solchen Formaten gehören das FASTA Format, das ursprünglich als Eingabeformat für den Suchalgorithmus FASTA entstanden ist. Ganz ähnlich aufgebaut ist die IntelliGenetics (IG) Formatbeschreibung. Im Gegensatz zu der FASTA Formatbeschreibung wird im IG-Format der Kommentar durch Semikolon angezeigt und nicht durch das Größer-Zeichen „>“. Die Literaturdatenbank MEDLINE besitzt ebenfalls ein eigenes Format mit Feldbezeichnungen am linken Rand. Das Format hat aber im Gegensatz zum Flat-File Format hinter der Zweibuchstaben-Codierung ein Bindestrich mit einem Leerzeichen dahinter.

3. Ergebnisse

3.1. Zusammenstellung des Datenmaterials

Das Vorgehen zur Erstellung einer Datenbank über posttranskriptional wirksame Sequenzabschnitte und deren Struktur beinhaltete zunächst das Anlegen einer Dateibibliothek, in der Literaturangaben aus der Literaturdatenbank MEDLINE und Gensequenzen aus der DNA-Datenbank EMBL Data Library als Dateien in Verzeichnissen abgespeichert wurden, die jeweils nach den Gennamen benannt und geordnet sind. Für jedes Gen mit einer Postreg-Sequenz eines definierten Typs bzw. einer Klasse, das aus einem Organismus einer Spezies stammt, wird jeweils ein Sequenzeintrag abgespeichert, wenn nicht paraloge Gene oder Gene unterschiedlicher Allele verschiedene Postreg-Sequenzen aufweisen. So wird das gesamte Spektrum der Postreg-Sequenzvariationen eines Typs oder einer Klasse einer Art abgedeckt.

3.1.1. Zusammenstellung der PORD Literaturdaten

Ausgangspunkt der Zusammenstellung des Datenmaterials zum Aufbau der Signalstruktur-Datenbank sind Publikationen aus der biomedizinischen Literaturdatenbank MEDLINE, welche die Herkunft, die Sequenz- und die Struktureigenschaften sowie die Funktion von RNA Signalstrukturen beschreiben. Solche Publikationen sind in der Literaturdatenbank MEDLINE durch Datenbankabfragen recherchiert worden, denen jeweils zwei unterschiedliche Strategien zugrunde lagen. Einerseits wurden Publikationen durch Eingeben bestimmter Stichwörter recherchiert. Diese Stichwörter sind in der folgenden Tabelle dargestellt.

Abfrageziel	*Abfragewörter*
Sequenztyp	mrna, rna, messenger
Sequenzbereich	utr, untranslated region, ncr, noncoding region
Sequenzfunktion	motif, aktive region, element

Tabelle 15. Suchwörter für Literatursuche

Bei den gefundenen Publikationen, die posttranskriptional aktive Regionen in mRNAs dokumentieren, wurde die zusätzliche Möglichkeit genutzt, die mit der Publikation verknüpften „verwandten Publikationen“ aufzurufen. Diese automatische Funktion führt zu einer Auswahl von Veröffentlichungen in MEDLINE, deren Stichwortindex eine große Übereinstimmung zu dem Stichwortindex der jeweiligen Ausgangspublikation hat und in der deshalb ebenfalls Veröffentlichungen zu der Thematik posttranskriptional aktiver Domänen in mRNAs zu finden sind. Allerdings muss die Anzahl der verwandten Publikationen überschaubar sein, um eine manuelle Sichtung durchzuführen. Bei allen diesen Datenbankabfragen wurde der MESH-Index einge-

setzt. Der MESH-Index ist ein von der National Library of Medicine erstellter und laufend aktualisierter „Medicine Subject Headings“-Index, der den jeweils neu erschienenen Publikationen zugeordnet wird. Er spiegelt den Inhalt der Publikation durch eine globale Kategorisierung der Artikel wider. Die Kategorien sind dabei hierarchisch gegliedert. Die Liste der Kategorien der obersten Ebene und die Liste der eingesetzten MESH Begriffe sind ist in den folgenden Tabellen dargestellt.

MESH Wurzelkategorien
1. Anatomy [A] 2. Organisms [B] 3. Diseases [C] 4. Chemicals and Drugs [D] 5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E] 6. Psychiatry and Psychology [F] 7. Biological Sciences [G] 8. Physical Sciences [H] 9. Anthropology, Education, Sociology and Social Phenomena [I] 10. Technology and Food and Beverages [J] 11. Humanities [K] 12. Information Science [L] 13. Persons [M] 14. Health Care [N] 15. Geographic Locations [Z]

Tabelle 16. Liste der Ausgangskategorien im MESH Index

MESH Headings
RNA, Messenger [D13.444.735.544]
Untranslated Regions [D13.444.735.544.875]
Consensus Sequence [G05.331.599.056.160]
Sequence Homology, Nucleic Acid [G05.331.599.110.791]
Regulatory Sequences, Nucleic Acid [G05.331.599.110.689] +
Response Elements [G06.184.599.110.689.700]

Tabelle 17. Genutzte MESH-Indexeinträge

Die Suchwörter und die MESH-Indexeinträge wurden über Booleansche Operatoren miteinander verknüpft. Dieses schon ältere Verfahren für Datenbankabfragen ist durch die HTML-Technologie mittlerweile sehr vereinfacht worden, da das Wissen um die korrekte Syntax einer Abfrage nicht mehr benötigt wird, sondern in Form von HTML Links unter dem PubMed-Service bereits implementiert ist.

Die aus den Datenbankabfragen resultierenden Veröffentlichungen führten zu etwa 120 dokumentierten Postreg Motiven, wobei natürlich zu jedem Motiv unterschiedliche Anzahlen an Publikationen vorhanden waren. Entsprechend dem Umfang der Dokumentation konnten die Annotationen in der eigenen Datenbank erstellt werden.

3.1.2. Zusammenstellung der PORD Sequenzdaten

Für die Zusammenstellung der Sequenzen waren die Sequenzbeschreibungen in den Publikationen der Ausgangspunkt. Diese Beschreibungen sind allerdings unterschiedlich präzise. Dies hat mehrere Gründe. Zwar ist einerseits durch die Einführung eines einheitlichen Schlüssels, der sogenannten „Accession number“, für Datenbankeinträge zwischen den Sequenzdatenbanken EMBL Data Library, DDBJ und GenBank und andererseits durch die Vereinbarung vieler Herausgeber von Fachzeitschriften bei der Publikation von Sequenzdaten die Nennung der dazugehörigen „Accession number“ zwingend vor zu schreiben ein direkter Zugriff von einer Publikation auf die dazu gehörige Sequenz möglich. Aber solche Vereinbarungen sind noch relativ jung und deshalb noch nicht generell üblich. Deshalb muss man bei den meisten Publikationen zusätzlich eine Recherche nach der entsprechenden „Accession number“ durchführen. Dies geschieht mittels der Angaben in der Publikation zu Organismus, Zell- oder Gewebetyp und des Gennamens. Für die Abfrage der Sequenzeinträge wurde das bereits erwähnte SRS-System benutzt. SRS bietet für Fragen unter den Bezeichnungen „Standard“ und „Extended“ (Siehe Abbildung 20.) zwei Modi an. Der Modus „Extended“ bietet die Möglichkeit, für jedes Feld der EMBL Datenbank separate Kriterien in einer Abfrage anzugeben und war daher am besten geeignet, die Angaben aus der Literatur für die Suche nach Sequenzen aufzuschlüsseln. SRS bietet mittlerweile zusätzlich die Möglichkeit, Postreg-Sequenzen direkt als Teilsequenz eines kompletten EMBL-Eintrags zu suchen. Diese Option hängt natürlich davon ab, ob Postreg-Sequenzen zu dem Zeitpunkt bereits bekannt waren, als der Eintrag eingegeben wurde bzw. ob der Eintrag weiter gepflegt worden ist oder ob Recherchen zu einer nachträglichen Entdeckung von Postreg-Sequenzen geführt haben, die als Update des EMBL-Eintrags eingegeben wurden. Mittlerweile gibt es beispielsweise einen eigenen Eintrag im „Feature Table“ von EMBL für „Stemloops“ (siehe Tabelle 17.). Diese Verknüpfungen konnte die Literatursuche rückwirkend ergänzen und vervollständigen.

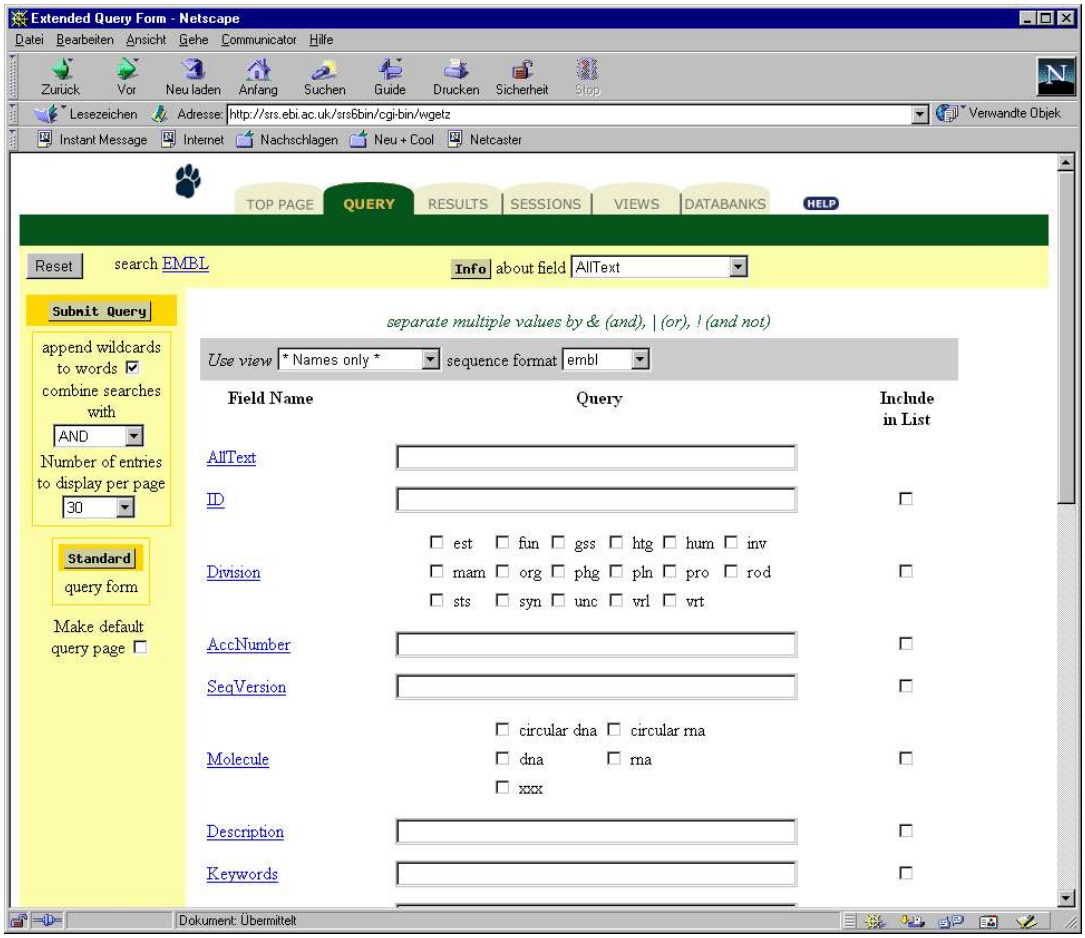


Abbildung 21. Extended Modus unter SRS

Beispiel für einen EMBL-Subeintrag für Stemloops
ID MEAJ4069_2; parent: MEAJ4069
AC AJ224069;
FT stem_loop 980..996
SQ Sequence 17 BP;
agccctttta agggcta 17
//

Tabelle 18. Histon H1 Stemloop Eintrag aus EMBL Data Library

Die so ermittelten Sequenzen wurden als Dateien in der Dateibibliothek abgespeichert.

3.1.2.1. Verfahren zur Selektion der Motivsequenzen

Die Postreg Sequenzen aus den vorhandenen Sequenzeinträgen, für die die Information über Position und Länge aus den entsprechenden Publikationen bekannt war, wurden zunächst separiert und im IG-Format abgespeichert. Wenn es mehrere Publikationen für eine Postreg-Sequenz aus demselben Organismus, Zelltyp und Gen gab, die sich in der Aussage über die Längen unterschieden, wurde die aktuellste Version als Vorlage benutzt. Dabei spielte auch eine Rolle, wie das experimentelle Vorgehen zu bewerten ist. So ist eine Aussage basierend auf einer kristallographischen oder spektroskopischen Analyse in der Zuverlässigkeit höher zu bewerten als ein Footprint Experiment. Solch eine Vorgehensweise der Bewertung von Experimenten ist durchaus üblich, und die Datenbank Transfac hat dafür sogar ein eigenes Bewertungsschema entwickelt. Eine Ressource für aufgeklärte Tertiärstrukturen bietet die PDB Datenbank, die - obwohl als Datenbank für Proteine bezeichnet - mittlerweile über 80 RNA-Einträge verfügt. In einem zweiten Schritt wurden die Publikationen herangezogen, aus denen nicht hervorging, welche konkrete Position das Postreg Element hat. Teilweise waren in solchen Publikationen nur die Primerpositionen aus einer mittels PCR durchgeführten Postreganalyse bekannt. Dieses Problem konnte meistens durch eine einfache Suche nach den Kernelementen einer Postreg Sequenzabfolge im IUB-IUPAC Code gelöst werden. Die Verwendung der IUB-IUPAC Codes und ein entsprechender einfacher Suchalgorithmus wurden neben der aufwendigeren Matrixsuche von der Software MATINSPECTOR angeboten und genutzt. In den seltenen Fällen, in denen die Sequenz der EMBL-Einträge besonders groß war und mehrere Matches anzeigten, dass die Consensussequenz einer Postreg-Sequenz nicht sicher eingegrenzt werden kann, wurde die Consensussequenz ergänzend mit Hilfe des Programms BLAST gesucht.

3.1.2.2. Verfahren zur Erstellung der Sekundärstrukturen

Bei Postreg-Motiven, in denen die Sekundärstrukturen Teil der Signalinformation sind und somit eine wichtige Rolle für die Protein-RNA Interaktion spielen, wurden die Sequenzen anschließend mit der Windows-Version des Programms MFOLD, RNASTRUCTURE, erstellt, welches das IG-Format verwendet. RNASTRUCTURE erstellt wie MFOLD aus einer Sequenz mehrere suboptimale Faltungsvarianten und berechnet mit einer Zuverlässigkeit von bis zu 85% die freie Energie dieser Strukturvarianten. Leitfaden für die Auswahl einer Faltungsvariante als Strukturrepräsentant der Signalstruktur in die Datenbank waren die publizierten Sekundärstrukturen. Um möglichst realistische Energiewerte zu erhalten, wurden die Sekundärstrukturen mit jeweils einer zusätzlichen freien Base stromauf- und stromabwärts der Sekundärstruktur erstellt, sofern die publizierte Struktur keine freien Enden - sogenannte „dangeling ends“ - besitzt. Damit soll vermieden werden, dass der endständigen Basenpaarung, die die Struktur „zusammenhält“, verhältnismäßig größere Bindungsenergie in der Berechnung zugemessen wird als einer normalen, nicht endständigen Basenpaarung. Der errechnete Gesamtwert der freien Energie würde dadurch verfälscht werden. Die gefalteten Strukturen wurden im MFOLD-eigenen „connect“- oder abgekürzt „ct“-Format in der Dateibibliothek zu den Medline-Einträgen abgespeichert.

3.2. RNA POSTREG Klassen in der Datenbank PORD

Die neu erstellte Datenbank hat die Bezeichnung PORD für „POsttranskriptional RNA Regulatory Region Database“. Das Datenmaterial der Datenbank basiert auf den gespeicherten Dateien der Dateibibliothek. Anhand der Literaturangaben zu einer PORD Klasse wird eine die Klasse beschreibende Annotation eingegeben, die auf Sequenzebene durch eine Consensusbeschreibung repräsentiert wird. Die Consensusbeschreibung besteht aus einer Matrix, die die Wahrscheinlichkeit des Vorkommens einer jeden Base an jeder Position innerhalb der Postreg Consensussequenz in der IUB-IUPAC-Notation angibt. Als zweites Element der Consensusangaben gibt es die Klassenbeschreibung der Sekundärstruktur im DESCRIPTOR-Format, wie es von dem Programm RNAMOT verwendet wird. Es enthält eine Beschreibung der Consensusstruktur nur der Postreg-Sequenzen, deren Struktur als Teil der RNA-Protein Bindungsinformation in der Literatur beschrieben ist. Jede einzelne neu importierte Motivsequenz aktualisiert die vorhandene Matrix der Klasse, indem sie automatisch entsprechend der neuen Sequenz verändert wird. So entspricht die Klassenbeschreibung immer exakt der Zahl der vorhandenen Einzelsequenzen, über die eine Klasse definiert wird. Jede einzelne Postreg-Sequenz wird zusammen mit der Sekundärstruktur und einem Teil der EMBL Annotationen wie z.B. der Taxonomie des Organismus in einem automatisierten Importschritt übernommen. Bei diesem Sequenzimport wird das Signalstruktur-Sequenzstück zusätzlich separat von den Sequenzen mit Bezeichnung der Transkriptregion (3'/5' UTR, Intron usw.) abgespeichert.

Wie bereits erwähnt, wurden die Annotationen nach dem Flat-File Formatkonzept der EMBL-Data Library aufgeschlüsselt. Soweit die Feldbezeichner der Annotationen sich auf Feldinhalte beziehen, die analog zu den Feldinhalten der EMBL Datenbank sind, wie z.B. in EMBL die Beschreibung der Taxonomie des Organismus, aus der das Gen mit der Postreg Sequenz stammt, wird die Annotation unter der analogen Feldbezeichnung in diesem Beispiel unter „OS“ für „Organismsource“ eingetragen. Dadurch soll eine möglichst hohe Interoperabilität der PORD Datenbanken mit vorhandenen molekularbiologischen Datenbanken erreicht werden. Dies ist ein anerkanntes Qualitätsmerkmal für alle molekularbiologische Datenbanken *(Karp et al., 1996)*.

3.2.1. Problem der Eindeutigkeit einer Klassenbeschreibung

Ziel der Datenbank ist es, nicht nur die bereits bekannten Postreg Sequenzen zu dokumentieren sondern auch diese Motivsequenzen so zusammen zu stellen und zu beschreiben, dass daraus neue Information entsteht und genutzt werden kann. Diese neue Information soll die Klassenbeschreibung sein, die verschiedene einzelne Motivsequenzen zu einer Klasse zusammenfasst. Ein Teil dieser Klassenbeschreibung ist die bereits erwähnte Consensusbeschreibung. Eine Consensusbeschreibung wird aus einzelnen Motivsequenzen erstellt, sofern diese sich durch sequenzspezifische, strukturelle und funktionelle Gemeinsamkeiten auszeichnen. Solche Gemeinsamkeiten sind:

- Sequenz- und Strukturhomologie
- Funktionale Übereinstimmung
- Herkunftsspezifische Übereinstimmung (Taxonomische Einordnung als auch Zelltyp)

Die Sequenzhomologie spiegelt in der Regel eine definierte genregulatorische Funktion innerhalb eines definierten zell- oder organismentypischen Zusammenhangs wider. Die meisten Beispiele posttranskriptional aktiver Sequenzmotive in mRNA Sequenzen zeigen eine eindeutige Sequenzhomologie, selbst wenn die Motivsequenz in verschiedenen Genen oder/und in verschiedenen Zellen oder Organsimen vorkommen und lassen sich daher eindeutig zu Klassen zusammenfassen. Die Sequenzhomologie ist aber selbst nur ein Durchschnittswert. Nur wenige Positionen in der Sequenzabfolge der Postreg Sequenzen lassen sich eindeutig einer Base zuordnen. Diese Sequenzabfolgen bezeichnet man als die sogenannten Kernregionen. In allen übrigen Positionen werden die Basen durch den IUB-IUPAC Code als begrenzt variierende Basen dargestellt. In einigen Beispielen aus der Literatur werden Signalstrukturen zu einer Klasse zusammengefasst, die sich aber aufgrund ihrer Funktion und Consensusbeschreibung in Sekundärstruktur und Sequenz weiter aufspalten lässt. Diese Tatsache kann in wenigen Fällen zu Problemen bei einer eindeutigen Klassendefinition führen. So hat das Element „IRE“ sowohl die Funktion, die Translation zu inhibieren als auch die, die mRNA zu stabilisieren, je nachdem ob es im 5' UTR des Ferritingens oder im 3'UTR des Transferringens vorkommt. Es besitzt homologe Sequenzabfolgen in den beiden Genen, aber die Sequenzhomologie zwischen den IREs der Trans-

ferringene und der Ferritingene untereinander weist noch höhere Übereinstimmung auf als zwischen allen zusammen. Beispielsweise sind die IRE Sequenzen in Transferringenen in der Regel länger und besitzen zusätzliche sogenannte flankierende Regionen *(Theil et al. 1998)* an ihren 5' und 3' Enden, die den IRE Elementen im Ferritingen fehlen. Ein weiteres Beispiel für eine Sequenzhomologie die sich weiter aufspalten lässt, ist das SECIS-Element. In diesem Fall ist es aber nicht die wechselnde Funktion sondern es sind verschiedene Organismenreiche, in denen die Sequenzhomologie innerhalb der jeweiligen Organismenreiche im Gegensatz zu der übergeordneten Sequenzhomologie noch verstärkt ist *(Low et al., 1996)*. In der Literatur werden solche Grenzfälle unterschiedlich behandelt. Während bei dem Beispiel des SECIS Elements der Vorschlag gemacht wurde, die SECIS Elemente in drei Subklassen einzuteilen, behandelt man die IRE Elemente als eine Klasse.

