Detektion funktioneller RNAs in Genomsequenzen

Dissertation
zur Erlangung des
mathematisch-naturwissenschaftlichen Doktorgrades
„Dr. rerum naturalium“
an der Georg-August-Universität Göttingen

vorgelegt von

Isabelle Heinemeyer

aus Prudnik (Polen)
Göttingen 2009

Referent: ..... Prof. Dr. Burkhard Morgenstern
Korreferent: ..... Prof. Dr. Stephan Waack
Tag der mündlichen Prüfung: ..... 15. April 2009

Abstract

Using approved tools a pipeline for fRNA prediction in complete genome sequences is developed. The program allows the prediction of new and until now unknown fRNAs as well as members of already known families. No expert knowledge about the used tools is necessary.

The fRNA prediction method is based on two pillars. The first pillar is a comparative approach which first identifies sequence similarities among related genomes and aligns them. In a second step the stability and conservation of the structures is evaluated. This approach allowes the prediction of new fRNAs since no information about known fRNA families is used. The second pillar is based on current knowledge about fRNA families. This knowledge is used to find homologies to known fRNA families. The predicted affiliation of a candidate to a specific family gives a hint for its function.

The pipeline that was developed in the course of this thesis was applied to several bacterial and archaeal genomes to predict the occurrence of fRNA coding genes and regulatory elements. The Pipeline is not only able to analyze complete genomes but can also be used on a selection of single sequences. This feature was used to predict structured regulatory elements in the mRNA of selected genes in *Saccharomyces cerevisiae* and related species.

Zusammenfassung

Unter Verwendung bewährter Werkzeuge wird eine Pipeline zur fRNA-Vorhersage in kompletten Genomsequenzen entwickelt. Das Programm ermöglicht sowohl die Vorhersage neuer, bisher unbekannter fRNAs, als auch die Vorhersage der Mitglieder bereits bekannter Familien. Hierbei ist kein Expertenwissen zu den einzelnen verwendeten Werkzeugen erforderlich.

Die fRNA-Vorhersage basiert auf zwei Säulen. Die erste Säule ist ein komparativer Ansatz, welcher zuerst Sequenzähnlichkeiten zwischen verwandten Genomen aufspürt und aligniert. Die Alignments werden dann auf das Vorkommen stabiler und konservierter Strukturen untersucht. Dieser Ansatz eignet sich vor allem zur Vorhersage neuer fRNAs, da keine Informationen über bekannte fRNAs benötigt werden. Die zweite Säule basiert auf aktuellem Wissen über fRNA-Familien. Dieses Wissen wird genutzt, um gezielt nach Homologien zu bekannten fRNA-Familien zu suchen. Die vorhergesagte Familienzugehörigkeit eines fRNA-Kandidaten gibt gleichzeitig einen Hinweis auf seine Funktion.

Die im Rahmen dieser Arbeit entwickelte Pipeline wurde auf mehrere Bakterien- und Archaeengenome angewandt, mit dem Ziel das Vorkommen von fRNA-kodierenden Genen und regulatorischen Elementen vorherzusagen. Die Pipeline ist nicht nur in der Lage komplette Genome zu untersuchen, sondern kann ebenfalls eine Auswahl einzelner Sequenzen behandeln. Sie wurde daher für die Suche nach möglichen regulatorischen Strukturelementen in der mRNA ausgewählter Gene in *Saccharomyces cerevisiae* und verwandten Spezies genutzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Symbole und Abkürzungen

$\mathcal{T}_A$	RNAz-Testmenge: Menge aller Homologiecluster zu einer Anfragesequenz A, die mit RNAz untersucht werden sollen, 40
$\mathcal{B}_A$	Menge aller BlastN-Treffer zu einer Anfragesequenz A, 37
$\overline{E}$	Mittelwert über die individuellen MFE aller Sequenzen in einem Alignment, 25
A	Anfragesequenz in einer BlastN-Untersuchung, 37
E_A	Konsensus-MFE eines Alignments, 25
`Rfam`	**R**NA-**Fam**ilien-Datenbank, 19
`SGD`	**S**accharomyces **G**enome **D**atabase, 96
bp	Basenpaar: Längenmaß einer Sequenz, 15
DNA	**D**eoxyribo**n**ucleic **A**cid, 1
DP	**D**ynamische **P**rogrammierung, 11
eIGR	**e**rweiterte **i**nter**g**enische **R**egion, 35
ESS	Datensatz: ***E****. coli*, ***S****. flexneri*, ***S****. enterica*, 55
EYK	Datensatz: ***E****. coli*, ***Y****. pestis*, ***K****. pneumoniae*, 55
fRNA	**f**unktionelle **RNA**, 1
IGR	**i**nter**g**enische **R**egion, 1
Intron	**Int**ervening **R**egi**on**, 1

IRES	**I**nternal **R**ibosomal **E**ntry **S**ite, 95
Mbp	Megabasenpaare: 1.000.000 bp, 55
MFE	**m**inimale **f**reie **E**nergie, 13
ORF	**O**pen **R**eading **F**rame, 2
Profile-HMM	**P**rofile **H**idden **M**arkov **M**odel, 20
RNA	**R**ibo**n**ucleic **A**cid, 1
SCFG	**S**tochastic **C**ontext-**F**ree **G**rammar, 20
SCI	**S**tructure **C**onservation **I**ndex: Mas für Strukturkonservierung, 25
SVM	**S**upport **V**ector **M**achine, 26
UTR	**Un**translated **R**egion, 4

Danksagung

Ich danke meinem Betreuer Herrn Prof. Morgenstern, dass er es mir ermöglicht hat an seinem Lehrstuhl zu promovieren. Er gab mir die Freiheit, meine eigenen Ideen zu verwirklichen. Herrn Prof. Waack danke ich für die Begleitung meiner Arbeit als Korreferent. Mein besonderer Dank gilt Herrn Dr. Liesegang, der meine Neugier für das Thema der fRNA-Vorhersage geweckt hat und mich damit zu dieser Arbeit inspiriert hat. Sein Interesse an meinen Ideen und den erzielten Ergebnissen hat mich stets aufs neue motiviert. Ich möchte ebenfalls Herrn Dr. Valerius und Nicole Rachfall für die angenehme und erfolgreiche Zusammenarbeit im Projekt *S. cerevisiae* danken.

Meinen lieben Kolleginnen und Kollegen danke ich für die nette Atmosphäre. Ohne sie hätte die Arbeit nur halb so viel Spaß gemacht. Insbesondere danke ich Anne-Kathrin für das intensive Korrekturlesen und die wertvollen Tipps beim Aufschreiben dieser Arbeit, sowie die ausgleichende Ablenkung mit Kaffee und Keksen. Sie war die beste Zimmerkollegin, die man sich vorstellen kann.

Ein herzlicher Dank geht ebenfalls an meine Familie und Freunde, die meine Arbeit immer mit Interesse verfolgt haben. Ganz besonders danke ich meiner Mutter. Sie hat mich in allen meinen Entscheidungen unterstützt und immer an mich geglaubt. Und schließlich bleibt mir nur noch meinem Mann Eric zu danken, meinem Kiesel in der Brandung.

Einleitung

Motivation

Lange Zeit wurde die Rolle der *Ribonukleinsäure* (RNA von *Ribonucleic Acid*) in der Regulation der Stoffwechselvorgänge in der Zelle unterschätzt. Dabei kann die RNA nicht nur in Form der mRNA die genetische Information übertragen. RNAs, die eine andere Funktion ausüben als die mRNA, werden unter dem Begriff der *funktionellen RNA* (fRNA) zusammengefasst. Dazu zählen sowohl eigenständige RNA-Moleküle, die wie Proteine durch ein eigenes Gen kodiert werden, als auch regulatorische Strukturelemente, die als Teilsequenzen der sogenannten *messenger RNA* (mRNA), eine für ihre Funktion spezifische Struktur ausbilden. Das letzte Jahrzehnt brachte durch neue experimentelle und bioinformatische Methoden eine Vielzahl bis dahin unbekannter fRNAs zu Tage. Wir beginnen erst eine Vorstellung davon zu entwickeln, welche vielfältigen und komplexen Aufgaben RNA-Moleküle übernehmen können, wie z. B. Regulation der Translation und Transkription, Katalyse chemischer Reaktionen und Transport. Dabei kann bis heute nicht allen neu entdeckten fRNAs eine Funktion zugeordnet werden.

Die Zahl neuer, bisher unbekannter fRNAs steigt kontinuierlich an. Es ist schwer abzuschätzen, wieviele fRNAs es insgesamt in den einzelnen Organismen gibt, und welche Funktionen sie übernehmen. Während regulatorische Strukturelemente vorwiegend Teil einer mRNA sind, werden eigenständige fRNA-Moleküle vor allem in *intergenischen Regionen* (IGR), aber auch als Teil von sogenannten ***int****ervening* ***regi****o****ns*** (Introns) kodiert. Intergenische Regionen und Introns bilden die nicht-Protein-kodierenden Regionen der *Desoxyribonukleinsäure* (DNA von *Deoxyribonucleic Acid*).

Mit zunehmender Komplexität des Organismus ist ein anwachsen der nicht-Protein-kodierenden Regionen in der DNA zu bemerken. Prokaryoten haben ein kom-

pakt gepacktes Genom. Weniger als 25% der DNA ist nicht-Protein-kodierend. Bei einfachen Eukaryoten sind es zwischen 25 und 50%, bei Pflanzen und Tieren bereits über 50% und beim Menschen sogar 98,5% [71]. Früher wurde vermutet, dass es sich bei der nicht-Protein-kodierenden DNA vorwiegend um „Abfall“ handelt, der keine weitere Bedeutung hat. Je mehr fRNAs in diesen Bereichen gefunden werden, umso mehr wird diese Vermutung angezweifelt [26, 80, 72, 70]. Nach wie vor ist jedoch nicht bekannt, wieviel der nicht-Protein-kodierenden Region tatsächlich für fRNA kodiert.

Die systematische Suche nach fRNA-kodierenden Regionen in vollständigen Genomsequenzen gewinnt, mit der in den letzten Jahren rasant zunehmenden Anzahl sequenzierter Genome, an Bedeutung. Neben dem menschlichen Genom wurden vor allem die Genome verschiedenster Archaeen und Bakterien sequenziert. Ihr Genom ist im Vergleich zu eukaryotischen Genomen deutlich kleiner und damit schneller zu sequenzieren. Die Sequenzierung bakterieller Genome wird aber vor allem durch ihre Rolle in der Medizin und Industrie vorangetrieben. Einige Bakterien sind z. B. in der Lage, Antibiotika zu produzieren [16], andere werden zur Herstellung von Waschmittelenzymen [92] oder Biodünger [17] verwendet. Um solche Organismen effektiv einsetzen zu können, ist eine detaillierte Kenntnis der entsprechenden Stoffwechselvorgänge und der daran beteiligten Regulatoren, wie fRNAs, notwendig.

Die Vorhersage neuer fRNA-kodierender Gene unterscheidet sich deutlich von der Vorhersage von Protein-kodierenden Genen. Die für fRNA-kodierenden Gene weisen, anders als Protein-kodierende Gene, keine gemeinsamen, statistisch signifikanten Signale in der Sequenz, wie z. B. die sogenannten ***o**pen **r**eading **f**rames* (*ORFs*), auf. Ein Ansatz zur fRNA-Vorhersage muss daher andere Informationsquellen verwenden. Eine mögliche Quelle ist die Struktur einer RNA. Die Funktion vieler fRNA-kodierender Gene und insbesondere regulatorischer Strukturelemente hängt vor allem von ihrer Struktur und nicht nur von ihrer Sequenz ab. Es ist daher zu erwarten, dass die Struktur stärker konserviert ist als die Sequenz und eine besondere Stabilität aufweist. Es gibt unterschiedliche Werkzeuge [85, 25, 22, 102], mit deren Hilfe die Strukturkonservierung und teilweise auch die Strukturstabilität mehrerer verwandter Sequenzen beurteilt werden kann. Die Untersuchung einer Genomsequenz erfordert jedoch viel Vorarbeit, da zuerst Sequenzhomologien identifiziert und in Alignments zusammengefasst werden müssen. Andere Werkzeuge können zwar mit einer kompletten Genomsequenz umgehen, sind aber meist auf einen fRNA-Typ

spezialisiert [64, 59, 60], oder benötigen die gemeinsamen Sequenz- und Strukturinformationen einer fRNA-Familie, um nach verwandten Sequenzen suchen zu können [30, 31]. Auf Grund ihrer Spezialisierung sind diese Programme nicht in der Lage, neue fRNAs zu finden.

Problemstellung und Lösungsansatz

Ein Ziel dieser Arbeit ist die Entwicklung und Anwendung einer Strategie, welche die Suche nach neuen fRNA-kodierenden Regionen in kompletten Genomsequenzen ermöglicht.

Unter Verwendung bewährter Werkzeuge, wie BlastN [4], ClustalW [98], RNAz [102] und INFERNAL [31], wurde ein Programm zur fRNA-Vorhersage in kompletten Genomsequenzen entwickelt. Das Programm ist in der Lage, sowohl neue, bisher unbekannte fRNAs, als auch Mitglieder bereits bekannter Familien aufzuspüren. Die Idee dazu basiert auf zwei Säulen.

- Die erste Säule ist ein komparativer Ansatz. Dabei werden zuerst Sequenzähnlichkeiten zwischen verwandten Genomen aufgespürt. Auch wenn die Struktur einer fRNA normalerweise stärker konserviert ist als die Sequenz, nehmen wir an, dass die Ähnlichkeit zwischen fRNA-kodierenden Sequenzen in verwandten Spezies ausreichend hoch ist, um eine sequenzbasierte Vorauswahl der Kandidaten treffen zu können. Diese Kandidaten werden zusammengefasst und aligniert. Die Alignments werden dann auf das Vorkommen stabiler und konservierter Strukturen untersucht. Dieser Ansatz eignet sich vor allem zur Vorhersage neuer fRNAs, da keine Informationen über bekannte fRNAs benötigt werden.

- Die zweite Säule basiert auf aktuellem Wissen über fRNA-Familien. Die Zahl und Größe von fRNA-Datenbanken wächst stetig an [45, 63, 78]. Wir nutzen dieses Wissen, um gezielt nach Ähnlichkeiten zu bekannten fRNA-Familien zu suchen. Mit der Familienzugehörigkeit eines fRNA-Kandidaten erhalten wir gleichzeitig einen möglichen Hinweis auf seine Funktion.

Neben der Entwicklung des Programms zur fRNA-Vorhersage, ist auch dessen Anwendung Teil dieser Arbeit. So haben wir mehrere Bakterien- und Archaeengenome auf das Vorkommen fRNA-kodierender Gene und regulatorischer Elemente

untersucht. Die erzielten Ergebnisse werden in dieser Arbeit und den folgenden Publikationen vorgestellt:

- X. H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W. R. Hess, O. Reva, H. Junge, B. Voigt, P. R. Jungblut, J. Vater, R. Süssmuth, H. Liesegang, A. Strittmatter, G. Gottschalk und R. Borriss. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium Bacillus amyloliquefaciens FZB42. *Nature Biotechnology*, 25: 1007-1014, September 2007.

- X. H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W. R. Hess, O. Reva, H. Junge, B. Voigt, P. R. Jungblut, J. Vater, R. Süssmuth, H. Liesegang, A. Strittmatter, G. Gottschalk, und R. Borriss. Genomanalyse eines phytostimulatorischen Bacillus-Stammes. *GenomXPress*, 3.07: 11-13, September 2007.

- C. Schmeisser, H. Liesegang, D. Krysciak, N. Bakkou, A. Le Quéré, A. Wollherr, I. Heinemeyer, B. Morgenstern, A. Pommerening-Röser, M. Flores, R. Palacios, S. Brenner, G. Gottschalk, R. A. Schmitz, W. J. Broughton, X. Perret, A. W. Strittmatter und W. R. Streit. Rhizobium sp. NGR234 possesses a remarkable number of secretion systems. Eingereicht bei *Applied and Environmental Microbiology* im März 2009.

Ein weiteres Projekt, das in diese Arbeit eingeht, ist die Suche nach möglichen regulatorischen Strukturelementen in der mRNA ausgewählter Gene in *Saccharomyces cerevisiae*, das auch als Bäckerhefe bekannt ist. Die Gene sind bei einer Proteomanalyse unter Aminosäuremangel durch eine erhöhte Translationsrate aufgefallen. Motiviert durch die Existenz regulatorischer Elemente in den sogenannten ***Un****traslated* ***R****egions* (UTRs), waren wir in diesem Fall nicht an einer Übersicht aller fRNAs im gesamten Genom interessiert, sondern nur an regulatorischen Strukturelementen in den UTR-Sequenzen, welche die Translation beeinflussen könnten. Unter Zuhilfenahme geeigneter Vergleichsdaten, haben wir unter anderem das bereits vorgestellte Programm zur fRNA-Vorhersage eingesetzt. Die erzielten Ergebnisse werden in dieser Arbeit vorgestellt. Parallel zu dem hier beschriebenen Ansatz werden experimentelle Untersuchungen durchgeführt. Es ist geplant, die Ergebnisse

beider Untersuchungen miteinander zu verglichen und in einem renommierten Journal zu veröffentlichen. Das methodische Vorgehen wird bereits im folgenden Artikel skizziert:

- N. Rachfall, I. Heinemeyer, O. Valerius. 5'-TRUE: Die wahre Translation? *BIOspektrum*. 02/2009. Im Druck.

Aufbau der Arbeit

In Kapitel 1 werden Grundlagen, die zum Verständnis dieser Arbeit notwendig sind, und Werkzeuge, die im Rahmen dieser Arbeit verwendet wurden, vorgestellt.

In Kapitel 2 wird die im Rahmen dieser Arbeit entwickelten Anwendung zur fRNA-Vorhersage in Genomsequenzen beschrieben.

In Kapitel 3 wird das Programm zur fRNA-Vorhersage auf den Musterorganismus *Escherichia coli*, Stamm K-12, angewandt und die erzielten Ergebnisse mit den aktuell bekannten fRNAs in diesem Organismus verglichen.

In Kapitel 4 werden die Ergebnisse der fRNA-Vorhersage auf neuen Datensätzen vorgestellt. Im Fokus der Untersuchungen standen die folgenden Organismen:

- *Bacillus amyloliquefaciens* FZB42,
- *Methanosarcina mazei* Go1,
- *Streptomyces coelicolor* A3(2),
- *Oceanobacillus iheyensis* HTE831,
- *Pyrococcus furiosus* DSM 3638.
- *Rhizobium* sp. NGR234

In Kapitel 5 wird die Suche nach regulatorischen Strukturelementen für ausgewählte Gene in *Saccharomyces cerevisiae* S288C vorgestellt.

In Kapitel 6 werden die wichtigsten Erkenntnisse dieser Arbeit zusammengefasst.

In Anhang A befinden sich zusätzliche Ergebnisse, die nur in zusammengefasster Form in dieser Arbeit auftreten.

1 Grundlagen

1.1 Aufbau der RNA

Eine RNA ist eine Kette von Nukleotiden. Jedes Nukleotid besteht aus einer von vier Basen, dem Zucker Ribose und einem Phosphatrest. Die DNA ist äquivalent aufgebaut, doch an Stelle der Ribose steht der Zucker Desoxyribose. Die vier Basen der RNA sind: Adenin (A), Guanin (G), Cytosin (C) und Uracil (U). In der DNA kommt an Stelle der Base Uracil die Base Thymin (T) vor. Die Nukleotide sind über Zucker-Phosphat-Bindungen aneinander gekoppelt. Die Abfolge der Nukleotide eines RNA-Moleküls wird als *Primärstruktur* bezeichnet.

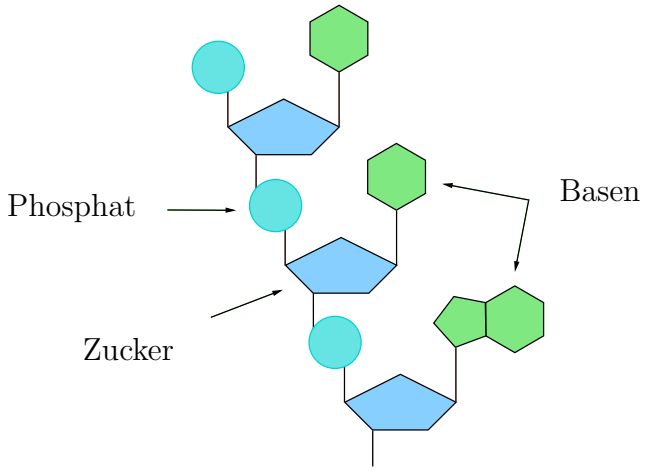


Abb. 1.1: Schematischer Aufbau einer RNA-Kette (Primärstruktur).

Zusätzlich zu den Zucker-Phosphat-Bindungen der Nukleotide können Basen untereinander Wasserstoffbrücken ausbilden. Theoretisch kann jede Base mit jeder anderen eine solche Bindung eingehen, jedoch nur komplementäre Basenpaare lassen sich problemlos aufeinander stapeln, wodurch regelmäßige Helices entstehen, die der Struktur Stabilität geben und daher bevorzugt gebildet werden. Die wichtigsten

komplementären Paarungen sind die Watson-Crick-Paarungen: G-C und A-U, sowie die Wobble-Paarung: G-U (Abbildung 1.2).

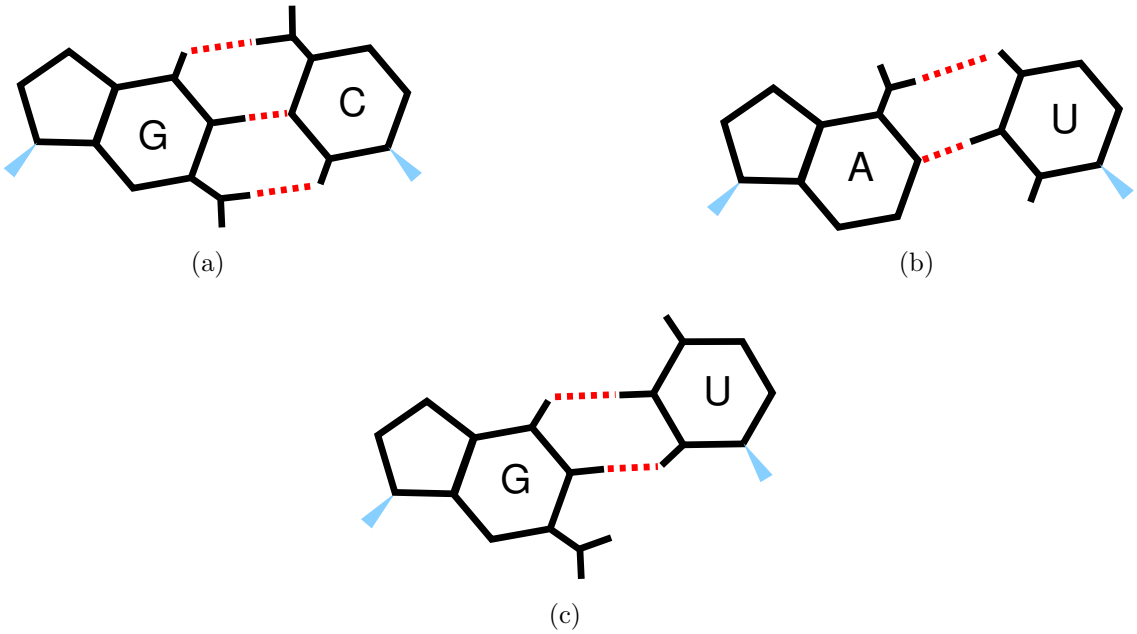


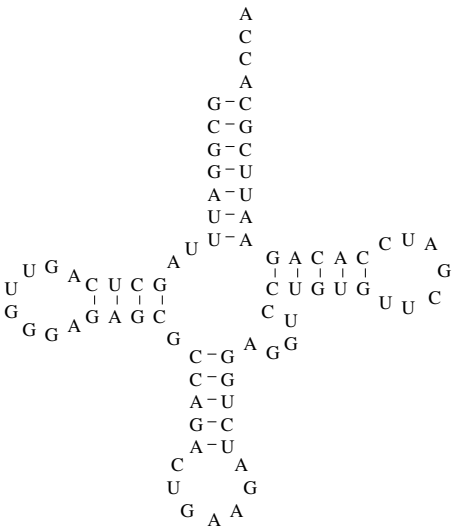
Abb. 1.2: Komplementäre Basenpaarungen: (a) Guanin mit Cytosin, (b) Adenin mit Uracil und (c) Guanin mit Uracil. Wasserstoffbrückenbindungen zwischen komplementären Basen werden mit einer gestrichelten roten Linie dargestellt.

Im Gegensatz zur doppelsträngigen DNA ist die RNA meistens einzelsträngig und kann durch Interaktionen mit sich selbst komplexe Strukturen ausbilden. Der erste Schritt zu einer drei-dimensionalen Struktur ist die Paarung komplementärer Basen derselben RNA. Diese Paarungen werden durch die sogenannte *Sekundärstruktur* beschrieben. Eine RNA-Sequenz ist meistens nicht über die gesamte Länge zu sich selbst komplementär, so dass gepaarte Regionen (*Stamm*) durch ungepaarte Regionen, welche in Form von *Schleifen* oder *Ausbuchtungen* auftreten können, unterbrochen werden. Die dadurch entstehenden Sekundärstrukturelemente werden in Abbildung 1.5 dargestellt. Die räumliche Anordnung der Basen einer RNA wird schließlich *Tertiärstruktur* genannt. Die Abbildung 1.3 zeigt eine mögliche Darstellung der drei Strukturzustände am klassischen Beispiel einer transfer-RNA (tRNA).

Die Strukturbildung, auch *RNA-Faltung* genannt, kann als ein hierarchischer Prozess verstanden werden. Es bilden sich zuerst die Sekundärstrukturelemente aus und im Anschluss daran entsteht die Tertiärstruktur, ohne die Sekundärstruktur

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

(a)



(b)

(c)

Abb. 1.3: Mögliche Darstellungen der (a) Primär-, (b) Sekundär- und (c) Tertiärstruktur am Beispiel von *Yeast* tRNA^{Phe}. (Quelle: [52], PDB Eintrag 1EVV [9]; 3-D Bild: Visualisierung mit Jmol [1])

stark zu verändern. Die Stabilität der drei-dimensionalen Struktur resultiert vor allem aus der Stabilität der Sekundärstruktur, d. h. der Wasserstoffbrückenbindungen zwischen komplementären Basen und den Wechselwirkungen zwischen benachbarten Basenpaaren. Daher kann die Sekundärstruktur, die mit bioinformatischen Mitteln einfacher vorherzusagen ist als die Tertiärstruktur, für den Vergleich von RNAs untereinander und die Untersuchung ihrer Funktion, herangezogen werden.

1.2 Sekundärstruktur-Vorhersage

Die RNA-Sekundärstruktur spielt eine entscheidende Rolle bei der Vorhersage funktioneller RNAs und ist grundlegend für das Verständnis der in dieser Arbeit verwendeten Methoden. Daher stellen wir in diesem Abschnitt Ansätze zur Vorhersage der RNA-Sekundärstruktur vor, die direkt oder indirekt in dieser Arbeit verwendet werden.

1.2.1 Definition der Sekundärstruktur

Es sei R eine RNA-Sequenz der Länge N, d. h.

$$R = r_1 \dots r_N$$

mit $r_i \in \{\text{A,G,C,U}\}$ für $i = 1, \dots, N$ und

$$R_{k,l} = r_k \dots r_l \quad \text{für} \quad 1 \leq k \leq l \leq N$$

die Teilsequenz vom k-ten bis zum l-ten Nukleotid in R. Die Nummerierung der Nukleotide erfolgt vom $5'$- zum $3'$-Ende. Man beachte, dass nur die folgenden Basenpaare erlaubt sind: A-C, G-C und G-U. Die Paarung zweier Basen r_i und r_j aus R wird durch

$$r_i : r_j \quad \text{oder einfach} \quad i : j \quad \text{für} \quad 1 \leq i < j \leq N$$

beschrieben. Eine Sekundärstruktur S der RNA-Sequenz R ist eine Menge von Basenpaaren in R. Für die Basenpaare gelten dabei folgende Bedingungen:

1. Jedes Basenpaar $i : j$ wird durch mindestens drei ungepaarte Basen getrennt,

d. h. $|j - i| \geq 4$.

2. Zwei beliebige Basenpaare $i : j$ und $k : l$ sind entweder identisch, d. h. $i = k$ und $j = l$, oder es gilt, dass sowohl $i \neq k$, als auch $j \neq l$ ist.

3. Pseudoknoten sind untersagt, d. h. für zwei Basenpaare $i : j$ und $k : l$, für die $i < k$ ist, muss entweder $i < j < k < l$ oder $i < k < l < j$ gelten.

Die erste Bedingung garantiert einen physikalisch sinnvollen Abstand zwischen zwei miteinander paarenden Basen. Die zweite Bedingung verbietet das Vorkommen von Bindungen zwischen drei oder mehr Basen. Eine solche Bindung kann auftreten, wird aber der Tertiärstruktur zugeordnet. Die dritte Bedingung verbietet das Vorkommen von sogenannten *Pseudoknoten*, auch wenn diese Strukturelemente in einigen RNA-Strukturen auftreten. Sie erschweren jedoch die Vorhersage der Sekundärstruktur erheblich und werden daher von den meisten Algorithmen nicht berücksichtigt. Wir schließen uns dieser Vorgehensweise an. In Abbildung 1.4 wird das mögliche Verhältnis von zwei Basenpaaren zueinander dargestellt.

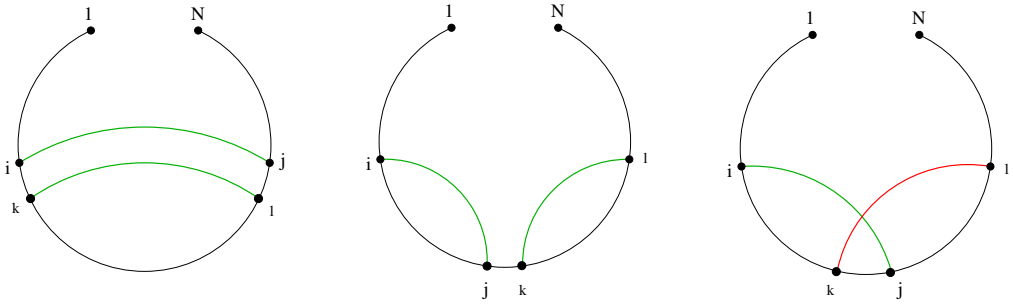


Abb. 1.4: Dargestellt ist das Verhältnis von zwei Basenpaaren in einer Sekundärstruktur. Die ersten beiden Fälle sind konform mit der Definition der Sekundärstruktur. Zwei unterschiedliche Basenpaare dürfen sowohl ineinander verschachtelt sein, wie in der ersten Abbildung, als auch hintereinander auftreten, wie in der zweiten Abbildung. Im dritten Fall überkreuzen sich die Basenpaarbindungen, was zu einem Pseudoknoten führt und in dieser Arbeit ausgeschlossen wurde.

1.2.2 Struktur-Vorhersage durch Maximierung der Basenpaaranzahl

Die ersten Ansätze zur Vorhersage der Sekundärstruktur basierten auf der Maximierung der Anzahl der Basenpaarbindungen. Dieses Problem wurde zuerst 1978 von

Nussinov *et al.* [76] gelöst. Die Berechnung erfolgt rekursiv. Es werden zuerst die Lösungen der kleinsten Teilprobleme bestimmt und daraus die Lösung des jeweils nächstgrößeren Problems zusammengesetzt. So wird immer weiter verfahren, bis die Lösung des Gesamtproblems bekannt ist. Ein solcher Lösungsansatz wird auch *dynamische Programmierung* genannt. Man spricht daher von einem *DP-Algorithmus*.

Die Idee des Algorithmus von Nussinov basiert darauf, die „optimale“ Sekundärstruktur als die Sekundärstruktur mit der maximalen Anzahl an gültigen Basenpaarungen (siehe Abschnitt 1.2.1) zu bestimmen. Dabei gibt es nur vier Möglichkeiten, die optimale Struktur für eine gegebene Teilsequenz $R_{i,j}$ von R aus den optimalen Strukturen ihrer Teilsequenzen zu bestimmen. Die vier Möglichkeiten werden verglichen und diejenige ausgewählt, welche die maximale Anzahl an Basenpaaren für $R_{i,j}$ liefert:

(i) Nehme die optimale Struktur für die Teilsequenz $R_{i+1,j-1}$ und füge das Basenpaar $i : j$ hinzu, falls die Basen r_i und r_j komplementär sind, sonst bleiben beide Basen ungepaart.

(ii) Nehme die optimale Struktur für die Teilsequenz $R_{i,j-1}$ und die Base j bleibt ungepaart.

(iii) Nehme die optimale Struktur für die Teilsequenz $R_{i+1,j}$ und die Base i bleibt ungepaart.

(iv) Für $i < k < j$ nehme die optimalen Strukturen der Teilsequenzen $R_{i,k}$ und $R_{k+1,j}$, für die die Summe paarender Basen maximal ist.

Formal wird für jede Teilsequenz $R_{i,j}$ einer RNA-Sequenz R der Länge N die maximale Anzahl paarender Basen $M(i,j)$ bestimmt. Für $|i-j| < 4$ ist $M(i,j) = 0$, da der Mindestabstand für zwei paarende Basen sonst nicht eingehalten werden kann. Für $|i-j| \geq 4$ gilt:

$$M(i,j) = \max \begin{cases} M(i+1, j-1) + \delta_{ij}, \\ M(i, j-1), \\ M(i+1, j), \\ \max\limits_{i<k<j} [M(i,k) + M(k+1,j)] \end{cases}$$

wobei $\delta_{ij} = 1$ ist, falls die Basen an den Positionen i und j miteinander paaren können und $\delta_{ij} = 0$ falls nicht. Die Schritte werden solange wiederholt bis $M(1, N)$ bestimmt wurde. Sind alle Zwischenergebnisse bekannt, kann die Struktur mit der maximalen Anzahl an Basenpaaren für die gesamte Sequenz durch das sogenannte *Backtracking* daraus rekonstruiert werden. Die Zeitkomplexität des Algorithmus beträgt $O(N^3)$ und die Speicherkomplexität $O(N^2)$ [76].

Dieser Algorithmus berücksichtigt keine energetischen Eigenschaften der RNA-Moleküle. Er kann aber entsprechend angepasst werden. Dabei wird jeder Basenpaarbindung ein bestimmtes Gewicht, z. B. die negative Paarungsenergie, zugeordnet. Anstatt dann die Anzahl der Basenpaare für eine Sequenz zu maximieren, wird die Gesamtenergie der paarenden Basen minimiert. Sind die Gewichte für alle Teilprobleme bestimmt, kann die Struktur mit der minimalen Paarungsenergie wiederum durch Backtracking rekonstruiert werden.

1.2.3 Struktur-Vorhersage durch Minimierung der freien Energie

Die Grundlage für die Vorhersage thermodynamisch optimaler Sekundärstrukturen, d. h. Strukturen mit *minimaler freier Energie* (MFE), legten Zuker und Stiegler in ihrem 1981 veröffentlichten Artikel [113]. Anders als beim Algorithmus von Nussinov wird beim Algorithmus von Zuker und Stiegler die freie Energie keinen Bindungen zugeordnet, sondern den durch diese Bindungen gebildeten Strukturelementen (Schleifen und Stapeln). Diese können am besten beschrieben werden, wenn die RNA-Sekundärstruktur als ein Graph aufgefasst wird. Nukleotide bilden dabei die Knoten im Graphen und Bindungen zwischen Nukleotiden entsprechen Kanten. Da wir zwei Typen von Bindungen unterscheiden, unterscheiden wir auch zwei Typen von Kanten. *Externe Kanten* entsprechen den Zucker-Phosphat-Bindungen zwischen Nukleotiden, wohingegen *interne Kanten* den Bindungen zwischen komplementären Basen entsprechen.

Ein Strukturelement, welches von höchstens einer internen Kante eingeschlossen ist, wird als

- *Haarnadel-Schleife*, siehe Abbildung 1.5(a).

bezeichnet. Elemente mit zwei internen Kanten werden in drei Kategorien unterteilt:

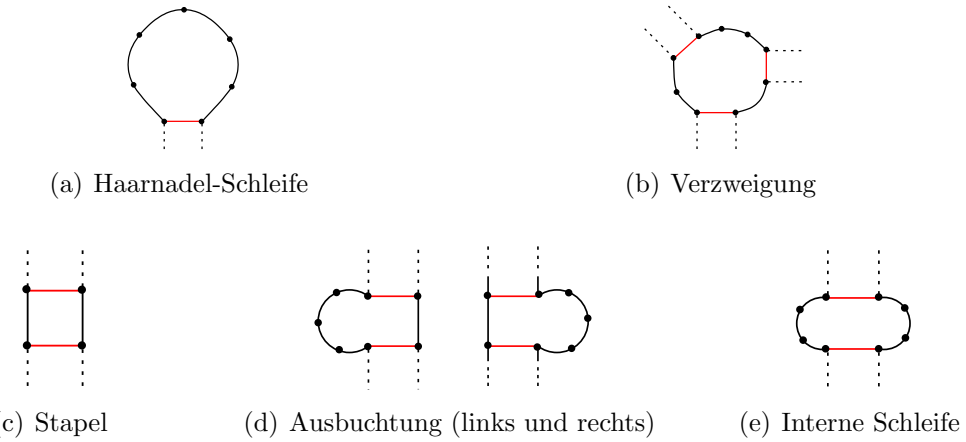
- *Stamm* oder *Stapel*: die internen Kanten sind auf jeder Seite durch genau eine

externe Kante getrennt, siehe Abbildung 1.5(c).

- *Ausbuchtung*: auf einer Seite sind mehrere externe Kanten und nur eine einzige auf der anderen Seite, siehe Abbildung 1.5(d).
- *Interne Schleife*: auf beiden Seiten sind mehrere externe Kanten, siehe Abbildung 1.5(e).

Ein Strukturelement, das von mehr als zwei internen Kanten umschlossen ist, heißt

- *Verzweigung* oder *multiple Schleife*, siehe Abbildung 1.5(b).



(a) Haarnadel-Schleife (b) Verzweigung

(c) Stapel (d) Ausbuchtung (links und rechts) (e) Interne Schleife

Abb. 1.5: Strukturelmente der MFE-Struktur-Vorhersage: rote Linien stehen für interne Kanten (Wasserstoffbrücken), schwarze Linien stehen für externe Kanten (Zucker-Phosphat-Bindungen). Ein Strukturelement wird nur von durchgezogenen Linien beschrieben.

Jedem der Strukturelemente wird, je nach Art und Anzahl der beteiligten Basen, eine freie Energie zugeordnet. Diese thermodynamischen Parameter wurden experimentell bestimmt, bzw. aus den experimentell bestimmten Werten extrapoliert [69]. Die freie Energie der Gesamtstruktur einer RNA wird additiv aus den freien Energien der einzelnen Elemente zusammengesetzt. Das Problem der Vorhersage einer MFE-Struktur kann mit Hilfe des DP-Algorithmus von Zuker und Stiegler gelöst werden, der große Ähnlichkeiten zum DP-Algorithmus von Nussinov (siehe Abschnitt 1.2.2) aufweist. Anstatt einer Matrix mit der maximalen Anzahl an Basenpaarungen für jede Teilsequenz einer Sequenz R, werden zwei Matrizen (V und W) berechnet. Dabei ist $W(i, j)$ die Energie der besten Struktur für die Teilsequenz

$R_{i,j}$, und $V(i,j)$ die Energie der besten Struktur für $R_{i,j}$ in dem Speziallfall, dass i und j eine Paarung eingehen. Der letzte berechnete Wert $W(1,N)$ ist die minimale Energie der Struktur der N Nukleotide langen RNA-Sequenz R. Die MFE-Struktur kann schließlich aus den Matrizen W und V durch Backtracking berechnet werden. Die Zeitkomplexität des Algorithmus beträgt $O(N^3)$ und die Speicherkomplexität $O(N^2)$ für eine Sequenz der Länge N. Eine genaue Beschreibung des Algorithmus ist in [113] zu finden.

Zu den meistgenutzten Programmen zur Sekundärstruktur-Vorhersage für RNA gehören mfold von Zuker [111, 112] und RNAfold aus dem Vienna-Paket von Hofacker *et al.* [49, 48]. Beide Programme verwenden die thermodynamischen Parameter von Mathews *et al.* [69].

1.2.4 Lokale MFE-Strukturen in langen Sequenzen

Soll nach kleinen MFE-Strukturen in einer langen Sequenz gesucht werden, so gibt es die Möglichkeit, Teilsequenzen einer konstanten Länge mit einer der Methoden aus Abschnitt 1.2.3 zu falten. Um jedoch alle optimalen Teilstrukturen zu erhalten, müssten alle Teilsequenzen einer vorgegebenen Länge mit einer Schrittweite von einem *Basenpaar* (bp) gefaltet werden. Bei langen Sequenzen führt das zu einem erheblichen Rechenaufwand. Um den Rechenaufwand zu reduzieren, kann die Schrittweite vergrößert werden. Auf diese Weise könnte aber ein Teil der lokalen MFE-Strukturen übersehen werden.

Eine bessere Lösung für dieses Problems bietet der Algorithmus von Hofacker *et al.* [50], welcher auf dem Algorithmus von Zuker und Stiegler (Abschnitt 1.2.3) basiert und als Teil des Vienna-Packets im Programm RNALfold implementiert ist. Dieses Programm berechnet alle lokalen, thermodynamisch optimalen Strukturen einer Sequenz der Länge N mit Hilfe eines DP-Algorithmus. Dabei kann der maximale Abstand paarender Basen L vorgegeben werden. Die Zeitkomplexität des Algorithmus beträgt $O(NL^2)$ und die Speicherkomplexität $O(N+L^2)$ [50].

1.2.5 Konsensus-Sekundärstruktur Vorhersage

Viele fRNAs bilden eine, für ihre Funktion charakteristische Struktur aus. Diese Struktur ist evolutionär konserviert und wird auch als *Konsensus-Sekundärstruktur*

der dazugehörigen Sequenzen bezeichnet. Es gibt mehrere Strategien, diese Struktur für eine Schar von RNA-Sequenzen vorherzusagen.

Eine Möglichkeit ist, gleichzeitig nach sequenziellen und strukturellen Gemeinsamkeiten zu suchen. In diesem Fall wird das Alignment- und das Faltungsproblem simultan mit Hilfe des von Sankoff *et al.* [90] vorgeschlagenen DP-Algorithmus gelöst. Der Nachteil dieses Ansatzes ist jedoch ein hoher Rechenaufwand. Die Zeitkomplexität für N Sequenzen, wobei die längste Sequenz die Länge L hat, beträgt $O(L^3K^N)$. K ist dabei eine ganzzahlige Konstante.

Eine andere Möglichkeit ist, die Suche nach sequenziellen und strukturellen Gemeinsamkeiten nacheinander durchzuführen. Dabei wird zuerst ein multiples Sequenzalignment gebildet und auf dessen Grundlage die Konsensus-Sekundärstruktur bestimmt. Werden beide Schritte nacheinander ausgeführt, führt das zu einem deutlichen Zeitvorteil gegenüber dem ersten Ansatz. Ein besonders häufig verwendeter Algorithmus zur Vorhersage einer Konsensus-Sekundärstruktur für alle Sequenzen eines Alignments wurde von Hofacker *et al.* [47] vorgeschlagen und im Programm RNAalifold implementiert. Er ist eine Variante des DP-Algorithmus von Zuker und Stiegler [113] unter besonderer Berücksichtigung konsistenter Mutationen. Dabei wird eine Mutation als *konsistent* bezeichnet, wenn eine oder beide Basen zweier paarender Basen mutieren, ohne dass die Basenpaarbindung dadurch gelöst wird. Das kann z. B. auftreten, falls das G im GU-Paar zu A mutiert. Da A und U komplementäre Basen sind, können auch A und U paaren. Ein Spezialfall der konsistenten Mutation ist die *kompensatorische Mutation*. Dabei handelt es sich um die Mutation beider Basen, die ein Paar gebildet haben und die trotz der Mutation die Eigenschaft zur Paarung beibehalten haben. So kann z. B. das Paar AU zu GC mutieren, ohne dass die Paarungseigenschaft an den entsprechenden Positionen in der Sequenz verloren geht. Werden vor allem Mutationen beobachtet, die Paarungseigenschaften erhalten, spricht das für eine funktionelle Bedeutung der Struktur dieser Sequenzen. Daher werden solche Mutationen von Hofacker *et al.* stärker gewichtet als einfache konservierte Basenpaare im Alignment.

Um die Idee des Algorithmus von Hofacker *et al.* zu verdeutlichen, wird er hier unter einer vereinfachenden Annahme beschrieben. Wir gehen davon aus, dass Energien einer Basenpaarung zugeordnet werden, wie beim Nussinov-Algorithmus (Abschnitt 1.2.2), und nicht gesamten Strukturelementen, wie im Falle des Zuker-Stiegler-Algorithmus (Abschnitt 1.2.3). Der Algorithmus basiert darauf, dass die Spalten

eines Alignments wie einzelne Basen einer Sequenz behandelt werden. So wird zuerst eine Paarungsmatrix berechnet, um zu entscheiden, welche Alignmentspalten miteinander „paaren“ können, d. h. ob die Basen an den entsprechenden Positionen in den einzelnen Sequenzen miteinander paaren können. Um mögliche Fehler im Alignment zu berücksichtigen, wird die Paarungsfähigkeit zweier Spalten nach speziellen Kriterien beurteilt. Es müssen also nicht alle Basen der beiden Alignmentspalten paaren können, damit die Spalten als paarungsfähig eingestuft werden.

An die Stelle der Berechnung der minimalen freien Energie bei der Sekundärstruktur-Vorhersage in einer Sequenz, tritt im Falle der Vorhersage der Konsensus-Sekundärstruktur eines Alignments die Berechnung eines Scores, welcher auch als *Konsensus-MFE* bezeichnet wird. Es handelt sich dabei aber nicht ausschließlich um einen Energiewert wie bei Einzelsequenzen. Für eine einzelne Sequenz wird jedem Basenpaar eine Paarungsenergie in Abhängigkeit vom Typ des Basenpaares zugeordnet. Die optimale Sekundärstruktur wird dann als diejenige Struktur mit der minimalen Gesamtenergie aller beteiligten Basenpaare bezeichnet. Im Falle eines Alignments wird jedem Spaltenpaar analog ein Gewicht zugeordnet. Dieses setzt sich aus dem durchschnittlichen Gewicht für die Energie der paarenden Basen in den einzelnen Sequenzen und einem Kovarianzterm, der konsistente und kompensatorische Mutationen belohnt, zusammen. Die Berechnung der Konsensus-MFE-Struktur eines Alignments erfolgt dann analog zur Berechnung der MFE-Struktur einer einzelnen Sequenz.

Der im Programm RNAalifold implementierte Algorithmus weist Energien nicht einzelnen Basenpaaren wie beim Nussinov-Algorithmus zu, sondern gesamten Strukturelementen wie im Falle des Zuker-Stiegler-Algorithmus.

1.3 Funktionelle RNA

In diesem Abschnitt werden mögliche Aufgabengebiete von fRNAs beschrieben und eine fRNA-Datenbank vorgestellt, die im Rahmen dieser Arbeit verwendet wurde.

1.3.1 Aufgaben der fRNA

Protein-kodierende Gene werden in die mRNA transkribiert, welche wiederum in Proteine übersetzt wird. Es gibt aber auch Gene, die zwar transkribiert, jedoch nicht

translatiert werden, sondern ihre Funktion direkt als RNA ausüben. Solche RNAs werden z. B. als *non-(protein-)coding RNA* (ncRNA) bezeichnet und gehören zu der Gruppe der fRNAs. Neben eigenständigen ncRNA-Molekülen, gibt es eine Vielzahl an *strukturbasierten regulatorischen cis-Elementen*, kurz *regulatorischen Elementen*, die ebenfalls als fRNAs bezeichnet werden. Sie sind Teil einer mRNA oder *prä-mRNA*, einer Vorstufe der mRNA, und können in UTRs und Introns auftreten und sich teilweise sogar mit Protein-kodierenden Bereichen der mRNA überschneiden. Auf der Grundlage ihrer Struktur regulieren sie die Expression angeschlossener Gene.

Es gibt verschiedene Typen von fRNA, die ebenso unterschiedliche Aufgaben übernehmen können. Die bekanntesten Vertreter ihrer Art sind die *transfer RNA* (tRNA) und die *ribosomale RNA* (rRNA). Beide sind an der Translation beteiligt, ebenso wie die *transfer-messanger RNA* (tmRNA). Bleibt ein Ribosom bei der Translation einer mRNA hängen, weil z. B. die mRNA fehlerhaft ist, so springt die tmRNA ein. Sie ermöglicht das Ablösen des Ribosoms und markiert gleichzeitig die unvollständige Proteinkette, so dass diese als nicht funktionstüchtig erkannt und zerstört werden kann [10].

Sogenannte *Ribozyme* (RNA-Enzyme) übernehmen katalytische Aufgaben bei chemischen Reaktionen, wie Spaltung und Ligation von RNA-Molekülen. Ihre Funktionsweise ähnelt der von proteinbasierten Enzymen, teilweise agieren sie auch in Verbindung mit einem Proteinkomplex. *Ribonuclease P* (RNase P) ist z. B. an der Weiterverarbeitung verschiedener Transkripte beteiligt [51]. Einige Ribozyme sind sogar in der Lage, ihre eigene Sequenz aus einem größeren Transkript eigenständig zu spalten, um so ihre endgültige Form und Funktion anzunehmen, wie z. B. das *Hammerhead Ribozym* [27].

Ein weiteres großes Aufgabenfeld der fRNAs ist die Regulation der Genexpression. Dabei sind vor allem *Mikro RNAs* (miRNAs) zu erwähnen. Sie kommen vorwiegend in Eukaryoten vor und sind komplementär zu einem Abschnitt einer oder mehrerer mRNAs, welche sie regulieren. Eine miRNA bindet an ihre Ziel-mRNA und blockiert die Translation oder ermöglicht die Zersetzung der mRNA durch ein Proteinkomplex, dem die miRNA als Wegweiser dient [5, 62, 108, 23]. In Prokaryoten hat die sogenannte *antisense RNA* (aRNA) eine vergleichbare Rolle. Sie kann die Expression eines Gens herunterregulieren oder sie überhaupt erst aktivieren [101].

Zu der Klasse der strukturbasierten cis-regulatorischen Elemente gehören z. B. die sogenannten *Riboswitches*. Sie sind Teil von bestimmten mRNAs und bilden

eine spezifische Struktur aus, an die ein ganz spezielles Molekül binden kann. Diese Bindung verändert wiederum die Struktur des Riboswitches. Abhängig vom Typ des Riboswitches und des Zielmoleküls, kann eine Bindung die Translation blockieren oder aber erlauben [73, 91, 75].

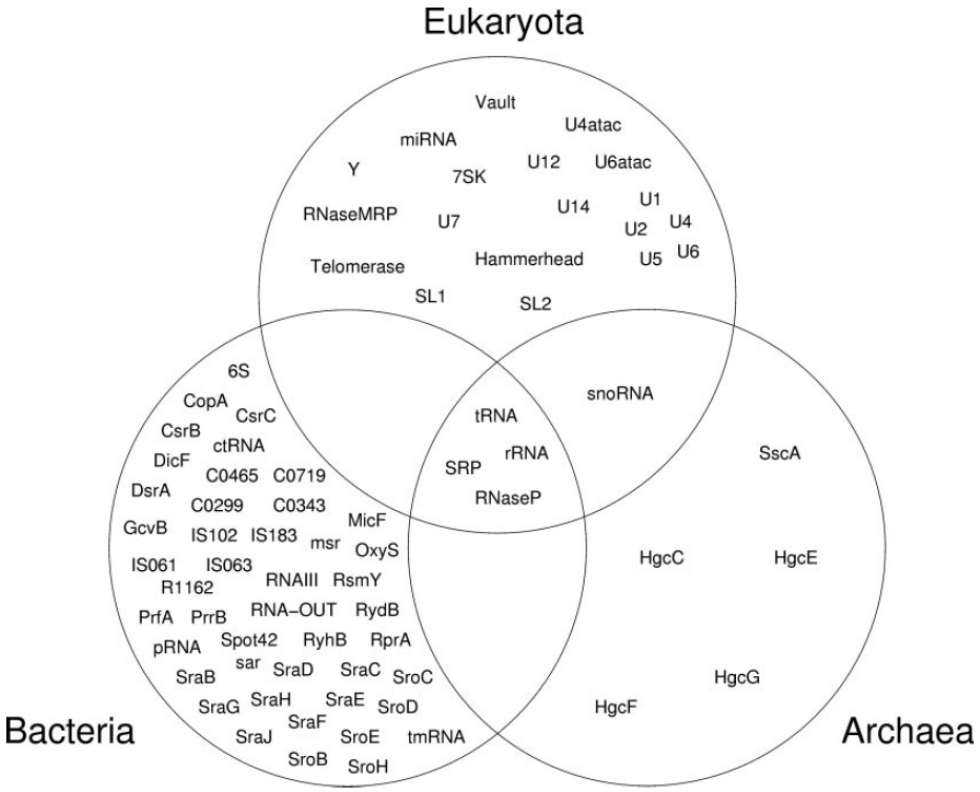


Abb. 1.6: Taxonomische Verteilung einiger fRNA-Familien in `Rfam`. (Abbildung aus [45])

Funktionelle RNAs sind in allen Domänen des Lebens vertreten. Einige Familien sind spezifisch für eine Domäne, andere wiederum sind in allen Domänen anzutreffen (Abbildung 1.6). Die Länge der fRNAs variiert zwischen 22 bp bei sogenannten kleinen fRNAs, wie z. B. miRNAs, und mehreren Tausend bp bei sogenannten langen fRNAs bei Eukaryoten, wie z. B. *Xist* [29]. Der größte Teil, der bisher entdeckten fRNAs, weist jedoch eine vergleichbar kurze Sequenz von bis zu 300 bp auf. Ein besonderes Merkmal funktioneller RNAs ist ihre hohe Strukturkonservierung innerhalb einer Familie. Da diese in den meisten Fällen der Schlüssel zu ihrer spezialisierten Funktion ist, ist sie sogar stärker konserviert als die Sequenz.

1.3.2 `Rfam`: RNA-Familien-Datenbank

Die **R**NA-**Fam**ilien-Datenbank (kurz: `Rfam`) [45, 40] ist eine Sammlung von fRNA-Familien. Sie werden grob in drei Klassen unterteilt:

1. eigenständige RNA-Moleküle, die durch ein eigenes Gen kodiert werden,

2. regulatorische cis-Elemente, die Teil einer DNA- oder mRNA-Sequenz sind und die Expression der Gene auf dem gleichen Strang regulieren und

3. Introns, die zwischen kodierenden Bereichen in der prä-mRNA eingeschlossen sind, selbst aber nicht für ein Protein kodieren.

Die fRNA-Familien werden durch multiple Sequenzalignments repräsentiert. Sogenannte *Seed*-Alignments enthalten jeweils bekannte, repräsentative Mitglieder einer Familie. Diese Alignments wurden per Hand nachbearbeitet [44] und lassen daher eine hohe Qualität erwarten. Neben sequenziellen Informationen ist zu jedem Alignment ebenfalls die Konsensus-Sekundärstruktur für die enthaltenen Sequenzen angegeben.

Wir nutzen diese Datenbank mit Hilfe eines geeigneten Programms (siehe Abschnitt 1.4.5), um in neuen Sequenzen nach Ähnlichkeiten zu den angegebenen fRNA-Familien zu suchen (siehe Kapitel 2).

1.4 Bekannte Ansätze zur fRNA-Detektion

Es gibt vielfältige Ansätze zur fRNA-Detektion, die auf unterschiedlichen Informationsquellen basieren. Je nachdem, nach welchen fRNAs gesucht wird: ob es sich dabei um Vertreter bekannter Familien handelt oder nicht, ob in dieser Familie die Sequenzinformationen oder die Strukturinformationen stärker konserviert sind oder ob ganz neue fRNAs vorhergesagt werden sollen, empfiehlt es sich, eine andere Strategie zu verwenden.

Im diesen Abschnitt werden zunächst die wichtigsten Strategien und ihre Anwendungsgebiete vorgestellt. Im Anschluss daran werden zwei spezielle Werkzeuge: RNAz und INFERNAL, die im Rahmen dieser Arbeit verwendet wurden, genauer beschrieben.

1.4.1 Homologiebasierte Suche nach bekannten fRNAs

Wird in der DNA nach Vertretern einer bekannten fRNA-Familie mit einer stark konservierten Sequenz gesucht, wie z. B. rRNA, so kann dies mit einer lokalen Sequenzhomologie-basierten Suche mit Hilfe von BlastN [4] oder FASTA [79] realisiert werden. Dabei wird eine bekannte fRNA-Sequenz als Referenz verwendet. Der große Vorteil dieser Algorithmen ist ihre Schnelligkeit. Sie sind jedoch nicht geeignet, wenn die Sequenzkonservierung nicht in ausreichendem Maße gegeben ist, was bei vielen fRNA-Familien der Fall ist.

Sollen zusätzlich zu Sequenzinformationen auch Strukturinformationen berücksichtigt werden, so kann RSEARCH [56] verwendet werden. Dieser auf sogenannten ***S****tochastic* ***C****ontext-****F****ree* ***G****rammar* (SCFG) ([28], Kap. 9) basierende Algorithmus ermöglicht es, eine Sequenzdatenbank nach Homologen zu einer fRNA-Sequenz mit gegebener Sekundärstruktur zu durchsuchen. Die Methode ist zeit- und speicherintensiv, weist jedoch für fRNA-Sequenzen eine höhere Sensitivität auf als BlastN oder FASTA [37].

Unterscheiden sich die Mitglieder einer bekannten fRNA-Familie in ihrer Sequenz oder Struktur stärker voneinander, so ist es sinnvoll, diese evolutionären Unterschiede zu berücksichtigen und nur nach gemeinsamen Merkmalen zu suchen, anstatt nach Ähnlichkeiten mit einem Vertreter der fRNA-Familie. Solche Anwendungen, die Informationen mehrerer Familienmitglieder ausnutzen, können in zwei Gruppen unterteilt werden: erstens, die familienspezifischen Anwendungen, wie z. B. tRNAscan-SE [64] für tRNAs, miRseeker [59] und MiRscan [61] für miRNAs, BRUCE [60] für tmRNAs und zweitens, Ansätze, die auf beliebige Familien anwendbar sind, da sie zuerst auf der jeweiligen Familie trainiert werden, wie HMMER [30] oder INFERNAL [31]. Programme, die nach Mitgliedern einer bestimmten Familie suchen, nutzen familienspezifische Heuristiken, wodurch die Rechenlaufzeit minimiert und die Sensitivität und Spezifität optimiert wird. Die charakteristischen Kriterien einer Familie können jedoch nicht auf eine andere Familie übertragen werden. Ein familienunabhängiges Werkzeug kann dagegen für jede bekannte Familie verwendet werden und ist damit universell einsetzbar, aber in vielen Fällen auch langsamer und ungenauer.

Während HMMER nicht speziell für die fRNA-Detektion entwickelt wurde und daher nur die konservierte Sequenzinformation eines Alignments mit Hilfe eines so-

genannten *Profile Hidden Markov Model* (Profile-HMM) ([28], Kap. 5) modelliert, ist INFERNAL in der Lage, gleichzeitig die konservierte Sequenzinformation und Konsensus-Sekundärstruktur eines Alignments zu modellieren. Dies geschieht mit Hilfe sogenannter *Profile-SCFG*, auch Kovarianzmodelle genannt. INFERNAL zeichnet sich auf fRNA-Sequenzen sowohl durch eine besonders hohe Sensitivität als auch Spezifität aus. Im Vergleich zu HMMER fällt es jedoch durch den hohen Rechenaufwand negativ auf [37]. In Abschnitt 1.4.5 wird INFERNAL genauer beschrieben, da es im Rahmen dieser Arbeit verwendet wird.

HMMER und INFERNAL ziehen die charakteristischen Informationen einer fRNA-Familie eigenständig aus einem gegebenen Alignment. Im Gegensatz dazu ermöglichen spezielle Sprachen, wie HyPAL [43] oder die RNAMotif [66] zu Grunde liegende Sprache, das gesuchte Motif selbst zu definieren und dann danach zu suchen. Neben Sequenz- und Strukturelementen können mit HyPAL thermodynamische Spezifikationen angegeben werden, während RNAMotif ein benutzterdefiniertes Bewertungssystem unterstützt. Die Nutzung dieser Methoden erfordert ein fundiertes Expertenwissen, um ein sinnvolles Motif definieren zu können.

In einer Studie von Freyhult *et al.* [37] wurde die Performanz der meistbekannten, homologiebasierten Werkzeuge verglichen. Danach gehörten BlastN und FASTA zu den schnellsten Algorithmen. Wurde hingegen die Genauigkeit betrachtet, so gehören INFERNAL, RSEARCH und HMMER zu den Besten. Bei der Studie wurde außerdem festgestellt, dass die 1-Sequenz-basierten Methoden wie z. B. BlastN, FASTA und RSEARCH schlechter abschneiden als Methoden, die Informationen mehrerer Sequenzen verwenden wie INFERNAL oder HMMER. Programme, die nur eine fRNA als Referenz nutzen, reagieren empfindlich auf zu große evolutionäre Unterschiede in der Sequenz und/oder Struktur. Wird hingegen ein statistisches Modell einer Schar von Mitgliedern einer fRNA-Familie verwendet, so können auch evolutionäre Unterschiede innerhalb der Familie berücksichtigt werden.

1.4.2 Lernbasierte Suche nach familienübergreifenden Merkmalen

Soll eine einzelne Sequenz auf das Vorkommen von fRNAs untersucht werden, so können Informationen wie Nukleotid- bzw. Dinukleotidgehalt der Sequenzen, spezielle Sequenzmotive oder die minimale freie Energie als Unterscheidungskriterium

zwischen fRNAs und sonstigen Sequenzen verwendet werden [14, 93]. Die Entscheidungsmerkmale werden zuerst an einem Datensatz von Positiv- und Negativbeispielen trainiert. Dabei dienen bereits bekannte fRNAs als Positivbeispiele und sonstige Sequenzen oder künstlich erzeugte Sequenzen als Negativbeispiele.

Der Erfolg einer solchen Strategie hängt stark vom untersuchten Organismus ab. Zum Beispiel in AT-reichen Hyperthermophilen, d. h. Lebewesen, die hohe Temperaturen von 80-120°C bevorzugen, wurden fRNAs alleine anhand des GC-Gehalts ihrer kodierenden Sequenz identifiziert [57]. Der Ansatz ist nur für diese spezielle Form der Organismen erfolgreich, da sie zur Stabilisierung der fRNA-Struktur bei den hohen Umgebungstemperaturen einen höheren GC-Gehalt aufweisen. Im Allgemeinen ist das jedoch nicht der Fall. In [86] wurde gezeigt, dass die thermodynamische Stabilität einer Sekundärstruktur als alleinige Information nur selten ein signifikantes Kriterium ist, um fRNAs in kompletten Genomsequenzen zu identifizieren.

1.4.3 Komparative Suche nach evolutionär konservierten Merkmalen

Kleine Änderungen in der Sequenz können eine große Auswirkung auf die Struktur haben. Damit eine Struktur erhalten bleibt, muss eine Sequenz in ihrer Basenabfolge aber nicht konserviert sein. Sogenannte konsistente Mutationen (siehe Abschnitt 1.2.5) erhalten Basenpaarungen und damit die gesamte Struktur. Sie sind daher ein Hinweis auf eine evolutionär konservierte und damit möglicherweise funktionelle Struktur. Im Gegensatz zur Stabilität der Sekundärstruktur einer einzelnen Sequenz, ist eine in mehreren verwandten Organismen konservierte Struktur ein deutlicher Hinweis auf eine fRNA. Programme wie QRNA [85], MSARi [22], ddbRNA [25] und RNAz [102] nutzen unter anderem diese Information aus, indem sie die Konsensus-Sekundärstruktur in einem multiplen Sequenzalignment bewerten.

Einer der ersten Ansätze mit dieser Strategie wurde in QRNA implementiert. Mit Hilfe von paarweisen HMM und paarweisen SCFG wurden drei evolutionäre Modelle für „kodierende Sequenzen“, „strukturelle RNAs“ und „andere Sequenzen“ erstellt. Das Programm analysiert ein paarweises Sequenz-Alignment und weist Teilsequenzen dieses Alignments dem wahrscheinlichsten der drei Modelle zu.

MSARi ist für Alignments mit zehn bis fünfzehn Sequenzen optimiert und konzentriert sich ganz auf das Auffinden kompensatorischer Mutationen. Dabei können

kleine Fehler im Alignment, bei denen ein konserviertes Basenpaar um bis zu zwei Positionen verschoben ist, ausgeglichen werden. Einer vergleichenden Untersuchung zufolge liefert das Program für eine kleine Anzahl an Sequenzen nur mäßige Ergebnisse [7].

ddbRNA sucht nach kompensatorischen Mutationen in konservierten, paarenden Regionen. Ihre Anzahl wird im Vergleich zur durchschnittlichen Anzahl kompensatorischer Mutationen in zufälligen Alignments, welche durch Permutation der Spalten aus dem ursprünglichen Alignment erzeugt wurden, bewertet.

RNAz ist in der Lage, Alignments mit zwei bis sechs Sequenzen zu untersuchen. Es vereint ein Maß für Strukturkonservierung, in das eine Bewertung kompensatorischer Mutationen eingeht, mit einem Maß für thermodynamische Strukturstabilität, das aus den MFE-Werten der einzelnen Sequenzen und des MFE-Wertes der Konsensus-Sekundärstruktur gebildet wird.

Eine Voraussetzung für die Anwendung dieser Methoden ist die Existenz eines Sequenzalignments. Soll eine einzelne Sequenz auf das Vorkommen von fRNAs untersucht werden, müssen zuerst geeignete Vergleichssequenzen bestimmt werden. Dabei hängt der Erfolg dieser Strategien von einer ausreichend hohen Sequenzkonservierung zwischen den betrachteten Sequenzen ab. Erst die Sequenzkonservierung ermöglicht das Erstellen eines korrekten Alignments. Der große Vorteil dieser Verfahren ist, dass sie keine Informationen über die gesuchten fRNAs benötigen, da die fRNA-Vorhersage auf der Grundlage evolutionär konservierter Strukturen und teilweise auch der Stabilität der Strukturen erfolgt. Damit sind sie generell einsetzbar und können sogar Mitglieder bisher unbekannter fRNA-Familien entdecken. Funktionelle RNAs ohne eine konservierte Struktur können damit jedoch nicht gefunden werden.

Sollen evolutionär konservierte Strukturen untersucht werden, so liefert ein paarweises Alignment wenig Informationen. QRNA kann aber nur solche Alignmnets verarbeiten. MSARi ist wiederum für Alignments mit zehn bis fünfzehn Sequenzen optimiert. Es ist jedoch schwierig, auf der Suche nach neuen fRNAs, so viele verwandte Sequenzen zu finden. RNAz und ddbRNA sind in Bezug auf die Anzahl der Sequenzen im Alignment flexibler. RNAz ist im Gegensatz zu den meisten anderen verfügbaren Methoden in der Lage, nicht nur die Strukturkonservierung zwischen den Sequenzen des Alignments zu bewerten, sondern auch die Strukturstabilität. Dabei erwarten wir, dass fRNAs deren Funktion von ihrer Struktur abhängt, eine

besonders stabile Struktur aufweisen, da diese nicht durch beliebige äußere Faktoren beeinflussbar sein darf, um funktionsfähig zu bleiben. Daher wird in dieser Arbeit unter anderem RNAz verwendet, um nach neuen fRNAs zu suchen. In Abschnitt 1.4.4 wird das Programm genauer erklärt.

1.4.4 RNAz

Basierend auf der Annahme, dass die Funktionalität vieler fRNAs von einer konservierten und stabilen Struktur abhängt, entwickelten Washietl *et al.* einen komparativen Ansatz zur Identifikation von fRNAs, der im Programm RNAz [102, 103] implementiert wurde. Sie kombinierten ein Maß für die thermodynamische Stabilität einer Sekundärstruktur mit einem Maß für die Strukturkonservierung in einem Alignment. Das ermöglicht eine effiziente Identifikation von fRNAs in multiplen Alignments mit nur wenigen Sequenzen. Die Laufzeit der Methode ist $O(N \times n^3)$, dabei ist N die Anzahl der Sequenzen und n die Länge des untersuchten Alignments.

Im Folgenden sind die Kriterien beschrieben, nach denen RNAz Sequenzen in strukturbasierte fRNA und nicht-fRNA klassifiziert werden.

Maß für Strukturkonservierung

Evolutionär konservierte Sekundärstrukturen werden mit Hilfe der MFE der Konsensus-Sekundärstruktur eines Alignments (Konsensus-MFE, siehe Abschnitt 1.2.5) bewertet. Bei der Berechnung der Konsensus-Sekundärstruktur und damit der Konsensus-MFE spielen insbesondere konsistente und kompensatorische Mutationen (siehe ebenfalls Abschnitt 1.2.5) eine wichtige Rolle. Das Besondere an diesen Mutationen ist, dass trotz einer Änderungen in der Sequenz, die ursprüngliche Struktur weiterhin ausgebildet werden kann. Werden solche Mutationen in paarenden Basen der Konsensus-Sekundärstruktur verwandter Sequenzen beobachtet, so ist das ein Hinweis auf die funktionelle Bedeutung der Struktur.

Es ist schwierig, den absoluten Wert der MFE zu interpretieren, da dieser unter anderem von der Basenzusammensetzung und der Länge der Sequenz bzw. des Alignments im Fall der Konsensus-MFE abhängt. Daher wird die Konsensus-MFE in Relation zu den MFE der einzelnen Sequenzen im Alignment betrachtet. Dazu wird zuerst die Konsensus-MFE des Alignments, E_A, mit RNAalifold [47] berechnet. Danach wird der Mittelwert über die MFE aller einzelnen Sequenzen im Alignment, $\overline{E}$

bestimmt. Die einzelnen Sequenzen werden jeweils mit Hilfe von RNAfold [49] gefaltet. Aus beiden Größen wird ein Maß für die Strukturkonservierung, der sogenannte ***S**tructure **C**onservation **I**ndex* (SCI) berechnet:

$$\mathrm{SCI} = \frac{E_A}{\overline{E}}.$$

Ein SCI nahe bei null bedeutet, dass RNAalifold keine gemeinsame Sekundärstruktur für alle Sequenzen im Alignment gefunden hat. Zu 100 % konservierte Strukturen ergeben hingegen einen SCI ≈ 1. Die Konsensus-MFE eines Alignments, E_A, ist kein reiner Energiewert, sondern beinhaltet einen Bonus für kompensatorische und/oder konsistente Mutationen. Das kann dazu führen, dass $|E_A| > \left|\overline{E}\right|$ ist, obwohl die Struktur in allen Sequenzen perfekt konserviert ist. Damit kann der SCI sogar einen Wert annehmen, der größer ist als eins.