3.2.2. Klassendefinition in PORD

Diese Verschiedenheit in der Definition von PORDklassen muss für das Datenbankvorhaben vereinheitlicht werden. Es muss für eine konsistente Klassendefinition entschieden werden, ob eine Klassenbeschreibung jeweils nur die kleinste mögliche Gemeinsamkeit der Sequenz- und Strukturinformation aller Signalstrukturen ausdrücken soll oder ob sie darüber hinaus auch den funktionellen, sequenzregionspezifischen, zellulären und taxonomischen Hintergrund enthält, sofern sich dieser in einer abweichenden Consensusbeschreibung nieder schlägt. Wählt man die erste Variante, dann hat man eine Consensusbeschreibung einer Klasse, die aufgrund der stärker eingeschränkten Anzahl der Sequenz und Sekundärstrukturmerkmale einerseits größere Variabilität enthält, andererseits einfacher zu warten wäre. Eine stärkere Eingrenzung der Klassendefinition durch Aufspaltung in mehrere Klassen hingegen schränkt die Variabilität der Consensusbeschreibung ein und führt zu geringerer Variabilität und höherer Informationsdichte, die die Eignung der Consenusbeschreibung als Vorlage für die Suche nach neuen homologen Signalstrukturen der Klasse verbessert. Zusätzlich kann bei letzterem Vorgehen antizipiert werden, dass eine neu gefundene Signalstruktur einer Klasse den gleichen funktionalen oder taxonomischen Hintergrund hat, wie er in der Klassenbeschreibung ausgedrückt ist. Dadurch sind die Vorhersagen über die Eigenschaften neu gefundener Klassenbeispiele leichter zu treffen. Aus dieser Erwägung heraus wurde hier das Vorgehen gewählt, die Klassen wie im Beispiel des SECIS Elements in weitere Klassen aufzuspalten, wenn die Consensusbeschreibung in der Literatur es erlaubt. Deshalb ist eine Klasse für die Datenbank PORD folgendermaßen definiert:

Ausgangspunkt einer PORD-Klasse ist die gemeinsame Consensussequenz und -struktur der einzelnen RNA-Beispiele, die mit einer genregulatorischen Funktion und einer eingrenzbaren zell- und organismenspezifischen Herkunft und einem bindenden Kofaktor korreliert. Eine Funktion ist dabei definiert als die Bindung verschiedener Kofaktoren, sei es einer anderen RNA Sequenz oder eines Proteins, durch die eine für diese Bindung spezifische zelluläre Veränderung im posttranskriptionalen Mechanis-

mus und Verhalten des Transkripts ausgelöst wird. Ist die Consensussequenz modifiziert und fällt dies zusammen mit einer Änderung in einer der drei genannten Kriterien Funktion oder Herkunft oder Kofaktorspezifität, dann begründet dieser Wechsel eine neue Klasse. Tritt ein solcher Wechsel auf, ohne dass sich dies in einer neuen Consensusequenz oder Sekundärstruktur niederschlägt, wird keine neue Klasse angelegt.

Entsprechend diesem Klassenkonzept ist die Datenbank PORD aufgebaut. Sie dokumentiert jede einzelne Motivsequenz und die daraus erstellte Consensussequenz, die zusätzlich mit Annotationen über die Herkunft und die Funktion der Motivklasse versehen ist.

3.2.3. Klassendefinition modularer RNA-Signalstrukturen

Ein weiteres Problem für eine Klassenbeschreibung ist der modulare Aufbau einer Signalstruktur. Das „Bicoid Localization Element“ BLE besteht beispielsweise aus drei in 5'-3' Richtung hintereinander liegenden Stemloops. Jedes der Strukturelemente vermittelt einen separaten Schritt in der gesamten Lokalisierung des Transkripts zum anterioren Pol einer Eizelle der Fruchtfliege (Ferrandon et al., 1997). Das Fehlen eines Strukturelements beeinträchtigt den Lokalisierungsvorgang bis hin zu seiner vollständigen Unterbindung. Daher kann man hier von einem modularen Aufbau sprechen, da einerseits alle drei Stemloops funktionell zusammengehören, aber andererseits separate Teilschritte des Lokalisierungsvorgangs vermitteln. Solche Signalstrukturen können natürlich komplett in der Klassenbeschreibung repräsentiert werden, würden dann aber in Widerspruch zu obiger Klassendefinition stehen und zweitens auch dem Konzept wiedersprechen, dass die Klassenbeschreibung möglichst aus einfachen Stemloopbeschreibungen aufgebaut sein soll. Gelöst wird dieses Problem, indem jeweils separat für jeden Stemloop die Consensusbeschreibung in einem separaten Klasseneintrag angelegt wird und der modulare Zusammenhang durch eine Datenbankreferenz auf die jeweils anderen Klasseneinträge angezeigt wird (siehe Tabelle 19.)

DR	PORD; M:C00004; BL1$FLY;
DR	PORD; M:C00005; BL2$FLY;
DR	PORD; M:C00006; BL3$FLY;

Tabelle 19. Beispiel von Datenbankreferenzen innerhalb von PORD, die das komplette Modul (M:) aus drei Stemloops beschreiben.

3.2.4. Relationales Konzept der Datenstruktur von PORD

Die Datenbank setzt sich hauptsächlich aus 4 Relationen bzw. Tabellen zusammen. Die Literaturtabelle „Literature“ enthält die aus MEDLINE importierten Datensätze, wie sie bereits in der Dateibibliothek vorliegen. In der Tabelle „Site“ werden jeweils

die einzelnen Postreg Sequenzstücke bzw. Signalstruktur, Sequenz und Annotationen dazu, die entweder aus EMBL übernommen oder aus vorgegebenen Einstellungen beim Import stammen (siehe 3.3.), gespeichert. Die Tabelle „Gene“ enthält die Beschreibung des Gens, in der die Signalstruktur enthalten ist, und in der Tabelle „Class“ wird die Sequenz- und Strukturinfomation zu den einzelnen Signalstruktur-Sequenzen einer Klasse zusammengefasst und eine Consensusbeschreibung erstellt.

3.2.4.1. Die PORD Literaturrelation

Die Literaturdaten werden in einer separaten Relation dokumentiert. Sie ist aus folgenden Feldern aufgebaut:

Feldbezeichner	*Feldname*	*Datentyp*	*Feldinhalt*
RX	Medline no.*	Integer	Medline Eintragsnummer
RN	Number	Integer	Zahl der Einträge zu einer Sequenz
RN	Number	Integer	Zahl der Einträge zu einer Klasse
RT	Titel	String	Referenztitel
RA	Authors	String	Referenzautoren
RL	Source	String	Referenz-“link“
RL	Year	Date	
RL	Volume	Integer	
RL	No	Integer	
RL	Page	String	
	Accession no.**	String	Eintragsnummer der Tabelle „Site“
	Accession no.**	String	Eintragsnummer der Tabelle „Consensus“

Tabelle 20. Literaturrelation

Die Felder "Accession no.**" repräsentieren die Verknüpfung der Tabelle zu der Signalstrukturrelation (s.u.) und zu der Klassenbeschreibung und sind als Sekundärschlüssel mit zwei Sternen gekennzeichnet. Der Primärschlüssel ist mit einem Stern markiert. Der Sekundärschlüssel dient zur Verknüpfung der Tabelle mit der Sequenztabelle „Site“ als N:N Verknüpfung. Um eine N:N Verknüpfung zu realisieren,, wird die Tabelle nicht direkt, sondern über eine zwischengeschaltete Sequenz-Medline Indextabelle referenziert.

3.2.4.2. Die PORD-Sequenzrelation

Für die Dokumentation der einzelnen Sequenzen der Postreg-Motive, aus denen die Consensussequenzen und -strukturen erstellt werden, wird eine separate Relation benötigt, die die Herkunft, Position und die Eigenschaften der einzelnen Sequenzeinträge beschreibt.

Feldbezeichner	*Feldname*	*Datentyp*	*Feldinhalt*
AC	Accession no.*	String	Eintragsnummer
ID	Identifier	String	Eintragsbezeichner
DT	Date	DateString	Eintragsdatum (Erstellungsdatum)
DT	Editor	String	Namenskürzel Überarbeiter
DT	Update	Date	Aktualisierungsdatum
DT	Editor	String	Namenskürzel Überarbeiter
SQ	Sequence	String	Motivsequenz
RG	Region	String	Sequenzbereich 3'/5'UTR, Intron, Exon
SF	Begin	Integer	Anfangsposition der Motivsequenz[4]
ST	End	Integer	Endposition der Motivsequenz
MM	Method	String	Experimentelle Methodik
DR	DB-Refenrence	String	Datenbankverweis (i.d.R. EMBL)
FE	Folding Energy	Integer	Faltungsenergie (ΔG)
FD	Foldingdescription	Memo	Sekundärstruktur im Deskriptorformat
	Accession no**	String	Zugangsnummer von „Gene“
	Medline no**	Integer	Zugangsnummer von „Literature“

Tabelle 21. Relation der Postreg-Sequenzen

Über die Sekundärschlüssel (**) wird die Literaturtabelle als N:N Verknüpfung und die Gentabelle als 1:N Verknüpfung referenziert. Die Felder "Structure" und "Description" enthalten die statistische Information über die Sekundärstrukturfaltung von MFOLD und die daraus ableitbare Sekundärstrukturbeschreibung.

4 Von der Startstelle der Transkription gezählt. Die Nummer bezeichnet die erste Base des Motivs.

3.2.4.3. Die PORD Genrelation

Die Beschreibung des Gens, in der die Postreg-Sequenz enthalten ist, beruht im Wesentlichen auf importierten, gefilterten Daten aus den Feldern OS, OC und FT eines Eintrags der EMBL Data-Library. Die EMBL-Daten werden so importiert und

Feldbezeichner	*Feldname*	*Datentyp*	*Feldinhalt*
AC	Accession no.*	String	Eintragsnummer
ID	Identifier	String	Eintragsbezeichner
DT	Date	DateString	Eintragsdatum (Erstellung&Update)
DT	Editor	String	Namenskürzel Überarbeiter
DT	Update	Date	Aktualisierungsdatum
DT	Editor	String	Namenskürzel Überarbeiter
OS	Organism	String	Organismus
OC	Organismkingdom	String	Taxonomie
DE	Description	String	Genname
SD	Short Description	String	Abgekürzter Genname
SO	Tissue type	String	Gewebetyp
SO	Type of Cell	String	Zellentyp
AC	Accession no.**	String	Zugangsnummer von „SITE“

Tabelle 22. Felder der Genrelation

aufgeschlüsselt, dass eine separate Recherche nach Gennamen oder Organismus möglich ist. Der angegebene Sekundärschlüssel bezieht sich auf die PORD-Sequenztabelle zu der eine 1:N Verknüpfung existiert, weil in einer Gensequenz mehrere Postreg-Sequenzen enthalten sein können. Die Werte in den drei Feldern "Type of Cell", "Tissue type" und "Organism" sollen wiederum zur taxonomischen und zelltypischen Spezifizierung als Klassifikatoren dienen, um die Postreg-Sequenzen nach ihrer Herkunft sortieren zu können. Sofern zu einem Gennamen Synonyme existieren, werden diese in einer zusätzlichen Tabelle mit dem Namen „Synonyms“ gespeichert, die als 1:N Verknüpfung mit der Gentabelle verbunden ist, da für einen Gennamen mehrere Synonyme existieren können.

Feldbe-zeichner	*Feldname*	*Datentyp*	*Feldinhalt*
	No.*	Integer	Fortlaufende Eintragsnummer
SY	Synonyme	String	Genname
SY	Short Synonyme	String	Abgekürzter Genname
	Accession no.**	String	Zugangsnummer von „Gene“

Tabelle 23. Relation für Synonyme von Gennamen oder deren wissenschaftliche Kurzbezeichnung

3.2.4.4. Die PORD Klassenrelation

Zentrale Bedeutung hat die Relation, die die Daten zur Consensusbeschreibung der Signalstrukturklassen enthält und die Bezeichnung „Class“ hat. Die regulatorische Information der Signalstruktur kann nur in der Sequenzabfolge oder auch in der Sequenzabfolge und der Sekundärstruktur enthalten sein. Beide Informationen sind in Form einer Repräsentation im IUPAC-IUB Code zugänglich, die die gemeinsamen Sequenz- und Struktureigenschaften aller Einzelbeispiele der RNA PORD-Sequenztabelle als Klasse darstellt.

3.2. RNA POSTREG Klassen in der Datenbank PORD

Feldbe-zeichner	*Feldname*	*Datentyp*	*Feldinhalt*
AC	Accession no.*	String	Eintragsnummer
ID	Identifier	String	Eintragsbezeichner
DT	Date	Date	Eintragsdatum (Erstellungsdatum)
DT	Editor	String	Namenskürzel Bearbeiter
DT	Update	Date	Aktualisierungsdatum
DT	Editor	String	Namenskürzel Bearbeiter
EL	Elementname	String	Name der Signalstruktur
SE	Short Elementname	String	Namenskürzel der Signalstruktur
FQ	Feature Qualifier	String	Funktionsbeschreibung
CS	Consensussequenz	String	Consensussequenz in IUPAC Code
CD	Consensusdescrip-tion	Memo	Consensus-Sekundärstrukturbe-schreibung
CC	Comments	Memo	Kommentar zur Klassenbeschrei-bung
DR	Datenbankreferenz	String	Datenbankreferenz
	Medline no.**	Integer	Eintragsnummer von „Literature“

Tabelle 24. Relation für die Klassenbeschreibung

Wenn das regulatorische RNA-Motiv Teil einer modular aus mehreren Haarnadel-strukturen oder Sequenzmotiven aufgebauten Gesamtstruktur ist, kann diese Tatsache über eine interne Verknüpfung in der Relation mit einem anderen Eintrag in PORD durch eine Referenz deutlich gemacht werden (Datenbankreferenz). Die Qualifier Felder enthalten Feldwerte nach einer festgelegten Nomenklatur, die die Struktur und Funktion der Postreg-Sequenz charakterisieren sollen (siehe folgende Tabelle). Da-durch lassen sich die Signalstrukturen später nach funktionellen und strukturellen Merkmalen ordnen.

Funktionalität	*Nomenklatur*	*Bedeutung*
Struktur	SS	Einzelstrang
	DS-Trans	Doppelstrang *in trans*
	DS-Cis	Doppelstrang *in cis*
	SSM	Einzelstrangmodul
	DSM	Doppelsstrangmodul

Tabelle 25. Strukturbezeichner und ihre Bedeutung

Funktionalität	*Nomenklatur*	*Bedeutung*
Funktion	Translation:Inhibition	Inhibierung der Translation
	Translation:Enhancing	Verstärkung der Translation
	Stability:Increase	Zunahme der RNA-Stabilität
	Stability:Decrease	Abnahme der Stabilität
	PolyAProcessing:Enhancing	Unterstützung der Poly-A Prozessierung
	PolyAProcessing:Inhibition	Inhibierung der Poly-A Prozessierung
	Transport:Nucleus	Transport der mRNA aus dem Nukleus
	Transport:Cytoplasma	Transport der mRNA zu Zellkompartimenten im Cytoplasma
(nur für RNA-Genome)	Transcription:Activation	Aktivierung der Transkription einschließlich Prolongierung
	Transcription:Inhibition	Unterbindung der Transkription

Tabelle 26. Bezeichnungen in der Nomenklatur für Funktion und Struktur

3.2.5. Das PORD Datenbankformat

Für die interne Darstellung und Handhabung der Daten wird ein definiertes Format genutzt, das seine Formateigenschaften zum Teil aus vorhandenen molekularbiologischen Datenbanken wie EMBL Data-Library und TRANSFAC ableitet und auf die vorhandenen Daten angepasst worden ist. Das allgemeine Formatmerkmal richtet sich in der Codierungsform nach dem bei molekularbiologischen Datenbanken weithin üblichen "Flat File" Format. So sind die globalen Merkmale der Eigenschaften wie die

Abstände zwischen Feldbezeichnung und Feldinhalte der EMBL- Formatbeschreibung entnommen. Dieses Format dient als „working draft“ und ist im Gegensatz zur EMBL Datenbank nicht in dieser Form, sondern in der oben bereits beschriebenen relationalen Form in der Datenbank abgespeichert. Dieses Format wird dynamisch erstellt, wenn Daten aus der Datenbank PORD dargestellt oder ausgegeben und extern genutzt werden sollen.

3.2.5.1. Formatdefinition

Die Felddefinitionen in der Liste der Formatbeschreibung werden durch einen Zweibuchstabencode wieder gegeben, wie er bereits oben in der Tabellenbeschreibung vorgestellt wurde. Im Folgenden wird das Format anhand eines Beispiels aus der PORD Sequenztabelle erläutert.

Beispiel: Eintrag aus PORD Sequenzrelation

```
AC    R00001
XX
ID    FER$IRE_01
XX
DT    13.12.98 (created); wsa
XX
DE    Ferritin light chain; FLH; G00001
SY    FTL
XX
RG    3'UTR
XX
SQ    CTGTCTCTTGCTTCAACAGTGTTTGGACGGAACAGA
XX
EL    Iron responsive Element; IRE; C0001
XX
SF    1
XX
ST    39
XX
OS    Homo sapiens; human
OC    Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia;
OC    Eutheria; Primates;Catarrhini; Hominidae; Homo
XX
MM    Rnase H footprinting
XX
DR    Embl: M97164; HSFERRL
XX
FE    -4,1
XX
```

DR	Embl: M97164; HSFERRL
XX	
FD	S1 H1 s2 H2 s3 H2 H1 s1
FD	
FD	H1 4:4 0 CTGT:ACAG
FD	H2 3:3 0 UCU:GGA
FD	H3 5:5 0 UUCAA:UUGGA
FD	s1 1:1 C
FD	s2 3.3 UGC
FD	s3 6:6 CAGTGT
FD	s4 1:1 C

Tabelle 27. "Flat file" Format von PORD

Der Feldbezeichner „AC“ ist der Schlüssel des Sequenzeintrags, bestehend aus dem Buchstaben „R“ und fünf Nummern, „DT“ enhält das Datum der Eingabe und der Aktualisierung des Eintrags sowie das Kürzel des Autors. Die Felder „DE“ und „SY“ enthalten die Bezeichnung des Gens und vorkommender Synonyme. Es folgen die Felder „RG“ für „Region“, das die Bezeichnung für die Sequenzregion enthält, in der sich die Postreg-Sequenz befindet, „SQ“ für die Sequenzabfolge, „EL“ für den Namen des PORD-Motivs, „SF“ und „ST“ für die erste und letzte Base der Sequenzabfolge stromaufwärts von der Transkriptionsstartseite, gezählt exklusive der ersten Base der Startseite und inklusive der ersten und letzten Base der Postreg-Sequenz, und schließlich „MM“ für die Kurzbeschreibung der Analysemethode. Das Feld „FE“ enthält die freie Energie der RNA-Struktur, sofern vorhanden, die in dem Feld „FD“ als Sekundärstrukturbeschreibung im DESCRIPTOR Format für die Nutzung durch das Programm RNAMOT bei einem Signalstrukturscreening vorliegt. Die ausgewiesenen HTML Hyperlinks (unterstrichen) verweisen auf die übergeordneten Einträge in der Gentabelle und der Klassentabelle. Der Eintrag wird mit Feldern aus übergeordneten Tabellen vervollständigt. So stammt der Eintrag von den Feldern „DE, SY, OS, OC“ aus der Gentabelle und von „EL“ aus der Klassentabelle.

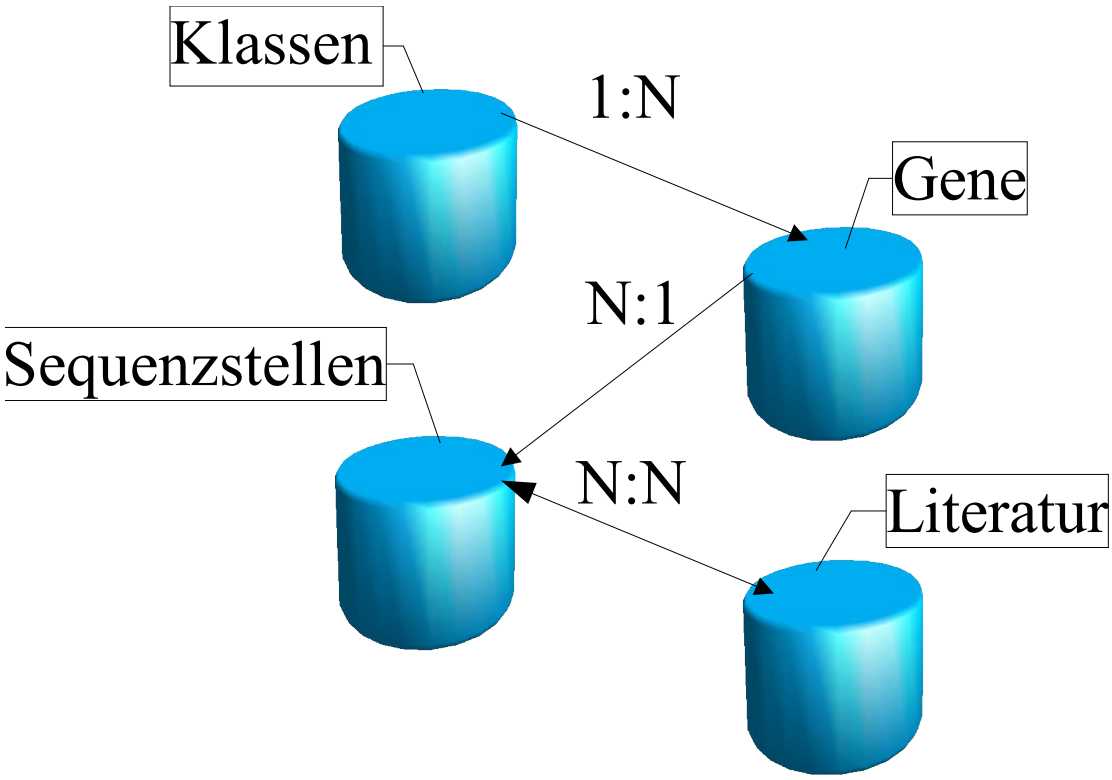


Abbildung 22. Übersicht über das relationale Konzept von PORD

3.3. Softwareaufbau von Postregfinder

Mit „Postregfinder“ wird die Arbeitsumgebung der Datenbank bezeichnet. Mit dieser Bezeichnung soll die Funktion der Software benannt werden, Informationen über Sequenz- und Sekundärstruktureigenschaften von posttranskriptional wirksamer, regulatorischer RNA bzw. Signalstrukturen bereit zu stellen, die es erlauben, weitere Analysen von nicht charakterisierten Sequenzen aus molekularbiologischen Datenbanken durch zu führen oder Besonderheiten von Signalstrukturen spezifisch für bestimmte Organismen aufzufinden. Die Funktionen, die die Arbeitsumgebung zur Verfügung stellt, lassen sich folgenden Kategorien zuordnen: Datenbankmanagement der Datenbank PORD, Signalstrukturrecherche, Datenanzeige und Auswahl der Daten nach Taxonomie, Gennamen und Genregion sowie Funktion der Postreg-Sequenz.