Maß für thermodynamische Stabilität

Die thermodynamische Stabilität der Sekundärstruktur einer Sequenz S wird mit Hilfe des z-Scores

$$z = \frac{m - \mu}{\sigma} \tag{1.1}$$

berechnet. Dabei ist m der MFE-Wert der Sekundärstruktur von S. Er wird verglichen mit den MFE-Werten der Strukturen zufälliger Sequenzen gleicher Länge und Nukleotid- bzw. Dinukleotidzusammensetzung wie S. Dazu wird der Mittelwert μ und die Standardabweichung σ der MFE-Werte der Sekundärstrukturen der zufälligen Sequenzen berechnet. Da die Parameter μ und σ Funktionen der Länge und Basenzusammensetzung der Sequenzen sind, werden sie jeweils mit Hilfe eines Regressionsmodells berechnet. Die Regressionsmodelle wurden wiederum mit der sogenannten ***S**upport **V**ector **M**achine* (SVM) [95] trainiert. Dabei wurde die SVM Bibliothek LIBSVM [15] verwendet.

Binäre Klassifikation mit SVM

Zur binären Klassifikation der Sequenzen in einem Alignment als fRNA oder nicht-fRNA nutzt RNAz ebenfalls die SVM Bibliothek LIBSVM. Der binäre SVM-Klassifikator wurde auf Alignments trainiert, die aus verwandten Sequenzen bekannter fRNA-Familien zusammengestellt wurden. Als Negativbeispiele dienten Alignments,

die durch Spaltenpermutation aus den Positivbeispielen erzeugt wurden. Dabei wurden die Spalten zwar zufällig, aber nicht beliebig permutiert, um lokale Eigenschaften des Alignments wie z. B. den Grad der Sequenzkonservierung zu bewahren. Für eine Beschreibung der Permutationsprozedur siehe [102]. Als Klassifikationsparameter werden neben dem SCI des Alignments und des Mittelwerts der z-Scores aller Sequenzen im Alignment auch die mittlere paarweise Sequenzidentität im Alignment und die Anzahl der beteiligten Sequenzen verwendet. Obwohl die SVM auf bekannten fRNA-Familien trainiert wurde, wurden keine speziellen Sequenz- oder Strukturmotive „gelernt“. Die SVM dient zur Beurteilung des SCI und des z-Scores. Diese beiden Werte beinhalten jedoch keine Informationen, die nur für eine fRNA-Familie typisch sind.

Die von der SVM geschätzte Klassenwahrscheinlichkeit wird als Signifikanzmaß verwendet. Wir bezeichnen sie fortan als RNAz-Score. Der RNAz-Score kann Werte zwischen 0 und 1 annehmen. Je höher dieser Wert ist, umso sicherer konnten die Sequenzen als fRNA klassifiziert werden. Ab einem Wert von 0,5 können die Sequenzen des gegebenen Alignments als fRNA-Kandidaten angesehen werden. In Abbildung 1.7 ist die Klassifikation am Beispiel von zwei bekannten fRNA-Familien visualisiert. Es werden Alignments der fRNA-Familien mit Alignments, die durch zufälliges Permutieren der Spalten der Originalalignments entstanden sind, verglichen.

1.4.5 Das INFERNAL-Paket

Es gibt eine wachsende Anzahl an bekannten fRNA-Familien. Eine Familie weist in den meisten Fällen neben gemeinsamen Sequenzmotifen auch gemeinsame Sekundärstrukturelemente auf. Eddy *et al.* [31, 74] entwickelten einen Ansatz, der diese Informationen ausnutzt, um nach ähnlichen Sequenzen zu suchen. Der Ansatz wurde im Programmpaket INFERNAL implementiert. Dabei werden im ersten Schritt die gemeinsamen Sequenz- und Struktur-Informationen einer gegebenen fRNA-Familie in einem sogenannten *Kovarianzmodell* zusammengefasst, um dann im zweiten Schritt eine Datenbank nach Sequenzen zu durchsuchen, welche zu diesem Kovarianzmodell homolog sind.

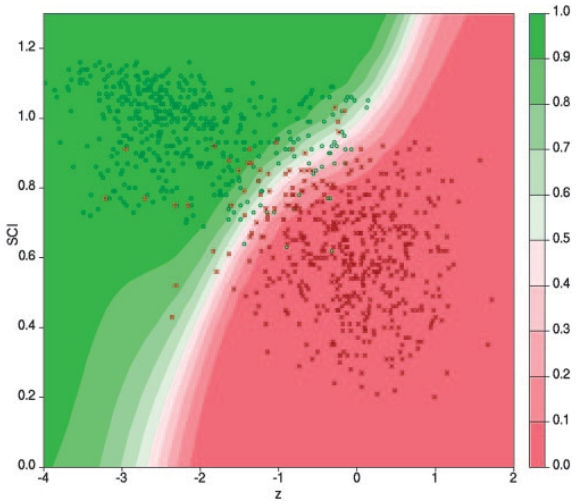


Abb. 1.7: SVM-Klassifikation auf der Grundlage des z-Scores und des SCI-Scores am Beispiel von tRNA- und 5S rRNA-Alignments mit zwei bis vier Sequenzen pro Alignment. Grüne Punkte repräsentieren Originalalignments, während rote Punkte für negative Kontrollalignments stehen. Der Farbverlauf des Hintergrunds, von grün zu rot, beschreibt die RNA-Klassenwahrscheinlichkeit in der z-Score - SCI-Score - Ebene. Bildquelle: [102].

Von einer fRNA-Familie zum Kovarianzmodell mit cmbuild

Ein bewährtes Mittel zum Modellieren von Protein- oder DNA-Sequenz-Familien sind Profile-HMMs [28, Kap. 5]. Sie sind gut geeignet, um die Konsensus-Sequenz eines Alignments zu beschreiben, können aber keine RNA-Sekundärstrukturen modellieren. Damit eignen sie sich nur bedingt für die Beschreibung der Gemeinsamkeiten einer fRNA-Familie. Für diesen Fall wurde eine spezielle, auf SCFG [28, Kap. 9] basierende Architektur, das Kovarianzmodell ([28, Kap.10], [34]), entwickelt. Während Profile-HMMs eine lineare Architektur zu Grunde liegt, weisen Kovarianzmodelle eine baumähnliche Architektur auf, die das Modellieren der Konsensus-Sekundärstruktur eines RNA-Alignments ermöglicht.

Das Aufstellen eines Kovarianzmodells erfolgt mit dem Programm cmbuild aus dem INFERNAL-Paket. Dieser Schritt muss für eine Referenzdatenbank von fRNA-Familien nur ein Mal durchgeführt werden. Als Referenzdatenbank empfiehlt es sich, z. B. `Rfam` (siehe Abschnitt 1.3.2) zu nehmen, da in dieser Datenbank eine große Anzahl an bekannten fRNA-Familien und regulatorischen RNA-Elementen in Form von Alignments mit gegebener Konsensus-Sekundärstruktur bereitgestellt wird.

Datenbanksuche mit cmsearch

Eine Anfragesequenz wird mit Hilfe von cmsearch aus dem INFERNAL-Paket nach Sequenzabschnitten abgesucht, die zur gegebenen Referenz-fRNA-Familie homolog sind. Dabei wird das sogenannte *Maximum-Likelihood-Alignment* des Kovarianzmodells, das die Konsensus-Sekundärstruktur der Referenz-fRNA-Familie modelliert, und der Anfragesequenz ohne bekannte Sekundärstruktur bestimmt. Der Ansatz ist eine Variante des *Cocke-Younger-Kasami Algorithmus* (CYK-Algorithmus) [20, 107, 53], eines DP-Algorithmus, der das Alignmentproblem einer SCFG zu einer Anfragesequenz löst. In der ursprünglichen Form des Algorithmus beträgt die Zeitkomplexität $O(LN^3)$ und die Speicherkomplexität $O(N^3)$ für eine Anfragesequenz der Länge L und ein Referenz-Alignment der Länge N. Durch die Einführung einer sogenannten *Divide-And-Conquer*-Variante des CYK-Algorithmus konnte die Speicherkomplexität jedoch auf $O(N^2 \log(N))$ reduziert werden [31].

INFERNAL kann sowohl im globalen als auch im lokalen Modus verwendet werden. Der lokale Modus kann zu einem, auf der Sequenzebene unterbrochenem, in der Struktur jedoch zusammenhängendem, Alignment führen. Dieser Ansatz ist vor allem zum Aufspüren solcher fRNAs geeignet, bei denen nur ein Teil der fRNA-Struktur eine funktionelle Bedeutung hat, und die „überflüssige“ Sequenz durch Mutationen, Deletionen oder Insertionen verändert werden kann. In Abbildung 1.9 wird das Phänomen an einem Beispiel dargestellt.

Da INFERNAL nach Ähnlichkeiten zu gegebenen fRNA-Familien sucht, liefert es nicht nur die Position eines Kandidaten in der Anfragesequenz, sondern auch seine Familienzugehörigkeit, und damit seine mögliche Funktion. In dieser Arbeit verwenden wir `Rfam 8.1` (Abschnitt 1.3.2) als Referenzdatenbank für fRNA-Familien. Dabei haben wir ausschließlich Seed-Alignments benutzt, da wir uns von ihnen eine höhere Genauigkeit erhoffen.

INFERNAL im Vergleich

In einem aktuellen Vergleich von Methoden zur Homologiesuche bei fRNAs [37], in welchem sowohl sequenzbasierte, als auch strukturbasierte Methoden und Profile-HMM-Methoden verglichen wurden, hat INFERNAL mit einer sehr hohen Sensitivität und Spezifität überzeugt. Der einzige Kritikpunkt war die Zeitkomplexität, die im Vergleich zu sequenzbasierten Methoden bzw. Profile-HMM-Methoden deut-

lich höher ausfällt. Eine Beschleunigung der Anwendung ist jedoch das Ziel aktueller Forschung [74].

Ausgabeformat

INFERNAL liefert detaillierte Informationen zu gefundenen Treffern. Das Ausgabeformat des Alignments wird hier an zwei Beispielen in Abbildung 1.8 und Abbildung 1.9 beschrieben. Die erste Ausgabe wurde im globalen Modus, die zweite im lokalen Modus von INFERNAL erstellt.

In der ersten Zeile steht die für die Anfragesequenz vorhergesagte Struktur. Paarende Basen werden durch komplementäre Klammern verschiedenen Typs: (), <>, [], {} dargestellt. Damit kann die Verschachtelung einer Struktur visualisiert werden. Ungepaarte Basen werden ebenfalls mit verschiedenen Zeichen dargestellt, je nachdem, an welcher Position in der Struktur sie sich befinden: „_“ in Haarnadelschleifen, „-“ in internen Schleifen und Ausbuchtungen, „,“ in multiplen Schleifen und „:“ außerhalb paarender Basen, d. h. an Sequenzenden. Ist ein Treffer in der Anfragesequenz auf Sequenzebene unterbrochen, da er aus einer lokalen Suche stammt, wird der fehlende Bereich in der dazugehörigen Struktur mit „~“-Zeichen gekennzeichnet. Insertionen in der Anfragesequenz werden durch das „.“-Zeichen beschrieben.

In der zweiten Zeile steht jeweils die Konsensus-Sequenz. Dabei handelt es sich um diejenige Sequenz, welche mit Hilfe des Kovarianzmodells den Treffer in der Anfragesequenz mit dem höchsten Score modelliert. Dabei werden hochkonservierte Residuen des Modells in Großbuchstaben angegeben. Kleinbuchstaben stehen für wenig oder gar nicht konservierte Positionen.

Die dritte Zeile gibt die Zusammensetzung des Alignmentscores wieder. Falls ein beobachtetes Basenpaar den bestmöglichen Score bezüglich des Basenpaares in der Konsensus-Sequenz hat, werden beide an der Paarung beteiligten Residuen in Großbuchstaben angegeben. Falls ein Paar einen Score, der größer als null ist, aufweist, werden beide Residuen mit dem „:“-Zeichen annotiert. Ein Leerzeichen bedeutet, dass dieses Basenpaar einen negativen Beitrag zum Alignmentscore liefert. Für ungepaarte Residuen gilt äquivalent: ist das Residuum, dasjenige mit dem höchsten Score, wird es als Großbuchstabe angegeben. Ein positiver Scorebeitrag wird durch das „+“-Zeichen gekennzeichnet und ein Leerzeichen steht für einen negativen Beitrag zum Alignmentscore. Die vierte Zeile gibt schließlich die Treffersequenz an, d. h.

den zum Kovarianzmodell homologen Abschnitt der Anfragesequenz.

Wird INFERNAL im lokalen Modus gestartet, kann das zu Treffern führen, die auf der Seqenzebene unterbrochen sind, in der Struktur jedoch zusammenhängen, wie in Abbildung 1.9 dargestellt ist. Die geklammerten Zahlen `*[32]*` in der Referenzsequenz und `*[18]*` in der Anfragesequenz geben an, wieviele der Residuen an der entsprechenden Position ausgelassen wurden, da sie keine ausreichende Ähnlichkeit aufweisen.

```
GCCCCGCUGAUGAGGUCAG.GGAAGACCGAAAGUGUCGACUCUACGGGGC
AUCCAACUGACCAGUCGGA.AAUUGGACGAAACGC....GCGUCCUGGAU
GCACUGCUGAGGAGUCCCACAAUAGGACGAAACGA....CCGUCCAGUAC
AUCCAGCUGACGAGUGCCA.AAUAGGACGAAAUGC....GCAUCCUGGAU
<<<<<<.......<<<<.......>>>>...<<<<....>>>>.>>>>>>

  ((((((,,,,,,,<<<<______>>>>,,,<<<<>>>>,))))))
1 gccCcGcUGAugAGgcCaaaauAGgcCGAAacggccguaCgGggc 45
  ::C: :CUGA GAG::: AAAUA:::CGAAAC :: GU+: :G::
58 AUCUUUCUGACGAGUUUCAAAUAGGACGAAACGCGUGUCAUGGAU 102
```

Abb. 1.8: Beispiel für die Anwendung von INFERNAL. Oben: Aus dem Referenzalignment des sogenannten *Hammerhead*-Ribozyms mit gegebener Konsensus-Sekundärstruktur (Quelle: `Rfam`, Eintrag: RF00163) wurde mit cmbuild ein Kovarianzmodell erstellt. Unten: Ergebnis des Vergleichs einer Anfragesequenz mit dem Kovarianzmodell des Referenzalignments. Die Suche wurde mit cmsearch durchgeführt.

```
AUCUGUUCACAUCUUACCGGGCAGCAUUAGAUCCUUUUU.....AUAACUCUAAUACUGUCAGGUAAAGAUGU.CGUCCGU
AGCCGUGGCCAUCUUACUGGGCAGCAUUGGAUAGUGUC......UGAUCUCUAAUACUGCCUGGUAAUGAUGA.CGGCGGA
CUCUGGGUCCAUCUUCCAGUGCAGUGUUGGAUGGUUGAAGUAUGAAGCUCCUAACACUGUCUGGUAAAGAUGGCCCCCGGG
AGGGGCCCUCGUCUUACCCAGCAGUGUUUGGGUGCUGGUUGG..GAGUCUCUAAUACUGCCGGGUAAUGAUGGAGGCCCCU
CUCUGUGGGCAUCUUACCGGACAGUGCUGGAUUUCUUGGC....UUGACUCUAACACUGUCUGGUAACGAUGUUCAAAGGU
AGCCGUGGCCAUCUUACUGGGCAGCAUUGGAUGGAGUC......AGGUCUCUAAUACUGCCUGGUAAUGAUGA.CGGCGGA
CCCUGGGUCCAUCUUCCAGUACAGUGUUGGAUGGUCUAAUUGUGAAGCUCCUAACACUGUCUGGUAAAGAUGGCUCCCGGG
GGGGGCCCUCGUCUUACCCAGCAGUGUUUGGGUGCGGUUGG...GAGUCUCUAAUACUGCCGGGUAAUGAUGGAGGCCCCU
CCCUGUGAGCAUCUUACCGGACAGUGCUGGAUUUCCCAGC....UUGACUCUAACACUGUCUGGUAACGAUGUUCAAAGGU
.....<<.<<<<<<<<<...<<<<<<<<<<.....................>>>>>>>>>>...>>>>.>>>>>.>>.....


   :::::<<-<<<<<<<<<<---<<~~~~~~>>--->>>>->>>>>->>:::::
 1 aucuGcgucCaUCUUACcGggC*[32]*GcCuGGUAAaGAuGgacgacgGu 77
    + UGCG+CC:: UU:C+GG::       ::CUGG:AA  ::GGACG+C G
78 GGAUGCGACCUGGUUGCUGGAU*[18]*AUCUGGCAA-GCAGGACGCCCGG 17
```

Abb. 1.9: Beispiel für die Anwendung von INFERNAL mit der Option `--local`. Oben: Referenzalignment von *mir-8* (Quelle: `Rfam`, Eintrag: RF00241). Unten: Ergebnis des lokalen Vergleichs mit einer Anfragesequenz.

2 Ansatz zur fRNA-Detektion in Genomsequenzen

Anders als Protein-kodierende Gene weisen fRNA-kodierende Gene keine gemeinsamen, statistisch signifikanten Signale in der Sequenz auf, wie z. B. ORFs, die als Grundlage für eine Vorhersagemethode verwendet werden könnten. Erschwerend kommt hinzu, dass nicht nur eigenständige fRNA-Moleküle, die durch ein eigenes Gen kodiert werden, sondern auch strukturbasierte regulatorische Elemente zur Gruppe der fRNAs zählen und ebenfalls vorhergesagt werden sollen. Sie werden nicht in ein eigenständiges Molekül übersetzt, sondern sind Teil regulatorischer, vorwiegend nicht-Protein-kodierender Bereiche der mRNA oder prä-mRNA. Ein weiteres Problem stellt die teilweise schlecht konservierte Sequenzinformation innerhalb einer fRNA-Familie dar, welche die sequenzbasierte Homologiesuche nach neuen Kandidaten deutlich erschwert. Dennoch weisen fRNAs häufig eine wichtige Eigenschaft auf. In vielen Fällen ist ihre Struktur grundlegend für ihre Funktion und dadurch innerhalb einer fRNA-Familie stark konserviert.

In diesem Kapitel wird ein Ansatz für die systematische Suche nach fRNA-kodierenden Genen und strukturbasierten regulatorischen Elementen in kompletten Genomsequenzen vorgestellt. Es ist schwierig exakte Start- und Endpositionen der fRNA-kodierenden Regionen vorherzusagen, daher konzentrieren wir uns auf die Suche nach Sequenzbereichen, die zu signifikanten fRNA-Kandidaten gehören. Der Ansatz ist in zwei Teile gegliedert. Im ersten Schritt, dem *komparativen Ansatz*, wird eine vergleichende Sequenzanalyse mit der Analyse der konservierten Sekundärstruktur kombiniert. Dabei werden keine a priori Informationen über bekannte fRNA-Familien benötigt. Im zweiten Schritt, dem *Kovarianzmodell-basierten Ansatz*, wird eine Anfragesequenz auf Ähnlichkeiten mit bekannten fRNA-Familien hin abgesucht. Die einzelnen Schritte des Ansatzes sind in der Abbildung 2.1 graphisch zusammengefasst.

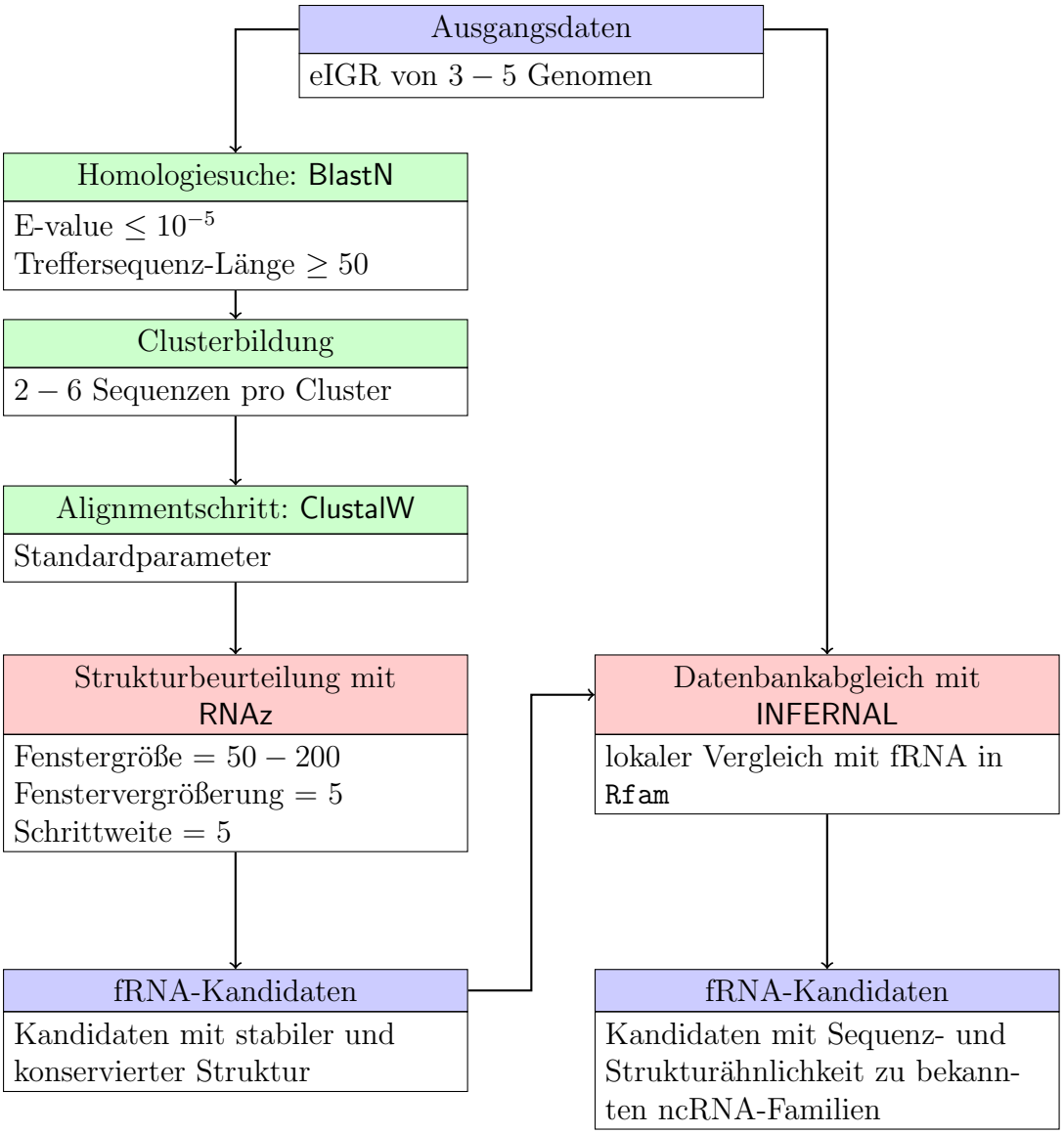


Abb. 2.1: Kurzübersicht über den Ansatz zur Vorhersage von fRNA-kodierenden Genen und strukturbasierten regulatorischen Elementen mit Angabe der Standardparameter.

Im ersten Schritt nehmen wir an, dass homologe fRNAs in verwandten Organismen nicht nur strukturelle Gemeinsamkeiten aufweisen, sondern auch auf der Sequenzebene ein ausreichend hohes Maß an Ähnlichkeiten besitzen. Mit Hilfe eines Vergleichs der Genomsequenzen treffen wir eine Vorauswahl an Kandidaten anhand ihrer Sequenzkonservierung untereinander. Ein deutlicher Hinweis auf eine fRNA in einer Sequenz ist eine stabile und konservierte Sekundärstruktur. Da die Strukturkonservierung einer Sequenz ebenfalls nur im Vergleich mit anderen Sequenzen untersucht werden kann, bewerten wir die Strukturen der Sequenzen mit einem ausreichend hohen Maß an Sequenzähnlichkeit gemeinsam. Diese Herangehensweise soll vor allem Kandidaten für bisher unbekannte fRNAs aufzeigen, da keine a priori Informationen über bekannte fRNA-Familien verwendet werden. Ein ähnlicher Ansatz wurde bereits von Axmann *et al.* [6] erfolgreich zur Vorhersage von fRNAs in Cyanobakterien angewandt. Im Gegensatz zu Axmann *et al.*, die alignierte Sequenzen ausschließlich auf der Grundlage des z-Scores beurteilen, wenden wir RNAz (Abschnitt 1.4.4) an, um nicht nur den Konservierungsgrad, sondern auch die Stabilität der Sekundärstrukturen zu beurteilen.

Im zweiten Schritt werden die gegebenen Sequenzen nach Sequenz- und Strukturähnlichkeiten zu bekannten fRNA-Familien abgesucht. Dieser Vergleich liefert detaillierte Informationen über die erhaltenen Kandidaten und ermöglicht sogar Rückschlüsse auf die Funktion der potentiellen fRNAs.

2.1 Auswahl der Daten

Nur wenn in den zu untersuchenden Sequenzen zu einer fRNA mindestens eine zweite homologe fRNA existiert, haben wir eine Chance, diese mit Hilfe des komparativen Teils der Strategie zu finden. Sollen fRNA-Kandidaten in einem Genom identifiziert werden, so ist es daher empfehlenswert, mindestens zwei verwandte Genome zum Vergleich heranzuziehen. Damit erhöhen wir die Wahrscheinlichkeit, dass es zu jeder fRNA, auch wenn diese nur ein einziges Mal in einem Genom auftritt, mindestens eine homologe fRNA im gesamten Datensatz gibt. Die Anzahl der Vergleichsgenome sollte nicht zu groß sein, um den Rechenaufwand zu beschränken. Wir empfehlen mindestens zwei aber maximal fünf Vergleichsgenome zu verwenden.

Wir erwarten, dass fRNA in *intergenischen Regionen* (IGR), d. h. DNA-Regionen

zwischen (bekannten) Genen, vorkommen und beschränken die Suche nach fRNA-Kandidaten daher auf diese Regionen. An dieser Stelle unterscheiden wir nicht zwischen Protein-kodierenden Genen und fRNA-kodierenden Genen. Für tRNA- und rRNA-kodierende Gene gibt es häufig zuverlässige Annotationen. Wir können die bekannten kodierenden Bereiche von unserer Suche ausschließen und den Suchraum damit weiter einschränken.

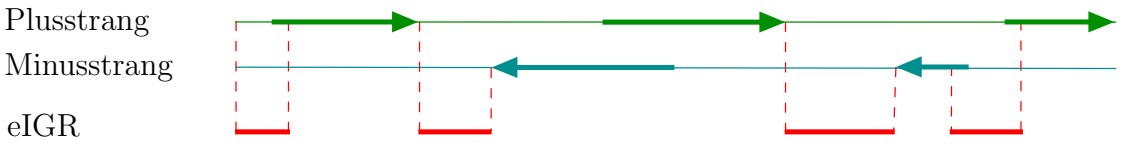


Abb. 2.2: Darstellung der eIGRs am Beispiel eines DNA-Abschnittes. In den ersten beiden Zeilen sind Gene auf dem jeweiligen Strang (Plus- bzw. Minusstrang), auf dem sie kodiert sind, abgebildet. In der dritten Zeile sind die daraus resultierenden eIGRs dargestellt. Die roten gestrichelten Linien geben an, von welchem Genbeginn oder -ende die jeweilige eIGR abgeleitet ist. Dabei ist zu beachten, dass die eIGRs mit einer konstanten Anzahl an bp in die kodierenden Regionen der Gene hineinragen.

Regulatorische Elemente wie z. B. Riboswitches [75, 99] können sich teilweise mit der kodierenden Region eines Gens überlappen. Um sie in unseren Suchraum aufzunehmen, erweitern wir die intergenischen Regionen um eine konstante Anzahl an bp (Voreinstellung = 50 bp) in diese Region hinein und erhalten so *erweiterte intergenische Regionen* (eIGR). Siehe dazu die Abbildung 2.2. Die eIGR-Sequenzen aller zu untersuchenden Genome werden in einer Datenbank zusammengefasst. Sie sind der Ausgangspunkt unserer Suche.

Die im Folgenden vorgestellte Methode ist zwar für die Untersuchung von Genomsequenzen konzipiert worden, kann aber auch genausogut auf eine Auswahl beliebiger DNA-Sequenzen angewandt werden, in denen homologe fRNAs vermutet werden. Eine mögliche Anwendung ist in Kapitel 5 beschrieben.

2.2 Komparativer Ansatz

2.2.1 Suche nach homologen Sequenzen

Wir nutzen `NCBI-BlastN` [4], um nach verwandten Sequenzabschnitten zwischen Sequenzen in der eIGR-Datenbank zu suchen. Jede eIGR-Sequenz wird dabei mit jeder

anderen verglichen, egal zu welchem Genom sie gehört. So wird sichergestellt, dass nicht nur Ähnlichkeiten zwischen verschiedenen Organismen, sondern auch eventuelle Mehrfachvorkommen einer fRNA in einem Genom gefunden werden können. Um die Qualität der Ergebnisse zu sichern, werden die Treffer anhand des sogenannten *Expect value* (E-Value) beurteilt, der von BlastN berechnet wird. Der E-Value eines Treffers gibt die Anzahl derjenigen Treffer mit gleichem Score an, die erwartungsgemäß per Zufall in einer Datenbank der gegebenen Größe gefunden werden. Wir verwenden nur Treffer, deren E-Value kleiner ist als ein vorgegebener Wert (Voreinstellung: 10^{-5}) und die eine Mindestlänge von 50 bp erreichen.

Eine Sequenzähnlichkeit zwischen eIGRs verwandter Organismen ist noch kein ausreichender Hinweis auf eine fRNA. Es könnte sich dabei um zufällige Ähnlichkeiten handeln, oder um Teile von ursprünglich Protein-kodierenden Sequenzen, welche im Laufe der Evolution so stark mutiert sind, dass sie nicht mehr exprimiert werden. Um diese Wahrscheinlichkeit zu minimieren, untersuchen wir die Konsensus-Sekundärstruktur, derjenigen von BlastN als homolog identifizierten Sequenzen, mit dem Programm RNAz (Abschnitt 1.4.4). Nur falls eine gemeinsame, stabile und konservierte Sekundärstruktur in allen Sequenzen gefunden wird, werden diese als fRNA-Kandidaten angesehen.

In Abschnitt 2.2.2 bis Abschnitt 2.2.6 wird beschrieben, wie aus den Ergebnissen einer BlastN-Untersuchung multiple Sequenzalignments gebildet werden. Das Vorgehen wird für eine Anfragesequenz beschrieben. Alle Schritte müssen daher für alle Sequenzen in der eIGR-Datenbank wiederholt werden.

2.2.2 Homologiecluster

Die RNAz-basierte fRNA-Vorhersage beruht auf der Bewertung der Sekundärstrukturen aller Sequenzen eines Alignments. Es ist zu erwarten, dass neben der Auswahl der Sequenzen auch ihre Anzahl im Alignment bei der fRNA-Identifikation eine Rolle spielt. Je mehr Sequenzen miteinander verglichen werden, umso deutlicher treten konsistente und kompensatorische Mutationen (Abschnitt 1.2.5) hervor, die ein Zeichen für die Konservierung und damit die funktionelle Bedeutung der Struktur einer RNA sind. Daher ist es unser Ziel, möglichst viele homologe Sequenzen zusammenzufassen und zu alignieren. Da RNAz jedoch höchstens sechs Sequenzen auf einmal bearbeiten kann, setzt das eine obere Schranke für die Anzahl der Sequenzen eines

Alignments.

Sequenzähnlichkeiten werden auf der Grundlage der Ergebnisse eines BlastN-Vergleichs einer Anfragesequenz mit allen Sequenzen der eIGR-Datenbank zusammengefasst. Diese Ähnlichkeiten werden als *BlastN-Treffer* oder vereinfacht Treffer X bezeichnet und mit den folgenden Merkmalen identifiziert: (1) mit dem E-Value e^X, mit dem der Treffer von BlastN bewertet wurde, und (2) den zueinander alignierten Sequenzen

$$A^X = a_i^X, \ldots, a_j^X,$$

einer Teilsequenz der Anfragesequenz A und

$$T^X = t_k^X, \ldots, t_l^X,$$

einer Teilsequenz der Treffersequenz T, die zu A^X aligniert wurde (siehe Abbildung 2.3). Ein Treffer beschreibt damit zwei zueinander ähnliche Sequenzen.

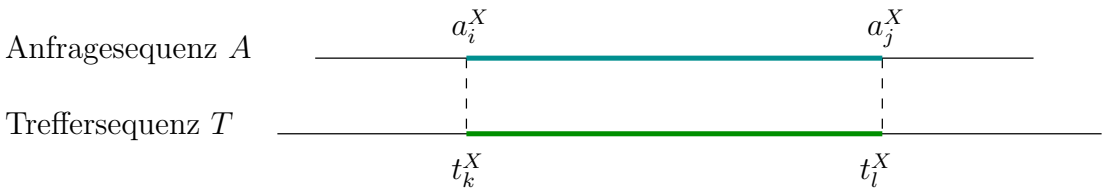


Abb. 2.3: Zusammensetzung eines BlastN-Treffers X.

Um Ähnlichkeiten zwischen mehr als zwei Sequenzen zu identifizieren, nutzen wir aus, dass mehrere Treffersequenzen zur gleichen Teilsequenz einer Anfragesequenz aligniert sein können. Im Folgenden sei $\mathcal{B}_A$ die Menge aller BlastN-Treffer zu einer Anfragesequenz A. Vergleichen wir zwei Treffer X und $Y \in \mathcal{B}_A$ über die dazugehörigen Teilsequenzen der Anfragesequenz $A^X = a_i^X, \ldots, a_j^X$ und $A^Y = a_k^Y, \ldots, a_l^Y$, so erwarten wir, dass die zu $A^X \cap A^Y$ alignierten Bereiche der Treffersequenzen von X und Y auch untereinander ähnlich sind.