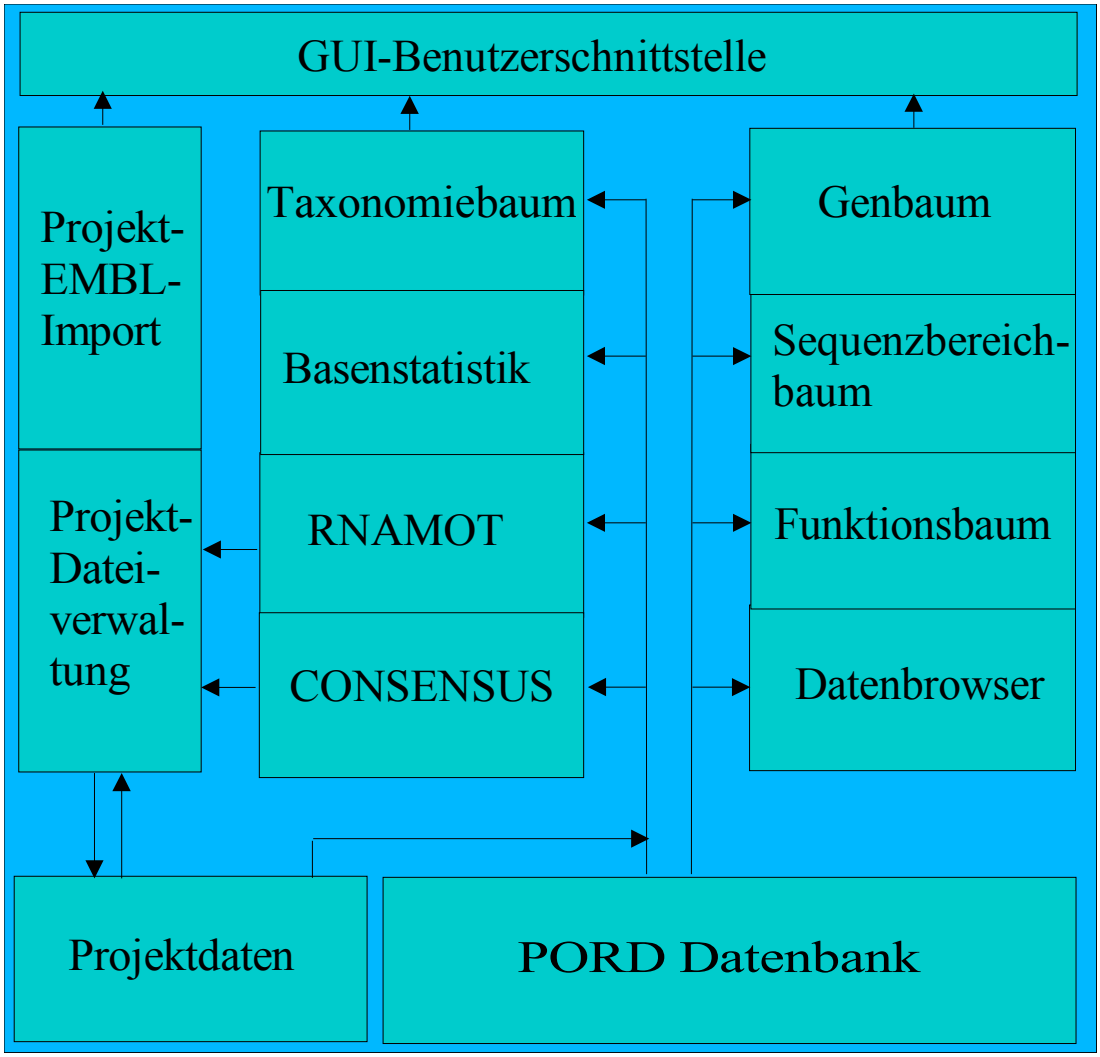


Abbildung 23. Schematischer Softwareaufbau von Postregfinder

3.3.1. Verwaltung genomischer Sequenzdaten

Die Benutzeroberfläche ist als „Single Document Interface“ (SDI) angelegt. Das Hauptfenster von Postregfinder unterteilt sich in vier Fensterbereiche. Im kleineren oberen Kopfbereich des Hauptfensters sind globale Einstellungen über den aktuellen Status der Software und Anzeigemodi zugänglich. Die Datenanzeige lässt sich hier zwischen Projektmodus und PORD-Modus umschalten, und die Datenauswahl kann im gefilterten und ungefilterten Modus dargestellt werden. Der prozentuale Anteil der ausgewählten Datensätze vom aktuellen Gesamtdatensatz wird in einem Informationsfenster angezeigt. Außerdem sind zwei Fensterbereiche vorhanden, die ausgewählte Datensätze als Liste von Subsets, sortiert nach den Quellen PORD Datenbank oder Projekt, vorhalten. Die weiteren drei größeren Fensterbereiche sind

funktionell gegliedert nach Projektverwaltung, Navigationsfenster für das Navigieren in Klassenbäumen und Browserfenster für die Datenanzeige.

3.3.1.1. Sequenzdatenhaltung in Projekten

Das Projektfenster enthält eine Liste von angelegten Projekten. Jedes Projekt beruht auf einer Datendatei, in der ein oder mehrere EMBL-Sequenzeinträge enthalten sind, die nach dem Vorkommen von neuen Instanzen der PORD-Klassen durchsucht werden sollen. Zu jedem Projekt wird ein Unterverzeichnis angelegt, in dem die Dateien mit den EMBL-Sequenzdaten als auch die Suchergebnisse abgelegt werden. Über den Inhalt jedes geöffneten Projekts informiert ein kleines Informationsfenster am unteren Rand der Projektliste. Als zusätzliche Option kann auch aus den zu analysierenden EMBL-Sequenzen des Projektes ein taxonomischer Klassenbaum dynamisch nach den in den EMBL Einträgen vorliegenden Daten erstellt werden, sobald ein Projekt geöffnet oder neu erstellt wird.

3.3.1.2. Repräsentation der Sequenzdaten in Klassenbäumen

Um eine Datenauswahl nach biologischen Kriterien zu ermöglichen, gibt es einen Baumgenerator im mittleren Fensterteil. Die darin abrufbaren Klassenbäume dienen dazu, anhand der Einträge in den Verzweigungen definierte Untermengen der aktuellen Datenmenge nach Funktion, biologischer Herkunft und Struktureigenschaft zu bilden. Die Quelle für den Aufbau der Klassenhierarchie in der Baumstruktur stammt aus der festgelegten Nomenklatur der Gen- und Funktionsklassen, wie sie als Annotation zu den Consensuseinträgen im Importprozess eingegeben wird und somit in der Klassenbeschreibung vorliegt. Bei den herkunftsspezifischen, taxonomisch geordneten Klassen wird der Baum aus den Werten der Felder „Organismclass - OC“ und „Organismsource - OS“ der EMBL Datei eines Projekts oder aus den Annotionen in der Gentabelle von PORD generiert. Es kann daher ein Funktions- oder Genklassenbaum auch nur für PORD Daten erzeugt werden, während die Einträge für die Klassen des Taxonomiebaums aus den Angaben in der PORD Datenbank oder aus dem Projekt erstellt werden können. In einem Kontextmenü sind die Klassenbäume der Rubrik Gen, Funktion und Herkunft separat abrufbar. Der Inhalt der Bäume besteht jeweils aus folgenden Ebenen:

- ✔Eucaria / Bacteria / Archea
- ✔Klassen gemäß der Liste im Feld OC (Taxonomie)
- ✔Organismus gemäß des Feldinhalts von OS
- ✔Gewebetyp
- ✔Zelltyp
- ✔Gen
- ✔Region (5'/3' UTR, Intron, Exon)

Die Klassenbäume werden durch eine SQL-Abfragemodul jeweils dynamisch neu und auf dem aktuellen Stand der Daten erzeugt. Die Einträge werden hierarchisch aufgeschlüsselt. Durch Anklicken oder doppeltes Anklicken eines Eintrags kann entweder die nächste Ebene des Baumes eingeblendet werden oder der selektierte Eintrag als Filterkriterium zur Erzeugung eine Untermenge der aktuellen Datenquelle verwendet werden. Die Datensätze dieser Untermenge können dann als Subset gespeichert werden.

3.3.1.3. PORD und Projekt Datenbrowser

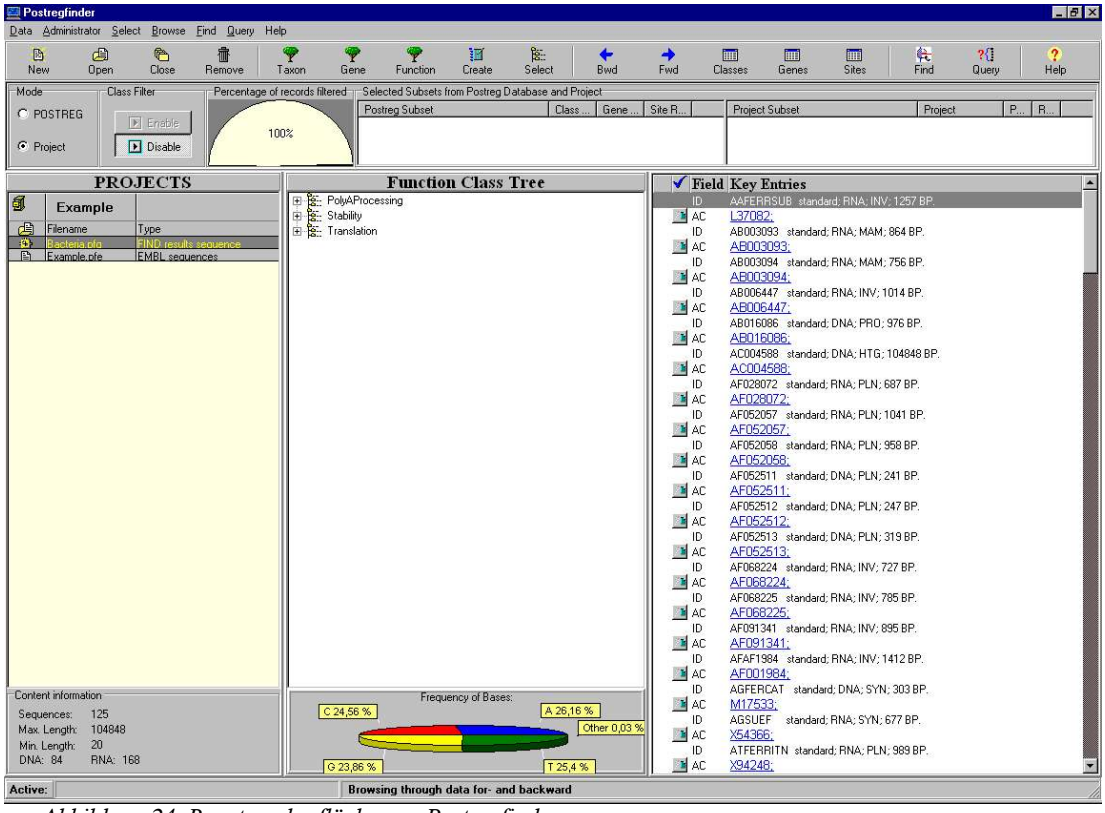


Abbildung 24. Benutzeroberfläche von Postregfinder

Die Daten in der Signalstrukturdatenbank PORD sollen als Informationsgrundlage dienen, um in noch nicht charakterisierten Sequenzdaten eines Projekts homologe Sequenzen der Signalstrukturen zu finden. Für diese Recherche sind alle Datensätze verwendbar, oder es müssen selektiv mehrere Datensätze ausgewählt werden, mit denen man in den Projektsequenzen nach deren Vorkommen sucht. Die Anzeige der aktuell ausgewählten Datensätze erfolgt über eine als HTML Browser konzipierte Datenanzeige im rechten Fensterteil, bei der die voreingestellte, ausgewählte Klasse berücksichtigt wird. Ausgewählte Daten können Projektdaten, PORD Daten, Untermengen beider Datenquellen oder Suchergebnisse sein. Die Anzeige ist in Ebenen

hierarchisch gegliedert. Die Wurzelkategorie der obersten Ebene ist ein HTML-Link auf ein Projekt oder auf die drei PORD-Tabellen, je nach dem gerade aktiven Modus. Die Funktion des Anzeigemoduls beinhaltet vor allem auch, die in den Tabellenindizes vorliegenden Verknüpfungen als „Flat File“ zu realisieren (siehe Beispiel in 3.2.2.1.). Die nächste Ebene zeigt einen Index der Dateneinträge an, der aus den Schlüsselwerten „ID - Identifier“ und „AC - Accession number“ aufgebaut ist. Die dritte Ebene zeigt dann die gesamten Einträge an. Die Sequenzdaten eines Projekts müssen dabei erst in eine Indexliste umgesetzt werden, die es erlaubt, den kompletten Dateneintrag über HTML Verknüpfungen anzuwählen. Auch die Ergebnisse einer Suche liegen, nach Signalstruktur Consensus sortiert, als Liste mit HTML-Links vor.

3.3.2. PORD Datenannotation - und Updateverwaltung

Für den Import und die Bearbeitung von Daten in der RNA-Signalstrukturdatenbank PORD gibt es ein separates Programm mit grafischer Arbeitsoberfläche. Dieses Programm besteht aus zwei Modulen in denen die Import- und Filterfunktionen für die Formate der Fremddaten und die Aktualisierung bereits importierter Daten in einem eigenen Menü steuerbar sind.

3.3.2.1. Annotationsmodul

Die grafische Benutzerschnittstelle dieses separaten Moduls besteht aus einem Editorfenster auf der linken Seite, in dem die Quelldaten geladen werden und einem Karteikartensystem auf der rechten Seite, auf dem sich editierbare Felder befinden. Vier Datenquellen werden als Datengrundlage für die PORD-Datenbank genutzt:

- ✔Die Sequenzeinträge aus der EMBL Data Library
- ✔Die Literatureinträge aus MEDLINE
- ✔Import von Sekundärstrukturdaten, die RNASTRUCTURE generiert
- ✔Manuelle Eingabe von Ergänzungen (z.B. Consensusangaben und Annotationen)

Der Annotationsvorgang verläuft in drei Schritten. Die zu importierenden Quelldaten werden aus einer Datei in den Editor geladen. Jeder Eintrag aus der Quelldatei wird einzeln eingelesen und in editierbare Datenfelder übertragen und mit weiteren Annotationen ergänzt. Anschließend werden die Annotationen aus den Editierfeldern in die Datenbank übertragen. Bei jedem Schritt sorgen Filteroperationen dafür, dass einerseits nur der benötigte Teil der vorhandenen Annotationen aus den Quelldaten übernommen wird, andererseits nur Einträge in die PORD Tabellen übernommen werden, die wirklich vollständig sind. Das Funktionieren der Filteroperationen beim Einlesen der Daten aus der Quelldatei beruht darauf, dass sie im korrekten Format vorliegen. Die MEDLINE Datensätze müssen beispielsweise in dem MEDLINE-spezifischen Datenformat vorliegen (siehe Datenformate), in dem die Feldinhalte, durch eine Feldcodierung getrennt, enthalten sind. Ein weiterer Filter ist für die Datensätze aus EMBL vorhanden aus denen die Klassifizierung des Organismus der Genname und die Sequenz herausgefiltert werden. Ergänzt werden die gefilterten Daten um Einträge

aus Auswahllisten über Herkunfts- und Funktionsangaben der Signalstrukturen, die der bereits beschriebenen Nomenklatur entsprechen. Die in den Feldern enthaltenen Inhalte werden durch den Import separat in die Tabellen für die Beschreibung der PORD-Klasse des Gens und der Motivsequenz abgelegt. Mit den Motivsequenzen werden auch die Sekundärstrukturinformationen gespeichert. Deshalb wird die optimale Faltungsvariante der Sequenz in das Sekundärstrukturfeld übernommen und dabei das „CONNECT“ Format von RNASTRUCTURE in das DESCRIPTOR Format automatisch umformatiert und in die Datenbank zu den entsprechenden Motivsequenzen abgelegt.

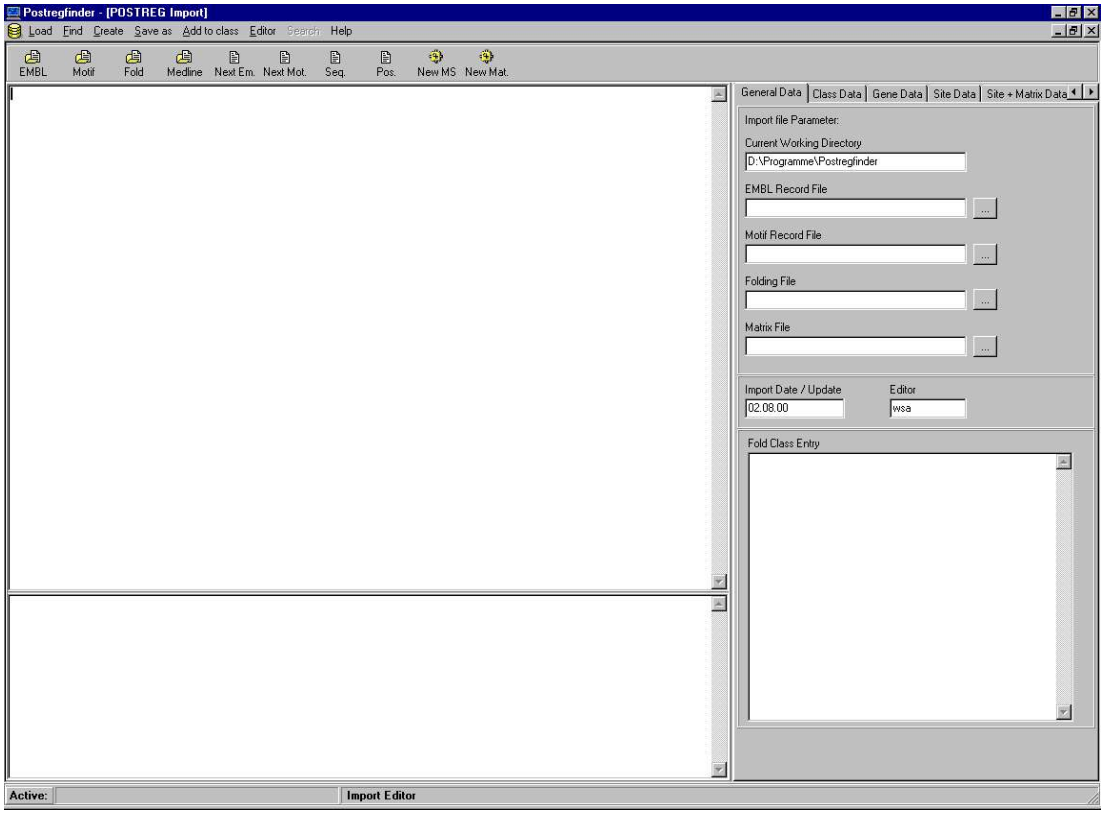


Abbildung 25. Benutzeroberfläche des POSTREG Annotationsmoduls. Im oberen linken Editor können EMBL Sequenzen und im untern MEDLINE Einträge zu dem Annotationsprozess geladen werden.

3.3.2.2. Menügesteuerte Annotationsfunktionen

Die Menüfunktion für das Öffnen der formatierten Quelldaten ist „Load“, in dem sich Untermenüs für den Zugriff auf MEDLINE- und EMBL-Daten sowie zum Öffnen von RNA-Faltungen als „*.ct“ Datei als auch von Motivsequenzen im IG-Format befinden. In dem zweiten Menüpunkt unter „Data“ befinden sich die Funktionen für das schrittweise Importieren von Einträgen für Quelldateien mit mehreren Einträgen, die Berechnung der Sequenzposition einer Motivsequenz in einem EMBL-Eintrag und die

Berechnung der Position von der Transkriptionsstartstelle aus. In weiteren Untermenüpunkten des Menüs „Create“ sind Funktionen für die Erstellung von Matrices und Sekundärstrukturbeschreibungen abrufbar. „New matseq“ erstellt eine Matrix Quelldatei für die Berechnung einer Matrix von Motivsequenzen einer Klasse, „New matrix“ führt die Berechnung der Matrix durch und „New description“ zeigt eine Gesamtansicht aller vorhandenen Sekundärstrukturbeschreibungen, um daraus eine Consensusbeschreibung zu entwickeln. Die Funktion „Skip matrix“ sorgt bei der automatischen Durchführung einer neuen Matrixberechnung bei jedem Import einer einzelnen Motivsequenz dafür, dass die Matrixberechnung übersprungen wird, wenn z.B. die Einzelsequenz die erste Sequenz einer Klasse ist. Der Menüpunkt „Save as“ enthält die Optionen für die Speicherung der gefilterten Einträge in die Tabellen. Im Menü „Add to Class“ kann man dem gerade aktuellen Eintrag in der Klassentabelle eine neue oder aktualisierte Consensusbeschreibung der Sekundärstruktur oder der Sequenz hinzufügen. Unter dem Menü „Editor“ kann man jeweils zwischen dem Update- und Importmodul wechseln.

3.3.2.3. Updatemodul

Das Updatemodul ermöglicht es, gespeicherte ganze Datensätze zu bearbeiten oder Werte einzelner Felder zu aktualisieren. Dafür ist ein Tabelleneditor vorhanden, der die Tabelleninhalte als Übersicht darstellt. Die Werte jeder Tabelle werden auf einer eigenen Karteikarte angezeigt.

3.3.2.4. Menügesteuerte Updatefunktionen

Man kann in dem Editor gezielt zu bestimmten Einträgen der Tabellen springen oder nach bestimmten Werten in den Tabellen suchen. Diese Suchfunktion ist unter dem Menüpunkt „Search“ abrufbar.

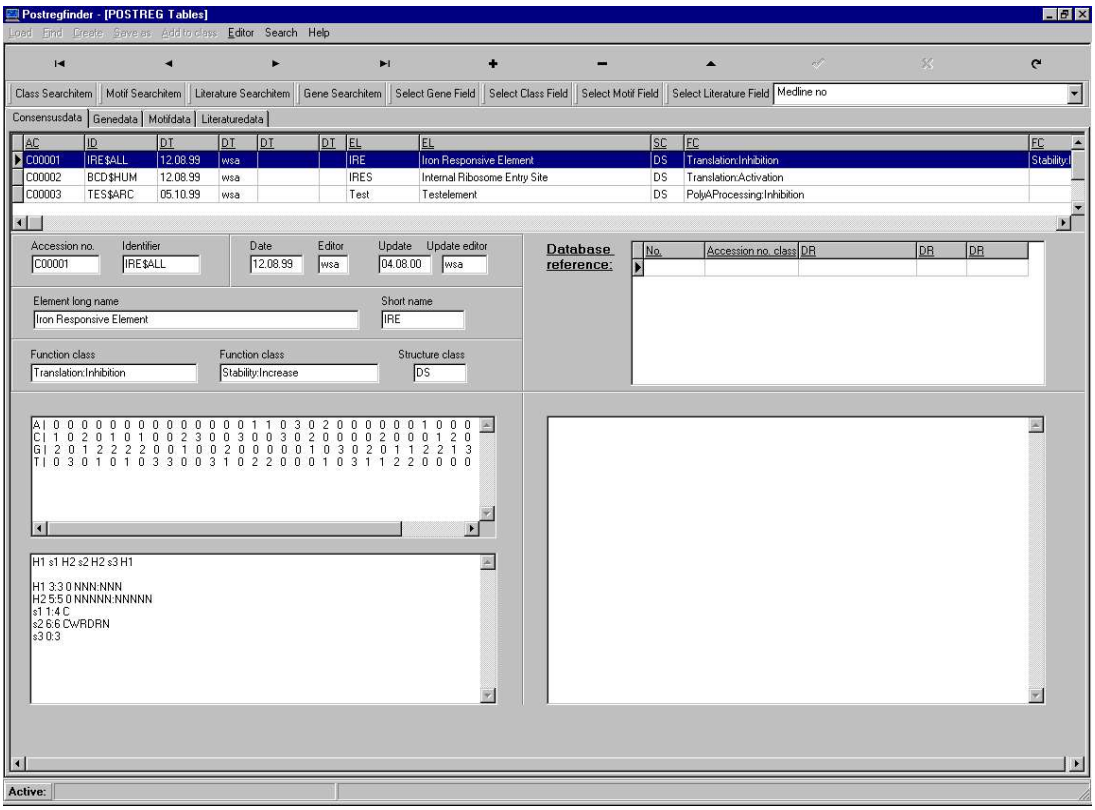


Abbildung 26. Updateschnittstelle von Postregfinder

3.3.3. Funktionen zur POSTREG Analyse genomischer Sequenzdaten

Die oben genannten Funktionen sind vom Hauptmenü im Hauptfenster der Benutzerschnittstelle steuerbar. Die Menüsteuerung ist objektorientiert, das bedeutet, dass jeweils immer nur die Menüpunkte und Funktionen sichtbar und abrufbar sind, die mit einem aktiven Fenster in Zusammenhang stehen. Das Hauptmenü enthält folgende Menüpunkte: Projektverwaltung im Menüpunkt „Project“, Auswahl von einzelnen Datensätzen im Menü „Datasets“, Auswahl einzelner Klassenbäume, die die Klassen von Signalstrukturen repräsentieren, aufgeschlüsselt nach Gennamen und –region, Funktion, und Taxonomie unter dem Menüpunkt „Classes“, Datenanzeige- und auswahl sowie Browserfunktionen unter dem Menüpunkt „Browse“ sowie die Rechercheoptionen und das Startkommando einer Recherche in EMBL-Dateien nach homologen Postreg-Sequenzen im Menü „Find“. Außerdem gibt es den Menüpunkt „Query“, der es erlaubt, über ein separates Fenster komplexere Anfragen an die PORD Datenbank zu erstellen, in denen Suchwörter mit Hilfe Booleanscher Operatoren verknüpft werden können. In der vorliegenden Version ist außerdem der Zugang zu den Annotations- und Aktualisierungsmodulen der PORD Datenbank über den Menüpunkt „Administrator“ erreichbar. Dieser Zugang ist durch Passwörter geschützt. Der

Menüpunkt „Hilfe“ ist zwar als Schnittstelle für eine Online-Hilfe vorhanden, eine Online-Hilfe ist in der aktuellen Version allerdings noch nicht verfügbar.