Definition 2.1. (Homologiecluster)
Es sei $\mathcal{X} = \{X_1, \ldots, X_n\} \subset \mathcal{B}_A$ mit $n \in \mathbb{N}$. Dann ist ein Homologiecluster *H von $\mathcal{X}$ die Menge aller Teilsequenzen der Treffersequenzen von $X_1, \ldots, X_n$, die aligniert sind zu*

$$A^H := A^{X_1} \cap \ldots \cap A^{X_n},$$

inklusive der gemeinsamen Teilsequenz der Anfragesequenz A^H. Das Niveau *von H ist der maximale E-Value aller Treffer in $\mathcal{X}$.*

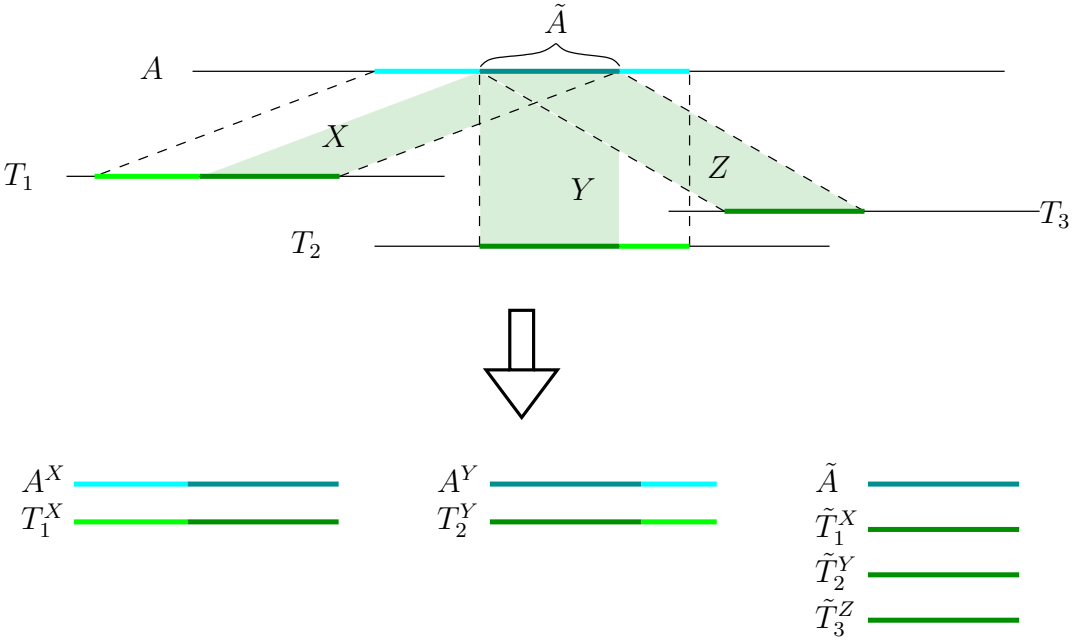


Abb. 2.4: Zu sehen ist eine Anfragesequenz A mit drei Treffersequenzen T_1, T_2, T_3. Die BlastN-Treffer X, Y und Z, d. h. die zueinander alignierten Teilsequenzen sind mit schwarzen, gestrichelten Linien verbunden und in der Sequenz farblich hervorgehoben. Der dunkelgrüne Abschnitt ist in allen Treffersequenzen homolog zur gleichen Teilsequenz der Anfragesequenz, wir bezeichnen sie hier mit $\tilde{A}$. Je nachdem, welche Treffer miteinander verglichen werden, enthält der dazugehörige Homologiecluster unterschiedliche Sequenzabschnitte der gleichen Treffersequenzen. Die unten angegebenen Beispielcluster zeigen nur einen Teil der möglichen Homologiecluster, die aus den gegebenen Treffern gebildet werden können. Der erste und zweite Cluster enthält nur die Teilsequenz der Anfrage- und der Treffersequenz des entsprechenden Treffers und bildet somit Homologiecluster mit der minimalen Anzahl an Sequenzen. Ganz rechts ist der Homologiecluster mit der maximalen Anzahl an Sequenzen, der aus den gegebenen Treffern X, Y und Z erzeugt werden kann, wobei $\tilde{T}_1^X$, $\tilde{T}_2^Y$ und $\tilde{T}_3^Z$ die zu $\tilde{A}$ alignierten Teilsequenzen der Treffersequenzen T_1, T_2 und T_3 sind.

Die Abbildung 2.4 zeigt Beispiele für mögliche Homologiecluster zu gegebenen BlastN-Treffern. Es wird deutlich, dass die Länge der Sequenzen in einem Homologiecluster von der Auswahl der Treffer abhängt, aus denen der Homologiecluster gebildet wird. Der Homologiecluster kann sogar leer sein, falls zwei der Treffer keine

gemeinsame Teilsequenz der Anfragesequenz haben. Im Folgenden suchen wir Homologiecluster mit maximal sechs Sequenzen, wobei die Anfragesequenz darin bereits berücksichtigt ist. Das entspricht der maximalen Sequenzanzahl, die in Form eines Alignments mit RNAz beurteilt werden kann. Die Menge der Homologiecluster für eine Anfragesequenz A, die letztendlich aligniert und mit RNAz untersucht werden, bezeichnen wir fortan mit RNAz-Testmenge $\mathcal{T}_A$. Die Testmenge wird in zwei Schritten zusammengesetzt, die in Abschnitt 2.2.4 und Abschnitt 2.2.5 beschrieben sind. Nicht alle Trefferkombinationen, die aus den Treffern zu einer Anfragesequenz gebildet werden können, gehen in Form eines Homologieclusters in die RNAz-Testmenge ein. Im folgenden Abschnitt beschreiben wir die Kriterien, nach denen solche Trefferkombinationen beurteilt und in die Testmenge aufgenommen werden. Nur eine Trefferkombination, die neue Informationen über die Anfragesequenz preisgeben könnte, d. h. Informationen, die noch nicht in $\mathcal{T}_A$ enthalten sind, wird aufgenommen.

2.2.3 Kriterien für die Aufstellung der RNAz-Testmenge

Um den Rechenaufwand bei der RNAz-Untersuchung nicht unnötig zu vergrößern, werden nur bestimmte Homologiecluster in die RNAz-Testmenge $\mathcal{T}_A$ aufgenommen. Dazu wird ein neuer Homologiecluster mit allen Homologieclustern, die bereits in $\mathcal{T}_A$ enthalten sind, verglichen. Um entscheiden zu können, welcher Homologiecluster aufgenommen wird und welcher nicht, vergleichen wir zwei Homologiecluster I und J bezüglich

1. $A^I = a_i^I, \ldots, a_j^I$ und $A^J = a_k^J, \ldots, a_l^J$,
2. der in ihnen enthaltenen Teilsequenzen der Anfragesequenz A,
3. der Gesamtanzahl der enthaltenen Sequenzen und
4. ihres Niveaus, d. h. des höchsten E-Values aller Treffer eines Homologieclusters.

Die Abschnitte der Anfragesequenzen verschiedener Homologiecluster unterscheiden sich teilweise nur um wenige bp, und dieser Unterschied fällt bei unserem Vergleich kaum ins Gewicht. Wir führen daher eine Konstante $c > 0$ ein. Unterscheiden sich A^I und A^J maximal um diese konstante Anzahl an bp, d. h.

$$\left|a_i^I - a_k^J\right| \leq c \qquad \text{und} \qquad \left|a_j^I - a_l^J\right| \leq c,$$

dann behandeln wir A^I und A^J als ob sie gleich wären.

Vergleich von Homologieclustern

Für zwei Homologiecluster kann gelten, dass sie (a) gleichwertig sind, (b) einer besser ist als der andere, oder (c) sie nicht vergleichbar sind:

(a) Zwei Homologiecluster I und J sind *gleichwertig* falls gilt:

1. A^I und A^J unterscheiden sich maximal um eine konstante Anzahl $c > 0$ an bp.
2. I und J weisen die gleiche Sequenzanzahl auf.
3. I und J weisen das gleiche Niveau auf.

(b) Ein Homologiecluster I ist *besser* als ein Homologiecluster J, falls gilt:

1. Die Abschnitte ihrer Anfragesequenzen, A^I und A^J, unterscheiden sich maximal um eine konstante Anzahl $c > 0$ an bp, oder A^J ist in A^I enthalten.
2. I enthält mindestens soviele Sequenzen wie J.
3. Das Niveau von I ist höchstens so hoch wie das Niveau von J.

In mindestens einer der Bedingungen darf keine Gleichheit herrschen.

(c) In allen anderen Fällen werden zwei Homologiecluster als *nicht vergleichbar* angesehen.

Die Vergleichskriterien zweier Homologiecluster sind deshalb so vorsichtig gewählt, damit möglichst keine relevanten Sequenzzusammenhänge verloren gehen. Wir versuchen an dieser Stelle, ein Gleichgewicht zwischen der Vielfalt der Informationen und der Optimierung der Rechenlaufzeit zu erreichen.

Wann wird ein Homologiecluster I zur RNAz-Testmenge $\mathcal{T}_A$ hinzugefügt:

Falls $\mathcal{T}_A = \emptyset$:

- füge I zu $\mathcal{T}_A$ hinzu.

Sonst: Für alle $J \in \mathcal{T}_A$, vergleiche I und J:

Falls J besser ist als I, oder J und I gleichwertig sind:

- verwerfe I.

Sonst, falls I besser ist als J:

- entferne J aus $\mathcal{T}_A$ und
- füge I zu $\mathcal{T}_A$ hinzu.

Sonst

- füge I zu $\mathcal{T}_A$ hinzu.

2.2.4 Trefferklassen und ihre Komprimierung

Der Vergleich einer Anfragesequenz mit den Sequenzen der gesamten eIGR-Datenbank kann viele Treffer hervorbringen, deren Treffersequenzenabschnitte zu einem sehr ähnlichen Abschnitt der Anfragesequenz aligniert werden. Es können aber nur jeweils sechs dieser Sequenzen mit Hilfe von RNAz in einem Alignment untersucht werden. Wir verlieren keine fRNA-Kandidaten, wenn wir nur die besten Treffer mit der Anfragesequenz vergleichen, da die Treffersequenzen selbst ein Mal als Anfragesequenzen dienen und somit im Mittelpunkt der Untersuchung stehen. Um den Rechenaufwand zu reduzieren, vergleichen wir daher alle Treffer untereinander und fassen sie in Gruppen mit besonders starker Ähnlichkeit zusammen. Diese Gruppen können dann auf einige repräsentative Vertreter reduziert werden.

Definition 2.2. (X-Trefferklasse)
Für einen Treffer $X \in \mathcal{B}_A$ und eine Konstante $c > 0$ ist die X-Trefferklasse *eine Menge, die aus genau den Treffern $Y \in \mathcal{B}_A$ besteht, deren Anfragesequenzabschnitt A^Y sich um maximal c bp von dem Anfragesequenzabschnitt A^X unterscheidet.*

Die X-Trefferklasse wird immer bezüglich eines Treffers X aufgestellt (siehe Abbildung 2.5). Wird die Qualität der Treffer einer X-Trefferklasse an Hand ihrer E-Values verglichen, so hat nicht notwendigerweise der namensgebende Treffer X den besten, d. h. niedrigsten E-Value. Daher die folgende Definition:

Definition 2.3. (Repräsentant einer X-Trefferklasse)
Als Repräsentant *einer X-Trefferklasse wird derjenige Treffer aus dieser Trefferklasse bezeichnet, welcher den niedrigsten E-Value aufweist.*

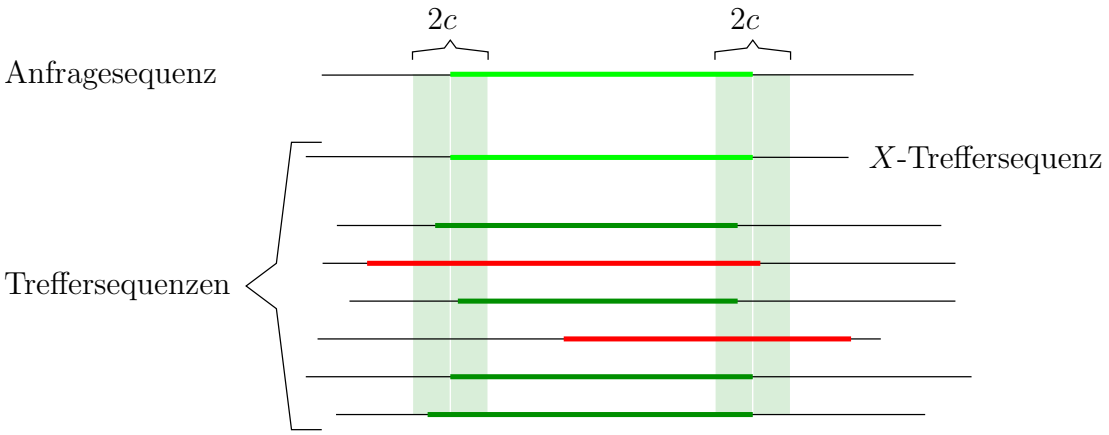


Abb. 2.5: Beispiel einer X-Trefferklasse: In hellgrün sind die zueinander homologen Teilsequenzen der Anfrage- und Treffersequenz des Treffers X dargestellt. Alle dunkelgrünen Sequenzen gehören zu Treffern der X-Trefferklasse laut Definition 2.2. Die vertikalen Balken markieren den Sequenzabschnitt, in dem alle Treffer dieser Trefferklasse beginnen bzw. enden müssen, wobei die Treffersequenzen bezüglich ihres zur Anfragesequenz homologen Bereichs ausgerichtet sind. Die roten Sequenzen gehören zu Treffern, die nach Definition kein Element der hier dargestellten X-Trefferklasse sind.

Es ist zu erwarten, dass alle Kombinationen von Treffern, die zu derselben Trefferklasse gehören, ein vergleichbares Ergebnis bei der Untersuchung mit RNAz liefern. Wir reduzieren daher Trefferklassen mit mehr als fünf Elementen auf die fünf besten Treffer. Dabei beurteilen wir die Qualität der Treffer einer Trefferklasse an Hand ihres E-Values.

Algorithmus zur Behandlung von Trefferklassen:

Initialisierung:
$\mathcal{B} \leftarrow \mathcal{B}_A$

Methodenrumpf:

1. Sortiere alle Treffer in $\mathcal{B}$ aufsteigend nach ihrer Startposition in der Anfragesequenz A.
2. $X \leftarrow$ Erster Treffer aus $\mathcal{B}$ mit der kleinsten Startposition in A.
3. Bestimme die X-Trefferklasse in $\mathcal{B}$ für eine gegebene Konstante $c > 0$.
4. Falls die X-Trefferklasse mehr als fünf Treffer enthält:

- Bestimmte die fünf Treffer der X-Trefferklasse, welche den niedrigsten E-Value aufweisen.
- Entferne alle anderen Treffer aus $\mathcal{B}$ und der X-Trefferklasse.

5. Falls die X-Trefferklasse mindestens zwei Treffer enthält und der gemeinsame Abschnitt der Anfragesequenz mindestens 50 bp lang ist:
   - Bestimme den Homologiecluster der reduzierten X-Trefferklasse und füge ihn zur RNAz-Testmenge $\mathcal{T}_A$ hinzu, falls kein besserer Homologiecluster bereits in $\mathcal{T}_A$ enthalten ist.
6. Entferne alle Treffer der X-Trefferklasse bis auf den Repräsentanten dieser Trefferklasse aus $\mathcal{B}$.
7. Wähle den Treffer, der X in $\mathcal{B}$ nachfolgt bzw. falls X entfernt wurde, dessen Platz eingenommen hat, und wiederhole die Schritte drei bis sieben, wobei X durch den neuen Treffer ersetzt wird.

Abbruchbedingung:
Der Abbruch erfolgt, wenn X keinen Nachfolger mehr in $\mathcal{B}$ hat.

Nach der Behandlung der Trefferklassen bleiben nur ihre Repräsentanten in $\mathcal{B}$ vertreten. Im nächsten Abschnitt wird beschrieben, wie diese Repräsentanten untereinander verglichen und ebenfalls in Form von Homologieclustern in die RNAz-Testmenge $\mathcal{T}_A$ aufgenommen werden, wenn sie gewisse vorgegebene Bedingungen erfüllen.

2.2.5 Behandlung der Repräsentanten aller Trefferklassen

Wir erwarten, dass BlastN eine konservierte, fRNA-kodierende Sequenz in Genomsequenzen verwandter Organismen identifizieren kann. Je nach Verwandtschaftsgrad der Organismen fällt die Ähnlichkeit der Genomsequenzen stärker oder schwächer aus. Das kann dazu führen, dass die mit BlastN ermittelten Treffer, welche im Kern die gleiche fRNA beschreiben, unterschiedlich lange Sequenzbereiche angeben. Entweder sind zwei der Organismen so nahe verwandt, dass auch die Umgebung der fRNA-kodierenden Sequenz in einen Treffer eingeht, was für andere Organismen vielleicht nicht der Fall ist. Oder sie sind evolutionär so weit voneinander entfernt, dass BlastN zwar noch eine Sequenzähnlichkeit findet, es sich dabei aber nicht um

die vollständige kodierende Sequenz der fRNA handelt. Da wir nach multiplen Sequenzähnlichkeiten suchen, betrachten wir daher alle Kombinationen aus den Repräsentanten der Trefferklassen, um gemeinsame Sequenzähnlichkeiten zu identifizieren. Aus jeder Trefferkombination von bis zu fünf Treffern wird ein Homologiecluster gebildet und in die RNAz-Testmenge hinzugefügt, falls kein besserer Homologiecluster dort bereits enthalten ist.

2.2.6 Alignment der Homologiecluster

Zum Alignieren homologer Sequenzen benutzen wir ClustalW [98], da RNAz auf ClustalW-Alignments trainiert wurde, und mit diesen Alignments die besten Ergebnisse zu erwarten sind. Bei einer mittleren Sequenzidentität von mehr als 55 % hat ClustalW in Benchmark-Tests außerdem gut abgeschnitten [39, 104]. In diesen Tests wurden unter anderem reine Sequenz-Alignment-Methoden mit Methoden, die sowohl Sequenz- als auch Strukturinformationen alignieren, verglichen. Ab einer durchschnittlichen Sequenzidentität von über 65 % haben beide Alignmentstrategien vergleichbar gut abgeschnitten. Daher ist es nicht notwendig, die langsameren Sequenz-Struktur-Alignmentmethoden zu verwenden, auch wenn wir die konservierte Struktur im Alignment untersuchen wollen. Das wurde auch in [102] bestätigt. Da die Sequenzen auf der Grundlage von BlastN-Alignments zusammengefasst werden, ist damit eine ausreichend hohe durchschnittliche Sequenzidentität gewährleistet.

2.2.7 Strukturbewertung mit RNAz

Jedes Alignment wird mit Hilfe von RNAz (siehe Abschnitt 1.4.4) auf das Vorkommen stabiler und konservierter Strukturen untersucht. Das Programm ist in der Lage, ein Alignment bis zu einer Länge von 400 Positionen zu behandeln. Ist das Alignment länger, kann es fensterweise betrachtet werden. Der Rechenaufwand steigt kubisch mit der Alignmentlänge an. Um den Aufwand zu beschränken, wird eine maximale Fensterlänge festgelegt. Ist das betrachtete Alignment länger und enthält es eine größere zusammenhängende Struktur, so besteht diese in der Regel aus kleineren Einzelkomponenten, die mit einem kleinen Fenster gefunden werden können. Dennoch kann ein zu kleines Fenster eine größere Struktur eventuell nicht erfassen. In einem zu großen Fenster kann das Signal kleiner stabiler Strukturen wiederum

im Rauschen der Umgebung untergehen. Daher untersuchen wir alle Alignments mit Fenstern unterschiedlicher Länge (Voreinstellung: $50 - 200$ bp) und vorgegebener Schrittweite (Voreinstellung: fünf bp). Nur die Sequenzabschnitte im Alignment, deren RNAz-Score über einer von uns vorgegebenen Schranke (Voreinstellung: 0,9) liegt, werden hier als fRNA-Kandidaten angesehen. In Abbildung 2.6 ist an einem Beispiel dargestellt, wie die Größe und Menge der fRNA-Kandidaten von der Wahl einer geeigneten Schranke abhängt.

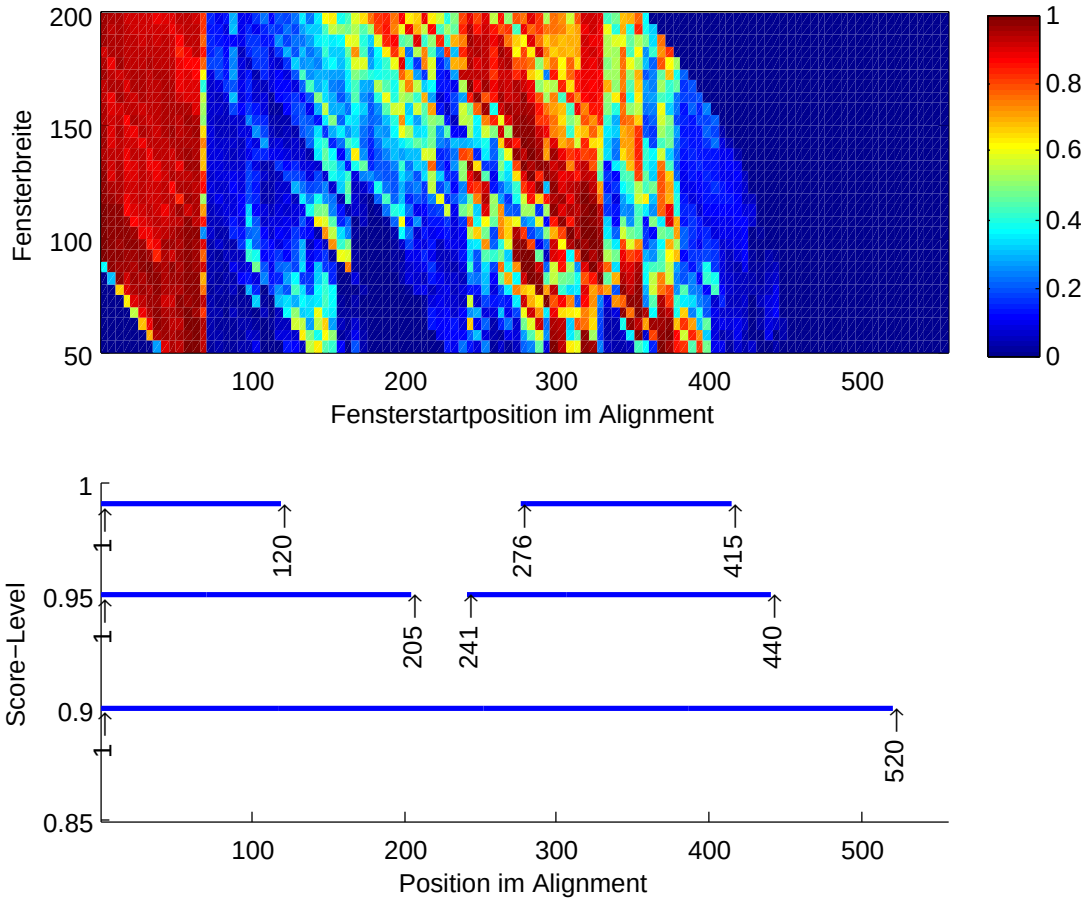


Abb. 2.6: Oben: Matrix mit Vorhersagescores von RNAz für ein Beispielalignment, das mit verschiedenen Fenstergrößen ($50 - 200$ bp) in Schrittweiten von fünf bp und an verschiedenen Startpositionen des Fensters im Alignment (Schrittweite = fünf bp) untersucht wurde. Unten: Als fRNA vorhergesagte Bereiche des Alignments in Abhängigkeit von vorgegebenen Scoreschranken von 0,9 sowie 0,95 und 0,99.

2.3 Kovarianzmodell-basierter Ansatz

Mit Hilfe des Programmpakets INFERNAL (siehe Abschnitt 1.4.5) suchen wir nach Ähnlichkeiten zu bekannten fRNA-Familien. Dabei werden sowohl Sequenz- als auch Strukturinformationen in Form eines Kovarianzmodells für die Suche verwendet. Jede fRNA-Familie, die durch ein Alignment und eine Konsensus-Sekundärstruktur repräsentiert wird, kann in ein Kovarianzmodell überführt werden. Im Rahmen dieser Arbeit haben wir fRNA-Familien-Informationen aus der `Rfam`-Datenbank (siehe Abschnitt 1.3.2) für die Suche herangezogen. Spielt der Rechenaufwand keine Rolle, so empfiehlt es sich, die gesamte eIGR-Datenbank zu untersuchen. Auf diese Weise können Kandidaten gefunden werden, deren Struktur eventuell nicht stark konserviert ist oder keine signifikante Stabilität aufweist und damit vom komparativen Ansatz nicht entdeckt werden kann. Zur Reduktion des Rechenaufwands kann INFERNAL aber auch direkt auf die fRNA-Kandidaten des komparativen Ansatzes angewandt werden. Damit wird die Signifikanz der Kandidaten überprüft und wir gewinnen zusätzliche Informationen über eine mögliche Funktion der fRNA. Es ist möglich, dass ein Kandidat aus dem komparativen Ansatz nur einen Teil einer fRNA beschreibt und mehrere, dicht aufeinanderfolgende Kandidaten zu einer fRNA gehören. Daher wird nicht nur der kodierende Bereich der Kandidaten in der DNA untersucht, sondern ebenfalls die nähere Umgebung (± 50 bp).

Jede Zielsequenz wird mit jeder fRNA-Familie aus der Datenbank `Rfam` verglichen. Da eine fRNA-kodierende Sequenz im Laufe der Zeit vielen Einflüssen unterworfen ist, kann sie unter Umständen nicht mehr global zum Kovarianzmodell einer Familie homologer fRNA aligniert werden. Wir führen daher einen lokalen Vergleich mit dem Parameter `--local` durch. Das lokale Alignieren eines Kovarianzmodells zu einer Zielsequenz kann zu mehreren Teilalignments führen, die auf der Sequenzebene unterbrochen sind, in der Struktur jedoch zusammenhängen. Eine Beschreibung des Phänomens ist in Abschnitt 1.4.5 zu finden.

2.4 Implementierung

Im Rahmen dieser Arbeit wurde der in Kapitel 2 beschriebene Ansatz ebenfalls implementiert. Die Anwendung ermöglicht eine fRNA-Vorhersage in langen Sequenzen, z. B. Genomsequenzen, ohne das notwendige Expertenwissen für die einzelnen

Werkzeuge aufbringen zu müssen. Die in sich abgeschlossenen Schritte, wie sie in Abbildung 2.1 dargestellt sind, wurden in der Programmiersprache PERL realisiert und werden mit Hilfe eines BASH-Skripts gestartet.

Die Implementierung wurde darauf ausgerichtet, dass Aufgaben, bei denen viele Datensätze unabhängig voneinander bearbeitet werden können, mit Hilfe der N1 GRID ENGINE 6-Software von SUN MICROSYSTEMS auf einem Cluster verteilt und parallel abgearbeitet werden. Dazu gehört unter anderem das Zusammenfassen multipler Sequenzähnlichkeiten für jede Anfragesequenz, das Alignieren zusammengefasster Sequenzen, die RNAz-Bewertung der Alignments, sowie die Suche nach Ähnlichkeiten zu bekannten fRNAs mit Hilfe von INFERNAL. Die Parallelisierung dieser Schritte bringt, je nach Anzahl der CPUs des verwendeten Clusters, einen enormen Zeitgewinn.

Die in dieser Arbeit vorgestellte Anwendung bindet die folgenden Werkzeuge ein:

- BlastN in der Version 2.2.15,
- ClustalW in der Version 1.83,
- RNAz in der Version 1.0
- und INFERNAL in der Version 0.81.

2.4.1 Einstellbare Parameter

Alle Parameter sind bereits voreingestellt, können aber vom erfahrenen Benutzer angepasst werden. Im Folgenden sind die wichtigsten Parameter zusammengefasst:

E-Value:

Der E-Value für die BlastN-Suche im komparativen Ansatz ist auf 10^{-5} voreingestellt. Je niedriger der Wert gewählt wird, umso weniger Treffer werden gefunden und umso signifikanter ist die Ähnlichkeit zwischen der Anfragesequenz und dem Treffer. Da dieser Schritt nur zur Vorauswahl interessanter Sequenzregionen dient, sollte der E-Value nicht zu niedrig und damit zu strikt gewählt werden.

Mindestanzahl der Sequenzen im Alignment:

Auf der Grundlage der BlastN-Treffer werden Ähnlichkeiten zwischen mehreren Sequenzen zusammengefasst und aligniert. Es ist möglich zu bestimmen, wieviele Sequenzen mindestens in einem Alignment enthalten sein müssen. Die Alignments werden in der letzten Phase des komparativen Ansatzes auf das Vorkommen stabiler und konservierter Strukturen untersucht. Je mehr Sequenzen ein Alignment enthält, umso genauer kann eine eventuell konservierte Struktur beurteilt werden. Diese Beurteilung wird mit Hilfe von RNAz durchgeführt. RNAz ist in der Lage, Alignments mit bis zu sechs Sequenzen zu behandeln. Beim Festlegen der Anzahl sollte bedacht werden, dass zu hohe Werte, wie z. B. sechs, dazu führen, dass keine fRNAs mehr gefunden werden, die in einer geringeren Stückzahl als sechs im gesamten Datensatz auftreten. Um keine Kandidaten zu übersehen, wird in der Voreinstellung nur die Mindestanzahl von zwei Sequenzen erwartet.

RNAz-Fenster:

Die Untersuchung der Strukturstabilität und -konservierung innerhalb eines Alignments erfolgt fensterweise. Da die daraus resultierenden Ergebnisse von der Fensterposition im Alignment und der Fenstergröße abhängen, wird die Untersuchung mit variablen Fenstergrößen durchgeführt. Es ist möglich, sowohl die minimale als auch die maximale Fenstergröße einzustellen. Die minimale Fenstergröße ist auf 50 bp voreingestellt. Kleiner darf ein Fenster nicht sein. Es kann aber bis zu einer Größe von 400 bp erweitert werden. Die maximale Fenstergröße ist auf 200 bp voreingestellt und kann auf einen Wert zwischen 50 und 400 bp eingestellt werden. Die maximale Fenstergröße muss mindestens so hoch sein wie die minimale Fenstergröße. Ebenso ist die Schrittweite für die Verschiebung der Fenster und die Fenstervergrößerung einstellbar. Beide Werte sind auf fünf bp voreingestellt, können aber in ein bp-Schritten verändert werden.

Persönliche Signifikanzschwelle für Scores:

Für die Ausgabe der Ergebnisse kann angegeben werden, ab welchem RNAz-Score die Ergebnisse des komparativen Ansatzes und ab welchem INFERNAL-Score die Ergebnisse des Kovarianzmodell-basierten Ansatzes angegeben werden sollen. Da-

mit legt man die persönliche Signifikanzgrenze für die Scores fest. Dieser Schritt ist unabhängig von der eigentlichen Vorhersage und kann mit verschiedenen Signifikanzgrenzen wiederholt werden. Die Signifikanzschwelle ist für den RNAz-Score auf 0,9 (Minimum = 0,5) und für den INFERNAL-Score auf 20 voreingestellt.

2.4.2 Format der Eingabedaten

Als Eingabedaten werden die Genomsequenzen im FASTA-Format erwartet. Eine Sequenz im FASTA-Format wird eingeleitet durch das „>“-Symbol, den Sequenznamen und eventuell weiteren Informationen, die alle in einer Zeile stehen. In den darauffolgenden Zeilen steht die Sequenz.

Beispiel 1 FASTA-Format

```
> Escherichia coli K12 MG1655, complete genome.
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTC
TGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGG
TCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTAC
ACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
...
```

Für die Bestimmung der eIGRs werden die Koordinaten der Protein-kodierenden Gene benötigt. Sie werden im folgenden Format angegeben:

```
Startkoordinate    Stoppkoordinate
```

Das Verhältnis der Koordinaten zueinander definiert den Strang, auf dem das aktuelle Gen liegt. Ist die Startkoordinate kleiner als die Stoppkoordinate, so befindet sich das Gen auf dem Plusstrang; ist die Startkoordinate größer, so befindet es sich auf dem Minusstrang. Diese Information wird benötigt, um intergenische Regionen in die kodierende Region ausdehnen zu können und somit die eIGRs zu erhalten.

Beispiel 2 Format der Koordinaten

```
6459       5683
926697     928418
1939659    1938337
2639853    2640866
...
```

Sind bereits einige fRNAs bekannt, wie z. B. tRNA- oder rRNA-kodierende Gene, können auch diese Koordinaten im gleichen Format angegeben werden. Diese Sequenzabschnitte werden ebenfalls aus den eIGRs herausgeschnitten. Die Reihenfolge der Koordinaten hat in diesem Fall keine Auswirkung. Die IGRs werden in diesem Fall nicht in die kodierenden Regionen ausgedehnt, da wir bereits wissen, dass es sich dabei um fRNA handelt.

2.4.3 Format der Ausgabedaten

Alle Zwischenergebnisse, wie BlastN-Ausgaben, ClustalW-Alignments, RNAz-Vorhersagescores und die Ausgabe von INFERNAL werden gespeichert und stehen für genauere Betrachtung zur Verfügung.

GFF-Format

Die fRNA-Kandidaten aus dem komparativen und dem Kovarianzmodell-basierten Ansatz werden jeweils im GFF-Format angegeben. Es ermöglicht die Angabe der wichtigsten Informationen zu den Kandidaten und erleichtert die Visualisierung der Ergebnisse mit einem geeigneten Werkzeug, wie z. B. Artemis [88]. Die Einträge des GFF-Formats sind in der Tabelle 2.1 zusammengefasst.

Darstellung der Kandidaten des komparativen Ansatzes

Da die Kandidaten mit Hilfe eines komparativen Ansatzes bestimmt werden, erhalten wir jeweils Gruppen ähnlicher fRNA-Kandidaten. Dieser Zusammenhang wird in der GFF-Datei unter dem Attribut `note` festgehalten. Zu jedem fRNA-Kandidaten wird dabei eine Liste von Sequenznamen mit Koordinaten angegeben. Sie beschreiben fRNA-Kandidaten in anderen Sequenzen, die eine Ähnlichkeit mit dem aktuellen Kandidaten aufweisen.

Beim komparativen Ansatz wird auf die Angabe der Strangorientierung verzichtet. Die fRNA-Vorhersage basiert bei diesem Ansatz auf der Beurteilung der gemeinsamen Struktur in einem Alignment. Da aber für komplementäre Sequenzen nahezu identische Strukturen vorhergesagt werden, können wir die Orientierung des fRNA-Kandidaten nicht festlegen. Daher werden überlappende Kandidaten zusammengefasst und ohne Strangorientierung angegeben.

Eintrag	Beschreibung
< Sequenzname >	Name der aktuellen Sequenz. In einer GFF-Datei können Einträge zu verschiedenen Sequenzen enthalten sein.
< Quelle >	Name der Datenbank, aus welcher dieser Eintrag stammt oder des Programms, mit dem es vorhergesagt wurde.
< Merkmal >	Name des Merkmaltyps, hier: ncRNA
< Start >	Startposition in gegebener Sequenz
< Ende >	Endposition in gegebener Sequenz
< Score >	Score, falls vorhanden, sonst '.'
< Strang >	'+' oder '-', sonst '.', falls nicht bekannt
< Leserahmen >	'0', '1' oder '2', sonst '.', falls nicht relevant
[Attribute]	Siehe Beschreibung bei `EMBL-EBI`.
[Kommentare]	Weitere, bisher nicht erfasste Informationen.

Tab. 2.1: Einträge des GFF-Formats. Die Einträge in spitzen Klammern müssen angegeben werden. Ist eine Eigenschaft nicht bekannt oder nicht relevant, so wird sie durch '.' als Platzhalter ersetzt. Einträge in eckigen Klammern sind optional. Standardspezifikationen der Merkmale und Attribute sind auf der Internetseite von `EMBL-EBI` zu finden [35].

Beispiel 3 Darstellung der Kandidaten des komparativen Ansatzes im GFF-Format. Für eine Beschreibung des Formats siehe Tabelle 2.1.

```
...
U00096    RNAz    ncRNA   160565  160639  0.9633  .   .   note  \\
  "U00096{[160565,160639]};AE014073{[151834,151908]};          \\
  AE014613{[221213,221287]};"
U00096    RNAz    ncRNA   160650  160744  0.9821  .   .   note  \\
  "U00096{[160650,160744]};AE014073{[151919,152013]};          \\
  AE014613{[221298,221391]};"
U00096    RNAz    ncRNA   164535  164779  1.0000  .   .   note  \\
  "U00096{[164535,164779][288458,288512][3237609,3237668]      \\
  [3957469,3957526]};AE014073{[155804,156048][3460618,3460724] \\
  [4395355,4395461]};AE014613{[225194,225368]};"
...
```

Darstellung der Kandidaten des Kovarianzmodell-basierten Ansatzes

Auch die mit INFERNAL vorhergesagten Kandidaten werden im GFF-Format ausgegeben. Jeder Kandidat wird durch seine Ähnlichkeit zu einer bekannten fRNA-

Familie lokalisiert. Diese Information wird unter dem Attribut `product` angegeben. Dabei wird der Name der fRNA-Familie mit der in `Rfam` verwendeten Kennung kombiniert. Im Gegensatz zum komparativen Ansatz, wird die Orientierung eines Kandidaten mit Hilfe von INFERNAL eindeutig festgelegt.

Beispiel 4 Darstellung der von INFERNAL vorhergesagten Kandidaten im GFF-Format. Für eine Formatbeschreibung siehe Tabelle 2.1.

```
...
U00096    Infernal    ncRNA    2753615    2753976 246.78  +    .    \\
  product=tmRNA_RF00023;
U00096    Infernal    ncRNA    2069339    2069542 238.70  +    .    \\
  product=IS102_RF00124;
U00096    Infernal    ncRNA    2702036    2702245 235.01  -    .    \\
  product=rncO_RF00552;
...
```

Visualisierung der GFF-Dateien mit Artemis

Die Ergebnisse sind so aufbereitet, dass sie mit Artemis [88] dargestellt werden können. Artemis ist ein Sequenz-Visualisierungs- und Annotations-Werkzeug. Es ermöglicht die Betrachtung der fRNA-Kandidaten im Genomkontext und ist besonders für die Analyse kompakter Genome von Bakterien, Archaeen und einfacher Eukaryoten geeignet. Abbildung 2.7 zeigt ein Beispiel für die Darstellung der Daten mit Artemis.

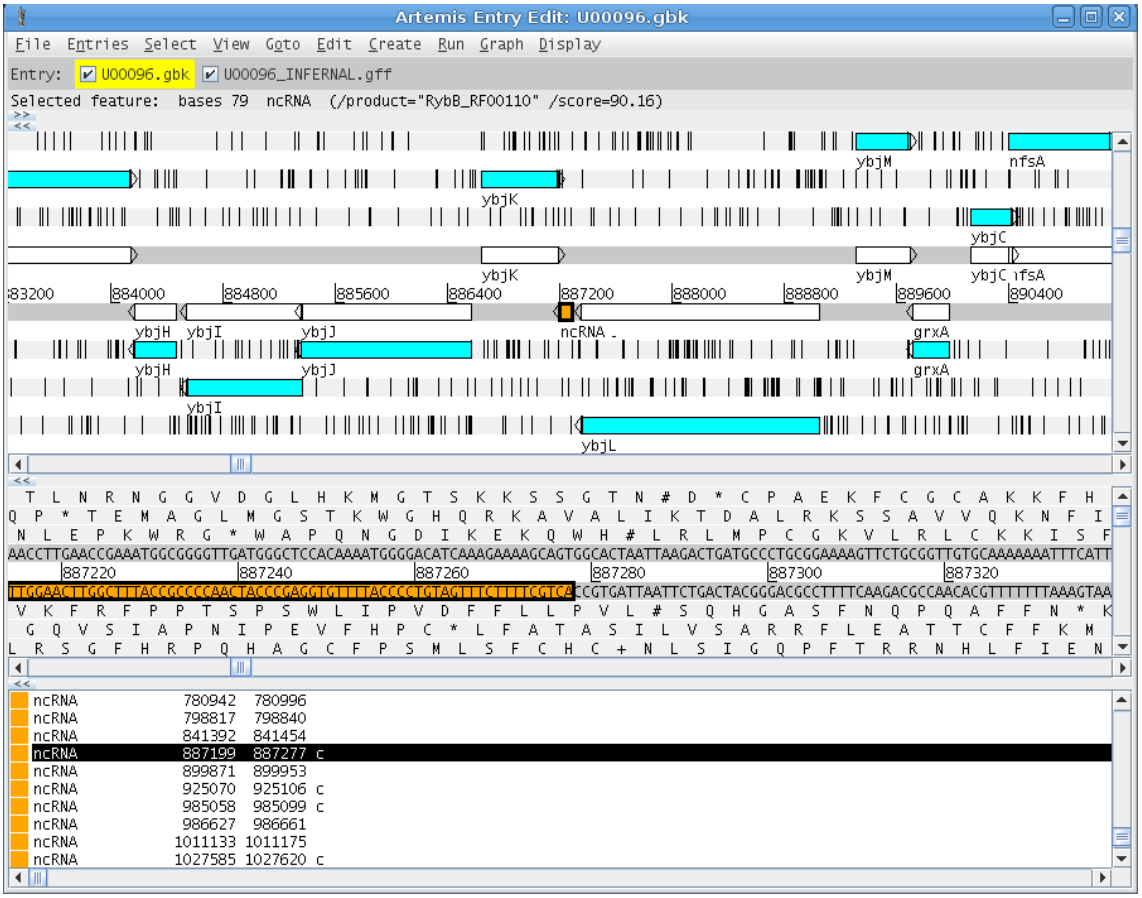


Abb. 2.7: Visualisierung der Ergebnisse im GFF-Format mit Artemis, dargestellt am Beispiel eines von INFERNAL vorhergesagten fRNA-Kandidaten. Oben: In orange ist der fRNA-Kandidat im Genomkontext abgebildet. Unten ist die gesamte Liste der an Artemis übergebenen fRNA-Kandidaten zu sehen. Durch das Auswählen eines Kandidaten wird der Fokus im oberen Fenster automatisch auf diesen Kandidaten und seine Umgebung gerichtet.

3 Test der fRNA-Detektion auf dem Musterorganismus Escherichia coli

In diesem Kapitel wird der Ansatz zur fRNA-Detektion aus Kapitel 2 am Beispiel von *Escherichia coli* [11] getestet. Da dieser Organismus in verschiedenen Studien zur fRNA-Detektion verwendet wurde, sind im Vergleich zu anderen Organismen relativ viele unterschiedliche fRNAs, zusätzlich zu tRNAs und rRNAs, bekannt. Das gibt uns die Möglichkeit, unsere Ergebnisse mit den gegebenen Daten zu vergleichen.

3.1 Auswahl der Vergleichsdaten

Wir nutzen die Sequenz von *Escherichia coli* K-12 MG1655 (kurz *E. coli*) [11] aus der `GenBank`-Datenbank [8]. Informationen zu Protein-kodierenden und bekannten fRNA-kodierenden Genen wurden aus `EcoGene` [87] in der Version 2.12 entnommen. Die Datenbank wird monatlich aktualisiert, weshalb wir eine hohe Aktualität erwarten.

Die Genomsequenz von *E. coli* wurde im Zusammenhang mit zwei unterschiedlichen Datensätzen untersucht. Dabei wurde der Einfluss der Vergleichsdaten auf die Anzahl und Qualität der Kandidaten betrachtet. Im ersten Durchlauf wurde *E. coli* mit *Shigella flexneri* 2a 2457T (kurz: *S. flexneri*) und *Salmonella enterica serovar Typhi* Ty2 (kurz: *S. enterica*) verglichen. Der Datensatz wird abgekürzt mit den Anfangsbuchstaben der Organismen: ESS. Im zweiten Durchgang erfolgte der Vergleich mit *Yersinia pestis* CO92 (kurz: *Y. pestis*) und *Klebsiella pneumoniae* MGH 78578 (kurz: *K. pneumoniae*). Dieser Datensatz wird analog abgekürzt mit EYK. Bis auf *E. coli* stammen alle Daten aus `GenBank`.

Die Tabelle 3.1 liefert einen Überblick über die Daten, ihre Quellen und die in `GenBank` verwendeten Kennungen. Außerdem ist die Anzahl der bekannten tRNAs,

	E. coli	*S. flexneri*	*S. enterica*	*Y. pestis*	*K. pneumoniae*
Quelle	GenBank/ EcoGene	GenBank	GenBank	GenBank	GenBank
Version	U00096.2	AE014073.1	AE014613.1	AL590842.1	CP000647.1
Größe (Mbp)	∼ 4,6	∼ 4,6	∼ 4,8	∼ 4,6	∼ 5,3
GC-Gehalt (%)	50,79	50,91	52,05	47,64	57,48
Protein-kod. Gene	4089	4068	4323	3885	4776
tRNAs+rRNAs	108	11	100	88	0
sonst. fRNAs	61	11	9	20	0

Tab. 3.1: Verwendete Daten und ihre Merkmale für die fRNA-Detektion in *E. coli*. Der ESS-Datensatz umfasst *E. coli*, *S. flexneri* und *S. enterica*, der EYK-Datensatz umfasst *E. coli*, *Y. pestis* und *K. pneumoniae*. Die Größe der Genome ist in Megabasenpaaren (Mbp) angegeben, wobei 1 Mbp = 10^6 bp ist.

rRNAs und sonstigen fRNAs, die den einzelnen Datensätzen zu entnehmen sind, aufgeführt. Wir nutzen die bekannten fRNAs im Folgenden als Referenzdaten für den Vergleich mit den von uns vorhergesagten fRNA-Kandidaten.

Die Datensätze wurden so zusammengestellt, dass die daraus resultierenden Alignments (siehe Abschnitt 2.2) eine unterschiedliche durchschnittliche Sequenzidentität aufweisen (siehe Abbildung 3.1). Die Berechnung der durchschnittlichen Sequenzidentität erfolgte mit der Funktion alistat aus dem Paket SQUID [33], einer Bibliothek von Funktionen zur biologischen Sequenzanalyse. Im Mittel betragen die durchschnittlichen Sequenzidentitäten der erzeugten Alignments 95,5 % im ESS-Datensatz und 91,82 % im EYK-Datensatz.

3.2 Ergebnisse

Um zu beurteilen, wie gut unser Ansatz tatsächliche fRNAs als fRNA-Kandidaten erkennen kann, bestimmen wir die *Sensitivität*. Sie wird berechnet mit Hilfe der Anzahl der *richtig positiven* Kandidaten, d. h. der richtig erkannten fRNAs, und der Anzahl der *falsch negativen* Ergebnisse, d. h. der nicht gefundenen fRNAs, und ist definiert als:

$$\text{Sensitivität} = \frac{\text{Anzahl der richtig positiven}}{\text{Anzahl der richtig positiven} + \text{Anzahl der falsch negativen}}.$$

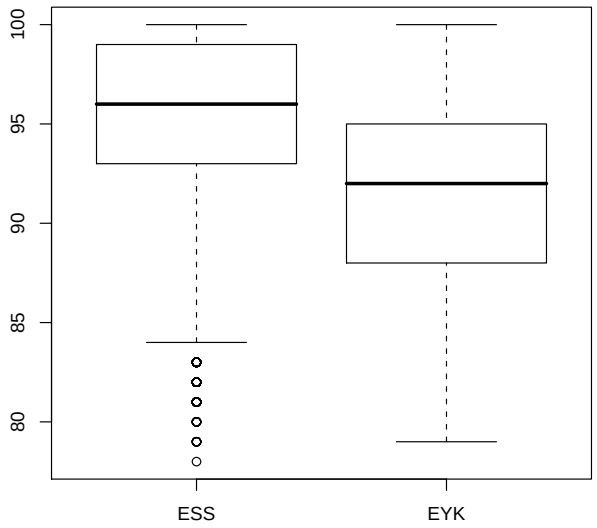


Abb. 3.1: Boxplot der durchschnittlichen Sequenzidentitäten aller untersuchten Alignments für die Datensätze ESS und EYK.

Eine Betrachtung der *Selektivität* und der *Spezifität* wäre ebenfalls wünschenswert, ist mit den gegebenen Referenzdaten jedoch schwierig. Die Selektivität ist definiert als:

$$\text{Selektivität} = \frac{\text{Anzahl der richtig negativen}}{\text{Anzahl der richtig negativen} + \text{Anzahl der falsch positiven}},$$

wobei diejenigen Sequenzen als *richtig negativ* bezeichnet werden, die keine fRNAs sind und als nicht-fRNA erkannt werden. *Falsch positive* Kandidaten sind hingegen Sequenzen, die keine fRNAs sind, aber als fRNA klassifiziert werden. Die Spezifität ist definiert als:

$$\text{Spezifität} = \frac{\text{Anzahl der richtig positiven}}{\text{Anzahl der richtig positiven} + \text{Anzahl der falsch positiven}}.$$

Auf Grund der nach wie vor wachsenden Anzahl an Entdeckungen von bisher unbekannten fRNAs ist zu vermuten, dass bislang kein vollständiges Bild der fRNAs in den gegebenen Organismen vorliegt. Somit sollten wir vorhergesagte Kandidaten

nicht als falsch positiv bzw. sonstige Sequenzen als richtig negativ einstufen.

In den vorliegenden Datensätzen kann nur *E. coli* eine vergleichbar große Anzahl von 61 bekannten fRNAs vorweisen (siehe Tabelle 3.1), wobei tRNAs und rRNAs nicht dazugerechnet wurden. Daher können wir die Sensitivität auf den Ergebnissen für *E. coli* berechnen. In den Vergleichsgenomen sind deutlich weniger fRNAs bekannt. Aus Gründen der Vollständigkeit wird die Sensitivität, soweit das möglich ist, dennoch auch für die Ergebnisse der anderen Organismen berechnet. Nur für *K. pneumoniae* entfällt die Sensitivitätsberechnung komplett, da keine Referenz-fRNAs vorliegen.

3.2.1 Komparativer Ansatz

Als letzte Stufe des komparativen Ansatzes zur fRNA-Vorhersage werden die alignierten Sequenzen mit Hilfe von RNAz auf das Vorkommen konservierter und stabiler Strukturen untersucht. Im Allgemeinen werden Sequenzen, die mit einem Score von mindestens 0,5 bewertet werden, als fRNA-Kandidaten betrachtet. Je höher der Score ist, umso sicherer ist die Vorhersage. Wird ein Mindestscore von 0,5 vorausgesetzt, erhalten wir z. B. eine zusammenhängende Sequenz als fRNA-Kandidat. Wird ein Mindestscore von 0,9 angesetzt, sind es eventuell zwei oder drei kürzere Teilsequenzen, wie am Beispiel eines Alignments in Abbildung 2.6 zu sehen ist. Die Höhe der von uns festgelegten Signifikanzgrenze des Scores beeinflusst damit sowohl die Anzahl, als auch die Länge der Kandidaten (siehe Abschnitt 2.2.7).

		RNAz-0,5	RNAz-0,9	RNAz-1
ESS	***E. coli***	**2585**	**2465**	**182**
	S. flexneri	2894	2786	346
	S. enterica	1233	945	116
EYK	***E. coli***	**829**	**684**	**211**
	Y. pestis	673	568	323
	K. pneumoniae	661	503	61

Tab. 3.2: Angegeben ist die Anzahl der fRNA-Kandidaten, die mit dem komparativen Ansatz im ESS- und EYK-Datensatz vorhergesagt wurden. Unter RNAz-0,5 sind z. B. alle Kandidaten angegeben, die von RNAz mit einem Score von mindestens 0,5 bewertet wurden.

Die Tabelle 3.2 gibt einen Überblick über die Anzahl der Kandidaten für vorge-

gebene Signifikanzschwellen für den RNAz-Score. Da sich die absolute Anzahl der fRNA-Kandidaten für einen höheren und damit signifikanteren Score sogar erhöhen kann, ist es nicht sinnvoll, diese Anzahl alleine zu betrachten. Wir geben ergänzend dazu in der Abbildung 3.2 den prozentualen Anteil der als fRNA klassifizierten Sequenzen an der Gesamtsequenz der einzelnen Genome an. Dazu wurden für RNAz-Scores von 0,5 bis 1 mit einer Schrittweite von 0,02 alle Kandidaten in einem Genom bestimmt, die über dem jeweiligen Score lagen. Die Längen der einzelnen Kandidatensequenzen wurden addiert und für diese Gesamtlänge der prozentuale Anteil an der Genomsequenz bestimmt. Da überlappende Kandidaten zusammengefasst wurden, sind keine Sequenzen doppelt gezählt worden.

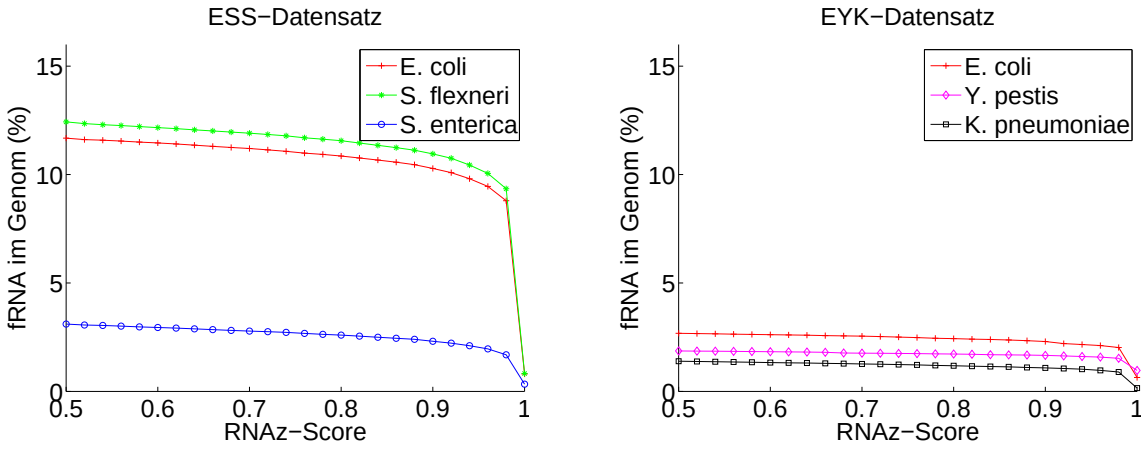


Abb. 3.2: Prozentualer Anteil der Genomsequenzen im ESS- und EYK-Datensatz, die mit Hilfe des komparativen Ansatzes als fRNA klassifiziert wurden.

Sowohl die absolute Anzahl der fRNA-Kandidaten, als auch der prozentuale Genomsequenzanteil aller betrachteter Organismen zeigt nur geringe Änderungen bis zu einem Score von 0,98 und danach einen deutlichen Abfall. Im ESS-Datensatz ist die Menge der in *S. enterica* vorhergesagten Kandidaten um mehr als zwei Drittel geringer als in *E. coli* und *S. flexneri*. Der quantitative Unterschied zwischen den Kandidaten des ESS-Datensatzes und des EYK-Datensatzes wird an den Vorhersagen für *E. coli* sichtbar. Für einen RNAz-Score von 0,9 wurden im ESS-Datensatz 10,28 % und im EYK-Datensatz 2,3 % der Genomsequenz von *E. coli* als fRNA vorhergesagt.

Da die in diesem Abschnitt betrachteten fRNA-Kandidaten mit Hilfe des komparativen Ansatzes bestimmt wurden, weisen jeweils Gruppen von Kandidaten, die

	Eigentreffer	*E. coli*	*S. flexneri*	*S. enterica*	Alle
E. coli	62	**2465**	2357	671	625
S. flexneri	325	2413	**2786**	645	597
S. enterica	153	727	697	**945**	632

Tab. 3.3: Anzahl der Kandidaten im ESS-Datensatz, die nur in einem Genom (Eigentreffer) oder ebenfalls in einem anderen Genom oder in allen Genomen gefunden wurden.

zusammen untersucht wurden, sequenzielle und strukturelle Gemeinsamkeiten auf. In der Tabelle 3.3 und der Tabelle 3.4 stellen wir den Zusammenhang der fRNA-Kandidaten für einen RNAz-Score von 0,9 zwischen den einzelnen Spezies dar. Daraus geht hervor, dass im ESS-Datensatz 62 der insgesamt 2465 Kandidaten als Eigentreffer in *E. coli*, d. h. über Vergleiche von Sequenzen, die ausschließlich aus *E. coli* stammen, gefunden wurden. Dahingegen wurden 625 der Kandidaten in Verbindung mit Sequenzen aus beiden Vergleichsgenomen, *S. flexneri* und *S. enterica*, identifiziert. Im EYK-Datensatz wurden 283 der insgesamt 684 Kandidaten durch Eigentreffer und 65 der Kandidaten in Verbindung mit Sequenzen aus beiden Vergleichsgenomen, *Y. pestis* und *K. pneumoniae*, lokalisiert.

	Eigentreffer	*E. coli*	*Y. pestis*	*K. pneumoniae*	Alle
E. coli	283	**684**	78	388	65
Y. pestis	479	91	**580**	78	68
K. pneumoniae	100	394	68	**503**	59

Tab. 3.4: Anzahl der Kandidaten im EYK-Datensatz, die nur in einem Genom (Eigentreffer) oder ebenfalls in einem anderen Genom oder in allen Genomen gefunden wurden.

Für die Berechnung der Sensitivität des komparativen Ansatzes werden alle vorhergesagten Kandidaten mit den gegebenen Referenzdaten verglichen. Ein Kandidat wird als richtig positiv eingestuft, wenn sich die entsprechende Sequenz im Genom zu mindestens 50 % mit der kodierenden Sequenz einer bekannten fRNA überlappt. Die Abbildung 3.3 zeigt die Sensitivität der Ergebnisse für beide Datensätze. Für *K. pneumoniae* aus dem EYK-Datensatz konnte keine Sensitivität berechnet werden, da

keine Referenz-fRNAs vorliegen. Mit Ausnahme der Sensitivität für *S. enterica* aus dem ESS-Datensatz bleiben die Werte bis zu einem Score von 0,94 weitesgehend stabil und sinken dann rasant ab. Die Sensitivität der Vorhersagen im EYK-Datensatz fällt im Vergleich zur Sensitivität im ESS-Datensatz auffallend niedriger aus.

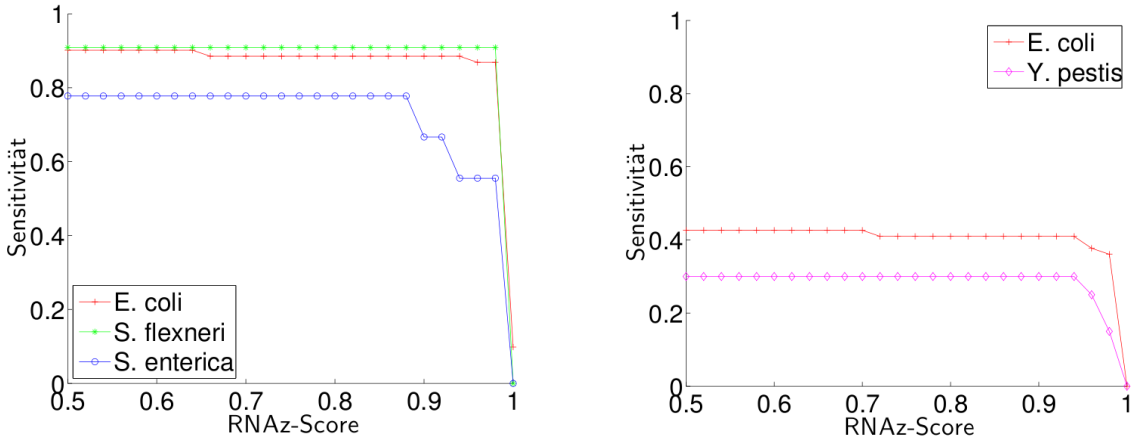


Abb. 3.3: Sensitivität des komparativen Ansatzes für die Vorhersagen in den Datensätzen ESS und EYK in Abhängigkeit vom RNAz-Score.

3.2.2 Kovarianzmodell-basierter Ansatz

Bei dem auf auf das Kovarianzmodell basierten Ansatz werden die fRNA-Kandidaten mit Hilfe von INFERNAL nicht nur lokalisiert, sondern gleichzeitig einer bekannten fRNA-Familie zugeordnet, die in Form eines Kovarianzmodells für die Suche verwendet wurde. Zur Berechnung der Sensitivität dieses Ansatzes wurden die vorhergesagten Kandidaten mit den bekannten fRNAs für die betrachteten Organismen verglichen. Ein Kandidat wurde als richtig positiv eingestuft, wenn er der korrekten fRNA-Familie zugeordnet wurde und sich zu mindestens 50 % mit der kodierenden Sequenz der entsprechenden bekannten fRNA überlappte. Die zweite Bedingung wurde eingeführt, da die Anwendung von INFERNAL im lokalen Modus dazu führen kann, dass nicht die vollständige fRNA lokalisiert wird. Mit der 50 %-Regel ist sichergestellt, dass der größere Teil der fRNA identifiziert werden konnte.

Nicht zu jeder Referenz-fRNA der hier betrachteten Organismen gibt es eine homologe Familie in `Rfam`. In *E. coli* gehören 17 fRNAs zu keiner der in `Rfam` vorkommenden Familien. In *S. flexneri* sind es 2 und in *Y. pestis* 13, in *S enterica* und K.

pneumoniae sind es jeweils 0. Diese fRNAs können mit INFERNAL nicht gefunden werden. Daher wird die Sensitivität jeweils auf der Grundlage von zwei unterschiedlich großen Datensätzen pro Organismus, die als Referenz dienen, berechnet. Ein Datensatz umfasst alle bekannten fRNAs zum jeweiligen Organismus. Der zweite Datensatz besteht ebenfalls aus den bekannten fRNAs, mit dem Unterschied, dass alle fRNAs, zu denen keine homologe Familie in `Rfam` präsent ist, entfernt wurden.

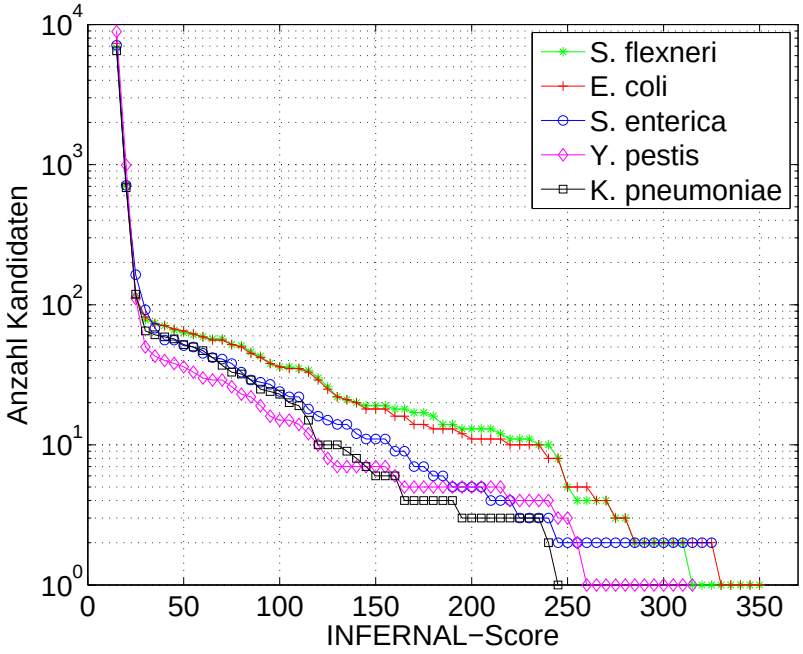


Abb. 3.4: Für jeden Score ist die Anzahl der mit INFERNAL lokalisierten fRNA-Kandidaten angegeben, die mindestens diesen Score erreichen. Insgesamt sind hier nur solche Kandidaten berücksichtigt, die mit einem Score von mindestens 15 bewertet wurden.

In der Abbildung 3.4 wurde zu jedem Score die Anzahl der Kandidaten aufgetragen, die mindestens diesen Score erreichen. Mit sinkender Scoregrenze steigt die Anzahl der Kandidaten exponentiell an.

Die Tabelle 3.5 fasst die Ergebnisse für drei unterschiedliche Mindestscores (20, 25 und 50) zusammen. Für jeden Score ist die Anzahl der Kandidaten, die mindestens diesen Score erreichen, und die Sensitivität der Vorhersage angegeben. Bei steigendem Score sinkt die Anzahl vorhergesagter Kandidaten. Besonders drastisch ist der Unterschied beim Sprung von Score 20 zu Score 25. Die Sensitivität bleibt beim Übergang von Score 20 zu Score 25 gleich und verändert sich nur minimal

	INFERNAL-20		INFERNAL- 25		INFERNAL-50	
	Anz.	Sens.	Anz.	Sens.	Anz.	Sens.
E. coli	**740**	**0,93 / 0,67**	**112**	**0,93 / 0,67**	**65**	**0,91 / 0,66**
S. flexneri	709	1,00 / 0,82	115	1,00 / 0,82	63	1,00 / 0,82
S. enterica	712	0,89 / 0,89	164	0,89 / 0,89	51	0,89 / 0,89
Y. pestis	992	1,00 / 0,35	111	1,00 / 0,35	36	1,00 / 0,35
K. pneumoniae	685	0,00 / 0,00	119	0,00 / 0,00	52	0,00 / 0,00

Tab. 3.5: In Abhängigkeit von einem Mindestscore (20, 25 und 50) ist jeweils die Anzahl der gefundenen fRNA-Kandidaten und die Sensitivität der INFERNAL-Vorhersagen angegeben. Die Sensitivität wurde auf zwei unterschiedlich großen Referenzmengen berechnet. Für den ersten Wert wurden alle fRNAs aus den Referenzdaten entfernt, zu denen keine homologe Familie in der `Rfam`-Datenbank existiert. Der zweite Wert gibt die Sensitivität in Bezug zur Gesamtmenge der Referenz-fRNAs für den jeweiligen Organismus an.

beim Übergang von Score 25 zu Score 50 für die Ergebnisse in *E. coli*. Die auf der reduzierten Menge der Referenz-fRNAs berechnete Sensitivität ist in allen Fällen, in denen sich die reduzierte Menge von der vollen Referenz-fRNA-Menge unterscheidet, höher.

3.2.3 Vergleich der Kandidaten aus beiden Ansätzen

In Tabelle 3.6 ist die Anzahl der sich überlappenden fRNA-Kandidaten aus dem komparativen Ansatz und dem Ansatz, der auf dem Kovarianzmodell basiert mit `Rfam` als Referenzdatenbank angegeben. Aus dem komparativen Ansatz werden nur Kandidaten mit einem RNAz-Score von mindestens 0,9 für den Vergleich herangezogen. Beim Kovarianzmodell-basierten Ansatz betrachteten wir drei Mengen von Kandidaten, indem die Signifikanzgrenze für den INFERNAL-Score auf einen Wert von 20, 25 und 50 gesetzt wird.

Für die INFERNAL-Scoregrenze von 20 wurde durchgehend die zwei- bis dreifache Menge an INFERNAL-Kandidaten im Vergleich zu den RNAz-Kandidaten vorhergesagt. Für einen INFERNAL-Score von mindestens 50, stimmt die Anzahl der sich überlappenden RNAz- und INFERNAL-Kandidaten nahezu überein. Mit steigendem INFERNAL-Score sinkt die Anzahl sich überlappender Kandidaten auf beiden Seiten. Auch hier ist die Änderung beim Übergang vom INFERNAL-Score von mindestens 20 zum INFERNAL-Score von mindestens 25 besonders ausgeprägt. Im Vergleich der

		RNAz-0,9 / INF.-20	RNAz-0,9 / INF.-25	RNAz-0,9 / INF.-50
ESS	*E. coli*	273 / 586	81 / 100	61 / 62
	S. flexneri	290 / 559	79 / 101	57 / 61
	S. enterica	94 / 158	50 / 60	41 / 41
EYK	*E. coli*	90 / 202	48 / 59	43 / 41
	Y. pestis	145 / 273	25 / 33	18 / 18
	K. pneumoniae	82 / 135	48 / 56	36 / 35

Tab. 3.6: Vergleich überlappender Kandidaten aus dem komparativen Ansatz (RNAz-Score $\geq$ 0,9) und dem Kovarianzmodell-basierten Ansatz (INFERNAL-Scores $\geq$ 20, 25 und 50).

Ergebnisse für *E. coli* aus dem ESS- und dem EYK-Datensatz, ist sowohl die Anzahl der Kandidaten aus dem komparativen Ansatz, als auch aus dem Kovarianzmodell-basierten Ansatz im EYK-Datensatz deutlich niedriger.

3.3 Diskussion

Im komparativen Ansatz könnte statt BlastN ein Werkzeug verwendet werden, das sowohl nach Sequenz- als auch nach Strukturähnlichkeiten sucht, wie z. B. RSEARCH (Abschnitt 1.4.1). Wir haben uns aus zwei Gründen dennoch für BlastN entschieden. Zum einen ist die Suche nach reinen Sequenzähnlichkeiten deutlich schneller, siehe [37]. Zum anderen wird für die Anwendung von RNAz, mit dem die endgültige Beurteilung eines Alignments erfolgt, ein Mindestmaß an Sequenzähnlichkeit im Alignment (mindesten 60 %) empfohlen [103]. Da wir Sequenzen verwandter Organismen vergleichen, erwarten wir eine ausreichend hohe Sequenzähnlichkeit zwischen fRNA-kodierenden Sequenzen, welche zu einer Familie gehören. Damit ist BlastN gut geeignet, um eine schnelle Vorauswahl an Kandidaten treffen zu können.

In den meisten Fällen liegen die kodierenden Regionen bekannter fRNAs innerhalb des von uns vorgegebenen Suchraums, den eIGRs. In jedem der hier betrachteten Genome gibt es jedoch Ausnahmen, bei denen sich die fRNA-kodierende Sequenz weniger als 50 % mit einer eIGR überlappt. In *E. coli* und *Y. pestis* betrifft das zwei, in *S. flexneri* und *S. enterica* jeweils eine Referenz-fRNA. Solche fRNAs können von unserem Ansatz nicht in ausreichendem Maße erfasst werden. Trotz einer Ausdehnung der IGRs in Protein-kodierende Regionen, können wir solche Fälle nicht ganz

ausschließen. Sie gehen als falsch negative Ergebnisse in die Sensitivitätsberechnung ein.

Mit dem komparativen Ansatz werden nur solche fRNAs identifiziert, die mindestens zweimal im gesamten Suchraum vertreten sind und eine ausreichend hohe Sequenzidentität vorweisen. Nur dann ist es möglich, die entsprechenden Sequenzregionen mit Hilfe von BlastN einander zuzuordnen und das notwendige Alignment zu erstellen, auf dessen Grundlage die gemeinsame Struktur mit RNAz bewertet wird. Wie aus der Tabelle 3.2 und Abbildung 3.2 hervorgeht, beeinflusst die Wahl der Vergleichsgenome zu einem Referenzgenom, in diesem Fall ist es *E. coli*, die Menge der vorhergesagten Kandidaten. Die Gesamtanzahl der Kandidaten in *E. coli* ist im ESS-Datensatz um ein Vielfaches höher als im EYK-Datensatz. Noch deutlicher wird der Einfluss beim Vergleich der zusammenhängenden Kandidaten (siehe ESS-Datensatz: Tabelle 3.3 und EYK-Datensatz: Tabelle 3.4). Die Anzahl der Kandidaten, die in jedem Genom mindestens einen ähnlichen Kandidaten ausweisen, ist im ESS-Datensatz deutlich höher als im EYK-Datensatz. Die Anzahl der Eigentreffer in *E. coli*, also der Kandidaten, die nur in *E. coli* lokalisiert wurden, ist im ESS-Datensatz im Gegensatz zum EYK-Datensatz hingegen deutlich geringer. Diese Unterschiede ergeben sich als Folge der unterschiedlichen evolutionären Distanzen der einzelnen Organismen zueinander. Daraus folgen wiederum die unterschiedlich hohen Sensitivitäten der Vorhersagen im ESS-Datensatz im Vergleich zum EYK-Datensatz, dargestellt in Abbildung 3.3.