3.3.3.1. Verwaltung von zu analysierenden Sequenzen in Projekten

Unter dem Menüpunkt zur Projekthandhabung von Dateien im EMBL-Format „Project“ sind untergeordnete Menüoptionen verfügbar, mit denen sich Projekte und Datensätze verwalten lassen. Ein Projekt basiert auf einer Datei, in der ein oder mehrere Sequenzeinträge der EMBL Data Library enthalten sind, sowie Dateien mit Ergebnissen aus Suchoperationen nach homologen Signalstrukturen aus der PORD Datenbank, wie sie mit Postregfinder durchgeführt werden können. Durch den Untermenüpunkt „New“ kann ein neues Projekt benannt und die neue EMBL Datei in die Projektverwaltung integriert werden. Es wird dadurch ein neues Unterverzeichnis für die Dateien des neuen Projekts eingerichtet, welches dann im Listenfenster "Projects" sichtbar wird. Zu den weiteren Operationen, die hier zur Verfügung stehen, gehören das Öffnen eines Projekts für die Datenanzeige als auch das Löschen ganzer Projekte. Angezeigte EMBL-Sequenzdaten im Datenbrowser können außerdem hier unter dem Untermenüpunkt „Save as...“ oder „Save“ wieder in separaten Dateien gespeichert werden. Da die Daten in dem Anzeigefenster im "Read-Only" Modus angezeigt werden, also nicht verändert werden können, sind die Menüpunkte "Save" und "Save as" vor allem für die Ergebnisse von Recherchen gedacht. Solche Ergebnisdateien werden in dem bereits vorhandenen Projektverzeichnis unter einem neuen Namen als separate Datei abgespeichert. Beim Öffnen eines Projektes werden die Daten in der EMBL Datei in im Datenbrowser im Flat-File Format angezeigt und ein taxonomischer Baum generiert, mit dem man sich organismen- oder zellspezifisch Datensätze aus dem gesamten Projektdatensatz auswählen kann. Auf einzelne Datensätze des Projekts kann im Datenbrowser über HTML-Verknüpfungen zugegriffen werden.

3.3.3.2. Selektion genomischer Sequenzdatensätze

Unter dem Menüpunkt „Datasets“ sind Kommandos vorhanden, mit denen man auf der aktuellen Datengrundlage und gewählten Filterkriterien eine neue Subauswahl von Datensätzen erzeugen und wieder löschen kann. Die Datengrundlage ist entweder die PORD Datenbank oder das aktuell geöffnete Projekt. Die Filterkriterien können entweder aus einem ausgewählten Klasseneintrag eines gerade aktiven Klassenbaumes stammen oder aus einer Datenbankabfrage der PORD Datenbank unter dem Menüpunkt „Query“. Jeder Satz von gefilterten Daten kann zu sogenannten Subsets zusammengefasst und separat temporär gespeichert werden. Bei der temporären Speicherung werden die Subsets ihrer Quelle nach sortiert, das heißt, je nach dem ob der Datensatz aus einem Projekt oder aus der PORD Datenbank stammt, werden sie in getrennten Fenstern im Kopfbereich angezeigt und können damit separat für Recherchen genutzt werden. Beispielsweise kann ein Subset aus der PORD Datenbank mit Motiven für die Poly-A Prozessierung gegen ein Subset von EMBL-Einträgen nur über Proteobacteria gescreent werden.

3.3.3.3. Auwahl der Datensätze über Klassenbaumrepräsentationen

Im Hauptmenüpunkt „Classes“ sind die drei verschiedenen Klassenbäume in den jeweiligen Untermenüs „Taxonomy“, „Genes“ und „Function“ abrufbar, die in dem mittleren Fensterteil als Klassenbäume angezeigt werden. Die Bäume werden über SQL Abfragen generiert, die entweder auf die PORD-Daten zugreifen oder auf die Datendatei eines geöffneten Projekts.

3.3.3.4. PORD-Einträge für die Analyse auswählen

Das Menü „Browse“ enthält die einzelnen Kommandos, um in den angezeigten Daten zu navigieren oder die Inhalte der einzelnen PORD Tabellen „Class“, „Gene“ oder „Site“ im rechten Browserfenster für die Datenansicht anzuzeigen bzw. zwischen den Ansichten zu wechseln. Außerdem kann unter dem Menüpunkt „Go To“ über ein neu geöffnetes Fenster nach Stichworten in den Einträgen gesucht werden. Man kann dabei jeweils in den nächsten angezeigten Eintrag mit dem gesuchten Stichwort springen. Dies ist vor allem deswegen nützlich, weil die ausgewählte Datenmenge den physikalisch sichtbaren Anzeigebereich häufig übersteigt.

3.3.3.5. Aufbau der Suchmaschine für RNA-Klassenbeschreibungen

Die Menüpunkte im Menü "Screening" dienen dazu, die Optionen für Suchprozesse einzugeben und sie zu starten. Das Verfahren für das Auffinden homologer Sekundärstrukturen und Sequenzen beruht, wie bereits erläutert, auf dem Sekundärstrukturalignment von RNAMOT und dem Algorithmus Patser. Die Suche wird als separater Thread gestartet und für jeden ausgewählten Datensatz durchgeführt.

RNA motif database search using RNAMOT, February 3, 1998 21:11

Database file: @brain2_52.txt

Descriptor file: ire.des

Complementary strand searched: yes

Filter out overlapping hits: yes

Results:

seq-f seq-t name description

246 223 EMEST10:HSC0FH071 Z42333 H. Sapiens partial cDNA sequence; clone c-0fh07. 9/95

218 242 EMEST10:HSC0HA052 F01887 H. Sapiens partial cDNA sequence; clone c-0ha05. 9/95

203 179 EMEST10:HSC0TF072 F02249 H. Sapiens partial cDNA sequence; clone c-0tf07. 9/95

319 296 EMEST10:HSC0XB071 Z42895 H. Sapiens partial cDNA sequence; clone c-0xb07. 9/95

125 102 EMEST10:HSC0ZF081 Z42974 H. Sapiens partial cDNA sequence; clone c-0zf08. 9/95

20 45 EMEST10:HSC12C071 F06255 H. Sapiens partial cDNA sequence; clone c-12c07. 9/95

Total number of bases scanned: 17160750

Tabelle 28 Beispiel einer RNAMOT Ausgabe (gekürzte Länge)

Unter dem Menüpunkt „Options“ wird ein Fenster geöffnet, in dem sich die Optionen eines „Screenings“ einstellen lassen und der Modus ausgewählt wird, ob eine Sekundärstruktursuche, eine Sequenzsuche oder beides parallel durchgeführt werden soll.

3.4. RNA-Strukturhomologiesuche mit Postregfinder

Die erste Aufgabenstellung bei der Durchführung einer Suche nach homologen RNA-Strukturmotiven ist die Aquirierung von genomischen Sequendaten, in denen die Suche durchgeführt werden soll, und die Erstellung eines neuen Projekts unter POSTREGFINDER. Wie in der Einleitung bereits dargestellt, sollen im Kontext dieser Arbeit Sequenzdaten von in Hirnzellen expremierten Genen untersucht werden.

Oft lassen sich die in der Evolution in verschiedenen Organismen konservierte, homologe Signalstrukturen durch eine von Dandekar *(Dandekar et al., 1995)* vorgestellte Filtertechnik so filtern, dass man die Homologie mit hoher Wahrscheinlichkeit feststellen kann. Um beispielsweise falsche positive Treffer aus einer Trefferliste nach einem Suchlauf zu entfernen, muss ein Filterverfahren angewendet werden, wie es bei der Suche nach homologen Strukturen schon publiziert worden ist (*Lescure et al., 1999)*. Das Verfahren beinhaltet die Entscheidung ob nach RNA oder DNA Strukturen gesucht wird, um die Sequenzen der Trefferliste entsprechend zu filtern. In dieser Filteroperation ist auch das Ausklammern von nichtcodierenden Sequenzen beinhaltet, wie etwa Pseudogenen oder genomische Spacersequenzen. In einem zweiten Schritt muss positiv geprüft werden, ob die Sequenzen überhaupt die Genregion beinhaltet, in der speziell die homologe Struktureigenschaft gesucht wird, wie etwa UTR-Region oder spezielle Intronbereiche.

Es werden im Folgenden nur Sequenzen berücksichtigt, die:

- mRNA Sequenz sind.
- ein entsprechende Genregion (3'/5'UTR,CDS etc.) besitzen.
- kein Pseudogen sind.

Dieses Filterverfahren lässt sich mittlerweile sehr gut vor einem Suchlauf durch entsprechende Voreinstellungen unter der Suchmaske von SRS durchführen.

Bei der Suche nach Sequenzeinträgen in der EMBL Data Library gibt es zwei Wege, über SRS nach Sequenzeinträgen zu suchen, die in Hirnzellen exprimierte Gene dokumentieren. Die eine Möglichkeit besteht in der Abfrage der einzelnen Hirnzelltypen, die andere in der Abfrage des Gewebetyps. Beide Abfragen nutzen Einträge im sogenannten „Feature Table FT“ der Sequenzeinträge. Die Trefferquote für hirnspezifische Sequenzen hängt dabei von dem Umfang der Annotationen ab. Sequenzen, die nicht entsprechend im „Feature Tabel“ einen Eintrag unter „/tissue type“ besitzen, können nicht als hirnspezifischer Eintrag identifiziert werden, obwohl sie möglicherweise eine hirnzellspezifische Sequenz dokumentieren. Sie können damit nicht zugeordnet werden. Dies ist ein generelles Problem bei der Nutzung der internationalen Datenbanken, bei denen im Wesentlichen die Autoren für die Vollständigkeit der Sequenzeinträge verantwortlich sind. Da bei beiden Abfragemöglichkeiten dieses Problem besteht, wurden die Sequenzeinträge wegen des einfacheren Vorgehens über den Gewebetyp selektiert. Die Abfrage unter dem Feature Qualifier „/tissue-type = brain“ ergab 200137 EMBL – Einträge. Die EMBL Datenbank lag in der Version 65 vor (siehe Einleitung). Mit dieser Datei als Ausgangsdaten wurde ein neues Projekt in POSTREGFINDER angelegt.

3.4.1. Einstellung der Suchparameter

Die Datei im neu erstellten Projekt kann mit den Strukturangaben der Klassenbeschreibungen aus PORD durchsucht werden. Dies wird hier am Beispiel der BICOID-Einträge von PORD gezeigt. Entsprechend der Klassendefinition von PORD besteht das Bicoid Localization Element (BLE) aus drei Stemloops und hat daher drei Klasseneinträge in PORD. Die Suche nach homologen Strukturen in der Projektdatei wird beispielhaft anhand der DESCRIPTOR-Strukturbeschreibung der Sekundärstruktur des ersten Stemloops durchgeführt. Sie ist der am besten beschriebene Hairpin des Strukturmoduls *(Ferrandon et al., 1997)*.

Descriptor	***Beschreibung***
H1 s1 H2 s2 H2 s3 H1 H1 22:30 6 H2 29:34 6 s1 0:5 s2 7:11 AAAGCCC s3 6:14 GGGCUU W 12 M 16	Der Stemloop besteht aus zwei helicalen Bereichen und drei einzelsträngigen Bereichen, gekennzeichnet mit H und s. Die konservierten Kernbereiche der Sequenz befinden sich in dem endständigen Loop s2 und dem Bulge s3. Die Stemregionen können sehr stark variieren. Deshalb wurde eine Mismatchzahl (M) von 16 und 12 Wobblepositionen (W) als Parameter eingegeben

Tabelle 29. Descriptorbeschreibung von Stem 1 (auch als Stem IV bezeichnet) des BICOID LOCALIZATION ELEMENT BLE

3.4. RNA-Strukturhomologiesuche mit Postregfinder

Zunächst wurden die Sequenzen der Einzelbeispiele (Tabelle 28) der Signalstruktur exportiert, um mit der Suche der Descriptorbeschreibung festzustellen, ob die Consensusstruktur ihr eigenes Signalstrukturkomplement wiederfindet (Tabelle 29).

Drosophila Art	*EMBL-Schlüssel (AC/ID)*	*Gen-Eintrag in PORD (AC/ID)*
D. heteroneura	M32125 / DHBCDA	G00058 / DRH$BCD
D. virilis	M32122 / DVBCDA	G00062 / DRV$BCD
D. melanogaster	X14460 / DMBCD16	G00059 / DRM$BCD
D. pseudoobscura	X55735 / DPBCD	G00060 / DRP$BCD
D. sechellia	M32124 / DSBCDA	G00061 / DRS$BCD

Tabelle 30. Einzelbeispiele des Stemloop 1 der Bicoid – Consensusstruktur

Es konnten mittels der Descriptorbeschreibung und RNAMOT alle Einzelbeispiele aus den Sequenzen identifiziert werden. Das Ergebnis dieser Überprüfung zeigt die Übersicht in Tabelle 31 anhand von *Drosophila heteroneura*. Dadurch wurde zunächst nachgewiesen, dass das technische Verfahren in der Lage ist, die jeweils eigenen komplementären Sequenzen zu finden, aus denen die Klassenbeschreibung gebildet wurde. Dies dient dazu, eventuelle Fehler in der Consensusbeschreibung aufzudecken und wird daher für alle Klassenbeschreibungen durchgeführt.

--- DROBCDA D.heteroneura bicoid (bcd) gene, 3' end. --- (939 bases)
|SCO: 1254.14|POS:290-442|MIS:11|WOB: 6|
|CCAUUAUUUGGAAACUUUAUUUGUAAAGCG|UUCUU|UGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAAUAU|GGGCUUUAAUUGAA|
CGCUUUACAGAAGAAGUUUAUAAAUACACA|
|SCO: 1155.14|POS:291-441|MIS:10|WOB: 6|
|CAUUAUUUGGAAACUUUAUUUGUAAAGCGU|UCUUU|GAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAAUA|UGGGCUUUAAUUGA|ACGCUUUACAGAAGAAGUUUAUAAAUACAC|
|SCO: 502.06|POS:292-440|MIS:10|WOB: 6|
|AUUAUUUGGAAACUUUAUUUGUAAAGCGUU|C|UUUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAAU|AUGGGCUUUAAUUG|AACGCUUUACAGAAGAAGUUUAUAAAUACA|
|SCO: 403.06|POS:293-439|MIS: 8|WOB: 6|
|UUAUUUGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUUUAAUU|GAACGCUUUACAGAAGAAGUUUAUAAAUAC|
|SCO: 354.06|POS:294-438|MIS: 7|WOB: 6|
|UAUUUGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUUUAAU|GAACGCUUUACAGAAGAAGUUUAUAAAUA|
|SCO: 355.06|POS:295-437|MIS: 7|WOB: 6|
|AUUUGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUUUAA|GAACGCUUUACAGAAGAAGUUUAUAAAU|
|SCO: 356.06|POS:296-436|MIS: 7|WOB: 6|
|UUUGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUUUA|GAACGCUUUACAGAAGAAGUUUAUAAA|
|SCO: 357.06|POS:297-435|MIS: 7|WOB: 6|
|UUGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUUU|GAACGCUUUACAGAAGAAGUUUAUAA|
|SCO: 358.06|POS:298-434|MIS: 7|WOB: 6|
|UGGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUU|GAACGCUUUACAGAAGAAGUUUAUA|
|SCO: 359.06|POS:299-433|MIS: 7|WOB: 6|
|GGAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUU|GAACGCUUUACAGAAGAAGUUUAU|
|SCO: 360.06|POS:300-432|MIS: 7|WOB: 5|
|GAAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUU|GAACGCUUUACAGAAGAAGUUUA|
|SCO: 311.06|POS:301-431|MIS: 6|WOB: 5|
|AAACUUUAUUUGUAAAGCGUUC|U|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUU|GAACGCUUUACAGAAGAAGUUU|
|SCO: 361.07|POS:302-430|MIS: 7|WOB: 5|
|AACUUUAUUUGUAAAGCGUUCU|-|UUGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGAA|UAUGGGCUU|UGAACGCUUUACAGAAGAAGUU|
|SCO: 412.07|POS:303-429|MIS: 8|WOB: 5|
|ACUUUAUUUGUAAAGCGUUCUU|-|UGAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACGA|AUAUGGGCUU|UUGAACGCUUUACAGAAGAAGU|
|SCO: 413.07|POS:304-428|MIS: 8|WOB: 5|
|CUUUAUUUGUAAAGCGUUCUUU|-|GAUCUCAACGCUGUCUGGCUGGACAUUUG|CCAAAGCCCA|
UGAAUGCCCAACCAGACACUGUUGAGACG|AAUAUGGGCUU|AUUGAACGCUUUACAGAAGAAG|
|SCO: 463.08|POS:305-427|MIS: 9|WOB: 5|
|UUUAUUUGUAAAGCGUUCUUUG|-|AUCUCAACGCUGUCUGGCUGGACAUUUGC|CCAAAGCCCA|
AUGAAUGCCCAACCAGACACUGUUGAGAC|GAAUAUGGGCUU|AAUUGAACGCUUUACAGAAGAA|

Total: 1 sequences, 939 nucleotides scanned, 378 matches, 16 actual

Tabelle 31. Liste der Matches im 3'UTR des Bicoid-Gens von Drosophila heteroneura. Das Consensusmotiv im endständigen Loop ist gelb unterlegt. Der Score in „SCO“ gibt die optimale Strukturvariante an (grün).

3.4.2. Auswertung der Suchergebnisse

RNAMOT verwendet einen Alignment-Algorithmus, der die mit dem niedrigsten Score bewerteten Sequenzen in eine Datei mit der Endung „*.sol“ für „Solution“ schreibt. Je niedriger der Score, desto größer ist die Übereinstimmung mit der Sequenz und Struktur, die als Suchparameter vorgegeben wurden. In den Score gehen die angegebenen Parameter in der folgenden Reihenfolge ein:

- Das Vorhandensein der eingegebenen Sequenzmuster
- Die Übereinstimmung der Längen der Helices (Je länger, desto kleiner der Score)
- Die Anzahl der Mismatches (Je weniger desto kleiner der Score)
- Die Übereinstimmung der freien Energien der Helices (Je mehr desto kleiner der Score)

Bei der Berechnung des Scores werden auch suboptimale Strukturvarianten berücksichtigt. Nur die Strukturvariante einer Postreg-Sequenz wird in die Ergebnismenge aufgenommen, die den niedrigsten Scorewert enthält. Mit einem speziellen Parameter kann man RNAMOT veranlassen, alternative Sekundärstrukturen der Sequenzen aus der Ergebnismenge in eine zweite Ergebnisdatei mit der Endung „*.alt“ abzuspeichern. Sie kann dazu genutzt werden, Sequenzen der Ergebnismenge darauf hin manuell zu überprüfen, ob alternative Sekundärstrukturen als falsche Negative bewertet worden sind. Die Tabelle 32 zeigt einige Beispiele mit sehr niedrigem Scorewert aus der Ergebnismenge von 1997 Einträgen. Es sind darin zwei Beispiele mit sehr niedrigem Scorewert aus dem resultierenden Datensatz angegeben.

Beispiel aus der Ergebnismenge
--- AF001462 Homo sapiens glypican-5 (GPC5) mRNA, complete cds. --- (2391 bases) |SCO: 300.04|POS:2072-2153|MIS: 6|WOB: 1| |AVRUBRAUAUUSUMMAMMAAUUHRARMAUSC|AUARRHNHMND| AHMRRNCBASSUAUUHRSANGAUAKSUUKMM|CAUDNGYHN| |SCO: 251.04|POS:2073-2153|MIS: 5|WOB: 1| |VRUBRAUAUUSUMMAMMAAUUHRARMAUSC|AUARRHNHMND| AHMRRNCBASSUAUUHRSANGAUAKSUUKM|MCAUDNGYHN| |SCO: 252.04|POS:2074-2153|MIS: 5|WOB: 1| |RUBRAUAUUSUMMAMMAAUUHRARMAUSC|AUARRHNHMND| AHMRRNCBASSUAUUHRSANGAUAKSUUK|MMCAUDNGYHN| |SCO: 253.04|POS:2075-2153|MIS: 5|WOB: 1| |UBRAUAUUSUMMAMMAAUUHRARMAUSC|AUARRHNHMND| AHMRRNCBASSUAUUHRSANGAUAKSUU|KMMCAUDNGYHN| |SCO: 204.04|POS:2076-2153|MIS: 4|WOB: 1| |BRAUAUUSUMMAMMAAUUHRARMAUSC|AUARRHNHMND| AHMRRNCBASSUAUUHRSANGAUAKSU|UKMMCAUDNGYHN|

Beispiel aus der Ergebnismenge
--- AF023449 Homo sapiens CHD2-42 Down syndrome cell adhesion molecule (DSCAM) --- (6600 bases) |SCO: 50.03|POS:6330-6415|MIS: 1|WOB: 0| |ANDMAHUBRURCHNNNYNSGANDKRNBRGRU|UDSCAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYNDRM|RGNANDSNVVDN| |SCO: 1.03|POS:6331-6415|MIS: 0|WOB: 0| |NDMAHUBRURCHNNNYNSGANDKRNBRGRU|UDSCAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYNDR|MRGNANDSNVVDN| |SCO: 2.03|POS:6332-6415|MIS: 0|WOB: 0| |DMAHUBRURCHNNNYNSGANDKRNBRGRU|UDSCAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYND|RMRGNANDSNVVDN| |SCO: 3.03|POS:6333-6415|MIS: 0|WOB: 0| |MAHUBRURCHNNNYNSGANDKRNBRGRU|UDSCAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYN|DRMRGNANDSNVVDN| |SCO: 4.03|POS:6334-6415|MIS: 0|WOB: 0| |AHUBRURCHNNNYNSGANDKRNBRGRU|UDSCAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSY|NDRMRGNANDSNVVDN| |SCO: 102.03|POS:6335-6415|MIS: 2|WOB: 1| |HUBRURCHNNNYNSGANDKRNBRGRUUDS|CAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYND|RMRGNANDSNVVDN| |SCO: 103.03|POS:6336-6415|MIS: 2|WOB: 1| |UBRURCHNNNYNSGANDKRNBRGRUUDS|CAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSYN|DRMRGNANDSNVVDN| |SCO: 104.03|POS:6337-6415|MIS: 2|WOB: 1| |BRURCHNNNYNSGANDKRNBRGRUUDS|CAMANVMMB| RUHMMUNGBUNSURAMYMASNADWNSY|NDRMRGNANDSNVVDN| |SCO: 154.03|POS:6338-6415|MIS: 3|WOB: 1| |RURCHNNNYNSGANDKRNBRGRUUDSC|AMANVMMBRUHM| MUNGBUNSURAMYMASNADWNSYNDRM|RGNANDSNVVDN|

Tabelle 32. Ausgewählte Einzelbeispiele aus der Ergebnismenge mit besonders niedrigem Scorewert für BLE Stemloop 1. (Stem IV).