Im Kovarianzmodell-basierten Ansatz wird in jeder Zielsequenz mit Hilfe von INFERNAL nach Sequenz- und Strukturähnlichkeiten zu bekannten fRNA-Familien gesucht. INFERNAL bewertet jede Ähnlichkeit zwischen einem Kovarianzmodell und der Zielsequenz mit einem positiven Score. Dabei werden auch zufällig auftretende Ähnlichkeiten lokalisiert. Die Anzahl der Treffer nimmt für einen fallenden Score ab einem Wert von 25 exponentiell zu, wie am Beispiel der untersuchten Sequenzen in der Abbildung 3.4 zu sehen ist. Um zu entscheiden, ob gegebene Treffer signifikant sind, kann INFERNAL neben einem Score auch einen E-Value berechnen. Der E-Value bewertet ein zufälliges Auftreten der Treffer, so wie es bei BlastN der Fall ist. Da dies aber den Rechenaufwand enorm vergrößert, ist es vorerst nicht zu empfehlen. Der Entwickler von INFERNAL, S. Eddy [32], empfiehlt, den Score in Abhängigkeit von der Länge der untersuchten Sequenz S zu bewerten. Als eine grobe Richtlinie

gibt er an, dass der Score eines Treffers signifikant ist, wenn

$$\text{Score} \geq \log_2(\,2 \cdot \text{Länge}(S))$$

ist. Es wird die doppelte Sequenzlänge verwendet, da jede Sequenz ein Mal in der Plus- und Minusorientierung untersucht wird. Wir vermuten, dass sich diese Abschätzung auf die Untersuchung von Genomsequenzen mit mehreren Mbp Länge, die in einem Stück abgescannt werden, bezieht. Die von uns untersuchten eIGRs sind teilweise nur wenige hundert bp lang. Nach eigenen Beobachtungen ist die von Eddy *et al.* empfohlene Scoregrenze für so kurze Sequenzen zu niedrig. Sie führt zu einer hohen Zahl sich überlappender partieller Ähnlichkeiten von sich deutlich voneinander unterscheidenden fRNA-Familien. Daher haben wir uns für eine manuell einstellbare Signifikanzgrenze des Score entschieden und die Ergebnisse für verschiedene Grenzen (20, 25, 50) in der Tabelle 3.5 gegenübergestellt. Trotz der deutlich sinkenden Anzahl gefundener Kandidaten bei steigender Scoregrenze, blieb die Sensitivität nahezu gleich. Fast alle bekannten fRNAs, zu denen entsprechende Informationen in `Rfam` verfügbar waren, konnten mit INFERNAL gefunden werden. Solche Kandidaten wurden zudem mit einem hohen Score bewertet. Nahm die Sensitivität nicht den maximalen Wert an, so kann das auf diejenigen bekannten fRNAs zurückgeführt werden, deren kodierende Sequenzen nicht in voller Länge in dem von uns festgelegten Suchraum, den eIGRs, enthalten sind.

Beim Vergleich der Kandidaten aus dem komparativen Ansatz und dem Kovarianzmodell-basierten Ansatz (Tabelle 3.6) wurden alle Kandidaten gezählt, die sich auf der Sequenzebene überlappen. Sind fRNAs in mehreren Organismen dicht hintereinander in der DNA kodiert, können sie im komparativen Ansatz als ein Kandidat behandelt werden. Auch die umgekehrte Variante ist denkbar. Falls die fRNAs nicht über die volle Länge, sondern nur abschnittsweise eine konservierte und stabile Struktur ausbildet, so werden nur diese Regionen als potentielle fRNAs angegeben. Für einen INFERNAL-Score von mindestens 20 ist es auffällig, dass sich mindestens doppelt so viele INFERNAL- wie RNAz-Kandidaten miteinander überlappen. Das Phänomen kann durch die höhere Gesamtanzahl der vorhergesagten INFERNAL-Kandidaten für niedrige Scores erklärt werden, die sich teilweise überlappen. In solchen Fällen weist ein Abschnitt einer Sequenz gleichzeitig lokale Ähnlichkeiten zu mehreren fRNA-Familien auf. Wird der Bereich auch durch den komparativen

Ansatz als fRNA klassifiziert, so überlappt sich dieser Kandidat gleich mit mehreren INFERNAL-Kandidaten. Für höhere INFERNAL-Scores ist die Anzahl sich überlappender Kandidaten auf beiden Seiten ausgeglichen.

Im ESS-Datensatz wurden nahezu alle INFERNAL-Kandidaten in *E. coli* mit einem Score von mindestens 50 durch den komparativen Ansatz bestätigt. So sind es in *E. coli* 62 von 65, in *S. flexneri* 61 von 63 und in *S. enterica* 41 von 51. Im EYK-Datensatz sind die Unterschiede größer. In *E. coli* sind 41 der 65 INFERNAL-Kandidaten bestätigt, in *Y. pestis* 18 von 36 und in *K. pneumoniae* 35 von 52. Wir vermuten, dass die Sequenzidentität innerhalb eines Datensatzes eine wichtige Rolle beim komparativen Ansatz spielt, aber nicht der einzige Grund für die unterschiedliche Anzahl dieser Kandidaten in *E. coli* im ESS- und EYK-Datensatz ist. Vergleichen wir die mit INFERNAL lokalisierten fRNA-Kandidaten für einen Score von mindestens 50, so stellen wir fest, dass *S. enterica*, *K. pneumoniae* und vor allem *Y. pestis* eine deutlich geringere Gesamtanzahl an Kandidaten hervorgebracht haben als *E. coli* und *S. flexneri*, die beide im ESS-Datensatz enthalten sind. Es besteht also die Möglichkeit, dass die Organismen im EYK-Datensatz trotz naher Verwandtschaft zu *E. coli* tatsächlich weniger fRNAs besitzen, oder diese anderen Familien angehören, die wiederum in *E. coli* nicht vertreten sind.

Mit Hilfe des komparativen Ansatzes werden sehr viele fRNA-Kandidaten vorhergesagt. Daher stellt sich die Frage, wieviele davon eventuell falsch positive Kandidaten sind. Es ist zu bedenken, dass der komparative Ansatz nicht nur fRNA-kodierende Gene, sondern ebenfalls strukturbasierte regulatorische Elemente aufspüren kann. Da außerdem sogar in einem Musterorganismus wie *E. coli* immer wieder bisher unbekannte fRNAs lokalisiert werden, können wir keinen unserer Kandidaten eindeutig als falsch positiv einstufen. Einen Aufschluss darüber kann nur eine experimentelle Verifikation der potentiellen fRNAs geben. Sollen nur die vielversprechendsten Kandidaten genauer untersucht werden, so empfiehlt es sich, nur diejenigen mit einem RNAz-Score über 0,9 zu betrachten. Als ein weiteres Auswahlkriterium empfehlen wir den Zusammenhang der Kandidaten in den verschiedenen Genomen. Das Auftreten einer in jedem Genom konservierten Sequenz mit entsprechend stabiler und konservierter Struktur deutet stark auf eine funktionelle Bedeutung dieser Sequenz hin. Der Zusammenhang der fRNA-Kandidaten lässt sich aus den als GFF-formatierten Ergebnisdateien problemlos ablesen.

Im Fall des Kovarianzmodell-basierten Ansatzes haben sich Kandidaten mit einem

hohen INFERNAL-Score von mindestens 50 als besonders vertrauenswürdig erwiesen. Fast alle dieser Kandidaten in *E. coli* konnten den Referenz-fRNAs in *E. coli* zugeordnet werden. Zusätzlich wurden 62 der 65 Kandidaten mit Hilfe des komparativen Ansatzes bestätigt. Wird die Signifikanzgrenze für den INFERNAL-Score weiter reduziert, steigt die Anzahl zufälliger, partieller Ähnlichkeiten zu bekannten fRNA-Familien. Auch hier gilt, dass experimentelle Verifikationen die endgültige Klarheit bringen.

Durch die Kombination des komparativen und Kovarianzmodell-basierten Ansatzes haben wir zwei komplementäre Ansätze zusammengeführt. Der komparative Ansatz kommt ohne a priori Informationen über bekannte fRNA-Familien aus und ist in der Lage, bisher unbekannte, strukturbasierte fRNAs aufzuspüren. Dahingegen werden im Kovarianzmodell-basierten Ansatz Informationen über bekannte fRNA-Familien verwendet. Dadurch werden fRNA-Kandidaten nicht nur vorhergesagt, sondern gleichzeitig einer bekannten fRNA-Familien zugeordnet. Die im Rahmen dieser Arbeit erfolgte Implementierung des kombinierten Ansatzes ermöglicht zudem eine einfache Anwendung des Verfahrens, da kein Expertenwissen zu den verwendeten Werkzeugen mehr notwendig ist. Wir haben diesen Ansatz auf mehrere neue Datensätze angewandt. Die erzielten Ergebnisse wurden teilweise bereits in renommierten Journalen veröffentlicht oder sind zur Veröffentlichung eingereicht (siehe [16, 17, 94]) und werden im folgenden Kapitel vorgestellt.

4 Ergebnisse der Anwendung der fRNA-Detektion

Die in diesem Kapitel beschriebenen Datensätze wurden in Kooperationen mit verschiedenen Instituten untersucht. Dabei haben wir den in Kapitel 2 beschriebenen Ansatz auf die gegebenen Genomsequenzen angewendet, um neue potentielle fRNAs zu finden. Einige der Ergebnisse, die in diesem Kapitel nur kurz vorgestellt werden, können im Anhang A nachgeschlagen werden. Im Abschnitt A.1 ist die Anzahl der Kandidaten aus dem komparativen Ansatz, dem Kovarianzmodell-basierten Ansatz und aus beiden Ansätzen für alle untersuchten Datensätze tabellarisch zusammengefasst. Im Abschnitt A.2 sind, basierend auf dem komparativen Ansatz, die Beziehungen zwischen fRNA-Kandidaten in verschiedenen Genomen angegeben.

4.1 Bacillus amyloliquefaciens

Pflanzenkrankheiten sind die Ursache für erhebliche Ertragseinbrüche in der Landwirtschaft und im Gartenbau. Ihre Bekämpfung mit Chemikalien belastet nicht nur die Umwelt, sondern auch die aus den behandelten Pflanzen erzeugten Lebensmittel. Um die gesundheitlichen Risiken zu minimieren und umweltverträgliche Landwirtschaft zu fördern, werden neue Strategien entwickelt. Eine Möglichkeit ist der Einsatz nützlicher Bakterien und Pilze. Sie können unter anderem das Pflanzenwachstum fördern und damit den Ertrag steigern. Allerdings ist ihre genaue Wirkungsweise bisher unzureichend verstanden, was zu schwankenden Anwendungserfolgen führt. Daher ist die Erforschung potentieller Kandidaten von großer Bedeutung.

Ein vielversprechender Organismus, der das Pflanzenwachstum fördert und gleichzeitig das Wachstum phytopathogener Organismen unterdrückt, ist *Bacillus amyloliquefaciens*, Stamm FZB42. Insgesamt 8,5 % seiner Genomsequenz stehen für die

Produktion von Antibiotika und vergleichbarer Stoffe zur Verfügung. Der Stamm FZB42 hat die natürliche Fähigkeit zur Aufnahme und zum Einbau von DNA. Das macht ihn besonders interessant als Modellobjekt für funktionelle Genomstudien.

Die Genomsequenz von *B. amyloliquefaciens* wurde an der Humboldt-Universität Berlin und dem Laboratorium für Genomanalyse an der Georg-August-Universität Göttingen entschlüsselt. In diesem Zusammenhang wurde der im Rahmen dieser Arbeit entwickelte Ansatz zur fRNA-Detektion auf die Genomsequenzen angewendet, und es entstanden die folgenden Publikationen: [16] und [17].

4.1.1 Daten

Wir haben die Genomsequenz von *Bacillus amyloliquefaciens* FZB42 (kurz *B. amyloliquefaciens*) [16] mit den Genomen von drei verwandten Bakterien verglichen: *Bacillus licheniformis* DSM 13 (kurz: *B. licheniformis*) [100], *Bacillus subtilis* (kurz: *B. subtilis*) [58] und *Bacillus anthracis* str. 'Ames Ancestor' (kurz: *B. anthracis*) [82]. Die Genomsequenzen und Informationen zu Protein-kodierenden Genen, sowie zu bereits bekannten tRNA- und rRNA-kodierenden Genen stammen aus `GenBank`.

Alle vier Organismen zählen zu den grampositiven Bakterien und dort zur Gattung der Bazillen, also der stäbchenförmigen Bakterien. *B. licheniformis* und *B. subtilis* werden bereits in der industriellen Produktion von z. B. Waschpulver genutzt, da sie in der Lage sind, große Mengen an Enzymen extrazellulär zu produzieren [92]. *B. anthracis* ist hingegen als der Erreger des Milzbrands bekannt.

	B. amyloliquefaciens	*B. licheniformis*	*B. subtilis*	*B. anthracis*
Quelle	`GenBank`	`GenBank`	`GenBank`	`GenBank`
Version	CP000560.1	AE017333.1	AL009126.2	AE017334.2
Größe (Mbp)	∼ 3,9	∼ 4,2	∼ 4,2	∼ 5,2
GC-Gehalt (%)	46,48	46,19	43,52	35,38
Protein-kod. Gene	3693	4196	4106	5309
tRNAs+rRNAs	116	93	117	128

Tab. 4.1: Merkmale der untersuchten Genomsequenzen.

Die Tabelle 4.1 gibt einen Überblick über die betrachteten Organismen und die wichtigsten Merkmale ihrer Genomsequenzen. Dabei fällt *B. anthracis* durch ein größeres Genom und einen um durchschnittlich 10 % geringeren GC-Gehalt auf.

4.1.2 Ergebnisse

Komparativer Ansatz

In der Abbildung 4.1 links ist die Gesamtmenge der vom komparativen Ansatz als fRNA klassifizierten Sequenzen dargestellt. Mit Hilfe dieses Ansatzes haben wir für einen RNAz-Score von mindestens 0,9 insgesamt 588 Sequenzen in *B. amyloliquefaciens* als potentielle fRNAs klassifiziert. Diese Sequenzen entsprechen zusammengenommen 1,65 % der gesamten Genomsequenz. Vergleichbar dazu haben wir in *B. subtilis* 635 Sequenzen (1,69 % der Genomsequenz), in *B. licheniformis* 265 Sequenzen (0,68 % der Genomsequenz) und in *B. anthracis* 198 Sequenzen (0,39 % der Genomsequenz) als fRNA vorhergesagt. Alle Kandidaten-Sequenzen sind zwischen 40 und 600 bp lang, wobei mehr als drei Viertel davon kürzer als 140 bp sind.

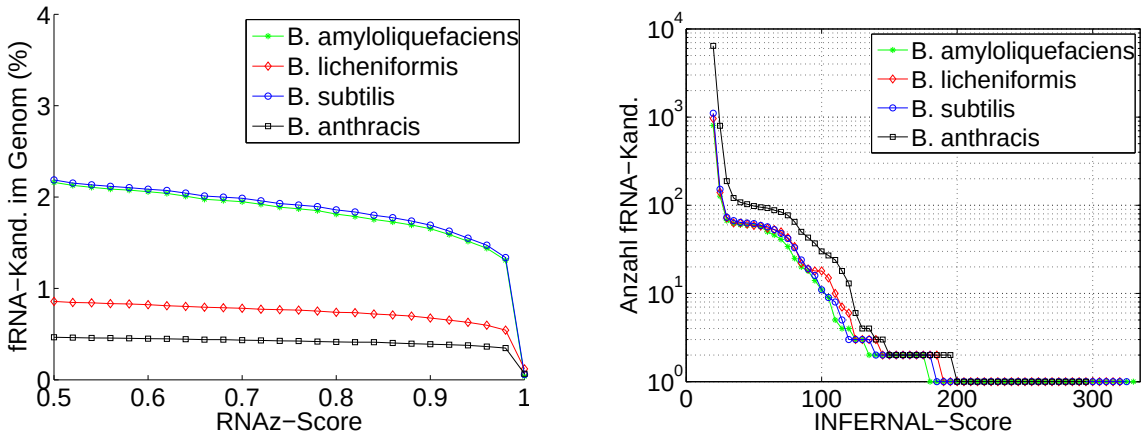


Abb. 4.1: Links: Prozentualer Genomsequenzanteil der fRNA-Kandidaten in Abhängigkeit vom RNAz-Score, die mit dem komparativen Ansatz vorhergesagt wurden. Rechts: Anzahl der mit INFERNAL vorhergesagten fRNA-Kandidaten (ab Score = 20). Für jeden Score ist die Anzahl der Kandidaten angegeben, die mit mindestens diesem Score bewertet wurden.

Da beim komparativen Ansatz jede Sequenz nicht einzeln sondern im Alignment mit anderen Sequenzen bewertet wird, weist ein fRNA-Kandidat immer strukturelle und sequenzielle Ähnlichkeiten zu den Kandidaten in den restlichen Sequenzen des Alignments auf. Wie in Abschnitt A.2 im Anhang zu sehen ist, haben wir in *B. amyloliquefaciens* 12 unterschiedliche Kandidaten gefunden, die in jedem der drei Vergleichsgenome mindestens einen ähnlichen fRNA-Kandidaten aufweisen. Fast alle der Kandidaten in *B. amyloliquefaciens* wurden im Zusammenhang mit Sequenzen

aus *B. subtilis* vorhergesagt. Das wird ebenfalls durch eine gleichhohe Gesamtmenge an Kandidaten in beiden Genomen wiedergespiegelt (siehe Abbildung 4.1 links). Ungefähr ein Viertel der Kandidaten in *B. amyloliquefaciens* steht in einer Beziehung zu Kandidaten in *B. licheniformis* und nur 14 zu den Kandidaten in *B. anthracis*. Die meisten der in *B. anthracis* gefundenen Kandidaten wurden fast ausschließlich als Eigentreffer, d. h. nur im Zusammenhang mit Sequenzen desselben Genoms, als potentielle fRNAs klassifiziert.

Kovariansmodell-basierter Ansatz

Im Gegensatz zum unterschiedlichen Verhältnis der Kandidatenmengen aus dem komparativen Ansatz, liefert INFERNAL in allen vier Genomen ähnlich hohe Kandidatenmengen (siehe Abbildung 4.1 rechts und Tabelle 4.2). Die Tabelle 4.2 gibt einen Überblick über die mit INFERNAL lokalisierten Kandidaten, die mindestens mit einem Score von 50 bewertet wurden. Sie ist nach der Gesamtanzahl der zu einer fRNA-Familie gehörenden Kandidaten absteigend sortiert. Dadurch ist leicht zu erkennen, welche fRNAs am häufigsten in den Sequenzen vorhergesagt werden. Insgesamt wurden nur wenige fRNA-kodierende Gene, dafür aber vergleichbar viele regulatorische Elemente gefunden. Besonders häufig kommt die T-Box-fRNA, auch T-Box-Leader genannt, und der SAM-Riboswitch (S-adenosylmethionin Riboswitch) vor. *B. anthracis* weist für fast jede fRNA-Familie die größte Anzahl an Kandidaten auf, im Falle der T-Box-fRNA und des SAM-Riboswitches sind es sogar fast doppelt so viele, wie in den anderen Genomen.

Von den 59 Kandidaten, die in *B. amyloliquefaciens* gefunden wurden, haben wir 49 ebenfalls mit Hilfe des komparativen Ansatzes vorhergesagt. In *B. licheniformis* wurden 40 der 59, in *B. subtilis* 51 von 62 und in *B. anthracis* 25 von 98 Kandidaten ebenfalls komparativ gefunden. Siehe dazu die Tabelle A.1.

4.2 Methanosarcina mazei

Das Archaeon *Methanosarcina mazei* und verwandte Spezies zählen zu den anaeroben methanogenen Organismen, d. h. Organismen, die Methan ohne Sauerstoffzufuhr produzieren können. Im Gegensatz zu anderen anaeroben methanogenen Spezies, sind sie als einzige in der Lage, Methan sogar auf drei verschiedene Arten,

fRNA-Name	Typ	*B. amyloliquefaciens*	*B. licheniformis*	*B. subtilis*	*B. anthracis*
T-box	cis-El.	18	18	20	41
SAM	cis-El.	10	10	11	17
TPP	cis-El.	6	4	5	7
Purine	cis-El.	5	4	5	6
PyrR	cis-El.	2	2	3	2
ydaO-yuaA	cis-El.	1	2	2	4
Lysine	cis-El.	2	2	1	4
6S	Gen	1	2	2	2
yybP-ykoY	cis-El.	1	2	1	2
FMN	cis-El.	1	2	1	2
ykkC-yxkD	cis-El.	2	2	2	-
L20_leader	cis-El.	1	1	1	1
RNaseP_bact_b	Gen	1	1	1	1
ylbH	cis-El.	1	1	1	1
SRP_bact	Gen	1	1	1	1
L10_leader	cis-El.	1	1	1	1
Glycine	cis-El.	2	-	1	1
tmRNA	Gen	1	1	1	1
ykoK	cis-El.	1	1	1	1
glmS	cis-El.	-	1	1	1
Cobalamin	cis-El.	-	1	-	1
tRNA	Gen	1	-	-	-
Intron_gpI	Intron	-	-	-	1

Tab. 4.2: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`). Neben dem fRNA-Familiennamen ist auch der fRNA-Typ angegeben: Gen, regulatorisches cis-Element oder Intron.

unter anderem aus Acetaten, herzustellen. Bis zu 60 % des Treibhausgases Methan wird aus Acetaten hergestellt, und die verschiedenen *Methanosarcina*-Spezies sind maßgeblich daran beteiligt [24].

An der Universität Kiel, in der Arbeitsgruppe von Prof. Ruth Schmitz-Streit, wird *Methanosarcina mazei* genauer untersucht. Dabei soll die Rolle regulatorischer RNAs im Hinblick auf Stickstoff-Stress und allgemeinen Stress betrachtet werden. Im Rahmen dieses Projekts erfolgt sowohl eine genomweite Identifikation funktioneller RNAs mit unterschiedlichen Ansätzen, als auch eine detaillierte Studie einzelner Kandidaten im Hinblick auf ihre Regulationsmechanismen. In diesem Zusammenhang haben wir den in dieser Arbeit vorgestellten Ansatz zur fRNA-Detektion auf einen Datensatz ausgewählter *Methanosarcina*-Spezies angewendet.

4.2.1 Daten

Für die Untersuchung von *Methanosarcina mazei* Go1 (kurz: *M. mazei*) [24] haben wir die folgenden Vergleichsorganismen gewählt: *Methanosarcina acetivorans* C2A (kurz: *M. acetivorans*) [38] und *Methanosarcina barkeri* Fusaro (kurz: *M. barkeri*) [68]. Die Genomsequenzen und Annotationen stammen aus `GenBank`.

	M. mazei	*M. acetivorans*	*M. barkeri*
Quelle	`GenBank`	`GenBank`	`GenBank`
Version	AE008384.1	AE010299.1	CP000099.1
Größe (Mbp)	∼ 4,1	∼ 5,8	∼ 4,8
GC-Gehalt (%)	41,48	42,68	39,28
Protein-kod. Gene	3371	4540	3607
tRNAs+rRNAs	67	10	72

Tab. 4.3: Merkmale der untersuchten Genomsequenzen.

In der Tabelle 4.3 sind die wichtigsten Merkmale der in diesem Abschnitt betrachteten Genomsequenzen zusammengefasst. Für *M. acetivorans* sind bei `GenBank` ausschließlich rRNA- und keine tRNA-Gene angegeben, was die geringe Gesamtanzahl dieser Gene in *M. acetivorans* im Vergleich zu den anderen Organismen erklärt. Die betrachteten Genome variieren deutlich in ihrer Größe, von 4,1 Mbp in *M. mazei* bis 5,8 Mbp in *M. acetivorans*. Sie zeichnen sich jedoch einheitlich durch einen niedrigen GC-Gehalt aus.

4.2.2 Ergebnisse

Komparativer Ansatz

Die Abbildung 4.2 links stellt die Menge der mit dem komparativen Ansatz vorhergesagten Kandidaten in Abhängigkeit vom RNAz-Score dar. Mit Hilfe dieses Ansatzes haben wir für einen Score von mehr als 0,9 die wenigsten Kandidaten, insgesamt 912, in *M. mazei* gefunden. Das entspricht 2,69 % der gesamten Genomsequenz. In *M. acetivorans* haben wir 1654 Kandidaten (3,95 % der Genomsequenz) und in *M. barkeri* 1491 Kandidaten (4,38 % der Genomsequenz) als fRNA-kodierend klassifiziert. Diese Kandidaten sind mindestens 40 bp lang. Maximal erreichen sie in *M. mazei* eine Länge von 750 bp, in *M. acetivorans* 1100 bp und in *M. barkeri* 1600 bp. Drei Viertel der Kandidaten in jeden der drei Organismen sind jedoch kürzer als 170 bp.

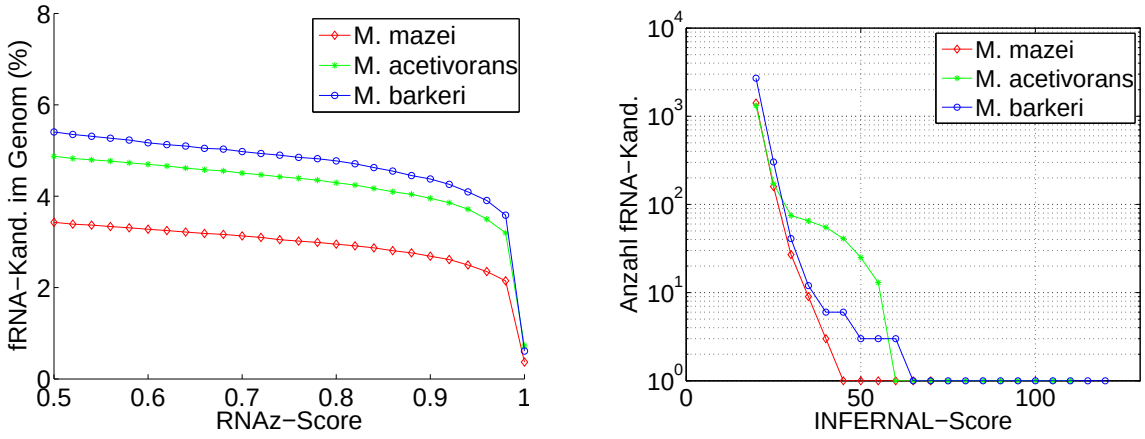


Abb. 4.2: Menge vorhergesagter fRNA-Kandidaten mit dem komparativen Ansatz (links) und dem Kovarianzmodell-basierten Ansatz (rechts).

Betrachten wir die Zusammenhänge zwischen den Kandidaten aus dem komparativen Ansatz, so haben wir in *M. mazei* 245 fRNA-Kandidaten gefunden, die in beiden Vergleichsgenomen mindestens einen ähnlichen Kandidaten aufweisen. Ungefähr zwei Drittel der Kandidaten im *M. mazei* wurden in Verbindung mit Sequenzen aus *M. acetivorans* und ein Drittel in Verbindung mit Sequenzen aus *M. barkeri* gefunden. Insgesamt haben wir 128 Kandidaten ausschließlich in *M. mazei* lokalisiert. In *M. acetivorans* sind über 600 und in *M. barkeri* sogar über 700 der Kandidaten Eigentreffer. Die hohe Anzahl der Kandidaten in *M. acetivorans* und

M. barkeri, welche ausschließlich in diesen Genomen auftritt, spiegelt sich in der höheren Gesamtanzahl der Kandidaten im Vergleich zu *M. mazei* wieder (siehe Abbildung 4.2 links). Eine Übersicht über die Anzahl zusammenhängender Kandidaten in den einzelnen Genomen, ist im Anhang in Tabelle A.4 zu finden.

Kovarianzmodell-basierter Ansatz

In der Tabelle 4.4 und dem rechten Teil der Abbildung 4.2 ist ein Überblick der mit INFERNAL vorhergesagten Kandidaten angegeben, die mindestens einen Score von 50 erreichen. Nur in *M. acetivorans* wurden tRNA-Gene gefunden. In *M. mazei* und *M. barkeri* waren sie bereits im Ausgangsdatensatz aus `GenBank` bekannt und wurden daher im Voraus aus dem Suchraum, den eIGRs, ausgeschlossen. Insgesamt haben wir in *M. mazei* nur eine, in *M. acetivorans* zwei (ohne tRNAs) und in *M. barkeri* drei fRNA-Kandidaten mit einem Score von über 50 gefunden. Mit Ausnahme eines Kandidaten in *M. barkeri* konnten alle mit Hilfe des komparativen Ansatzes bestätigt werden.

fRNA-Name	Typ	*M. mazei*	*M. acetivorans*	*M. barkeri*
tRNA	Gen	-	23	-
SRP_euk_arch	Gen	1	-	1
RNaseP_arch	Gen	-	1	1
RNAIII	Gen	-	-	1
Intron_gpII	Intron	-	1	-

Tab. 4.4: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`).

4.3 Streptomyces coelicolor

Streptomyces coelicolor ist ein grampositives Bodenbakterium, das in der Lage ist, Antibiotika zu produzieren. Damit hat es eine große biotechnologische Bedeutung und ist Gegenstand aktueller Forschung.

Die Arbeitsgruppe um Prof. Beatrix Süß, von der Goethe Universität Frankfurt, beschäftigt sich unter anderem mit der Charakterisierung neuartiger Klassen regulatorischer RNAs in *Streptomyces coelicolor*. Um neue fRNA-Kandidaten zu lokalisieren, haben wir den Ansatz aus Kapitel 2 auf dieses Genom angewendet.

4.3.1 Daten

Wir haben die folgenden verwandten Organismen miteinander verglichen: *Streptomyces coelicolor* A3(2) (kurz: *S. coelicolor*) [83], *Streptomyces avermitilis* MA-4680 (kurz: *S. avermitilis*) [77] und *Thermobifida fusca YX* (kurz: *T. fusca*) [65]. Die Sequenzen und Annotationen stammen aus `GenBank`.

	S. coelicolor	*S. avermitilis*	*T. fusca*
Quelle	`GenBank`	`GenBank`	`GenBank`
Version	AL645882.2	BA000030.3	CP000088.1
Größe (Mbp)	~ 8,7	~ 9,0	~ 3,6
GC-Gehalt (%)	72,12	70,72	67,50
Protein-kod. Gene	7769	7580	3110
tRNAs+rRNAs	78	86	64

Tab. 4.5: Merkmale der untersuchten Genomsequenzen.

In der Tabelle 4.5 sind die wichtigsten Merkmale der hier betrachteten Genomsequenzen zusammengefasst. Im Vergleich zum Genom von *S. coelicolor* und *S. avermitilis* ist das Genom von *T. fusca* um mehr als fünf Mbp kleiner. Alle Sequenzen zeichnen sich durch einen überdurchschnittlich hohen GC-Gehalt aus. Er variiert zwischen 67,50 % in *T. fusca* und 72,12 % in *S. coelicolor*.

4.3.2 Ergebnisse

Komparativer Ansatz

Im linken Teil der Abbildung 4.3 ist die Menge der Kandidaten dargestellt, die mit Hilfe des komparativen Ansatzes vorhergesagt wurden. Die Menge der fRNA-Kandidaten in *T. fusca* ist deutlich geringer, als in den Vergleichsgenomen. Betrachten wir nur Vorhersagen mit einem Score von mindestens 0,9, so ergeben sich in *T. fusca* insgesamt 27 Kandidaten, was 0,15 % der gesamten Genomsequenz entspricht. Die Längen der Kandidaten variieren zwischen 50 und 470 bp. In *S. coelicolor* konnten wir 966 fRNA-Kandidaten vorhersagen. Das entspricht 1,44 % der Genomsequenz. Diese Kandidaten sind zwischen 45 und 960 bp lang. In *S. avermitilis* wurden 916 Sequenzen bzw. 1,15 % der Genomsequenz als fRNA klassifiziert. Hier sind die Kandidaten zwischen 45 und 690 bp lang. Im Gegensatz zu *S. coelicolor* und

S. avermitilis, für die drei Viertel ihrer fRNA-Kandidaten kürzer sind als 160 bp, erreichen sie in *T. fusca* eine Länge von 350 bp. Im Vergleich der Kandidatenmengen untereinander, fällt *T. fusca* durch einen besonders kleinen Genomsequenzanteil, der als fRNA-kodierend klassifiziert wurde, auf.

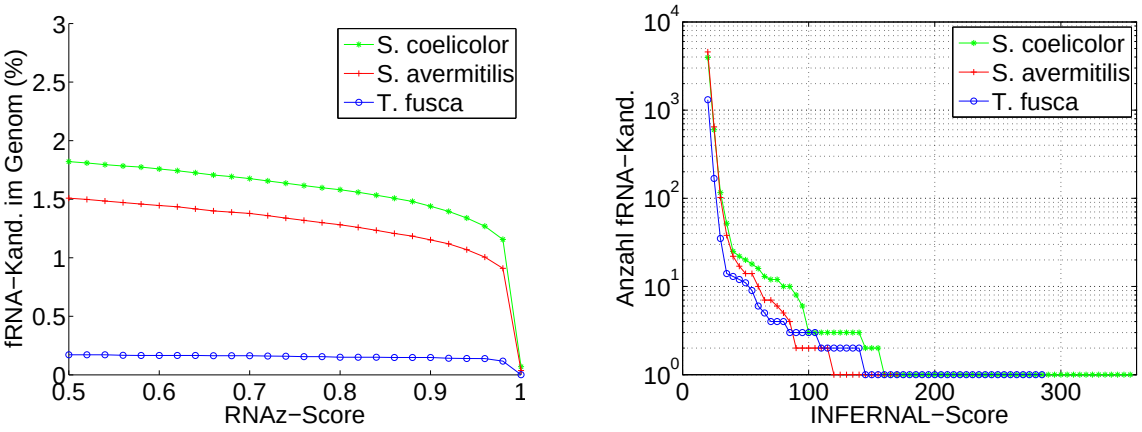


Abb. 4.3: Menge vorhergesagter fRNA-Kandidaten mit dem komparativen Ansatz (links) und dem Kovarianzmodell-basierten Ansatz (rechts).

Der größte Teil der Kandidaten in *S. coelicolor*, insgesamt 814 von 966, wurde im Zusammenhang mit Sequenzen aus *S. avermitilis* als fRNA klassifiziert. Dahingegen stehen nur zwei der Kandidaten aus *S. coelicolor* in einer Beziehung zu Kandidaten aus beiden Vergleichsgenomen. *S. coelicolor* zeichnet sich in diesem Datensatz durch die größte Anzahl an Eigentreffern, d. h. Kandidaten aus. Insgesamt 153 Kandidaten wurden ausschließlich in *S. coelicolor* gefunden, 96 ausschließlich in *S. avermitilis* und 24 ausschließlich in *T. fusca*. Die Kandidaten in *T. fusca* zeigen, bis auf drei Ausnahmen, keine Ähnlichkeit zu Kandidaten aus den anderen beiden Genomen und wurden ausschließlich mit Eigentreffern als fRNA klassifiziert. Die Gesamtübersicht über die Beziehungen zwischen den Kandidaten der einzelnen Genome ist im Anhang in der Tabelle A.5 dargestellt.