3.4.3. Evaluierung des Ergebnisses

Wie anhand der Tabelle 32 gezeigt, sind teilweise sehr niedrige Scorewerte vohanden aber die Sequenz enthält sehr viele unspezifische, im IUB-IUPAC angegebene Basen. Daher wird die Sekundärstruktur zu der Klassenbeschreibung sehr ähnlich sein, aber, wie die Beispiele zeigen, ist die notwendige Motivsequenz im Loop sehr wahrscheinlich nicht enthalten. Um diese Aussage zu verifizieren, wurde ein Profilalignment angelegt, bei dem die Motivsequenzen des Loops ein Alignmentprofil bilden, gegen

das alle übrigen Sequenzen aligniert werden. Dazu wurden die Sequenzen einfach aus PORD exportiert und mit CLUSTALW zu einem Profil aligniert.

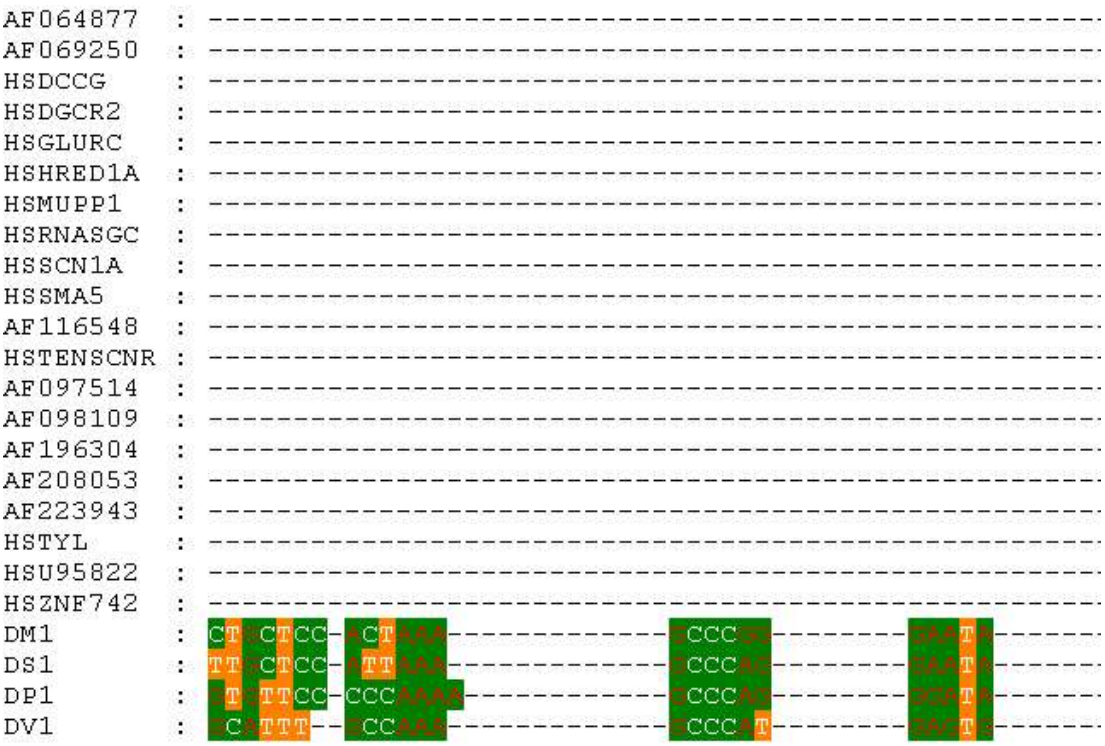


Abbildung 28. Auschnitt aus dem Profilalignment mit den gefilterten Sequenzen aus der Ergebnismenge. Die Abbildung zeigt das zu dem Anteil des Stemloop 1 von Bicoid, der die konservierten Sequenzanteile aufweist, keine Sequenzen aus der Ergebnismenge zugeordnet werden können.

Die Screenings von Stemloop 2 und 3 der modular aufgebauten Signalstruktur BLE zeigten ähnliche Ergebnisse. Es konnten zwar ähnliche Sekundärstrukturen gefunden werden, aber die wenigen Consensussequenzmotive die BLE enthält, waren darin nicht enthalten

Weitere Suchen mit RNAMOT wurden sowohl unter POSTREGFINDER als auch auf dem Rechner des GENIUSnet am DKFZ in Heidelberg als Kontrolle durchgeführt. Hier wurde aus den gefundenen Einträgen ein „file of filenames“ erzeugt, der die Dateinamen der Einträge als Liste enthält. Um die Suche im Batchverfahren durchzuführen, wurde eine Batchdatei mit einem UNIX Skript angelegt, in der die Abfolge des Suchverfahrens festgelegt wurde. Die Ergebnisse, bezogen auf das angelegte Projekt von in Hirenzellen exprimierten Genen, verlief jedoch analog zu Bicoid Stem 1.

Batchverfahren Kommandos	*Bedeutung*
fetch @brainhum.txt -out / tmp/brainhumgb.txt -nomon	Sucht die Einträge zu den Dateinamen in der „file of filenames“ -Liste und speichert sie in einem temporären Verzeichnis. Dabei wird die Monitorausgabe unterdrückt.
rnamot -s -s /tmp/brainhumgb.txt -d bcdstem1.des -o stem1 -t > stem1.tra	Startet die Suche mit RNAMOT im Suchmodus (-s) mit der Deskriptor-beschreibung und protokolliert sie in einer Trace-Datei (-t).
rm /tmp/brainhumgb.txt	Löscht die Daten im temporären Verzeichnis.

Tabelle 33. Liste der Kommandos für das Batchverfahren.

3.5. Postreganalyse mittels KDD Methoden

Ziel des Ansatzes ist, die Methode des „Relational Instance Based Learning“ RIBL aus dem Bereich der KDD Methoden zur Erkennung von Postreg Motiven anzuwenden. RIBL wurde als geeignete Methode eingesetzt, um neue Sequenzen und Klassen von Signalstrukturen aus genomischen Sequenzdaten herausfiltern und definieren zu können, ohne dass die Information über solche neuen Klassen vorher bekannt sind. Für diese Suche nach noch unbekannten neuen PORD-Klassen bedarf es eines Ähnlichkeitsmaßes für die Eigenschaften der einzelnen Instanzen einer PORD-Klasse, das geeignet ist, eine neue PORD-Klasse so zu charakterisieren, dass sie von genomischen Sequenzen ohne oder anderer Bedeutung unterschieden werden kann. Um solch ein Ähnlichkeitsmaß zu finden, müssen zunächst die vorhandenen Postreg-Sequenzen und deren Sekundärstrukturen in eine Repräsentation gebracht werden, die der Aufschlüsselung ihrer Eigenschaften ermöglicht und damit jedes Merkmal der Sequenzabfolge und Sekundärstruktur einer Auswertung zugänglich macht. In einem weiteren Schritt muss ein Ähnlichkeitsmaß anhand von vorhandenen Instanzen der Klassen gefunden werden, welches sich dann auf genomische Sequenzdaten anwenden lässt, um dort neue Klassen zu identifizieren.

3.5.1. Repräsentation von RNA-Signalstrukturen

Die Repräsentation wurde in einer Prädikatenlogik zusammengefasst, die einer kontextfreien Grammatik entspricht. Die dreidimensionale Struktur einer RNA ist aus einer Abfolge von Basenpaar- und Einzelstrangregionen aufgebaut, wodurch ihre Sekundärstruktur ein Muster mit unterscheidbaren Teilelementen erhält. Diese strukturgebenden Teilelemente der Sekundärstruktur sind „stacking regions“, „hairpin loops“, „internal loops“, „multibranch loops“, „bulge loops“, und „dangling ends“ *(Shapiro et al., 1990)*. Die als Strukturelemente bezeichneten Teilelemente können als Knotenpunkte eines Baumes betrachtet werden, in der die „multibranch loops“ jeweils

den Ausgangspunkt für den Baum einer Teilstruktur bildet. Für eine vollständige Beschreibung werden hauptsächlich drei Eigenschaften benutzt: Die Topologie der Sekundärsturktur, die in einer Baumbeschreibung mündet, sowie die Größe und die Basenabfolge der Strukturelemente. Ein Beispiel ist auf der folgenden Seite dargestellt.
In dem Beispiel ist jedes Objekt s einer Fallbeschreibung durch das Fakt repräsentiert:

structure(s; root(se 1 ; se 2 (se 3 ; se 4 (se 5 ; se 6 (se 7 ; se 8 (se 9); se 10)); se 11); se 12)) :

-wobei *s* eine Signalstruktur und *se 1;....; se 12* die Strukturelemente und die Knoten der Baumstruktur repräsentieren und selber Objekte darstellen. „*root*" ist als Hilfskonstrukt der Ausgangsknoten der Baumstruktur. Grundlegendes Fakt ist jedes Postreg-Element und wird durch das Prädikat beschrieben:

- *signal_structure_class(struct_id,* name*)*
  - ✔*signal_structure_class* ist der Prädikatenname für jede Signalstrukturklasse
  - ✔*struct_id* ist der Identifikator jeder Instanz einer Klasse

Die Instanz einer Strukturklasse und jedes Strukturelement wird repräsentiert durch die Prädikate:

- structure(*struct id* : name; *topology* : term)
  - ✔*topology* ist die Baumstruktur der Postreg-Instanz
- helical(*se_id* : name; *type* : constant; *size* : number; *bases* : list; *bases* : list)
- single(*se_id* : name; *type* : constant; *size* : number; *bases* : list)
  - ✔ „helical" und „single" ist definiert als Prädikatenname *se_type*, der den allgemeinen Typ des Strukturelements kennzeichnet
  - ✔*se_id* ist der Identifikator des Strukturelements
  - ✔*type* ist die Prädikatbezeichnungen des Typs der Strukturelemente hairpin, stem, bulge_5, bulge_3, single_5, single_3.
  - ✔*size* ist die Größe des Elements in bp
  - ✔*bases* gibt die Liste der Nukleinsäureabfolge des Strukturelements an

3.5.2. Klassifikation von RNA-Signalstrukturen

Aus der vorhandenen Dateibibliothek wurden Signalstrukturklassen ausgewählt, deren Funktion sowohl durch die Sequenzabfolge als auch durch die Sekundärstruktur codiert wird. Dabei wurden alle Instanzen der ausgewählten Klasse übernommen, also alle Instanzen, die verschiedene taxonomische Quellen repräsentieren, aber in der Literatur als einer Klasse zugehörig charakterisiert worden sind.

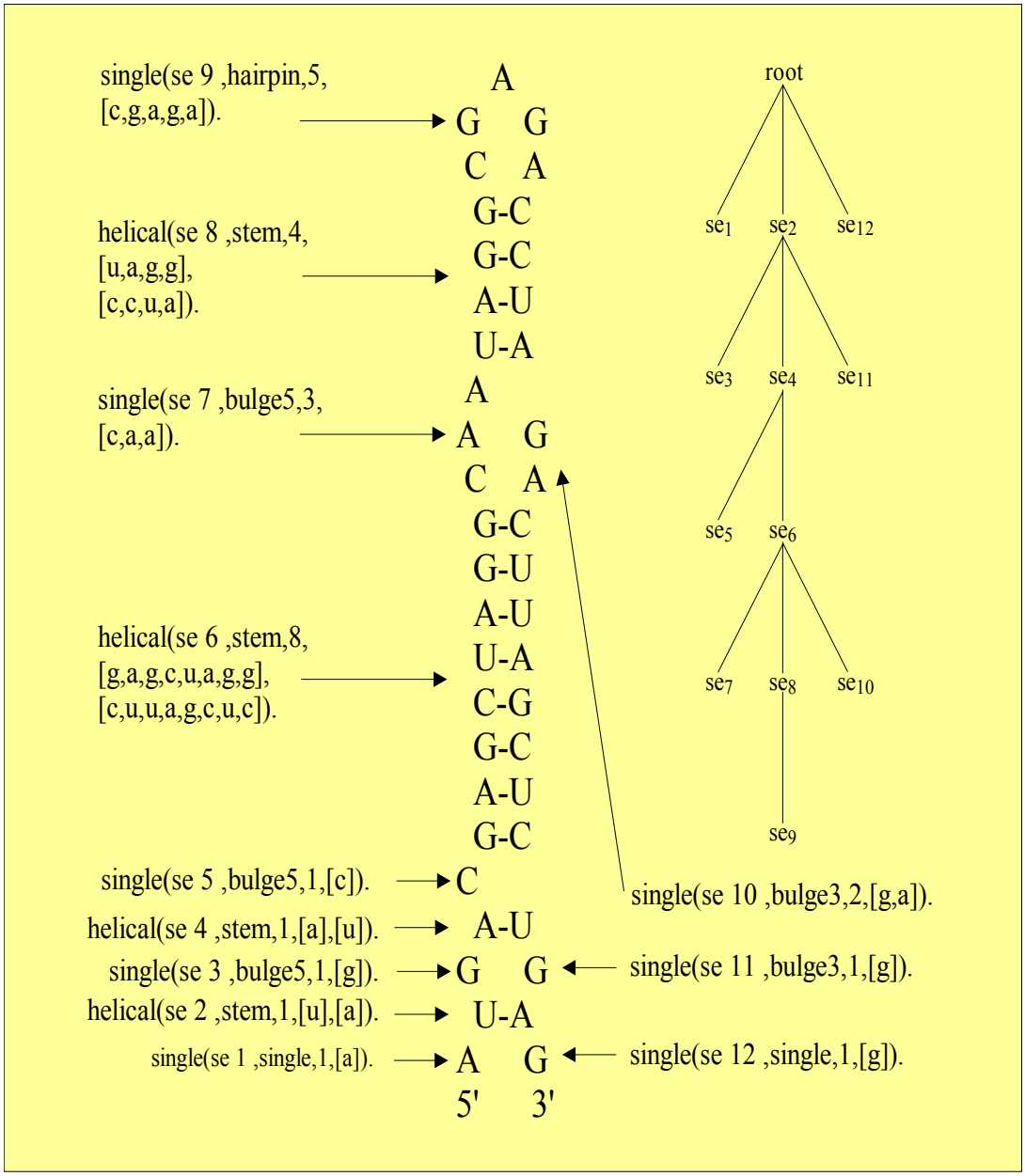


Abbildung 29. Beispiel für eine Baumrepräsentation (Secis E2)

Wie die obere Liste aufzeigt, haben die Eigenschaften „helical“ und „single“ jeweils eine unterschiedliche Anzahl von Prädikaten, um die verschiedenen Basenabfolgen zu dokumentieren.

3.5.2.1. Fallbasis und Parameter der Klassifikation

Die Fallbasis besteht aus 7 Postreg-Klassen: APO *(Margot et al., 1996)*, BICOID *(Macdonald et al., 1998)*, COXATP9 *(Dekker et al., 1992)*, HISTONE *(Hanson et al., 1996)*, IRE *(Hentze & Kuhn, 1996; Dandekar et al., 1998)*, SECIS 2 *(Low & Berry, 1996, Wilting et al., 1997)*, TAR *(Verhoef et al., 1997).* Der Datensatz besteht aus den gefalteten Strukturen im „connect" oder „ct" Format, die von einem Parserprogramm in die Grammatik umformatiert werden. Die Klassen enthalten 400 Instanzen mit Strukturvarianten aus 144 verschiedenen Sequenzen und 157 verschiedenen Stukturen. Die Ähnlichkeit zwischen den verschiedenen Instanzen basiert auf der Edit-Distanz. Die Edit-Distanz beruht auf den Kosten, die der Austausch (Editierung) von Basen oder Strukturelementen verursacht, um eine Instanz auf eine andere Instanz abzubilden. Die Klassifikation wurde durchgeführt anhand festgelegter Distanzwerte, die als Parameter eingegeben werden und vom Benutzer verändert werden können. Das von RIBL vorgegebene Intervall für Distanzwerte ist [0,1], so dass die Werte als Subtraktion von 1 untereinander transformiert werden können. Das Kostenmodell für die Edit-Distanz zwischen Basen des Alphabets {A,C,G,U} hat die Konstante 1. Das Kostenmodell für Editoperationen zwischen Strukturelementen ist in der Tabelle 29. enthalten.

	root	***single***	***bulge_3***	***bulge_5***	***hairpin***	***stem***
root	0,00	0,03	0,03	0,03	0,90	0,03
single		0,00	0,03	0,03	0,18	1,00
bulge_3			0,00	0,06	0,12	1,00
bulge_5				0,00	0,12	1,00
hairpin					0,00	1,00
stem						0,00

Tabelle 34. Kostenmodell für Editoperationen von Strurkturelementen

Die Distanzwerte sind heuristischen Ursprungs und spiegeln Erfahrungen im Umgang mit optimalen und suboptimalen Strukturen wider, aus denen sich folgende Kriterien ergeben haben: Die Änderung eines Strukturelements von einem Einzelstrang- in einen Doppelstrangtyp stellt beispielsweise eine substantielle Änderung dar und wurde mit entsprechend hohen Kosten versehen, weil die Abbildung solcher Strukturelemente unterdrückt werden sollte. Zweitens ist die endständige Schleife einer Haarnadelstruktur, der „hairpin loop", deshalb von entscheidender Bedeutung, weil die Existenz einer Haarnadelstruktur die endständige Schleife voraussetzt. Deshalb sind hier ebenfalls höhere Editkosten angesetzt.

3.5.2.2. Klassifikationsergebnisse

Die ersten beiden durchgeführten Experimente betrafen den Datensatz, der jeweils nur verschiedene Strukturen beinhaltet und deshalb als „Nonredundant“ bezeichnet wird. Der erste Teil eines jeden Experiments wurde ohne die Strukturrepräsentation durchgeführt, um zu testen, welchen zusätzlichen Beitrag die Sekundärstrukturinformationen für die Diskriminierung von Klassen leistet. Das zweite Experiment beruht auf dem redundanten Datensatz inklusive der Faltungsvarianten. Die Ergebnisse sind in Tabelle 35. zusammengefasst.

Representation	*Instanzen*	*Richtig klassifiziert (k 1,2,...10)*	*k*	*Genauigkeit (%)*
Sequenz $_{nonredundant}$	144	**134**,134,131,129,127,126,124,121,120,119	1	93,06
Struktur $_{nonredundant}$	157	**145**,145,141,143,142,142,138,138,133,131	1	92,36
Sequenz $_{redundant}$	400	**380**,380,377,375,367,366,364,361,360,359	1	95,00
Struktur $_{redundant}$	400	**385**,385,379,383,381,381,372,372,365,363	1	96,25

Tabelle 35.Ergebnisse des Klassifikationsexperiments

Die Ergebnisse zeigen, dass insgesamt eine sehr hohe Genauigkeit der Klassifikation von teilweise deutlich über 90% erzielt werden kann. Die Tatsache, dass fünf von zwölf Strukturen in dem „nonredundant“ Experiment falsch klassifiziert wurden, ist eine Folge der großen Ähnlichkeit zwischen den SECIS Klassen *(Low & Berry, 1996)* und IRE. In sechs Fällen wurde SECIS als IRE klassifiziert und eine falsche Einteilung erfolgte zwischen COXATP9 und TAR. Die hohe Klassifikationsgenauigkeit ist ein Resultat des Verhältnisses der relativ kurzen Sequenzen der Postreg-Strukturen zu den relativ großen Kernbindungsregionen, die darin enthalten sind. Daraus ergibt sich, dass der hohe Grad der Sequenzhomologie zwischen den Instanzen einer Klasse ausreicht, sie richtig zu klassifizieren.

In diesem ersten Experiment sind in dem Datensatz nur Instanzen von Signalstrukturklassen enthalten, die nachweislich zu einer Klasse gehören. In diesem ersten Experiment des Klassifikationsverfahrens fehlt daher noch der Nachweis, zwischen ähnlichen Sequenzen, die aber unterschiedliche Strukturen aufweisen, und ähnlichen Strukturen, die signifikant unterscheidbare Sequenzabfolgen aufweisen, zu diskriminieren, also falsche Positive zu erkennen. In einem zweiten Experiment soll daher die Vermutung überprüft werden, dass die Einbeziehung der Sekundärstrukturbeschreibung einen signifikanten Vorteil gegenüber der Berücksichtigung der bloßen Sequenzab-

folge bei der Klassifizierung bietet, insbesondere was die Erkennung falscher Positiver Beispiele im Datensatz anbelangt.

Für das zweite Experiment wurde ein Testdatensatz aus 24 IRE Signalstrukturen und 88 falschen Postitiven mit der Software Matinspector *(Quandt et al., 1995)* generiert. Die Sequenzen sind Resultat eines Sequenzscreenings mit einer Matrix aus dem kompletten Satz der IRE-Motivsequenzen und dem korrespondierenden Sequenzdatensatz aus EMBL-Sequenzeinträgen. Die Parametereinstellungen für das Screening (core similarity = 0:77 und matrix similarity = 0:72) waren so optimiert, dass alle IRE-Subsequenzen in den EMBL-Einträgen gefunden werden. Der EMBL-Datensatz wurde außerdem um 15 auf 24 Einträge verringert, weil nicht in allen EMBL Datensätzen mit den optimierten Parametereinstellungen die IRE Subsequenzen gefunden oder mit RNASTRUCTURE mit den Standardparametereinstellungen nicht alle Sequenzen korrekt gefaltet werden konnten. Die gefalteten Sequenzen ergaben 179 optimale und suboptimale Faltungen die für das Experiment eingesetzt wurden.

Alle Beispiele	112	179	112
Enthaltene IRE	24	32	24
Korrekt klassifizierte IRE	24	32	24
Klassifiziert als IRE	112	80	48
Enthaltene Nicht-IRE	88	147	88
Klassifiziert als Nicht-IRE	0	99	64
recall IRE	1	1	1
recall Nicht-IRE	0	0,67	0,73
Genauigkeit	0,21	0,73	0,79

Tabelle 36. Ergebnisse der Klassifikation mit IRE Instanzen und falschen Positiven. Der Ausdruck „recall" bezeichnet die Sensitivität mit der eine Klasse als Template die zu ihr gehörigen Beispiele erfasst und ist damit ein Maß für die Anzahl falsch negativ klassifizierten Beispiele. Je höher der Wert desto größer ist die Genauigkeit der Klassifikation.