Kovarianzmodell-basierter Ansatz

In der Tabelle 4.6 und dem rechten Teil der Abbildung 4.3 sind diejenigen fRNA-Kandidaten aufgeführt, welche mit Hilfe des Kovarianzmodell-basierten Ansatzes vorhergesagt wurden. Für einen INFERNAL-Score größer als 50, wurden vorwiegend regulatorische Elemente vorhergesagt. Während in *S. coelicolor* und *S. avermitilis*

gleich mehrere Vertreter der fRNA-Familie *ydaO-yuaA* gefunden wurden, weist *T. fusca* kein einziges Mitglied dieser Familie auf. Dahingegen wurde die höchste Anzahl an *TPP*-fRNAs in *T. fusca* lokalisiert. Ebenfalls auffällig ist die vergleichsweise hohe Anzahl an *Cobalamin*-fRNAs in *S. coelicolor*.

fRNA-Name	Typ	*S. coelicolor*	*S. avermitilis*	*T. fusca*
ydaO-yuaA	cis-El.	5	4	-
Cobalamin	cis-El.	5	2	1
TPP	cis-El.	1	2	3
Glycine	cis-El.	2	2	1
FMN	cis-El.	1	1	1
RNaseP_bact_a	Gen	1	1	1
tmRNA	Gen	1	1	1
SAM	cis-El.	1	1	1
T-box	cis-El.	1	-	1
5S_rRNA	Gen	1	-	-
SRP_bact	Gen	-	-	1
SSU_rRNA_5	Gen	1	-	-

Tab. 4.6: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`).

In *S. coelicolor* wurden 14 der insgesamt 20 Kandidaten mit einem INFERNAL-Score von mehr als 50 mit Hilfe des komparativen Ansatzes bestätigt. In *S. avermitilis* wurden 12 der insgesamt 14 Kandidaten und in *T. fusca* 2 der insgesamt 11 Kandidaten bestätigt.

4.4 Ozeanobacillus iheyensis

Ozeanobacillus iheyensis, das vom Meeresboden isoliert wurde, ist eine mit dem Bacillus verwandte Spezies, die ausgesprochen halotolerant ist, d. h. in stark salzhaltigen Lebensräumen überleben kann. Außerdem ist es ein alkaliphiler Organismus, d. h. er kann in einer Umgebung mit einem hohen pH-Wert leben [97]. Da er ebenso wie *Bacillus licheniformis* und *Bacillus subtilis* zur extrazellulären Enzymproduktion in der Lage ist, hat er eine wichtige technische Bedeutung für die Industrie, wie z. B. in der Waschmittelproduktion [92].

In der Abteilung für Allgemeine Mikrobiologie (Prof. Stülke) an der Georg-August-Universität Göttingen, wird unter anderem die Funktionsweise kleiner funktioneller RNAs untersucht. Im Rahmen einer Kooperation haben wir daher nach fRNA-Kandidaten in *Ozeanobacillus iheyensis* gesucht.

4.4.1 Daten

Zur komparativen Untersuchung von *Oceanobacillus iheyensis* HTE831 (kurz: *O. iheyensis*) [97] wurden *Bacillus licheniformis* DSM 13 (kurz: *B. licheniformis*) [100] und *Bacillus subtilis* (kurz: *B. subtilis*) [58] als Vergleichsorganismen herangezogen. Alle Sequenzen und Annotationsinformationen stammen aus `GenBank`.

	O. iheyensis	*B. licheniformis*	*B. subtilis*
Quelle	`GenBank`	`GenBank`	`GenBank`
Version	BA000028.3	AE017333.1	AL009126.2
Größe (Mbp)	∼ 3,6	∼ 4,2	∼ 4,2
GC-Gehalt (%)	35,68	46,19	43,52
Protein-kod. Gene	3496	4196	4106
tRNAs+rRNAs	91	93	116

Tab. 4.7: Merkmale der untersuchten Genomsequenzen.

Die Tabelle 4.7 gibt eine Übersicht über die betrachteten Organismen und ihre Sequenzmerkmale. *O. iheyensis* weist mit 3,6 Mbp das kleinste Genom in diesem Datensatz auf. Der GC-Gehalt aller Genome liegt unter 50 %. Der GC-Gehalt des *O. iheyensis*-Genoms ist um ungefähr 10 % niedriger, als der GC-Gehalt der Vergleichssequenzen.

4.4.2 Ergebnisse

Komparativer Ansatz

Im linken Teil der Abbildung 4.4 ist die Menge der vorhergesagten Kandidaten in Abhängigkeit vom RNAz-Score dargestellt. Mit Hilfe dieses Ansatzes haben wir für einen Score von mindestens 0,9 insgesamt 206 Sequenzen (1,06 % der gesamten Genomsequenz von *O. iheyensis*) als fRNA-Kandidaten klassifiziert. Diese Sequenzen sind zwischen 45 und 755 bp lang. In *B. licheniformis* wurden 250 Sequenzen

(0,63 % der Genomsequenz) als fRNA klassifiziert. Sie sind zwischen 45 und 350 bp lang. In *B. subtilis* haben wir 250 Sequenzen (0,64 % der Genomsequenz) als fRNA-Kandidaten vorhergesagt. Diese Sequenzen variieren zwischen 45 und 590 bp in der Länge. Drei Viertel der fRNA-Kandidaten in *B. licheniformis* und *B. subtilis* sind kürzer als 145 bp. In *O. iheyensis* sind drei Viertel der Kandidaten kürzer als 240 bp. Obwohl in *O. iheyensis* der größte Genomsequenzanteil als fRNA-kodierend klassifiziert wurde, ist die absolute Anzahl an Kandidaten um ein Fünftel niedriger als in den Vergleichsgenomen.

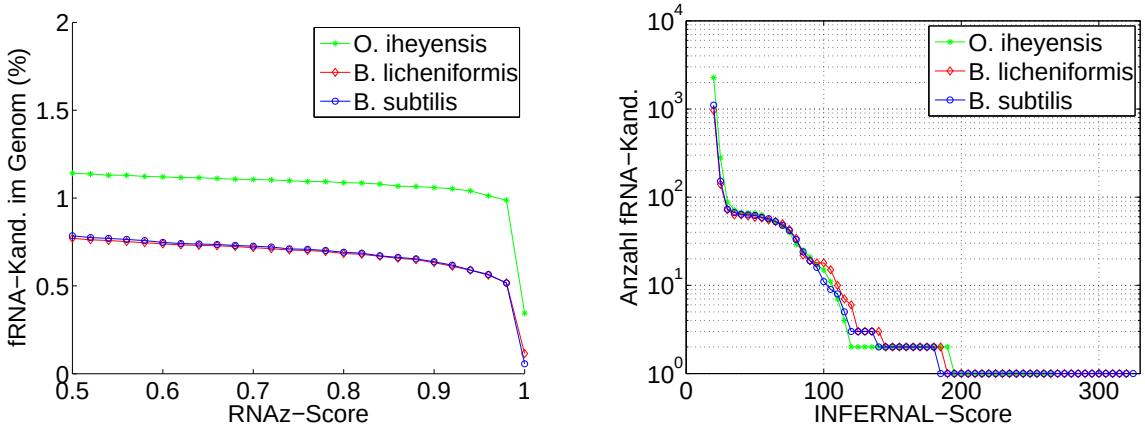


Abb. 4.4: Menge vorhergesagter fRNA-Kandidaten mit dem komparativen Ansatz (links) und dem Kovarianzmodell-basierten Ansatz (rechts).

In *O. iheyensis* stehen 9 der 209 fRNA-Kandidaten gleichzeitig in einer Beziehung zu Kandidaten aus beiden Vergleichsgenomen. Vergleichen wir die Genome paarweise, so ändert sich die Größenordnung der zusammenhängenden Kandidaten in *O. iheyensis* und *B. licheniformis* bzw. in *O. iheyensis* und *B. subtilis* kaum. Die meisten potentiellen fRNAs in *O. iheyensis*, insgesamt 196, wurden ausschließlich in diesem Genom lokalisiert. Im Gegensatz dazu haben wir über 200 der Kandidaten in *B. licheniformis* im Zusammenhang mit Sequenzen aus *B. subtilis* vorhergesagt. Der Anteil der Eigentreffer an allen Kandidaten in *B. licheniformis* und in *B. subtilis* beträgt weniger als ein Fünftel. Die Übersicht über die Beziehungen der Kandidaten zwischen allen Genomen dieses Datensatzes ist im Anhang in Tabelle A.6 zusammengefasst.

Kovarianzmodell-basierter Ansatz

Die Anzahl der Kandidaten, welche mit Hilfe des Kovarianzmodell-basierten Ansatzes gefundenen wurden, ist in allen drei Genomen vergleichbar hoch (siehe Abbildung 4.4 rechts). In der Tabelle 4.8 sind die Kandidaten mit einem INFERNAL-Score über 50 zusammengefasst. Ebenso wie im Datensatz von *B. amyloliquefaciens* (siehe Abschnitt 4.1), für den wir die Ergebnisse von *B. licheniformis* und *B. subtilis* bereits vorgestellt haben, konnten wir in *O. iheyensis* eine besonders hohe Anzahl an regulatorischen Elementen vorhersagen. Dabei haben wir auffallend viele zur *T-Box*- und der *SAM*-Riboswitch-Familie gehörende Kandidaten beobachtet.

fRNA-Name	Typ	*O. iheyensis*	*B. licheniformis*	*B. subtilis*
T-box	cis-El.	22	18	20
SAM	cis-El.	13	10	11
TPP	cis-El.	5	4	5
Purine	cis-El.	4	4	5
PyrR	cis-El.	3	2	3
6S	Gen	2	2	2
yybP-ykoY	cis-El.	2	2	1
FMN	cis-El.	2	2	1
ykkC-yxkD	cis-El.	1	2	2
ydaO-yuaA	cis-El.	1	2	2
Lysine	cis-El.	2	2	1
L20_leader	cis-El.	1	1	1
RNaseP_bact_b	Gen	1	1	1
ylbH	cis-El.	1	1	1
SRP_bact	Gen	1	1	1
L10_leader	cis-El.	1	1	1
glmS	cis-El.	1	1	1
tmRNA	Gen	1	1	1
Glycine	cis-El.	1	-	1
ykoK	cis-El.	-	1	1
Cobalamin	cis-El.	-	1	-
Intron_gpII	Intron	1	-	-

Tab. 4.8: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`).

Ein Vergleich der Kandidaten aus dem komparativen und dem Kovarianzmodell-basierten Ansatz hat ergeben, dass in *O. iheyensis* 6 der insgesamt 66 fRNA-Kandidaten, in *B. licheniformis* 35 der insgesamt 59 Kandidaten und in *B. subtilis* 35 von insgesamt 62 Kandidaten mit Hilfe des komparativen Ansatzes bestätigt wurden.

4.5 Pyrococcus furiosus

Am Archaeenzentrum der Universität Regensburg, in der Arbeitsgruppe von Prof. Thomm, wurde im Rahmen einer Diplomarbeit das Archaeon *Pyrococcus furiosus* untersucht. Im Fokus stand das Vorkommen und die Funktionsweise eines *Hfq*-ähnlichen Proteins. *Hfq* ist in Bakterien bekannt als ein fRNA-bindendes Protein. Es wird benötigt, damit diese speziellen fRNAs ihre Funktion ausüben können [110]. Im Rahmen einer Kooperation haben wir in *Pyrococcus furiosus* und geeigneten Vergleichsorganismen nach fRNA-Kandidaten gesucht.

4.5.1 Daten

Für die komparative Untersuchung von *Pyrococcus furiosus* DSM 3638 (kurz: *P. furiosus*) [67] haben wir *Pyrococcus abyssi* GE5 (kurz: *P. abyssi*) [21] und *Pyrococcus horikoshii* OT3 (kurz: *P. horikoshii*) [54] in den Vergleichsdatensatz aufgenommen. Die Genomsequenzen stammen aus `GenBank`, die Annotationsinformationen stammen aus `Genome Properties Database` von `TIGR` (**T**he **I**nstitut for **G**enomic **R**eacherch) [46].

	P. furiosus	*P. abyssi*	*P. horikoshii*
Quelle	`GenBank/ TIGR`	`GenBank/ TIGR`	`GenBank/ TIGR`
Version	AE009950.1	AL096836.1	BA000001.2
Größe (Mbp)	$\sim$ 1,9	$\sim$ 1,8	$\sim$ 1,7
GC-Gehalt (%)	40,77	44,71	41,88
Protein-kod. Gene	2065	2006	1941
tRNAs+rRNAs	50	50	51

Tab. 4.9: Merkmale der untersuchten Genomsequenzen.

In der Tabelle 4.9 sind die untersuchten Sequenzen und ihre wichtigsten Merkmale

aufgeführt. Alle drei Organismen weisen ein kleines Genom von unter 2 Mbp und ein GC-Gehalt zwischen 40 % und 45 % auf.

4.5.2 Ergebnisse

Komparativer Ansatz

Die Menge der Kandidaten, welche mit Hilfe des komparativen Ansatzes vorhergesagt wurden, ist im linken Teil der Abbildung 4.5 dargestellt. In *P. furiosus* konnten 99 Sequenzen mit einem RNAz-Score von mehr als 0,9 als fRNA klassifiziert werden. Diese Kandidaten entsprechen 0,67 % der gesamten Genomsequenz. In *P. abyssi* haben wir 95 Sequenzen (0,64 % der Genomsequenz) und in *P. horikoshii* 102 Sequenzen (0,70 % der Genomsequenz) als fRNA vorhergesagt. Die Kandidaten sind in allen Organismen mindestens 45 bp lang. In *P. furiosus* werden sie bis zu 500 bp lang, in *P. abyssi* bis zu 415 bp und in *P. horikoshii* bis zu 555 bp. Drei Viertel der Kandidaten weisen für jeden der drei Organismen jedoch eine Länge von weniger als 140 bp.

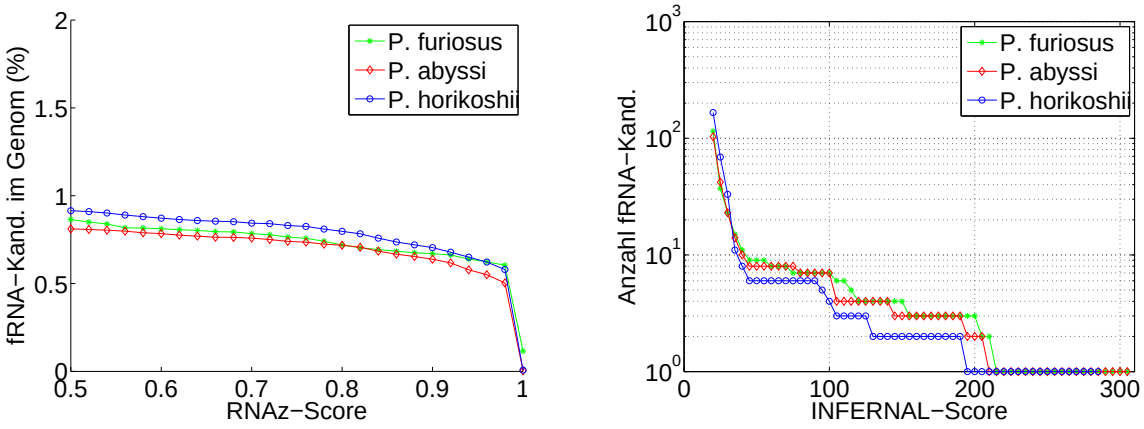


Abb. 4.5: Menge vorhergesagter fRNA-Kandidaten mit dem komparativen Ansatz (links) und dem Kovarianzmodell-basierten Ansatz (rechts).

Betrachten wir die Beziehungen der Kandidaten zueinander (siehe Anhang, Tabelle A.7), so stellen wir fest, dass in jedem der drei Genome mindestens 20 Kandidaten im Zusammenhang mit Sequenzen aus den beiden anderen Genomen als potentielle fRNA klassifiziert wurden. Betrachten wir die Genome paarweise, so weisen *P. abyssi* und *P. horikoshii* jeweils mit 70 bzw. 71 die größte Menge zusammenhängender

Kandidaten auf. Dahingegen haben wir in *P. furiosus* mit 30 die höchste Anzahl an Eigentreffern. Die Gesamtmenge der Kandidaten ist in allen drei Genomen jedoch ungefähr gleich groß (siehe Abbildung 4.5 links).

Kovarianzmodell-basierter Ansatz

Im rechten Teil der Abbildung 4.5 und der Tabelle 4.10 sind die Kandidaten aus dem Kovarianzmodell-basierten Ansatz mit einem INFERNAL-Score von mindesten 50 dargestellt. Es wurden ausschließlich Gene und keine cis-regulatorischen Elemente für diesen Mindestscore lokalisiert. Die Gene *HgcA*, *HgcC*, *HgcG*, *HgcE*, *HgcF*, *SscA* und *snoR9* wurden in AT-reichen Thermophilen, d. h. Lebewesen, die Temperaturen von 45-80° C bevorzugen, bzw. Hyperthermophile, d. h. Lebewesen, die Temperaturen von 80-120° C bevorzugen, zum ersten Mal von Klein *et al.* identifiziert [57]. Sie wurden an Hand ihres hohen GC-Gehalts mit Hilfe von QRNA vorhergesagt. Die meisten von ihnen konnten mit einer Northern-blot-Analyse und der RACE-PCR-Analyse experimentell bestätigt werden [57]. Bis auf *snoR9* ist die Funktion dieser fRNA unbekannt. *snoR9* ist an der Modifikation von *small nuclear RNAs* (snRNAs) beteiligt.

fRNA-Name	Typ	*P. furiosus*	*P. abyssi*	*P. horikoshii*
SRP_euk_arch	Gen	1	1	1
HgcF	Gen	1	1	1
SscA	Gen	1	1	1
HgcG	Gen	1	1	1
HgcE	Gen	1	1	1
snoR9	Gen	1	1	1
RNaseP_bact_a	Gen	1	1	-
RNaseP_arch	Gen	1	1	-
HgcC	Gen	1	-	-

Tab. 4.10: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`).

Bei dem Vergleich der Kandidaten, welche mit Hilfe des komparativen und des Kovarianzmodell-basieren Ansatzes bestimmt wurden, haben wir festgestellt, dass 6 der insgesamt 9 Kandidaten in *P. furiosus*, 6 der insgesamt 8 Kandidaten in *P. abyssi* sowie 5 der insgesamt 6 Kandidaten in *P. horikoshii* mit beiden Ansätzen

vorhergesagt wurden.

4.6 Rhizobium sp. NGR234

Rhizobium sp. NGR234 ist ein gramnegatives Bodenbakterium und lebt in Symbiose mit Pflanzen. Es ist in der Lage, Stickstoff zu fixieren und es dem Wirt zuzuführen. Bemerkenswert ist, dass das Genom von *Rhizobium* sp. NGR234 mehr unterschiedliche Sekretionssysteme als jede andere *Rhizobium*-Spezies kodiert, was es dazu befähigt, mit mehr Pflanzen in Symbiose leben zu können, als das bei jeder anderen *Rhizobium*-Spezies der Fall ist [12].

Das Genom von *Rhizobium* sp. NGR234 besteht aus drei Replikons: zwei Plasmiden (pNGR234a und pNGR234b) und einem Chromosom. Das pNGR234a-Plasmid wurde von Freiberg *et al.* [36] sequenziert. In einer Kooperation des Göttinger Genomlabors (G_2L) der Universität Göttingen und des Laboratoriums von Prof. Broughton und Dr. Perret von der Universität Genf (Schweiz) wurde das zweite pNGR234b-Plasmid [96] und das Chromosom sequenziert. Für weitere Untersuchungen des Genoms haben wir in Kooperation mit dem G_2L eine fRNA-Vorhersage für diesen Organismus und verwandte Spezies erstellt. Die erzielten Ergebnisse wurden bei *Applied and Environmental Microbiology* [94] zur Veröffentlichung eingereicht.

4.6.1 Daten

Die genomische Sequenz von *Rhizobium (Sinorhizobium)* sp. NGR234 (kurz: *R.* NGR234), ist insgesamt 6,9 Mbp lang und setzt sich aus einem Chromosom (chrNGR234) und zwei Plasmiden (pNGR234a [36], pNGR234b) zusammen. Für die komparative fRNA-Vorhersage in *R.* NGR234 haben wir den folgenden Vergleichsdatensatz zusammengestellt: *Rhizobium etli* CFN42 (kurz: *R. etli*) [42], *Agrobacterium tumefaciens* str. C58 (kurz: *A. tumefaciens*) [105], *Sinorhizobium medicae* WSM419 (kurz: *S. medicae*) und *Sinorhizobium meliloti* 1021 (kurz: *S. meliloti*) [13].

	R. NGR234			*A. tumefaciens*			
Quelle	G_2L - Universität Göttingen			GenBank			
Abkürzung	chrNGR234	pNGR234a	pNGR234b	cchrC58	lchrC58	pTi	pAt
Version	-	U00090.1	-	AE007869.2	AE007870.2	AE007871.2	AE007872.2
Größe (Mbp)	∼ 3,9	∼ 0,5	∼ 2,4	∼ 2,8	∼ 2,1	∼ 0,2	∼ 0,5
GC-Gehalt (%)	63,03	58,49	62,30	59,38	59,28	56,67	57,33
Protein-kod. Gene	3638	483	2351	2735	1851	197	542
tRNAs+rRNAs	61	1	0	48	21	0	0

	R. etli						
Quelle	GenBank						
Abkürzung	chrRetli42	p42a	p42b	p42c	p42d	p42e	p42f
Version	CP000133.1	CP000134.1	CP000135.1	CP000136.1	U80928.5	CP000137.1	CP000138.1
Größe (Mbp)	∼ 4,4	∼ 0,2	∼ 0,2	∼ 0,3	∼ 0,4	∼ 0,5	∼ 0,6
GC-Gehalt (%)	61,27	58,00	61,81	61,52	57,82	61,67	61,22
Protein-kod. Gene	4035	175	163	232	336	455	567
tRNAs+rRNAs	59	0	0	0	0	0	0

	S. medicae				*S. meliloti*		
Quelle	GenBank				GenBank		
Abkürzung	chrSMED	pSMED01	pSMED02	pSMED03	chrSM1021	pSymA	pSymB
Version	CP000738.1	CP000739.1	CP000740.1	CP000741.1	AL591688.1	AE006469.1	AL591985.1
Größe (Mbp)	∼ 3,8	∼ 1,6	∼ 1,2	∼ 0,2	∼ 3,7	∼ 1,4	∼ 1,7
GC-Gehalt (%)	61,50	61,46	59,89	60,06	62,73	60,37	62,42
Protein-kod. Gene	3529	1441	1094	149	3359	1290	1584
tRNAs+rRNAs	61	1	0	0	60	0	1

Tab. 4.11: Merkmale der untersuchten Genomsequenzen.

Ebenso wie das *R.* NGR234-Genom besteht die genomische Sequenz von *S. meliloti* aus einem Chromosom (chrSM1021) und zwei Plasmiden (pSymA und pSymB). Das *A. tumefaciens*-Genom setzt sich aus einem zirkulären Chromosom (cchrC58), einem linearen Chromosom (lchrC58) und zwei Plasmiden (pTi und pAt) zusammen. Das *S. medicae*-Genom setzt sich aus einem Chromosom (chrSMED) und drei Plasmiden (pSMED01, pSMED02 und pSMED03) zusammen. Besonders auffällig ist die genomische Sequenz von *R. etli*, die aus einem Chromosom (chrRetli42) und sechs Plasmiden (p42a, p42b, p42c, p42d, p42e und p42f) besteht.

In der Tabelle 4.11 sind alle Replikons zu jedem Genom mit den wichtigsten Sequenzmerkmalen aufgeführt. Die Sequenz- und Annotationsinformationen stammen vorwiegend aus `GenBank`. Das *R.* NGR234-Genom und die dazugehörigen Annotationsinformationen sind noch nicht öffentlich verfügbar und stammen von unseren Kooperationspartnern vom G_2L der Universität Göttingen.

4.6.2 Ergebnisse

Komparativer Ansatz

Der linke Teil der Abbildung 4.6 stellt die Menge der fRNA-Kandidaten dar, die mit dem komparativen Ansatz vorhergesagt wurden. Der Ansatz brachte für einen Mindestscore von 0,9 die meisten Kandidaten in *S. meliloti*, insgesamt 2310 Sequenzen (4,46 % der Genomsequenz), und in *S. medicae*, insgesamt 2058 Sequenzen (4,05 % der Genomsequenz) hervor. In *R.* NGR234 wurden halb so viele Kandidaten lokalisiert, insgesamt 1377 Sequenzen (2,35 % der Genomsequenz) und noch deutlich weniger in *R. etli*, insgesamt 489 Sequenzen (0,88 % der Genomsequenz) und *A. tumefaciens*, insgesamt 308 Sequenzen (0,61 % der Genomsequenz). Wie sich die Anzahl der Kandidaten und der prozentuale Sequenzanteil auf die einzelnen Replikons verteilt, ist der Tabelle A.2 im Anhang zu entnehmen.

Alle vorhergesagten Kandidaten sind mindestens 45 bp lang. Ihre maximale Länge variiert je nach Organismus. In *R.* NGR234 sind die Kandidaten maximal 590 bp lang, in *A. tumefaciens* 470 bp und in *R. etli* 670 bp. Die längsten Kandidaten haben wir in *S. medicae* mit einer Länge von 1605 bp in den Plasmiden pSMED02 und pSMED03 vorhergesagt. Die zweitlängsten Kandidaten mit einer Länge von 1345 bp haben wir in der Chromosomsequenz von *S. meliloti* gefunden. Mindestens drei Viertel der Kandidaten in allen Genomen sind kürzer als 225 bp.

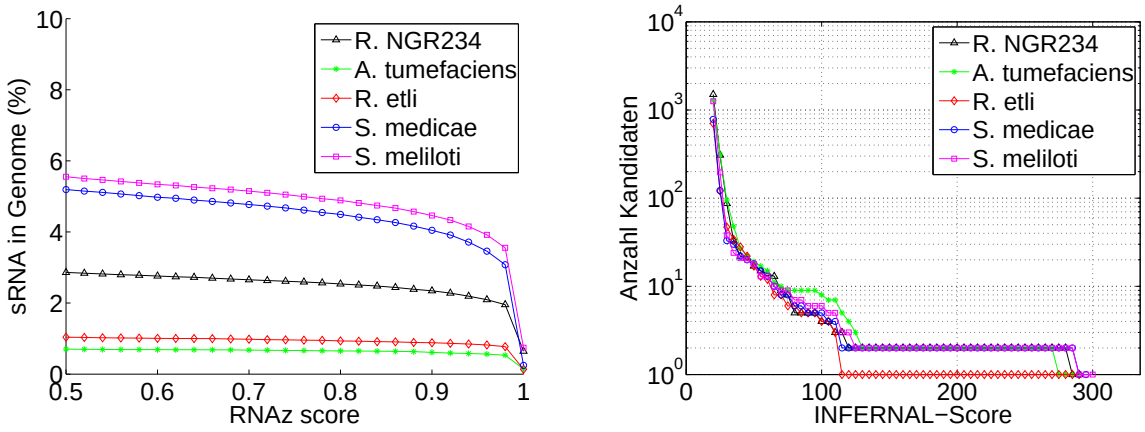


Abb. 4.6: Menge vorhergesagter fRNA-Kandidaten mit dem komparativen Ansatz (links) und dem Kovarianzmodell-basierten Ansatz (rechts).

In der Tabelle A.8 im Anhang sind die Beziehungen der Kandidaten, die aus dem komparativen Ansatz resultieren, zueinander, für alle Replikons zu jedem Genom aufgeführt. Besonders viele gemeinsame Treffer, sowohl in den Chromosomsequenzen als auch in den Plasmiden, weisen *S. medicae* und *S. meliloti* auf. Die Kandidaten in *R.* NGR234 stehen vor allem mit den Kandidaten aus diesen beiden Genomen in Beziehung. Betrachten wir nur den Zusammenhang in den Chromosomsequenzen der drei Genome (siehe Tabelle 4.12), so stellen wir fest, dass mehr als die Hälfte der Kandidaten in *R.* NGR234 eine Beziehung sowohl zu *S. medicae* als auch *S. meliloti* aufweist. In *S. medicae* und *S. meliloti* ist es jeweils ungefähr ein Drittel der gesamten Kandidaten.

	Eigentreffer	chrNGR234	chrSMED	chrSM1021	Alle
chrNGR234	192	1018	586	768	528
chrSMED	74	478	1231	1123	444
chrSM1021	198	744	1200	1586	556

Tab. 4.12: In Beziehung stehende Kandidaten in den Chromosomsequenzen von *R.* NGR234, *S. medicae* und *S. meliloti*.

Kovarianzmodell-basierter Ansatz

Im rechten Teil der Abbildung 4.6 ist die Menge der fRNA-Kandidaten, die mit Hilfe des Kovarianzmodell-basierten Ansatzes vorhergesagt wurden, in Abhängigkeit vom INFERNAL-Score dargestellt. Von den 17 Kandidaten in *R.* NGR234 haben wir 14 ebenfalls mit dem komparativen Ansatz vorhergesagt. In *A. tumefaciens* wurden 10 der 19 Kandidaten, in *R. etli* 9 der 17 Kandidaten, in *S. medicae* 18 der 18 Kandidaten und in *S. meliloti* 16 der 18 Kandidaten mit dem komparativen Ansatz bestätigt. In der Tabelle A.2 sind alle Kandidaten, die mit Hilfe des komparativen oder des Kovarianzmodell-basierten oder mit beiden Ansätzen vorhergesagt wurden, zusammengefasst. Es wird insbesondere angegeben, wieviele Kandidaten in welchem Replikon eines Genoms gefunden wurden.

In der Tabelle 4.13 sind alle Kandidaten angegeben, die mit mindestens einem Score von 50 bewertet wurden. Bis auf wenige Ausnahmen befinden sich diese Treffer in den Chromosomsequenzen der einzelnen Genome, bzw. besonders großen Plasmiden, wie pNGR234b in *R.* NGR234, pSMED01 in *S. medicae* und pSymB in *S. meliloti*. Wir haben überwiegend regulatorische cis-Elemente, aber auch einige Gene und eine Intron-Sequenz gefunden. Dabei sind die cis-Elemente *Cobalamin*, *TPP*, *SAM_alpha*, *FMN* und *ybhL*, sowie die Gene *RNaseP_bact_a* und *suhB* besonders hervorzuheben, da wir sie mindestens ein Mal in jedem der Genome lokalisiert haben.

fRNA-Name	Typ	*R.* NGR234			*A. tumefaciens*				*R. etli*							*S. medicae*			*S. meliloti*	
		chrNGR234	pNGR234a	pNGR234b	cchrC58	lchrC58	lchrC58	pAt	chrRetli42	p42a	p42b	p42c	p42d	p42e	p42f	chrSMED	pSMED01	pSMED02	pSMED03	chrSM1021	pSymA	pSymB
Cobalamin	cis-El.	**2**	-	**1**	**4**	**2**	-	-	**2**	-	-	-	-	-	-	**2**	**1**	-	-	**2**	-	**2**
TPP	cis-El.	**1**	-	**2**	**2**	**1**	-	-	**1**	-	**1**	-	-	-	-	**1**	**2**	-	-	**1**	-	**1**
SAM_alpha	cis-El.	**2**	-	-	**2**	-	-	-	**1**	-	-	-	-	-	-	**2**	-	-	-	**2**	-	-
FMN	cis-El.	**1**	-	-	**1**	-	-	-	**2**	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
RNaseP_bact_a	Gen	**1**	-	-	**1**	-	-	-	**1**	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
suhB	Gen	**1**	-	-	-	**1**	-	-	**1**	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
ybhL	cis-El.	**1**	-	-	**1**	-	-	-	**1**	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
speF	cis-El.	**1**	-	-	-	**1**	-	-	-	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
Glycine	cis-El.	-	-	-	**1**	-	-	-	**1**	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
S-element	cis-El.	-	-	-	-	-	-	-	-	-	-	-	**1**	-	-	-	-	-	**1**	-	**1**	-
SRP_bact	Gen	-	-	-	**1**	-	-	-	**1**	-	-	-	-	-	-	**1**	-	-	-	-	-	-
serC	cis-El.	**1**	-	-	-	-	-	-	-	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
6S	Gen	**1**	-	-	-	**1**	-	-	**1**	-	-	-	-	-	-	-	-	-	-	-	-	-
RNaseP_arch	Gen	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**1**	-	-	-	**1**	-	-
ykkC-yxkD	cis-El.	-	-	**1**	-	-	-	-	-	-	-	-	-	**1**	-	-	-	-	-	-	-	-
RNA-OUT	Gen	-	-	-	-	-	-	-	**1**	-	-	-	-	-	-	-	-	-	-	-	-	-
tRNA	Gen	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**1**	-	-
ROSE	cis-El	-	-	**1**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Intron_gpI	Intron	-	-	-	-	-	-	-	**1**	-	-	-	-	-	-	-	-	-	-	-	-	-

Tab. 4.13: Zuordnung der mit INFERNAL lokalisierten Kandidaten für einen Score von mehr als 50 zu einer der bekannten fRNA-Familien (nach `Rfam`).

4.7 Diskussion

In diesem Kapitel haben wir die von uns erzielten Ergebnisse der fRNA-Detektion mit Hilfe des Ansatzes aus Kapitel 2 auf verschiedenen Datensätzen vorgestellt. Die Organismen, im Speziellen der jeweilige Zielorganismus, wurden im Rahmen von Forschungsschwerpunkten einzelner Arbeitsgruppen an uns herangetragen.

Was wir bereits an den Ergebnissen für *E. coli* und verwandter Spezies (Kapitel 3) beobachtet haben, ist auch in diesem Kapitel sichtbar geworden. Die Auswahl der Vergleichsgenomen hat einen erheblichen Einfluss auf die Ergebnisse des komparativen Ansatzes. So haben wir sowohl in Abschnitt 4.1 als auch in Abschnitt 4.4 *B. subtilis* und *B. licheniformis* als Vergleichsgenome in die Untersuchung aufgenommen. Im ersten Fall wurde 1,69 % und im zweiten Fall 0,64 % der Genomsequenz von *B. subtilis* für einen RNAz-Score von mindesten 0,9 als fRNA-kodierend klassifiziert. Im ersten Fall wurden die meisten Kandidaten im Zusammenhang mit Sequenzen aus *B. amyloliquefaciens* und im zweiten Fall im Zusammenhang mit Sequenzen aus *B. licheniformis* gefunden. *B. subtilis* und *B. amyloliquefaciens* weisen deutlich mehr Gemeinsamkeiten auf als *B. subtilis* und *B. licheniformis*. Je enger die Organismen verwandt sind, umso mehr Sequenzähnlichkeiten werden im ersten Schritt des komparativen Ansatzes lokalisiert und können mit Hilfe von RNAz auf eine stabile und konservierte Struktur untersucht werden. Je höher die Sequenzähnlichkeit ist, umso höher ist aber auch die Wahrscheinlichkeit, dass die Sequenzen ebenfalls eine sehr ähnliche Struktur, d. h. eine hohe Strukturkonservierung, aufweisen. Solche Ähnlichkeiten können bei sehr eng verwandten Organismen zufällig auftreten. Aus diesem Grund geht auch die Strukturstabilität als ein Entscheidungskriterium bei RNAz ein. Dennoch ist nicht auszuschließen, dass je höher die Sequenzähnlichkeit zwischen Genomsequenzen ist, desto mehr falsch positive Vorhersagen werden auf Grund dieser Ähnlichkeit gemacht. Daher ist es empfehlenswert, ein Referenzgenom immer mit mehr als einem weiteren Genom zu vergleichen und die Vergleichsgenome nicht nur aus dem nächsten Umfeld des Referenzgenoms zu wählen. Hierdurch wird vor allem für Kandidaten, die im Zusammenhang mit Sequenzen aus mehr als einem weiteren Genom als fRNA vorhergesagt wurden, die Wahrscheinlichkeit einer zufälligen Ähnlichkeit verringert.