Der Schwellenwert für die maximale Distanz, der von RIBL angewendet wurde, um ein Beispiel als IRE zu klassifizieren, wurde aufgrund der folgenden Kalkulation ermittelt:

$$max(nn\,ire(ire\,1); nn\,ire(ire\,2); : : : ; nn\,ire(ire\,n)); k = 1$$

wobei nn ire (ire i) die Distanz des nächsten IRE Nachbarn von der IRE Instanz i ist. Das Ergebnis des Experiments ist in Tabelle 31. dargestellt und zeigt, dass die Anzahl der korrekt klassifizierten IRE's zunimmt, wenn die topologische Information in die Klassifikation mit eingeschlossen wird. Die Genauigkeit nimmt noch einmal zu, wenn die Strukturen wieder auf die Sequenzabfolge abgebildet werden, von denen sie Re-

präsentanten sind, d.h. ein gegebenes Sequenzbeispiel wird dann als IRE klassifiziert, wenn eins der korrespondierenden Faltungen als IRE klassifiziert worden ist.

4. Diskussion

Die in der Arbeit vorgestellten Suchverfahren werden in der Informatik in dem Begriff „Data Mining“ zusammengefasst. Data Mining ist mittlerweile als Verfahren in der Bioinformatik im Bereich der Genomforschung etabliert. Unter Data Mining versteht man die Anwendung unterschiedlicher, aufeinander abgestimmter Algorithmen, die im Baukastensystem zu einem Prozess zusammengebunden werden, um in großen Datenmengen neue Zusammenhänge und damit neues Wissen zu generieren *(DataMining in Biotechnology 2000)*. Die beiden hier entwickelten Data Mining Verfahren unterscheiden sich dadurch, dass einerseits direkt aus einer Datenbank von annotierten Klassenbeschreibungen homologe RNA-Signalstrukturen über Sequenz- und Strukturhomologien gesucht und andererseits über ein Ähnlichkeitsmaß komplett neue Klassen voneinander diskriminiert werden können. Letzteres geschieht mittels des maschinellen Lernverfahrens RIBL über eine Klassifikation, in der Beispiele einer Signalstrukturklasse einander zugeordnet und darüber die Merkmale „gelernt“ werden, die die Klassen voneinander unterscheiden.

4.1. Anwendung von Data-Mining Verfahren auf RNA-Signalstrukturen

Durch Testklassifikationen mit ausgewählten Datensätzen konnte festgestellt werden, dass bei redundanten Datensätzen, wie sie bei nicht ausgewählten Sekundärstrukturen zu erwarten sind, die Sekundärstrukturinformation die Klassifikationsgenauigkeit sichtbar erhöht (siehe Tabelle 23). Bei der Klassifizierung von RNA-Signalstrukturen, vor allem mit vielen redundanten Sekundärstrukturen, wird also durch die Einbeziehung der Sekundärstrukturinformation eine deutlich größere Trennschärfe zwischen den Klassen im Klassifikationsergebnis erreicht, während bei nicht redundanten Sequenzen natürlich die Sequenzabfolge verlässlicher für eine sichere Klassifizierung ist. Hier macht sich die Sekundärstruktur sogar als Störung der Klassifikation bemerkbar, weil sie das Kriterium der aufgrund ihrer Sequenzabfolge unterschiedenen Klassenbeispiele überlagert.
Eine Anwendung in Form eines Clusterings von den so sichtbar gemachten statistischen Eigenschaften von Signalstrukturklassen konnte im Rahmen der Kooperation mit der Informatik nicht erfolgen, wäre aber eine sinnvolle und notwendige Weiterentwicklung des Verfahrens. Das Clustering würde die Anwendung des aus der Klassifikation gewonnen Ähnlichkeitsmaßes auf genomische Sequenzdaten beinhalten.

4.2. POSTREGFINDER und PORD

Das in dieser Arbeit beschriebene Softwaresystem POSTREGFINDER ist das einzige existierende Softwaresystem, das die Annotation von RNA-Signalstrukturklassen und geeignete Screeningverfahren nach homologen Primär- und Sekundärstrukturen in

einer Plattform vereint. Es unterstützt die benutzergesteuerte Annotation und Eingabe neuer Signalstrukturenklassen sowie deren Auswahl und die Parametrisierung für Suchverfahren in genomischen Datenmengen. Dabei hängt die Laufzeit und Größe des Suchraums jeweils von der zur Verfügung stehenden Hardware ab.

4.2.1. Annotation und Importfunktion

Die Importfunktion sieht vor, dass die Einträge, die in die Tabellen „Gene“, „Sequence“ und „Literature“ importiert werden und eine vollständige RNA-Signalstrukturklasse charakterisieren und dokumentieren, sich aus verschiedenen Quellen zusammensetzen. Als Quellen dienen die Einträge aus der EMBL-Data-Library, Literaturangaben aus MEDLINE und Ausgabe des Programms MFOLD/ RNASTRUCTURE. Die Annotation von neuer Sequenzinformation ist, wie man an dem viel umfangreicheren humanen Genomprojekt sehen kann, eine zeitaufwändige Arbeit. Dies liegt daran, dass trotz einer Vielfalt von Algorithmen, die den Annotationsprozess unterstützen, manuelle Arbeitsschritte in der Regel nicht zu umgehen sind, weil bestimmtes Detailwissen über die zu annotierenden Sequenz wie beispielsweise ihre Herkunft nur von den Autoren der neuen Sequenzeinträge gewusst werden. Die Arbeitsgruppen der Institutionen, die die internationalen Sequenzdatenbanken betreuen und auf aktuellem Stand halten, haben für diesen Zweck der Annotation neuer Sequenzen internetbasierte Software wie beispielsweise WEBIN am EBI („European Bioinformatics Institute“, *Hingamp et al., 1999)* und PC Software wie SEQUIN *(Benson et al., 2002)* bereit gestellt. SEQUIN kann für die Eingabe von Sequenzeinträgen in allen drei Datenbanken GenBank, EMBL Data-Library und DDBJ genutzt werden. Der Autor eines neuen Sequenzeintrags wird durch automatische Verfahren, wie die Analyse des genetischen Codes, Suche nach ORF und Repeats, dabei unterstützt, wenn bestimmte Informationen, wie die taxonomische Einordnung der eingegebenen Sequenz, manuell eingegeben worden sind. In dem neuen Annotationsverfahren für RNA Primär- und Sekundärstrukturen in dem Annotations- und Importmodul von POSTREGFINDER, werden ebenfalls bestimmte manuelle Eingaben benötigt, bevor der weitere Prozess der Annotation durch Algorithmen unterstützt werden kann. Essentiell ist zunächst die Eingabe eines Klassenidentifikators. Als Klassenidentifikatoren existieren in PORD jeweils zwei Schlüsselwerte, wie es in den internationalen Sequenzdatenbanken gebräuchlich ist, ein „ID“ und „AC“ Schlüssel. Die „Accession number“ AC wird automatisch, während ein neuer Klasseneintrag angelegt wird, vergeben. Der zweite Klassenidentifikator, die ID, sollte in Anlehnung an die in der Literatur gebräuchlichen Bezeichner für die Klasse erstellt werden, beispielsweise „ID BL1$FLY“ für BLE Stem 1 in der Fruchtfliege „FLY“. Das Annotationsmodul unterstützt weiterhin das Anlegen eines modularen Klasseneintrags durch PORD-interne Verweise sowie die Filterung relevanter Informationen aus einem EMBL-Eintrag in die Gen- und Sequenzrelation wie Genname, UTR-Bereich. Sukzessive können im Annotationsmodul die Einträge in die Gen-und Sequenzrelation durch Laden der dafür notwendigen EMBL-Einträge hinzu gefügt werden. Dabei wird der Benutzer durch die automatische Neuberechnung von Sequenzmatrices beim Import von neuen

Motivsequenzen sowie die automatisierte Erstellung von Sekundärstrukturbeschreibungen im DESCRIPTOR Format beim Import von Sekundärstrukturangaben im „Connect“ Format von Mfold unterstützt. Bei der Aktualisierung der Sekundärstrukturbeschreibung der Klasse kann der Benutzer auf einen Editor und eine Liste aller zu einer Klasse zur Verfügung stehenden Sekundärstrukturbeschreibungen zurück greifen, um dann die Consensusstruktur zu erstellen.

Auch um Literaturangaben zu ergänzen, stehen Filteroperationen zu Verfügung, die die Angaben zu Autor, Titel, Zeitschrift automatisch aus MEDLINE-Einträgen geordnet in die PORD Literaturrelation übernehmen. Für das Einbeziehen von in der Literatur vermerkten Labormethoden, durch die ein neues Beispiel einer Signalstruktur ermittelt wurde, muss die Angabe zunächst manuell aus der Literatur übertragen werden. Parallel wird aber eine nicht redundante Liste aller Methoden zu jeder neuen Eingabe erstellt, aus der die zutreffende gewählt werden kann, so dass mit dem Anwachsen der Literaturdaten sukzessive ein Kompendium aller benutzten Methoden entsteht. Im Gegensatz zu der Datenbank TRANSFAC wird darauf verzichtet, die Labormethode einer Bewertung auf ihre Zuverlässigkeit hin zu unterziehen, indem man ihr einen Index zuordnet. Dabei besteht die Gefahr, den jeweils neu entstehenden Varianten einer Methode nicht gerecht zu werden. Insgesamt bietet die Software im Gegensatz zu den bekannten Annotationsverfahren als einzige Software eine breite Unterstützung für die Annotation von RNA-Signalstrukturen.

4.2.2. Erstellung der Signalstruktur- Klassen

Die Klassenbeschreibungen von PORD enthalten sowohl Matrices als auch Strukturbeschreibungen. Diese werden manuell aus den Strukturbeschreibungen der einzelnen Sequenzen erstellt. Sie werden mittels eines selbst entwickelten Algorithmus aus der Datei mit den Daten der Sekundärstruktur der Postreg-Sequenz, wie sie MFOLD generiert, geparst. Dadurch wird die Erstellung der Ausgangsdatenbasis für die Klassenbeschreibung beschleunigt. Bei der Sekundärstrukturbeschreibung im DESCRIPTOR-Format der Klasseneinträge werden nur die minimalen gemeinsamen Merkmale aller Einzelbeispiele übernommen. Wie das Testscreening gezeigt hat, wird dadurch gewährleistet, dass die Strukturangaben als Vorlage geeignet sind, das eigene Komplement zu finden und damit als Screeningvorlage tauglich sind. Jeder Klasseneintrag ist in dieser Weise geprüft worden. Es ist allerdings wünschenswert, dass die Consensusstruktur ebenfalls durch einen intelligenten Parser aus den Strukturbeschreibungen der jeweils einzelnen Sequenzen erstellt werden kann. Letztere sind Sturkturbeschreibungen im DESCRIPTOR Format einzelner Sequenzmotive in der Tabelle „site“ von PORD, die direkt aus der Faltung der einzelnen Motivsequenz erhalten werden. Die manuelle Erstellung komplexer Strukturangaben, wie bei dem Gen Bicoid das BLE Element, ist ganz eindeutig der „Flaschenhals“ im gesamten Annotationsprozess. Die Erstellung von Consensusstrukturen solcher RNA Signalstrukturen kann bis zu einem ganzen Tag und mehr beanspruchen, bis alle Strukturvarianten und Sequenzabschnitte innerhalb der Strukturelemente mit der Primärsturktur

richtig abgeglichen sind. Hier ist auch der Zusammenhang zu sehen, zu den Arbeiten mittels Data Mining in Kooperation mit der Informatik ein Ähnlichkeitsmaß für RNA Primär- und Sekundärstrukturen zu finden, das in der Lage wäre, diesen Schritt durchzuführen. Das Klassifikationsverfahren ist ein viel versprechender Schritt dahin, aber noch keine Lösung, die auf das genannte Problem anwendbar wäre (siehe 4.1.). Die so erstellten Klassenbeschreibungen bieten eine flexible, gut strukturierte Vorlage für die Suche nach neuen Instanzen einer Signalstrukturklasse.

4.2.3. Parametrisierung und Laufzeitverhalten einer Suche

Für die Erstellung der Matrices wurde der Algorithmus WCONSENSUS von G. Hertz *(1990, 1999)* verwendet. Alternativ wäre auch das Programm MATIND aus dem Programmpaket MATINSPECTOR *(Quandt et al., 1995)* geeignet gewesen, mit

Abbildung 30. Parameterdialog für das Screening. Links sind die Parameter für PATSER einzugeben, rechts für RNAMOT. Unten können die Daten für den aktuellen Suchvogang bestimmt werden.

dem eine erste Klassenbeschreibung für die Fallbasis der Klassifikation generiert wurde. Allerdings lag für das Programm kein Quellcode vor, während das Matrixerstellungsprogramm Wconsensus und das Matrixsuchprogramm Patser mit Quellcode

frei erhältlich waren. Beide sind hinsichtlich Laufzeitverhalten und Sensitivität vergleichbar. Die Programme sind wie das Programm RNAMOT als Dynamic Link Library in die POSTREGFINDER Software integriert und können über eine gemeinsame Schnittstelle, dem Parameterdialog, aufgerufen werden. Dadurch kann das Screeningverfahren einfach parametrisiert werden, indem die Daten der Datenbank PORD und des aktuell selektierten Projekts mit den Parametern zu einem neuen Suchprozess verknüpft werden.

Die Suche nach Sequenzabfolgen, die den eingegebenen Matrices entsprechen, benötigt wegen der Komplexität des Suchvorgangs naturgemäß erheblich weniger Zeit als ein Sekundärstrukturalignment mit RNAMOT. Die zeitliche Dimension ist bei einer Sequenzhomologiesuche in genomischen Datenmengen im Minutenbereich angesiedelt, während eine Suche mit RNAMOT ein bis mehrere Stunden dauern kann. Die benötigten Zeiten für einen Suchlauf für die 40 MB große Datei in dem Screeningbeispiel (siehe 3.4.) sind in der folgenden Tabelle dargestellt.

Signalstruktur	***Sequenzsuche mit den Matrices***		***Struktursuche mit DESCRIPTOREN***	
	Matrixgröße (Sequenzanzahl/Länge)	Minuten	Descriptorgröße (Strukturelemente/ Basen[5])	Minuten
IRE	15 Sequenzen 50 Basen	4	7 SE 29 Basen	190
BLE Stem 1	5 Sequenzen 115 Basen	11	14 SE 24 Basen	390

Tabelle 37. Laufzeitverhalten der Suchalgorithmen PATSER und RNAMOT

Es bietet sich daher als Möglichkeit an, die Suchläufe für das Auffinden von Strukturhomologien jeweils zu kombinieren, indem man den Suchraum bei großen Sequenzdatenmengen mittels einer vorgeschalteten Matrixsuche zunächst eingrenzt. Ein Problem dabei ist noch, dass das Ergebnis einer solchen Matrixsuche in der jetzigen Version von Postregfinder noch manuell umgesetzt werden muss in eine neue Sequenzdatenmenge, d.h. aus der Ergebnismenge müssen die „Accession"-Schlüssel heraus kopiert werden, um damit eine neue Sequenzdatei über SRS anzulegen, in der dann die zweite Suche mit RNAMOT erfolgen kann. Dieses Problem sollte in einer der nächsten Versionen von POSTREGFINDER gelöst werden.

5 Die Anzahl der Basen gibt an, wieviel Basenübereinstimmungen im Descriptor vorgeschrieben sind.

4.2.4. Anwendung von POSTREGFINDER und PORD

Die Erstellung von Klassenbeschreibungen mittels Matrices ist für Datenbanken eine geeignete Methode, Klassen von Sequenzmotiven bestehend aus mehreren Sequenzen zu dokumentieren und sie als Suchschema einzusetzen *(Quandt et al., 1995)*. Matrixanwendungen und Strukturalignmentprogramme sind erfolgreich eingesetzt worden, um z.B. Signalstrukturen von Promotoren in Genen und Genclustern von Insekten *(Pongjaroenkit et al., 2001)* zu finden. Auch das DESCRIPTOR - Format *(Gautheret et al, 1990)* hat sich für die Beschreibung von PORD-Sekundärstrukturklassen und als Suchmuster für die Suche nach homologen Sekundärstrukturen zum SECIS Element *(Lescure et al., 1999)* bewährt.
Allerdings sind diese Strukturbeschreibungen bisher immer für Einzelbeispiele erstellt worden. Durch die Sammlung der Strukturbeschreibungen in einer Datenbank, wie in dieser Arbeit vorgestellt kann jetzt parallel in einem Suchlauf nach Beispielen mehrerer Klassen gescreent werden.

Es gibt verschiedene Ansätze, Datenbanken für RNA-Signalstrukturen aufzubauen, wie die Datenbank TRANSTERM („TRANScription TERMination“) oder die Alignmentsammlung ACUTS *(„Ancient Conserved UnTranslated Sequences“ - Duret et al. 1993, Duret et al. 1997)*. Die Alignmentsammlung, die über das Verfahren des „phylogenetischen footprinting“ *(Tagle et al., 1988)* erstellt worden ist, kann nicht direkt für ein Screening genutzt werden, da aus einem Alignment erst eine Matrix extrahiert werden muss. Die Datenbank TRANSTERM enthält zwar Matrices, die auch mittels WConsensus von G. Hertz erstellt wurden, aber die Signalstrukturinformationen beschränken sich auf die sogenannten Koszak- Sequenzen – die kurzen Sequenzabschnitte stromauf- und stromabwärts der Start- und Stopcodons. In PORD können alle Typen von RNA-Signalstrukturen gespeichert werden. PORD hält im Gegensatz zu den bisherigen Ansätzen nicht nur die Sequenz, sondern auch die Strukturinformation vor. Dadurch, dass sowohl Sequenzmatrices als auch Strukturangaben in Form von deren strukturierter Beschreibung vorhanden sind, können alternative Suchstrategien gewählt werden und damit die unterschiedlichen Grade von Sequenz und Strukturkonservierung berücksichtigt werden.

Liegt ein positives Suchergebnis vor, kann die Klassenbeschreibung unmittelbar erweitert werden, indem der EMBL Eintrag mit der homologen Sequenz oder Sekundärstruktur über das Annotationsmodul als Bestandteil der Klasse annotiert werden kann. Dadurch wird ein komplettes Roundtrip Engeneering von Klassenbeschreibungen gewährleistet.

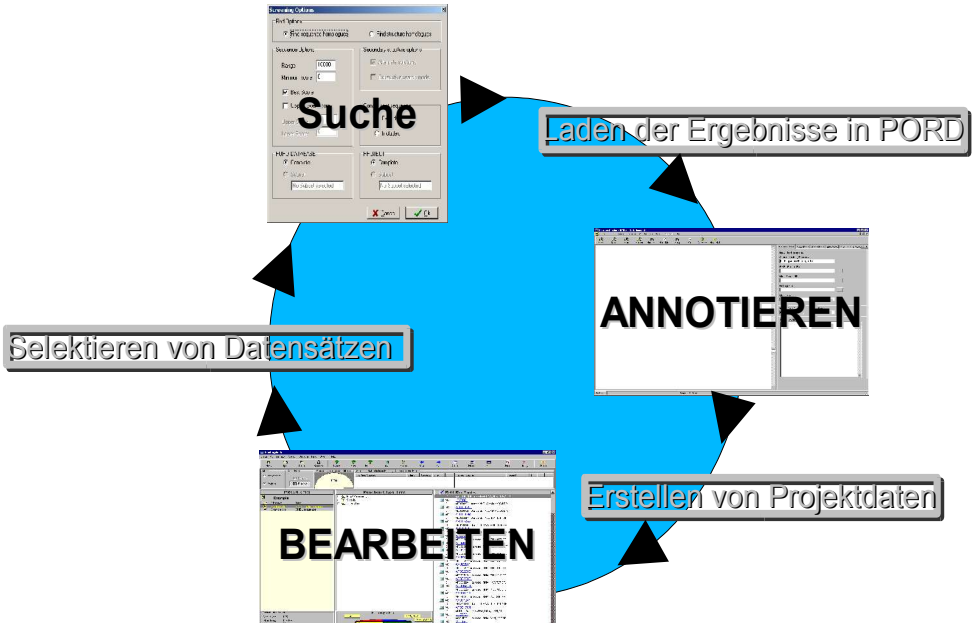


Abbildung 31. "Roundtrip Engineering" der Datenbankannotationen mit Hilfe positiver Screeningergebnisse

4.2.5. Weiterentwicklung der Datenbank PORD

Die Datenbank PORD liegt in der ersten Version als relationale Datenbank vor. Die Zahl von 12 Klasseneinträgen ist noch zu erweitern, um die komplette Anzahl der analysierten Postreg Sequenzen und Struturen vorzuhalten. Dazu kann die integrierte Schnittstelle für den Sequenzimport und -update eingesetzt werden. In der überwiegenden Anzahl interagieren die Signalstrukturen mit Proteinmotiven *(McCarthy et al. 1995),* die als Klassenbeschreibung der Proteinbindungssequenzen in der Datenbank PROSITE *(Hofmann et al. 1999)* dokumentiert sind. Es erscheint sinnvoll, die RNA-bindenden Motive der Datenbank PROSITE zu extrahieren und der Klassenbeschreibung der RNA-Bindungsstellen in der Datenbank PORD zuzuordnen, soweit Literaturangaben zu RNA-Protein Interaktionen eine eindeutige Zuordnung zulassen. Der Funktionsgewinn liegt in der dann gegebenen Möglichkeit, zu jeder PORD-Klasse über die entsprechende RNA-Bindungssequenz zu verfügen und damit eine Suche nicht nur nach homologen RNA-Beispielen, sondern auch nach deren Protein-Bindungspartnern durchzuführen. Dazu kann beispielsweise via Internet mit den Motivsequenzen der Proteine nach neuen Bindungssequenzen in den Proteindatenbanken PIR oder SWISS-PROT gesucht werden. Die dafür einsetzbaren Algorithmen sind durch die lokalen Alignmentprogramme FASTA und BLAST vorhanden, die sowohl für Nukleinsäure- als auch Proteinbindungssequenzen einsetzbar sind. Speziellere Screeningprogramme für Klassenbeschreibungen von Proteinbindungssequenzen können angewendet werden, wie beispielsweise PFAM *(Sonnhammer et al., 1997, Bateman et al., 2002).* Eine geeignete Datenstruktur für dieses Vorgehen wäre eine

neue Tabelle, die in N:N Relation zu der Tabelle für die Klassenbeschreibung der RNA-Bindungsmotive stehen würde. Daraus lässt sich ein RNA-Protein Interaktionsnetzwerk erstellen, das als Knowledge Discovery System Vorhersagen über das posttranskriptionale Verhalten eines Transkriptoms ableitet.

4.2.6. Weiterentwicklung von POSTREGFINDER

Um den Suchmodus zu erweitern, kann der RNAMOT Algorithmus für die Nutzung auch der einzelnen Motiv-Descriptoren eingesetzt werden. Dazu kann die Suchmaschine um eine Schnittstelle zum Suchen einzelner Strukturen ergänzt werden. In der vorliegenden Prerelease Version von POSTREGFINDER wird die Suche nach Strukturen nur über die Consensusbeschreibung unterstützt.

Als nächster Schritt wäre die Kombinierung des bisher separaten Primär- und Sekundärstrukturscreenings sinnvoll. Dadurch wird ein Arbeitschritt, der oft nach Sekundärsturktursuche erfolgt, nämlich ein Abgleich der Screeningergebnisse mit PATSER, um die Übereinstimmung der Sequenzen festzustellen, eingespart.

Die Weiterentwicklung von POSTREGFINDER erfordert eventuell auch eine Aktualisierung des RNAMOT Algorithmus. Während der Erstellung der Postregfinder Software ist auch der Algorithmus überarbeitet worden und firmiert unter dem neuen Namen „RNAmotif“ *(Macke et al., 2001, Lesnik et al., 2002)*. Dieser neue Algorithmus ist unter der Fragestellung zu evaluieren, welche Vorteile er für das Sekundärstrukturscreening bringt und welche Anpassungen von der Datenstruktur in PORD dafür notwendig sind.

5. Zusammenfassung

Die vorliegende Arbeit stellt die Software Postregfinder vor, die basierend auf Klassenbeschreibungen von RNA Signalstrukturen der Datenbank PORD in der Lage ist, in genomischen Sequenzdaten nach homologen Sequenzabschnitten und Teilstrukturen zu suchen. Dabei wurde insbesondere das Problem gelöst, eine Klassendefinition zu finden, die:

- einfach zu warten ist
- weitgehend automatisiert annotiert werden kann
- sowohl Sequenz- als auch Sekundärstrukturinformationen enthält
- alternative Suchstrategien erlaubt.

Die Klassenbeschreibung in PORD enthält sowohl einen Importmechanismus, der aus den einzelnen Sequenzbeispielen einen Klasseneintrag generiert, als auch die darin enthaltene Sequenzmatrix. Für eine alternative Struktursuche wird eine Strukturbeschreibung vorgehalten, die sich bereits für die Suche nach Sekundärstrukturen in genomischen Sequenzdaten bewährt hat. Über die Suchsoftware Postregfinder können PORD-Klassen und Sequenzdaten ausgewählt werden. Sie stellt einen intuitiven Zugang zu der Suchmaschine zur Verfügung, über die Suchalgorithmen parametrisiert und Suchprozesse gestartet werden können.

Ergänzend wurde in Kooperation mit dem Fachbereich Informatik der Universität Bremen ein Lernverfahren etabliert, das in der Lage ist, die in den Klassenbeschreibungen enthaltene Sequenz- und Strukturinformation für eine Diskriminierung der Klassen nutzen. Auf diese Weise wird nachgewiesen, dass eine erfolgreiche Diskriminierung von Signalstrukturklassen als Voraussetzung für die Suche nach neuen, noch unbekannten Signalstrukturklassen in genomischen Sequenzdaten möglich ist.

6. Verzeichnisse

Stichwortverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

1. Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., George R.A., Lewis S.E., Richards S., Ashburner M., Henderson S.N., Sutton G.G., Wortman J.R., Yandell M.D., Zhang Q., Chen
   The genome sequence of Drosophila melanogaster.
   Science. 2000 Mar 24;(287)5461 2185-95.
2. Agabian N.
   Trans splicing of nuclear pre-mRNAs.
   Cell. 1990 Jun 29;(61)7 1157-60.
3. Ahringer J., Kimble J.
   Control of the sperm-oocyte switch in Caenorhabditis eleganshermaphrodites by the fem-3 3' untranslated region.
   Nature. 1991 Jan 24;(349)6307 346-8.
4. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.
   Basic local alignment search tool.
   J Mol Biol. 1990 Oct 5;(215)3 403-10.
5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.
   Gapped BLAST and PSI-BLAST: a new generation of protein database searchprograms.
   Nucleic Acids Res. 1997 Sep 1;(25)17 3389-402.
6. Aly R., Argaman M., Halman S., Shapira M.
   A regulatory role for the 5' and 3' untranslated regions in differentialexpression of hsp83 in Leishmania.
   Nucleic Acids Res. 1994 Aug 11;(22)15 2922-9.
7. Aziz N., Munro H.N.
   Iron regulates ferritin mRNA translation through a segment of its 5'untranslated region.
   Proc Natl Acad Sci U S A. 1987 Dec;(84)23 8478-82.
8. Baas D., Puymirat J., Sarlieve L.L.
   Posttranscriptional regulation of oligodendroglial thyroid hormone (T3)receptor beta 1 by T3.
   Int J Dev Neurosci. 1998 Oct;(16)6 461-7.
9. Babitzke P., Yanofsky C.
   Reconstitution of Bacillus subtilis trp attenuation in vitro with TRAP,the trp RNA-binding attenuation protein.
   Proc Natl Acad Sci U S A. 1993 Jan 1;(90)1 133-7.
10. Bassett D.E., Eisen M.B., Boguski M.S.
    Gene expression informatics--it's all in your mine.
    Nat Genet. 1999 Jan;(21)1 Suppl 51-5.
11. Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L.
    The Pfam protein families database.
    Nucleic Acids Res. 2002 Jan 1;(30)1 276-80.
12. Baumstark T., Schroder A.R., Riesner D.
    Viroid processing: switch from cleavage to ligation is driven by a changefrom a tetraloop to a loop E conformation.
    EMBO J. 1997 Feb 3;(16)3 599-610.
13. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., Wheeler D.L.
    GenBank.
    Nucleic Acids Res. 2002 Jan 1;(30)1 17-20.
14. Benton D.
    Bioinformatics--principles and potential of a new multidisciplinary tool.
    Trends Biotechnol. 1996 Aug;(14)8 261-72.

15. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.
The Protein Data Bank.
Nucleic Acids Res. 2000 Jan 1;(28)1 235-42.
16. Bernal A., Ear U., Kyrpides N.
Genomes OnLine database (GOLD): a monitor of genome projects world-wide[In Process Citation]
Nucleic Acids Res. 2001 Jan 1;(29)1 126-7.
17. Bevilacqua J.M., Bevilacqua P.C.
Thermodynamic analysis of an RNA combinatorial library contained in ashort hairpin.
Biochemistry. 1998 Nov 10;(37)45 15877-84.
18. Bhat T.N., Bourne P., Feng Z., Gilliland G., Jain S., Ravichandran V., Schneider B., Schneider K., Thanki N., Weissig H., Westbrook J., Berman H.M.
The PDB data uniformity project.
Nucleic Acids Res. 2001 Jan 1;(29)1 214-8.
19. Biebricher C.K., Diekmann S., Luce R.
Structural analysis of self-replicating RNA synthesized by Q beta replicase.
J Mol Biol. 1982 Feb 5;(154)4 629-48.
20. Biebricher C.K., Luce R.
In vitro recombination and terminal elongation of RNA by Q beta replicase.
EMBO J. 1992 Dec;(11)13 5129-35.
21. Brazma A., Jonassen I., Eidhammer I., Gilbert D.
Approaches to the automatic discovery of patterns in biosequences.
J Comput Biol. 1998 Summer;(5)2 279-305.
22. Brion P., Westhof E.
Hierarchy and dynamics of RNA folding.
Annu Rev Biophys Biomol Struct. 1997;(26) 113-37.
23. Brown M., Wilson C.
RNA pseudoknot modeling using intersections of stochastic context freegrammars with applications to database search.
Pac Symp Biocomput. 1996;() 109-25.
24. Burks C., Lawton J.R., Bell G.I.
The LiMB database.
Science. 1988 Aug 19;(241)4868 888.
25. Burks C.
Molecular Biology Database List.
Nucleic Acids Res. 1999 Jan 1;(27)1 1-9.
26. Burland T.G.
DNASTAR's Lasergene sequence analysis software.
Methods Mol Biol. 2000;(132) 71-91.
27. Cao Y., Wilcox K.S., Martin C.E., Rachinsky T.L., Eberwine J., Dichter M.A.
Presence of mRNA for glutamic acid decarboxylase in both excitatory andinhibitory neurons.
Proc Natl Acad Sci U S A. 1996 Sep 3;(93)18 9844-9.
28. Chen Q.K., Hertz G.Z., Stormo G.D.
MATRIX SEARCH 1.0: a computer program that scans DNA sequences fortranscriptional elements using a database of weight matrices.
Comput Appl Biosci. 1995 Oct;(11)5 563-6.
29. Chetouani F., Monestie P., Thebault P., Gaspin C., Michot B.
ESSA: an integrated and interactive computer tool for analysing RNAsecondary structure.
Nucleic Acids Res. 1997 Sep 1;(25)17 3514-22.
30. Chicurel M.E., Singer R.H., Meyer C.J., Ingber D.E.
Integrin binding and mechanical tension induce movement of mRNA andribosomes to focal

adhesions.
Nature. 1998 Apr 16;(392)6677 730-3.
31. Chung S., Eckrich M., Perrone-Bizzozero N., Kohn D.T., Furneaux H.
The Elav-like proteins bind to a conserved regulatory element in the3'-untranslated region of GAP-43 mRNA.
J Biol Chem. 1997 Mar 7;(272)10 6593-8.
32. Claverie J.M.
Do we need a huge new centre to annotate the human genome? [letter;comment] [see comments]
Nature. 2000 Jan 6;(403)6765 12.
33. Cline T.W., Meyer B.J.
Vive la difference: males vs females in flies vs worms.
Annu Rev Genet. 1996;(30) 637-702.
34. Corpet F., Michot B.
RNAlign program: alignment of RNA sequences using both primary andsecondary structures.
Comput Appl Biosci. 1994 Jul;(10)4 389-99.
35. Dalphin M.E., Stockwell P.A., Tate W.P., Brown C.M.
TransTerm, the translational signal database, extended to include full coding sequences and untranslated regions.
Nucleic Acids Res. 1999 Jan 1;(27)1 293-4.
36. Dandekar T., Beyer K., Bork P., Kenealy M.R., Pantopoulos K., Hentze M., Sonntag-Buck V., Flouriot G., Gannon F., Schreiber S.
Systematic genomic screening and analysis of mRNA in untranslated regions and mRNA precursors: combining experimental and computational approaches.
Bioinformatics. 1998;(14)3 271-8.
37. Danon A., Mayfield S.P.
Light-regulated translation of chloroplast messenger RNAs through redoxpotential.
Science. 1994 Dec 9;(266)5191 1717-9.
38.
Data mining.
Nat Biotechnol. 2000 Oct;(18 Suppl) IT35-6.
39. Davis J.P., Janjic N., Javornik B.E., Zichi D.A.
Identifying consensus patterns and secondary structure in SELEX sequence sets.
Methods Enzymol. 1996;(267) 302-14.
40. Davis J.P., Janjic N., Pribnow D., Zichi D.A.
Alignment editing and identification of consensus secondary structures for nucleic acid sequences: interactive use of dot matrix representations.
Nucleic Acids Res. 1995 Nov 11;(23)21 4471-9.
41. Dekker P.J., Stuurman J., van Oosterum K., Grivell L.A.
Determinants for binding of a 40 kDa protein to the leaders of yeastmitochondrial mRNAs.
Nucleic Acids Res. 1992 Jun 11;(20)11 2647-55.
42. Delcher A.L., Kasif S., Fleischmann R.D., Peterson J., White O., Salzberg S.L.
Alignment of whole genomes.2369-76A new system for aligning whole genome sequences is described. Using anefficient data structure called a suffix tree, the system is able torapidly align sequences containing millions of nucleotides. Its use isdemonstrate
Nucleic Acids Res. 1999 Jun 1;(27)11 .
43. Dicks J.
Graphical tools for comparative genome analysis.
Yeast. 2000 Apr;(17)1 6-15.
44. Discala C., Benigni X., Barillot E., Vaysseix G.
DBcat: a catalog of 500 biological databases.
Nucleic Acids Res. 2000 Jan 1;(28)1 8-9.

45. Dominski Z., Marzluff W.F.
Formation of the 3' end of histone mRNA.
Gene. 1999 Oct 18;(239)1 1-14.
46. Draper D.E.
Themes in RNA-protein recognition.
J Mol Biol. 1999 Oct 22;(293)2 255-70.
47. Durand P., Canard L., Mornon J.P.
Visual BLAST and visual FASTA: graphic workbenches for interactiveanalysis of full BLAST and FASTA outputs under MICROSOFT WINDOWS 95/NT.
Comput Appl Biosci. 1997 Aug;(13)4 407-13.
48. Emerick V.L., Woodson S.A.
Self-splicing of the Tetrahymena pre-rRNA is decreased by misfolding during transcription.
Biochemistry. 1993 Dec 21;(32)50 14062-7.
49. Etzold T., Argos P.
SRS--an indexing and retrieval tool for flat file data libraries.
Comput Appl Biosci. 1993 Feb;(9)1 49-57.
50. Etzold T., Argos P.
SRS--an indexing and retrieval tool for flat file data libraries.
Comput Appl Biosci. 1993 Feb;(9)1 49-57.
51. Etzold T., Ulyanov A., Argos P.
SRS: information retrieval system for molecular biology data banks.
Methods Enzymol. 1996;(266) 114-28.
52. Etzold T., Ulyanov A., Argos P.
SRS: information retrieval system for molecular biology data banks.
Methods Enzymol. 1996;(266) 114-28.
53. Evers D., Giegerich R.
RNA movies: visualizing RNA secondary structure spaces.
Bioinformatics. 1999 Jan;(15)1 32-7.
54. Fayat G., Mayaux J.F., Sacerdot C., Fromant M., Springer M., Grunberg-Manago M., Blanquet S.
Escherichia coli phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identification of the gene for the ribosomal protein.
J Mol Biol. 1983 Dec 15;(3)171 239-61.
55. Ferrandon D., Koch I., Westhof E., Nusslein-Volhard C.
RNA-RNA interaction is required for the formation of specific bicoid mRNA3' UTR-STAUFEN ribonucleoprotein particles.
EMBO J. 1997 Apr 1;(16)7 1751-8.
56. Flamm C., Fontana W., Hofacker I.L., Schuster P.
RNA folding at elementary step resolution.
RNA. 2000 Mar;(6)3 325-38.
57. Florea L., Riemer C., Schwartz S., Zhang Z., Stojanovic N., Miller W., McClelland M.
Web-based visualization tools for bacterial genome alignments.3486-96With the increase in the flow of sequence data, both in contigs and wholegenomes, visual aids for comparison and analysis studies are becomingimperative. We describe three web-based too
Nucleic Acids Res. 2000 Sep 15;(28)18 .
58. Fontana W., Stadler P.F., Bornberg-Bauer E.G., Griesmacher T., Hofacker I.L., Tacker M., Tarazona P., Weinberger E.D., Schuster P.
RNA folding and combinatory landscapes.
PHYSICAL REVIEW. E. STATISTICAL PHYSICS, PLASMAS, FLUIDS, AND RELATED. 1993 Mar;(47)3 2083-2099.
59. Fox C.A., Sheets M.D., Wickens M.P.
Poly(A) addition during maturation of frog oocytes: distinct nuclear andcytoplasmic activities and

regulation by the sequence UUUUUAU.
Genes Dev. 1989 Dec;(3)12B 2151-62.
60. Frech K., Dietze P., Werner T.
ConsInspector 3.0: new library and enhanced functionality.
Comput Appl Biosci. 1997 Feb;(13)1 109-10.
61. Frech K., Quandt K., Werner T.
Finding protein-binding sites in DNA sequences: the next generation.
Trends Biochem Sci. 1997 Mar;(22)3 103-4.
62. Freier S.M., Kierzek R., Jaeger J.A., Sugimoto N., Caruthers M.H., Neilson T., Turner D.H.
Improved free-energy parameters for predictions of RNA duplex stability.
Proc Natl Acad Sci U S A. 1986 Dec;(83)24 9373-7.
63. Fresco J.R., Adams A., Ascione R., Henley D., Lindahl T.
Tertiary structure in transfer ribonucleic acids.
Cold Spring Harb Symp Quant Biol. 1966;(31) 527-37.
64. Frishman D., Heumann K., Lesk A., Mewes H.W.
Comprehensive, comprehensible, distributed and intelligent databases:current status.
Bioinformatics. 1998;(14)7 551-61.
65. Giegerich R., Haase D., Rehmsmeier M.
Prediction and visualization of structural switches in RNA.
Pac Symp Biocomput 1999;126-37.
66. Giegerich R.
A systematic approach to dynamic programming in bioinformatics.
Bioinformatics. 2000 Aug;(16)8 665-77.
67. Gotoh O.
Multiple sequence alignment: algorithms and applications.
Adv Biophys. 1999;(36) 159-206.
68. Grate L., Herbster M., Hughey R., Haussler D., Mian I.S., Noller H.
RNA modeling using Gibbs sampling and stochastic context free grammars.
Proc Int Conf Intell Syst Mol Biol. 1994;(2) 138-46.
69. Grate L.
Automatic RNA secondary structure determination with stochasticcontext-free grammars.
Proc Int Conf Intell Syst Mol Biol. 1995;(3) 136-44.
70. Gribskov M., McLachlan A.D., Eisenberg D.
Profile analysis: detection of distantly related proteins.
Proc Natl Acad Sci U S A. 1987 Jul;(13)84 4355-8.
71. Guex N., Diemand A., Peitsch M.C.
Protein modelling for all.
Trends Biochem Sci. 1999 Sep;(24)9 364-7.
72. Guigo R., Agarwal P., Abril J.F., Burset M., Fickett J.W.
An assessment of gene prediction accuracy in large DNA sequences
Genome Res. 2000 Oct;(10)10 1631-42.
73. Gultyaev A.P., van Batenburg F.H., Pleij C.W.
Dynamic competition between alternative structures in viroid RNAssimulated by an RNA folding algorithm.
J Mol Biol. 1998 Feb 13;(276)1 43-55.
74. Gultyaev A.P., van Batenburg F.H., Pleij C.W.
The computer simulation of RNA folding pathways using a genetic algorithm.
J Mol Biol. 1995 Jun 30;(250)1 37-51.
75. Gultyaev A.P.
The computer simulation of RNA folding involving pseudoknot formation.
Nucleic Acids Res. 1991 May 11;(19)9 2489-94.

76. Gupta S.K., Kececioglu J.D., Schaffer A.A.
Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment.
J Comput Biol. 1995 Fall;(2)3 459-72.
77. Hanson R.J., Sun J., Willis D.G., Marzluff W.F.
Efficient extraction and partial purification of thepolyribosome-associated stem-loop binding protein bound to the 3' end ofhistone mRNA.
Biochemistry. 1996 Feb 20;(35)7 2146-56.
78. Haq I., Jenkins T.C., Chowdhry B.Z., Ren J., Chaires J.B.
Parsing free energies of drug-DNA interactions.
Methods Enzymol. 2000;(323) 373-405.
79. Hawkins E.R., Chang S.H., Mattice W.L.
Kinetics of the renaturation of yeast tRNA3 leu.
Biopolymers. 1977 Jul;(16)7 1557-66.
80. Hecker R., Wang Z.M., Steger G., Riesner D.
Analysis of RNA structures by temperature-gradient gel electrophoresis:viroid replication and processing.
Gene. 1988 Dec 10;(72)1-2 59-74.
81. Hentze M.W., Kuhn L.C.
Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress.
Proc Natl Acad Sci U S A. 1996 Aug 6;(93)16 8175-82.
82. Hermann T., Patel D.J.
Stitching together RNA tertiary architectures.
J Mol Biol. 1999 Dec 10;(294)4 829-49.
83. Hertz G.Z., Hartzell G.W. .3.r.d., Stormo G.D.
Identification of consensus patterns in unaligned DNA sequences known to be functionally related.
Comput Appl Biosci. 1990 Apr;(6)2 81-92.
84. Hertz G.Z., Stormo G.D.
Identifying DNA and protein patterns with statistically significantalignments of multiple sequences.
Bioinformatics. 1999 Jul-Aug;(15)7-8 563-77.
85. Higgins D.G., Bleasby A.J., Fuchs R.
CLUSTAL V: improved software for multiple sequence alignment.
Comput Appl Biosci. 1992 Apr;(8)2 189-91.
86. Higgins D.G., Sharp P.M.
CLUSTAL: a package for performing multiple sequence alignment on amicrocomputer.
Gene. 1988 Dec 15;(73)1 237-44.
87. Hingamp P., van den Broek A.E., Stoesser G., Baker W.
The EMBL Nucleotide Sequence Database. Contributing and accessing data.
Mol Biotechnol. 1999 Oct;(12)3 255-67.
88. Hofmann K., Bucher P., Falquet L., Bairoch A.
The PROSITE database, its status in 1999.
Nucleic Acids Res. 1999 Jan 1;(27)1 215-9.
89. Jacobson A.B., Zuker M.
Structural analysis by energy dot plot of a large mRNA.
J Mol Biol. 1993 Sep 20;(233)2 261-9.
90. Jeanmougin F., Thompson J.D., Gouy M., Higgins D.G., Gibson T.J.
Multiple sequence alignment with Clustal X.
Trends Biochem Sci. 1998 Oct;(23)10 403-5.

91. Kable M.L., Seiwert S.D., Heidmann S., Stuart K.
RNA editing: a mechanism for gRNA-specified uridylate insertion into precursor mRNA
[published erratum appears in Science 1996Oct 4;274(5284):21]
Science. 1996 Aug 30;(273)5279 1189-95.
92. Karp P.D.
Database links are a foundation for interoperability.
Trends Biotechnol. 1996 Aug;(14)8 273-9.
93. Katz D.A., Theodorakis N.G., Cleveland D.W., Lindsten T., Thompson C.B.
AU-A, an RNA-binding activity distinct from hnRNP A1, is selective for AUUUA repeats and
shuttles between the nucleus and the cytoplasm.
Nucleic Acids Res. 1994 Jan 25;(22)2 . 238-46.
94. Khan J.Y., Rajakumar R.A., McKnight R.A., Devaskar U.P., Devaskar S.U.
Developmental regulation of genes mediating murine brain glucose uptake.
Am J Physiol. 1999 Mar;(276)3 Pt 2 R892-900.
95. King P.H., Levine T.D., Fremeau R.T., Keene J.D.
Mammalian homologs of Drosophila ELAV localized to a neuronal subset canbind in vitro to the
3' UTR of mRNA encoding the Id transcriptionalrepressor.
J Neurosci. 1994 Apr;(14)4 1943-52.
96. Kitano H.
Computational systems biology.
Nature. 2002 Nov 14;(420)6912 206-10.
97. Klemencic E., Todorovski L.
Analysis of free MEDLINE on the WWW.
Stud Health Technol Inform. 1999;(68) 553-6.
98. Kraemer E.T., Ferrin T.E.
Molecules to maps: tools for visualization and interaction in support of computational biology.
Bioinformatics. 1998;(14)9 764-71.
99. Kruger K., Grabowski P.J., Zaug A.J., Sands J., Gottschling D.E., Cech T.R.
Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA
Cell. 1982 Nov;(31)1 147-57.
100. Laferriere A., Gautheret D., Cedergren R.
An RNA pattern matching program with enhanced performance and portability.
Comput Appl Biosci. 1994 Apr;(10)2 211-2.
101. Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K.,
Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L.,
Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Mi
Initial sequencing and analysis of the human genome.
Nature. 2001 Feb 15;(409)6822 860-921.
102. Lawton J.R., Martinez F.A., Burks C.
Overview of the LiMB database.
Nucleic Acids Res. 1989 Aug 11;(17)15 5885-99.
103. Leclerc F., Cedergren R.
Modeling RNA-ligand interactions: the Rev-binding elementRNA-aminoglycoside complex.
J Med Chem. 1998 Jan 15;(41)2 175-82.
104. Lefebvre F.
An optimized parsing algorithm well suited to RNA folding.
Proc Int Conf Intell Syst Mol Biol. 1995;(3) 222-30.
105. Lengeler J.W.
Metabolic networks: a signal-oriented approach to cellular models
Biol Chem. 2000 Sep-Oct;(381)9-10 911-20.
106. Le S.Y., Chen J.H., Maizel J.V.
Prediction of alternative RNA secondary structures based on fluctuatingthermodynamic

parameters.
Nucleic Acids Res. 1993 May 11;(21)9 2173-8.
107. Lescure A., Gautheret D., Carbon P., Krol A.
Novel selenoproteins identified in silico and in vivo by using a conservedRNA structural motif.
J Biol Chem. 1999 Dec 31;(274)53 38147-54.
108. Lesnik E.A., Sampath R., Ecker D.J.
Rev response elements (RRE) in lentiviruses: an RNAMotif algorithm-basedstrategy for RRE prediction.
Med Res Rev. 2002 Nov;(22)6 617-36.
109. Lodmell J.S., Dahlberg A.E.
A conformational switch in Escherichia coli 16S ribosomal RNA duringdecoding of messenger RNA.
Science. 1997 Aug 29;(277)5330 1262-7.
110. Loss P., Schmitz M., Steger G., Riesner D.
Formation of a thermodynamically metastable structure containing hairpinII is critical for infectivity of potato spindle tuber viroid RNA.
EMBO J. 1991 Mar;(10)3 719-27.
111. Lowe T.M., Eddy S.R.
tRNAscan-SE: a program for improved detection of transfer RNA genes ingenomic sequence.
Nucleic Acids Res. 1997 Mar 1;(25)5 955-64.
112. Low S.C., Berry M.J.
Knowing when not to stop: selenocysteine incorporation in eukaryotes.
Trends Biochem Sci. 1996 Jun;(21)6 203-8.
113. Luck R., Steger G., Riesner D.
Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prionprotein.
J Mol Biol. 1996 May 24;(258)5 813-26.
114. Macauley J., Wang H., Goodman N.
A model system for studying the integration of molecular biologydatabases.
Bioinformatics. 1998;(14)7 575-82.
115. Macdonald P.M., Kerr K.
Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA.
Mol Cell Biol. 1998 Jul;(18)7 3788-95.
116. Macdonald P.M., Struhl G.
cis-acting sequences responsible for anterior localization of bicoid mRNAin Drosophila embryos.
Nature. 1988 Dec 8;(336)6199 595-8.
117. Macilwain C.
World leaders heap praise on human genome landmark [news]
Nature. 2000 Jun 29;(405)6790 983-4.
118. Macke T.J., Ecker D.J., Gutell R.R., Gautheret D., Case D.A., Sampath R.
RNAMotif, an RNA secondary structure definition and search algorithm.
Nucleic Acids Res. 2001 Nov 15;(29)22 4724-35.
119. Margot J.B., Williams D.L.
Estrogen induces the assembly of a multiprotein messenger ribonucleoprotein complex on the 3'-untranslated region of chicken apolipoprotein II mRNA.
J Biol Chem. 1996 Feb 23;(271)8 4452-60.
120. Martinez H.M.
An RNA folding rule.
Nucleic Acids Res. 1984 Jan 11;(12)1 Pt 1 323-34.
121. Martinez H.M.
Detecting pseudoknots and other local base-pairing structures in RNA sequences.
Methods Enzymol. 1990;(183) 306-17.

122. Mathews D.H., Sabina J., Zuker M., Turner D.H.
Expanded sequence dependence of thermodynamic parameters improvesprediction of RNA secondary structure.
J Mol Biol. 1999 May 21;(288)5 911-40.
123. Mathews D.H., Sabina J., Zuker M., Turner D.H.
Expanded sequence dependence of thermodynamic parameters improvesprediction of RNA secondary structure.
J Mol Biol. 1999 May 21;(288)5 911-40.
124. Mathews D.H., Turner D.H.
Experimentally derived nearest-neighbor parameters for the stability ofRNA three- and four-way multibranch loops.
Biochemistry. 2002 Jan 22;(41)3 869-80.
125. Mattaj I.W.
RNA recognition: a family matter?
Cell. 1993 Jun 4;(73)5 837-40.
126. Matzura O., Wennborg A.
RNAdraw: an integrated program for RNA secondary structure calculation andanalysis under 32-bit Microsoft Windows.
Comput Appl Biosci. 1996 Jun;(12)3 247-9.
127. McCarthy J.E., Kollmus H.
Cytoplasmic mRNA-protein interactions in eukaryotic gene expression.
Trends Biochem Sci. 1995 May;(20)5 191-7.
128. Mironov A.A., Lebedev V.F.
A kinetic model of RNA folding.
Biosystems. 1993;(30)1-3 49-56.
129. Monod J. Jacob F.
Genetic regulatory mechanisms in the synthesis of proteins
J Mol Biol. 1961;(3) 318-356.
130. Morgenstern B., Dress A., Werner T.
Multiple DNA and protein sequence alignment based on segment-to-segmentcomparison.
Proc Natl Acad Sci U S A. 1996 Oct 29;(93)22 12098-103.
131. Morgenstern B., Frech K., Dress A., Werner T.
DIALIGN: finding local similarities by multiple sequence alignment.
Bioinformatics. 1998;(14)3 290-4.
132. Morgenstern B.
DIALIGN 2: improvement of the segment-to-segment approach to multiplesequence alignment.
Bioinformatics. 1999 Mar;(15)3 211-8.
133. Nagai K.
RNA-Protein interactions
Curr Opin Struct Biol. 1992;(2) 131-137.
134. Nagel J.H., Gultyaev A.P., Gerdes K., Pleij C.W.
Metastable structures and refolding kinetics in hok mRNA of plasmid R1.
RNA. 1999 Nov;(5)11 1408-18.
135. Needleman S.B., Wunsch C.D.
A general method applicable to the search for similarities in the amino acid sequence of two proteins.
J Mol Biol. 1970 Mar;(48)3 443-53.
136. Nicholas K.B., Nicholas H.B.
GENEDOC: A tool for annotating multiple sequence alignments
Distributed by the author. ;() .

137. Nielsen D.A., Shapiro D.J.
Insights into hormonal control of messenger RNA stability.
Mol Endocrinol. 1990 Jul;(4)7 953-7.
138. Noller H.F.
Structure of ribosomal RNA.
Annu Rev Biochem. 1984;(53) 119-62.
139. Pan J., Thirumalai D., Woodson S.A.
Folding of RNA involves parallel pathways.
J Mol Biol. 1997 Oct 17;(273)1 7-13.
140. Pan T., Fang X., Sosnick T.
Pathway modulation, circular permutation and rapid RNA folding underkinetic control.
J Mol Biol. 1999 Feb 26;(286)3 721-31.
141. Pan W.
A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.
Bioinformatics. 2002 Apr;(18)4 546-54.
142. Pearson W.R.
Flexible sequence similarity searching with the FASTA3 program package.
Methods Mol Biol. 2000;(132) 185-219.
143. Pelletier J., Sonenberg N.
Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA.
Nature. 1988 Jul 28;(334)6180 320-5.
144. Perrotta A.T., Been M.D.
A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation.
J Mol Biol. 1998 Jun 5;(279)2 361-73.
145. Pongjaroenkit S., Jirajaroenrat K., Boonchauy C., Chanama U., Leetachewa S., Prapanthadara L., Ketterman A.J.
Genomic organization and putative promoters of highly conservedglutathione S-transferases originating by alternative splicing inAnopheles dirus.
Insect Biochem Mol Biol. 2001 Jan;(31)1 75-85.
146. Preiss T., Hentze M.W.
From factors to mechanisms: translation and translational control in eukaryotes.
Curr Opin Genet Dev. 1999 Oct;(9)5 515-21.
147. Putzer H., Gendron N., Grunberg-Manago M.
Co-ordinate expression of the two threonyl-tRNA synthetase genes inBacillus subtilis: control by transcriptional antitermination involving aconserved regulatory sequence.
EMBO J. 1992 Aug;(11)8 3117-27.
148. Quandt K., Frech K., Karas H., Wingender E., Werner T.
MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.
Nucleic Acids Res. 1995 Dec 11;(23)23 4878-84.
149. Rastinejad F., Blau H.M.
Genetic complementation reveals a novel regulatory role for 3'untranslated regions in growth and differentiation.
Cell. 1993 Mar 26;(72)6 903-17.
150. Rivas E., Eddy S.R.
A dynamic programming algorithm for RNA structure prediction includingpseudoknots.
J Mol Biol. 1999 Feb 5;(285)5 2053-68.

151. Rivas E., Eddy S.R.
The language of RNA: a formal grammar that includes pseudoknots.
Bioinformatics. 2000 Apr;(16)4 334-40.
152. Rodriguez-Tome P.
The BioCatalog.
Bioinformatics. 1998 Jun;(14)5 469-70.
153. Sakakibara Y., Brown M., Hughey R., Mian I.S., Sjolander K., Underwood R.C., Haussler D.
Stochastic context-free grammars for tRNA modeling.
Nucleic Acids Res. 1994 Nov 25;(22)23 5112-20.
154. SantaLucia J. .J.r., Turner D.H.
Measuring the thermodynamics of RNA secondary structure formation.
Biopolymers. 1997;(44)3 309-19.
155. SantaLucia J. .J.r.
A unified view of polymer, dumbbell, and oligonucleotide DNAnearest-neighbor thermodynamics.
Proc Natl Acad Sci U S A. 1998 Feb 17;(95)4 1460-5.
156. Schafer M., Kuhn R., Bosse F., Schafer U.
A conserved element in the leader mediates post-meiotic translation aswell as cytoplasmic polyadenylation of a Drosophila spermatocyte mRNA.
EMBO J. 1990 Dec;(9)13 4519-25.
157. Schaftenaar G., Cuelenaere K., Noordik J.H., Etzold T.
A Tcl-based SRS v. 4 interface.
Comput Appl Biosci. 1996 Apr;(12)2 151-5.
158. Schmitz M., Steger G.
Base-pair probability profiles of RNA secondary structures.
Comput Appl Biosci. 1992 Aug;(8)4 389-99.
159. Schultes E.A., Bartel D.P.
One sequence, two ribozymes: implications for the emergence of newribozyme folds.
Science. 2000 Jul 21;(289)5478 448-52.
160. Schuster P., Stadler P.F., Renner A.
RNA structures and folding: from conventional to new issues in structure predictions.
Curr Opin Struct Biol. 1997 Apr;(7)2 229-35.
161. Schuster P.
How to search for RNA structures. Theoretical concepts in evolutionary biotechnology.
J Biotechnol. 1995 Jul 31;(41)2-3 239-57.
162. Schuster S., Dandekar T., Fell D.A.
Detection of elementary flux modes in biochemical networks: a promisingtool for pathway analysis and metabolic engineering.
Trends Biotechnol. 1999 Feb;(17)2 53-60.
163. Searls D.B.
Linguistic approaches to biological sequences.
Comput Appl Biosci. 1997 Aug;(13)4 333-44.
164. Searls D.B.
The language of genes.
Nature. 2002 Nov 14;(420)6912 211-7.
165. Senger M., Flores T., Glatting K., Ernst P., Hotz-Wagenblatt A., Suhai S.
W2H: WWW interface to the GCG sequence analysis package.
Bioinformatics. 1998 Jun;(14)5 452-7.
166. Shapiro B.A., Wu J.C.
Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm.
Comput Appl Biosci. 1997 Aug;(13)4 459-71.

167. Shapiro B.A., Zhang K.Z.
Comparing multiple RNA secondary structures using tree comparisons.
Comput Appl Biosci. 1990 Oct;(6)4 309-18.
168. Shaw G., Kamen R.
A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation.
Cell. 1986 Aug 29;(46)5 659-67.
169. Smaglik P.
Researchers take a gamble on the human genome [news]
Nature. 2000 May 18;(405)6784 264.
170. Smith C.W., Valcarcel J.
Alternative pre-mRNA splicing: the logic of combinatorial control.
Trends Biochem Sci. 2000 Aug;(25)8 381-8.
171. Smith T.F., Waterman M.S.
Identification of common molecular subsequences.
J Mol Biol. 1981 Mar 25;(147)1 195-7.
172. Sonnhammer E.L., Eddy S.R., Durbin R.
Pfam: a comprehensive database of protein domain families based on seedalignments.
Proteins. 1997 Jul;(28)3 405-20.
173. Soukup G.A., Breaker R.R.
Engineering precision RNA molecular switches.
Proc Natl Acad Sci U S A. 1999 Mar 30;(96)7 3584-9.
174. Stebbins-Boaz B., Richter J.D.
Translational control during early development.
Crit Rev Eukaryot Gene Expr. 1997;(7)1-2 73-94.
175. Steller U., Kohls S., Muller B., Soller R., Muller R., Schlender J., Blohm D.H.
The RNA binding protein HuD: rat cDNA and analysis of the alternativespliced mRNA in neuronal differentiating cell lines P19 and PC12.
Brain Res Mol Brain Res. 1996 Jan;(35)1-2 285-96.
176. Stoesser G., Baker W., van d.e.n. .B.r.o.e.k. .A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Leinonen R., Lin Q., Lombard V., Lopez R., Redaschi N., Stoehr P., Tuli M.A., Tzouvara K., Vaughan R.
The EMBL Nucleotide Sequence Database.
Nucleic Acids Res. 2002 Jan 1;(30)1 21-6.
177. Stoesser G., Baker W., van Den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Lombard V., Lopez R., Parkinson H., Redaschi N., Sterk P., Stoehr P., Tuli M.A.
The EMBL nucleotide sequence database [In Process Citation]
Nucleic Acids Res. 2001 Jan 1;(29)1 17-21.
178. Sutcliffe J.G., Foye P.E., Erlander M.G., Hilbush B.S., Bodzin L.J., Durham J.T., Hasel K.W.
TOGA: an automated parsing technology for analyzing expression of nearly all genes.
Proc Natl Acad Sci U S A. 2000 Feb 29;(97)5 1976-81.
179. Tabaska J.E., Cary R.B., Gabow H.N., Stormo G.D.
An RNA folding method capable of identifying pseudoknots and base triples.
Bioinformatics. 1998;(14)8 691-9.
180. Tagle D.A., Koop B.F., Goodman M., Slightom J.L., Hess D.L., Jones R.T.
Embryonic epsilon and gamma globin genes of a prosimian primate (Galagocrassicaudatus).
Nucleotide and amino acid sequences, developmentalregulation and phylogenetic footprints.
J Mol Biol. 1988 Sep 20;(203)2 439-55.
181. The Arabidopsis Genome Initiative
Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.
Nature. 2000 Dec 14;(408)6814 .

182. The C. elegans Sequencing Consortium
Genome sequence of the nematode C. elegans: a platform for investigatingbiology. The C. elegans Sequencing Consortium [published errata appear inScience 1999 Jan 1;283(5398):35 and 1999 Mar 26;283(5410):2103 and 1999Sep 3;285(5433):1493]
Science. 1998 Dec 11;(282)5396 2012-8.
183. Theil E.C.
Targeting mRNA to regulate iron and oxygen metabolism.
Biochem Pharmacol. 2000 Jan 1;(59)1 87-93.
184. Theil E.C.
The iron responsive element (IRE) family of mRNA regulators. Regulation of iron transport and uptake compared in animals, plants, and microorganisms.
Met Ions Biol Syst. 1998;(35) 403-34.
185. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G.
The CLUSTAL_X windows interface: flexible strategies for multiple sequencealignment aided by quality analysis tools.4876-82CLUSTAL X is a new windows interface for the widely-used progressivemultiple sequence alignment program CLUSTAL W. The new system i
Nucleic Acids Res. 1997 Dec 15;(25)24 .
186. Tinoco I. .J.r., Bustamante C.
How RNA folds.
J Mol Biol. 1999 Oct 22;(293)2 271-81.
187. Tuerk C., Gold L.
Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.
Science. 1990 Aug 3;(249)4968 505-10.
188. van Helden J., Naim A., Mancuso R., Eldridge M., Wernisch L., Gilbert D., Wodak S.J.
Representing and analysing molecular and cellular function using the computer
Biol Chem. 2000 Sep-Oct;(381)9-10 921-35.
189. Vassella E., Den Abbeele .J.V., Butikofer P., Renggli C.K., Furger A., Brun R., Roditi I.
A major surface glycoprotein of trypanosoma brucei is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia.
Genes Dev. 2000 Mar 1;(14)5 615-26.
190. Verhoef K., Tijms M., Berkhout B.
Optimal Tat-mediated activation of the HIV-1 LTR promoter requires afull-length TAR RNA hairpin.
Nucleic Acids Res. 1997 Feb 1;(25)3 496-502.
191. von Ahsen U.
Translational fidelity: error-prone versus hyper-accurate ribosomes.
Chem Biol. 1998 Jan;(5)1 R3-6.
192. Wang J.H., Sun G.Y.
Ethanol Inhibits Cytokine-Induced iNOS and sPLA(2) in ImmortalizedAstrocytes: Evidence for Posttranscriptional Site of Ethanol Action.
J Biomed Sci. 2001 Jan;(8)1 126-133.
193. Wheeler D.L., Chappey C., Lash A.E., Leipe D.D., Madden T.L., Schuler G.D., Tatusova T.A., Rapp B.A.
Database resources of the National Center for Biotechnology Information.
Nucleic Acids Res. 2000 Jan 1;(28)1 10-4.
194. Wheeler D.L., Church D.M., Lash A.E., Leipe D.D., Madden T.L., Pontius J.U., Schuler G.D., Schriml L.M., Tatusova T.A., Wagner L., Rapp B.A.
Database resources of the national center for biotechnology information[In Process Citation]
Nucleic Acids Res. 2001 Jan 1;(29)1 11-6.

195. Wickens M.
Messenger RNA. Springtime in the desert.
Nature. 1993 May 27;(363)6427 305-6.
196. Willis A.E.
Translational control of growth factor and proto-oncogene expression.
Int J Biochem Cell Biol. 1999 Jan;(31)1 73-86.
197. Wilting R., Schorling S., Persson B.C., Bock A.
Selenoprotein synthesis in archaea: identification of an mRNA element ofMethanococcus jannaschii probably directing selenocysteine insertion.
J Mol Biol. 1997 Mar 7;(266)4 637-41.
198. Wingender E., Chen X., Hehl R., Karas H., Liebich I., Matys V., Meinhardt T., Pruss M., Reuter I., Schacherer F.
TRANSFAC: an integrated system for gene expression regulation.316-9
Nucleic Acids Res. 2000 Jan 1;(28)1 .
199. Wingender E.
ISB: Just another journal?
In Silico Biol.. 1998;(1)1 1-2.
200. Womble D.D.
GCG: The Wisconsin Package of sequence analysis programs.
Methods Mol Biol. 2000;(132) 3-22.
201. Wool I.G., Gluck A., Endo Y.
Ribotoxin recognition of ribosomal RNA and a proposal for the mechanism oftranslocation.
Trends Biochem Sci. 1992 Jul;(17)7 266-9.
202. Wuchty S., Fontana W., Hofacker I.L., Schuster P.
Complete suboptimal folding of RNA and the stability of secondarystructures.
Biopolymers. 1999 Feb;(49)2 145-65.
203. WuJu L., JiaJin W.
Prediction of RNA secondary structure based on helical regionsdistribution.
Bioinformatics. 1998;(14)8 700-6.
204. Zamora H., Luce R., Biebricher C.K.
Design of artificial short-chained RNA species that are replicated by Q beta replicase.
Biochemistry. 1995 Jan 31;(34)4 1261-6.
205. Zhang M.Q.
Large-scale gene expression data analysis: a new challenge tocomputational biologists [published erratum appears in Genome Res 1999Nov;9(11):1156]
Genome Res. 1999 Aug;(9)8 681-8.
206. Zuker M., Stiegler P.
Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.
Nucleic Acids Res. 1981 Jan 10;(9)1 133-48.
207. Zuker M.
Computer prediction of RNA structure.
Methods Enzymol. 1989;(180) 262-88.
208. Zuker M.
On finding all suboptimal foldings of an RNA molecule.
Science. 1989 Apr 7;(244)4900 48-52.
209. Zuker M.
Prediction of RNA secondary structure by energy minimization.
Methods Mol Biol. 1994;(25) 267-94.

7. Anlage zur Dissertation

7.1. Erklärung

Name:_____ Ort, Datum:_____

Anschrift_____

ERKLÄRUNG

gem. § 6 (5) Nr. 1 – 3 PromO

Ich erkläre, dass ich

1. die Arbeit ohne unerlaubte Hilfe angefertigt habe,
2. keine anderen, als die von mir angegebenen Quellen und Hilfsmittel benutzt habe und
3. die den benutzten Werken wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

_____

(Unterschrift)

7.2. Publikationsliste

1. **Bohnebeck U., Sälter W., Herzog O., Wischnewsky M., and Blohm D**.
„An Approach to mRNA Signalstructure Detection through Knowledge Discovery“
in D. Frishman and H. Mewes: *Proceedings of the German Conference on Bioinformatics*, (1997), 125-126.

2. **Boldt L., Gersdorf H., Niemeyer C.M., Holtkamp F., Bischoff R., Sälter W., Adler M., Kayser O., Wolf M., Jüptner W. and Blohm D.**
„A nanotiterplate - based DNA Array applied for cDNA - Detection of Hepatitis C Virus“
BIOSENSORS '98, 5th World Congress on Biosesnsors, Berlin, June 3-5, 1998.

3. **Bohnebeck U., Horvath T., Sälter W., Wrobel S.**
„Klassifikation von mRNA Signalstrukturen durch Relationales Lernen aus Beispielen“
in Wysotzki F., Geibel P. und Schadler K., *Beiträge zur GI-Fachtagung Maschinelles Lernen FGML'98*, Serie Technischer Bericht, Universität Berlin, 11, 1998, 114-118.

4. **Bohnebeck U., Sälter W., Horvath T., Wrobel S., Blohm D.**
„Measuring similarity of RNA structures by relational instance-based learning: A first step toward detecting RNA signal structures in silico“
in O. Zimmermann and D. Schomburg: *Proceedings of the German Conference on Bioinformatics*, 1998, 38-49.

5. **Gersdorf, H., Boldt, L., Niemeyer, C. M., Bischoff, R., Bohnebeck, U., Sälter, W., Wolf, M., Blohm, D. (1999)**

7.2. Publikationsliste

„Hepatitis-C-Nachweis mit Hilfe von DNA-Chips“

DECHEMA Statusseminar "Chiptechnologie" Frankfurt, 25.-26.01.1999.