Die aus der Anwendung des komparativen Ansatzes resultierenden Kandidaten sind in allen betrachteten Genomen meistens kürzer als 200 bp. Das entspricht den

typischen Längen funktioneller RNAs in der Datenbank `Rfam` (Abschnitt 1.3.2). Dennoch ist der Ansatz ebenfalls in der Lage, lange Kandidaten vorherzusagen. So wurden z. B. in *M. mazei* Kandidaten mit bis zu 750 bp, in *M. acetivorans* mit bis zu 1100 bp und in *M. barkeri* sogar mit bis zu 1600 bp lokalisiert.

Da der Kovarianzmodell-basierte Ansatz mit INFERNAL auf jedes Genom einzeln angewendet wird, ist dieser Schritt unabhängig von der Auswahl der Vergleichsgenome. Dafür hängt er vom Umfang und der Qualität der verwendeten Referenz-fRNA-Datenbank ab. In unseren Untersuchungen haben wir `Rfam` verwendet. Viele fRNAs sind typisch für eine Gruppe verwandter Organismen und kommen in einer anderen Gruppe eventuell gar nicht vor. Werden also Organismen wie *B. amyloliquefaciens* betrachtet, die zu besonders bekannten und gut untersuchten Spezies wie z. B. *E. coli* oder *B. subtilis* verwandt sind, so können mit Hilfe des Kovarianzmodell-basierten Ansatzes vergleichbar viele Treffer mit einem signifikanten Score erzielt werden (siehe Abschnitt 4.1). In bisher wenig untersuchten Organismenklassen, wie das bei *M. mazei* der Fall ist, können entsprechend wenige signifikante Treffer lokalisiert werden (siehe Abschnitt 4.2). Wir konnten nur einen einzigen Kandidaten mit einem INFERNAL-Score von über 50 in *M. mazei* finden. Der komparative Ansatz kommt hingegen ohne a priori Informationen über bekannte fRNA-Familien aus (siehe Abschnitt 1.4.4). Mit Hilfe dieses Ansatzes konnte eine große Menge fRNA-kodierender Regionen vorhergesagt werden. Gerade in solchen Fällen ist die Anwendung des komparativen Ansatzes besonders wichtig, denn er bietet die Möglichkeit, völlig neuen fRNA-Familien auf die Spur zu kommen.

Ebenso wie wir bei den INFERNAL-Ergebnissen zum ESS- und EYK-Datensatz beobachtet haben, nimmt auch bei den in diesem Kapitel vorgestellen Ergebnissen die Anzahl der Kandidaten mit sinkendem Score exponentiell zu. Daher haben wir nur eine Auswahl an Kandidaten mit einem unserer Meinung nach besonders signifikanten Score von mehr als 50 vorgestellt.

5 Translationskontrolle unter Aminosäuremangel in Saccharomyces cerevisiae

Die 5'-Cap-Struktur ist eine Modifizierung des 5'-Endes der mRNA, die vor allem bei Eukaryoten auftritt. Sie ist ein wichtiger Faktor bei der Initiation der Translation, indem sie die Bindung des Ribosoms an die mRNA ermöglicht. Die 5'-Cap-Struktur ist jedoch nicht die einzige Möglichkeit, um eine Translation zu initiieren. So kann z. B. die sogenannte ***i****nternal* ***r****ibosomal* ***e****ntry* ***s****ite* (IRES) ebenfalls die Bindung von Ribosomen an die mRNA vermitteln. IRES ist eine Teilsequenz der 5'-UTR einiger mRNAs und zeichnet sich durch eine spezielle Struktur aus. Sie kann eine wichtige Rolle bei der Initiation der Translation spielen, wenn die kanonische, 5'-Cap-initiierte Translation unter Umweltbedingungen wie z. B. Stress, reduziert wird [41].

In der Arbeitsgruppe von Prof. Braus am Institut für Mikrobiologie und Genetik der Universität Göttingen wird unter anderem Aminosäuremangel als Stressfaktor in der Genexpression in *Saccharomyces cerevisiae* untersucht. Mit Hilfe einer *Proteomanalyse*, d. h. der Untersuchung des gesamten Proteinprofils einer Zelle unter bestimmten Bedingungen, wurde eine Gruppe von Proteinen identifiziert, deren Produktion unter Aminosäuremangel hochreguliert wird. Gleichzeitig konnte mit Hilfe einer *Transkriptomanalyse*, d. h. der Untersuchung aller unter bestimmten Bedingungen in einer Zelle hergestellten RNA-Moleküle, für die Hälfte der dazugehörigen mRNA ein konstantes oder sogar sinkendes Niveau beobachtet werden. Dies lässt vermuten, dass die Hochregulierung der jeweiligen Proteinmenge post-transkriptionell und somit möglicherweise während der Translation erfolgt.

Motiviert durch die Existenz von Elementen wie z. B. IRES [41] und einer Studie von Ringnér *et al.* [84], der zufolge die Strukturstabilität der 5'-UTRs in *S. cere-*

visiae mit der Translationsrate der entsprechenden mRNAs korreliert, wollten wir herausfinden, ob die mRNAs der auffälligen Gene regulatorische Strukturelemente (siehe Abschnitt 1.3) in der 5'- bzw. 3'-UTR aufweisen.

In diesem Kapitel widmen wir uns der bioinformatischen Suche nach solchen regulatorischen Sekundärstrukturelementen. Im ersten Schritt haben wir die UTR-Sequenzen untereinander verglichen und nach möglichen Sequenzähnlichkeiten gesucht. Um eventuelle Ähnlichkeiten zu bekannten fRNAs zu finden, haben wir die Sequenzen mit der `Rfam`-Datenbank verglichen. Anschließend haben wir die Stabilität möglicher Strukturen in den einzelnen UTR-Sequenzen bestimmt und bewertet. Zuletzt haben wir jede UTR-Sequenz mit geeigneten Vergleichsdaten auf das Vorkommen stabiler, konservierter Strukturen mit Hilfe der komparativen fRNA-Detektion aus Kapitel 2 untersucht.

Eine Beschreibung des methodischen Vorgehens wurde im Artikel Rachfall *et al.* [81] veröffentlicht. Der bioinformatische Ansatz der Methode wird in diesem Kapitel ausführlich beschrieben. Die Ergebnisse der bioinformatischen und der experimentellen Untersuchung werden zur Veröffentlichung vorbereitet.

5.1 Daten

5.1.1 Post-transkriptionell regulierte Gene

Die hier untersuchten Sequenzen von *Saccharomyces cerevisiae* (kurz: *S. cerevisiae*), dem auch als Bäckerhefe bekannten Eukaryoten, stammen aus der ***S****accharomyces* ***G****enome* ***D****atabase* (kurz: `SGD`) [2]. Wir betrachteten 85 Gene in *S. cerevisiae*, die eine post-transkriptionell regulierte Genexpression vermuten lassen. Insgesamt 34 der Kandidaten sind bei einer Proteomuntersuchung aufgefallen. Unter Zugabe einer Droge (3-Amino-1,2,4-Triazol, 3AT), die einen Aminosäuremangel bewirkt, erhöhte sich bei diesen Kandidaten die produzierte Proteinmenge deutlich, obwohl die Menge der entsprechenden mRNAs nicht im gleichen Maße gestiegen ist. Für 16 der Kandidaten wurde unter Zugabe der Droge sogar ein Absinken der mRNA-Menge bei gleichzeitigem Anstieg der Proteinmenge beobachtet. Da bei einer Proteomanalyse Gene, die zu einer vergleichsweise niedrigen Proteinmenge führen, nicht beobachtet werden können, stammen alle weiteren Kandidaten aus der Literatur oder theoretischen Überlegungen.

Auf der Suche nach Faktoren der post-transkriptionellen Regulation, konzentrierten wir uns auf die Betrachtung der nicht translatierten Regionen der mRNAs. Die Sequenzen variieren stark in der Länge. Die 5'-UTR sind zwischen wenigen bp und 1200 bp lang, die 3'-UTR-Sequenzen sind ebenfalls zwischen wenigen bp und 568 bp lang. Da es schwierig ist, die Strukturstabilität in besonders kurzen Sequenzen zu beurteilen, betrachten wir im Folgenden nur Sequenzen, mit einer Mindestlänge von 50 bp. Damit verbleiben **67 5'-UTR-Sequenzen** mit einer durchschnittlichen Länge von 198,3 bp und **60 3'-UTR-Sequenzen** mit einer durchschnittlichen Länge von 140 bp. Im Vergleich zu einem GC-Gehalt von 37,27 % des gesamten *S. cerevisiae*-Genoms, beträgt der durchschnittliche GC-Gehalt der 5'-UTR-Sequenzen 33,91 % und der 3'-UTR-Sequenzen 29,10 %.

5.1.2 Potentiell orthologe Gene

Um die komparative fRNA-Detektion (siehe Kapitel 2) anwenden zu können, haben wir verwandte Spezies zum Vergleich herangezogen. Zu jedem ORF in *S. cerevisiae* bietet `SGD` die Sequenzen potentiell orthologer Gene in verwandten Spezies (siehe Tabelle 5.1) an. Nicht alle Gene sind in allen Spezies verfügbar. Mögliche Gründe hierfür können evolutionäre Unterschiede zwischen den Spezies sein. Ein anderer Grund könnte eine eventuell unvollständige Sequenzabdeckung sein, die aus der sogenannten *Shotgun*-Sequenzierung der betrachteten Genome resultiert.

Abkürzung	Spezies	Verwandschafts-verhältnis	Quelle der Sequenzen
MIT_Spar	*S. paradoxus*	eng	Kellis *et al.* [55]
MIT_Smik	*S. mikatae*	eng	Kellis *et al.* [55]
WashU_Smik	*S. mikatae*	eng	Cliften *et al.* [18]
MIT_Sbay	*S. bayanus*	eng	Kellis *et al.* [55]
WashU_Sbay	*S. bayanus*	eng	Cliften *et al.* [18], Kellis *et al.* [55]
WashU_Skud	*S. kudriavzevii*	eng	Cliften *et al.* [18]
WashU_Scas	*S. castellii*	entfernt	Cliften *et al.* [18]
WashU_Sklu	*S. kluyveri*	entfernt	Cliften *et al.* [18]

Tab. 5.1: Zu *S. cerevisiae* verwandte Spezies, welche in `SGD` als Vergleichsorganismen verfügbar sind.

Die UTR-Sequenzen zu den hier betrachteten mRNAs wurden unabhängig voneinander untersucht. Die Datensätze für die komparative fRNA-Detektion bestanden jeweils aus einer 5'-UTR-Sequenz und Sequenzen von 1000 bp Länge stromaufwärts der potentiell orthologen ORFs. Für die Untersuchung der 3'-UTR haben wir entsprechend Sequenzen von 1000 bp Länge stromabwärts dieser ORFs verwendet (siehe Abbildung 5.1).

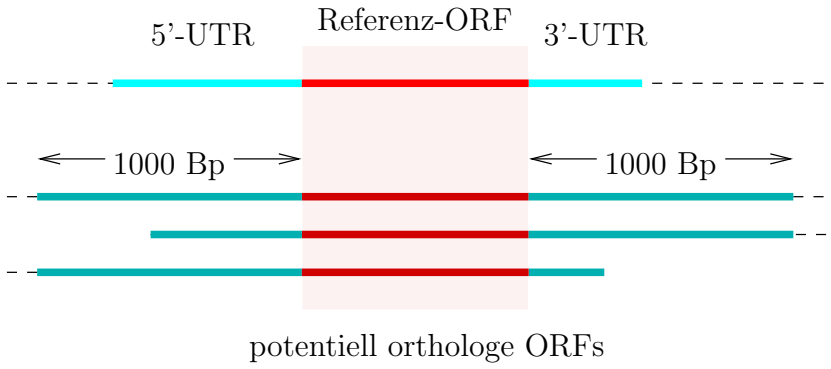


Abb. 5.1: Datenzusammenstellung für komparative fRNA-Detektion in 5'- bzw. 3'-UTR in *S. cerevisiae*. Die 5'-UTR-Sequenz wird in Verbindung mit Sequenzen von bis zu 1000 bp Länge stromaufwärts von potentiell orthologen ORFs bzw. von gegenseitig besten Treffern untersucht. Entsprechend wird die 3'-UTR-Sequenz mit Sequenzen von bis zu 1000 bp stromabwärts dieser ORFs verglichen.

5.2 Methoden

5.2.1 Suche nach Sequenzähnlichkeiten

Unter der Annahme, dass die Translation der hier betrachteten mRNAs über ähnliche UTR-Elemente reguliert wird, haben wir die entsprechenden 5'- bzw. 3'-UTR-Sequenzen zuerst untereinander verglichen und nach sequenziellen Gemeinsamkeiten gesucht. Dazu haben wir die 5'-UTR mit Hilfe von BlastN miteinander verglichen. Ebenso wurde mit den 3'-UTR verfahren. Nur Ergebnisse mit einem E-Value kleiner als 10^{-5} wurden betrachtet.

5.2.2 Vergleich mit `Rfam`

Auf der Suche nach Ähnlichkeiten zu bekannten regulatorischen Elementen, haben wir die 5'- und 3'-UTR-Sequenzen mit bekannten fRNA-Familien aus `Rfam` (siehe Abschnitt 1.3.2) verglichen. Für den Vergleich haben wir INFERNAL (siehe Abschnitt 1.4.5) mit dem Parameter `--local` verwendet und jede UTR-Sequenz mit allen zur Verfügung stehenden fRNA-Familien verglichen. Da im lokalen Vergleich sehr viele zufällige Ähnlichkeiten mit einem niedrigen Score gefunden werden, haben wir nur Ergebnisse mit einem INFERNAL-Score von mindesten 20 genauer betrachtet.

5.2.3 Berechnung der Sekundärstrukturstabilität

Hängt die Regulation der Translation von einer Struktur in der 5'- oder 3'-UTR ab, so darf diese nicht beliebig beeinflussbar sein, sondern nur auf vorbestimmte Faktoren reagieren, um ihre Funktion erfüllen zu können. Wir haben daher erwartet, dass es sich bei solchen Strukturen um besonders stabile Strukturen handelt und infolgedessen die Strukturstabilität der UTR-Sequenzen betrachtet.

Die thermodynamische Stabilität der Sekundärstruktur kann theoretisch aus der minimalen freien Energie (MFE) der Struktur-Vorhersage abgelesen werden. Es gilt dabei der Grundsatz: Je niedriger die MFE ist, umso stabiler ist die Struktur. Die MFE hängt jedoch unter anderem von der Länge und der Dinukleotidzusammensetzung der betrachteten Sequenz ab [106]. Dies erschwert eine Beurteilung des absoluten MFE-Werts. Um die Strukturstabilität dennoch beurteilen zu können, berechnen wir den z-Score der MFE (Abschnitt 1.4.4: Gleichung 1.1). Dabei wird die Stabilität der Struktur einer Zielsequenz mit der Stabilität von Strukturen zufälliger Sequenzen mit gleicher Länge und Dinukleotidzusammensetzung wie die Zielsequenz verglichen.

Um die „zufälligen" Vergleichssequenzen mit den gewünschten Eigenschaften zu erhalten, kann das Programm Dishuffle [19] verwendet werden. Das Programm basiert auf dem Permutationsalgorithmus von Altschul und Ericson [3]. Es ist in der Lage, eine Sequenz zu permutieren und dabei die Dinukleotidzusammensetzung der Sequenz zu bewahren.

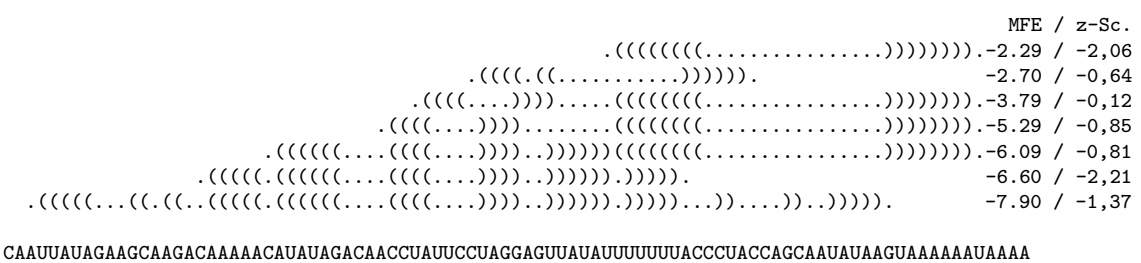


Abb. 5.2: 5'-UTR von *ILV5* mit lokal-optimalen MFE-Strukturen, berechnet mit RNALfold. Am Ende jeder Struktur ist ihr MFE-Wert und der dazugehörige z-Score angegeben.

Strukturstabilität in kompletten UTR-Sequenzen

Um den z-Score für eine komplette UTR-Sequenz berechnen zu können, haben wir diese zuerst mit Dishuffle 100-mal permutiert. Alle Sequenzen wurden danach mit RNAfold (siehe Abschnitt 1.2.3) und der Option `-noLP` gefaltet. Die Option `-noLP` (von: *no lonly pairs*) verhindert die Ausbildung isolierter Basenpaare. Für die von RNAfold berechneten MFE-Werte der zufälligen Sequenzen haben wir den Mittelwert und die Standardabweichung bestimmt und daraus nach Gleichung 1.1 aus Abschnitt 1.4.4 den z-Score ermittelt.

Strukturstabilität lokaler Strukturen

In einem zweiten Schritt haben wir die Stabilität lokaler Sekundärstrukturen der UTR-Sequenzen betrachtet. Die Teilstrukturen wurden nicht beliebig gewählt. Es handelt sich dabei um lokale, thermodynamisch optimale Strukturen, mit einer vorgegebenen maximalen Größe. Um alle diese Strukturen berechnen zu können, ohne die UTR-Sequenz fensterweise falten zu müssen, haben wir RNALfold (siehe Abschnitt 1.2.4) verwendet. Das Ergebnis sind Teilstrukturen, die sich überlappen oder sogar ineinander enthalten sein können, wie am Beispiel der 5'-UTR von *ILV5* in Abbildung 5.2 zu sehen ist.

RNALfold wurde mit den Optionen `-noLP` und `-L 100` auf alle UTR-Sequenzen angewandt. Die erste Option verhindert wieder die Ausbildung isolierter Basenpaare, während die zweite den maximalen Abstand paarender Basen festlegt. Da wir hier nur an lokalen Strukturen interessiert sind, haben wir den Abstand des äußersten Basenpaares auf 100 bp festgelegt. Andererseits ist es nicht sinnvoll, zu kleine Struk-

turen zu betrachteten, da die Strukturstabilität in solchen Fällen nur eine geringe Aussagekraft hat. Daher beschränken wir uns auf Teilstrukturen mit einer zugrunde liegenden Sequenz von mindestens 50 bp Länge.

Die Strukturstabilität einer jeden Teilstruktur wurde separat untersucht. Wir haben die dazugehörige Sequenz 100-mal mit Dishuffle permutiert und die permutierten Sequenzen mit RNAfold, über die komplette Länge, gefaltet. Aus dem Mittelwert und der Standardabweichung der MFE-Werte und dem MFE-Wert der Originalsequenz haben wir den z-Score bestimmt (Bsp. siehe Abbildung 5.2).

5.2.4 Komparative fRNA-Detektion

In Abschnitt 5.2.3 haben wir uns ausschließlich auf die Strukturstabilität, als einen Hinweis für eine regulatorische Struktur, konzentriert. Ein weiterer Hinweis für eine funktionelle Bedeutung einer Struktur, ist ihre Konservierung zwischen verwandten Spezies. Eine geeignete Methode, um solche Strukturen aufzuspüren, ist der komparative Ansatz zur fRNA-Detektion aus Kapitel 2.

Da wir nur an ausgewählten UTR-Sequenzen interessiert sind, haben wir nicht das vollständige *S. cerevisiae*-Genom nach fRNA-Kandidaten abgesucht, wie in Kapitel 2 beschrieben, sondern uns auf diese Sequenzen konzentriert. Als Vergleichsdaten für den komparativen Ansatz haben wir Sequenzen potentiell orthologer Gene verwendet (siehe Abschnitt 5.1.2).

Wir haben zuerst mit Hilfe von BlastN nach Sequenzähnlichkeiten der jeweiligen UTR-Sequenz und ihrer Vergleichssequenzen gesucht. Nur Treffer mit einem E-Value von höchstens 10^{-5} und einer Länge von 50 bp haben wir weiter betrachtet. Die von BlastN erzeugten paarweisen Alignments dienten uns als Grundlage, um signifikante Ähnlichkeiten in mehr als zwei Sequenzen aufzuspüren und zusammenzufassen (Abschnitt 2.2.2 - Abschnitt 2.2.5). Diese Sequenzen wurden schließlich mit ClustalW (Standardparameter) aligniert.

Als letzten Schritt haben wir alle Alignments mit Hilfe von RNAz fensterweise auf das Vorkommen von stabilen und konservierten Strukturen untersucht. Da die Wahl der Fenstergröße das Vorhersageergebnis beeinflusst, haben wir das Fenster zwischen 50 und 200 bp variiert, wobei es jeweils um 5 bp vergrößert wurde. Das Alignment wurde mit jeder Fenstergröße, in einer Schrittweite von 5 bp, durchsucht. Jedem untersuchten Alignmentabschnitt wurde ein Score zugeordnet. Liegt der Score über

einer von uns vorgegebenen Schwelle, werden die Sequenzen in diesem Bereich des Alignments als fRNA-Kandidaten angesehen. Überlappende positive Vorhersagen, fassen wir als einen Kandidat auf.

Der RNAz-Score kann als Wahrscheinlichkeit für das Auftreten einer fRNA-kodierenden Region innerhalb eines untersuchten Bereichs interpretiert werden. Ab einem Score von 0,5 werden die untersuchten Sequenzen als fRNA-Kandidaten angesehen. Je höher der Score ist, umso signifikanter ist die Vorhersage. In Abschnitt 5.3.4 betrachten wir nur Vorhersagen mit einem RNAz-Score von mindestens 0,9. Die vollständige Liste der positiven Vorhersagen, d. h. die Vorhersagen mit einem RNAz-Score von mindestens 0,5, ist im Anhang Abschnitt A.3.3 zu finden.

5.3 Ergebnisse

5.3.1 Gemeinsamkeiten zwischen 5'- bzw. 3'-UTR-Sequenzen

Der BlastN-Vergleich der UTR-Sequenzen untereinander ergab keine signifikanten Treffer. Für einen E-Value von weniger als 10^{-5} haben wir zwischen den 5'-UTR-Sequenzen überhaupt keine Ähnlichkeiten gefunden. Zwischen drei der insgesamt 60 3'-UTR-Sequenzen gab BlastN zwar Treffer an, es handelt sich dabei jedoch um kurze Sequenzabschnitte, mit einer maximalen Länge von 32 bp. Sie zeichnen sich außerdem durch eine besonders niedrige Sequenzkomplexität aus, da sie fast ausschließlich aus Wiederholungen der Sequenz 'TATA' bestehen.

5.3.2 Ähnlichkeiten zu bekannten fRNAs aus `Rfam`

Die UTR-Sequenzen wurden mit Hilfe von INFERNAL mit bekannten fRNA-Familien aus `Rfam` verglichen. Da die Qualität der INFERNAL-Treffer mit sinkendem Score abnimmt, betrachten wir nur solche Treffer, die mindestens einen Score von 20 erreicht haben. Alle diese Treffer sind im Anhang in der Tabelle A.9 aufgelistet.

Insgesamt weisen eine 5'-UTR und acht 3'-UTR partielle Ähnlichkeiten zu bekannten fRNA-Familien auf. In den meisten Fällen sind die Teilsequenz einer UTR mehreren fRNA-Familien zugeordnet. Der INFERNAL-Score ist jedoch nie höher als 26.

Bei den meisten Treffer-Familien handelt es sich um fRNA-Gene, die in ein ei-

Anfragesequenz: 3'-UTR von *HAP4*
Treffer-fRNA: *L20_leader*

```
     ::::::::::<<<<<<__~~~~~~>>>>>>,<~~~~~~>--------------------
   1 uaAaaaAAGaAGAAGCAc*[ 3]*GCUUCUCa*[27]*uaaaaauauaaaauuaauauau 77
     +AAAAAAAG+A:AA::AC      ::UU:U A      U++A+ +++AAAA +AA +UA+
 144 AAAAAAAAGUAAAAAAAC*[13]*UUUUUUAA*[27]*UUUAUCAUAAAAAAAAAGUUAA 231

     -------------<<<<<~~~~~~>>>>>:::::::::::::
  78 aauauuuauuaaGaGCGG*[14]*CCGCuuUUUUUUuuuuuu 127
     AAU++++A+ AA  ::GG      CC::  UUUUUUUUUU+U
 232 AAUGAAAAACAAAUAUGG*[10]*CCAUUAUUUUUUUUUUAU 277
```

Anfragesequenz: 3'-UTR von *SKN7*
Treffer-fRNA: *RbcL_stabil*

```
     <<<<<<<<<<<<<<-<<_~~~~~>>->>>>>>>->>>>>>>:::::
   1 aaauguauuuaaaaauuu*[6]*aauuuuuaaauuauauuuccGGA 46
     A:AU UAU:UA:A:AU:U     :AU:U:UA:A UA AU:U  G A
 257 AUAUAUAUAUAUAUAUAU*[1]*UAUAUAUAUAGUAAAUAUAAGAA 216
```

Abb. 5.3: Ähnlichkeiten der 3'-UTR von *HAP4* und *SKN7* zu regulatorischen Elementen in `Rfam`. Der Vergleich erfolgte mit INFERNAL. Treffer in der 3'-UTR von *HAP4*: fRNA-Familie = *L20_leader*, Score = 20,66, Strang = '+' und in der 3'-UTR von *SKN7*: fRNA-Familie = *RbcL_stabil*, Score = 21,53, Strang = '-'. Für eine Beschreibung des INFERNAL-Ausgabeformats siehe Abschnitt 1.4.5.

genständiges Transkript umgewandelt werden. Keine der 5'-UTR weist hingegen eine deutliche Ähnlichkeit (INFERNAL-Score von mehr als 20) zu einer fRNA-Familie vom Typ „regulatorisches Element" auf. Nur in zwei der 3'-UTR wurde ein solcher Treffer gefunden. Die 3'-UTR von *HAP4* weist eine lokale, mehrmals auf Sequenzebene unterbrochene Ähnlichkeit zu der *L20_leader*-Familie mit der `Rfam`-ID: RF00558 auf. Die 3'-UTR von *SKN7* weist eine Ähnlichkeit zu der *RbcL_stabil*-Familie mit der `Rfam`-ID: RF00197 auf (Abbildung 5.3).

Beim *L20_leader* handelt es sich um eine potentielle regulatorische Struktur im Leader des ribosomalen Protein-kodierenden Gens *L20* [109]. Sie wurde in *B. subtilis* und anderen grampositiven Bakterien mit einem niedrigen GC-Gehalt gefunden. Normalerweise ist dieses Element Teil der 5'-UTR einer mRNA und beeinflusst die

Translation der folgenden Sequenz, welche für die ribosomalen Proteine *L35* und *L20* kodiert.

Der Name der *RbcL_stabil*-Familie ist abgeleitet vom *RbcL*-Protein-kodierenden Gen, dass unter anderem in der Grünalge *Chlamydomonas reinhardtii* auftritt [89]. Das regulatorische Element ist Teil der 5'-UTR der *RbcL*-mRNA. Durch die spezielle Struktur trägt es zur Stabilität des Transkripts bei. Modifikationen dieser Struktur durch Änderungen in der entsprechenden Sequenz führen zu einer 50-fach geringeren Stabilität dieses Transkripts.

5.3.3 Stabilität der Sekundärstrukturen

Um die Stabilität der Sekundärstrukturen der hier betrachteten UTR-Sequenz beurteilen zu können, haben wir den z-Score der MFE berechnet. Im ersten Schritt wurden die Sequenzen in voller Länge untersucht. Im zweiten Schritt haben wir uns auf lokale MFE-Strukturen mit einer Sequenz zwischen 50 und 100 bp Länge konzentriert.

Die Untersuchung der vollständigen Sequenzen liefert für die 5'-UTR z-Scores zwischen -2,95 und +2,53. Im Mittel erreicht der z-Score einen Wert von ungefähr +0,1. Von den insgesamt 67 Sequenzen haben 31 einen negativen z-Score. Die 3'-UTR-Sequenzen weisen einen z-Score zwischen -3,62 und +2,61 auf. Im Mittel erreichen sie einen Wert von ungefähr +0,29. Von den insgesamt 60 3'-UTR weisen 21 einen negativen z-Score auf (Abbildung 5.4 oben).

Die Suche nach lokalen MFE-Strukturen mit RNALfold ergab, abhängig von der Sequenzlänge, zwischen 1 und 60 Teilstrukturen pro UTR-Sequenz. Für jede dieser Teilstrukturen haben wir den z-Score berechnet. Für Teilstrukturen der 5'-UTR erhielten wir Werte zwischen -5,34 und +2,12. Im Mittel beträgt der z-Score -0,72. Für Teilstrukturen der 3'-UTR erhielten wir Werte zwischen -5,44 und +2,37. Im Mittel beträgt der z-Score -0,5. Mehr als 70 % der Strukturen in den 5'-UTR und mehr als 60 % der Strukturen in den 3'-UTR weisen einen negativen z-Score auf (Abbildung 5.4 unten).

Ein z-Score, der kleiner als -4 ist, kann als signifikant angesehen werden [86]. In der 5'-UTR von *ABF1*, *ADE16*, *ADR1*, *MXR1* und *PCL5*, sowie in der 3'-UTR von *HAP4*, *YHI9* und *SKN7* haben wir Strukturen mit einem solchen z-Score gefunden (siehe Anhang: Tabelle A.10, Tabelle A.11).

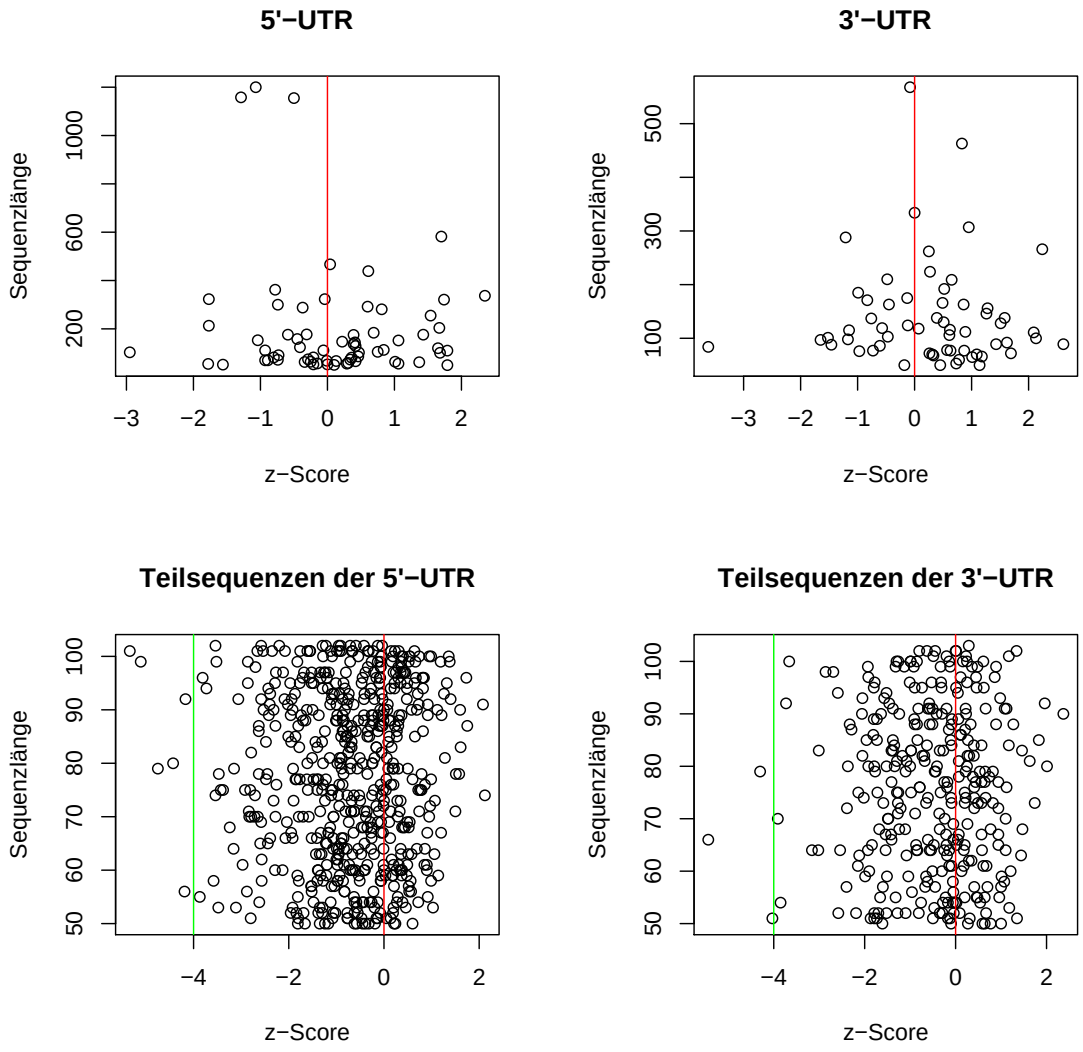


Abb. 5.4: z-Score der MFE für die Strukturen der kompletten UTR-Sequenzen (oben) und für Teilstrukturen mit einer lokal-optimalen MFE (unten). Die rote Linie markiert einen z-Score von 0. Die grüne Linie markiert die Schwelle für einen signifikanten z-Score. Markierungen links von der grünen Linie stehen für Sequenzen mit einer signifikant stabilen Struktur.

5.3.4 fRNA-Detektion mit Hilfe potentiell orthologer Gene

Um passende Vergleichsequenzen für eine komparative fRNA-Detektion in den UTR zu erhalten, haben wir die 5'- bzw. 3'-UTR mit Hilfe von BlastN mit Sequenzen stromaufwärts bzw. stromabwärts von potentiell orthologen ORFs verglichen. Dabei haben wir, bei einem E-Value von maximal 10^{-5}, für 39 der 67 5'-UTR und für 43 der 60 3'-UTR ähnliche Sequenzen in verwandten Spezies gefunden, die mindestens 50 bp lang sind.

Insgesamt konnten wir für 26 mRNAs stabile und konservierte Strukturen mit einem RNAz-Score von mindesten 0,9 als Kandidaten für regulatorische Elemente in den UTR vorhersagen. Für drei mRNAs zu: *ADR1*, *NDD1* und *STE12*, haben wir sowohl in der 5'-UTR als auch in der 3'-UTR eine solche Struktur gefunden. In 15 Fällen haben wir nur in der 5'-UTR und in acht Fällen nur in der 3'-UTR eine signifikante Struktur entdeckt.

potentiell-regulatorischen Struktur in der 5'-UTR, die mit einem hohen RNAz-Score von mindestens 0,9 vorhergesagt wurden, sind in der Tabelle 5.2 zusammengefasst. Die entsprechenden Vorhersagen für 3'-UTR-Sequenzen sind in der Tabelle 5.3 zu finden. Ausführliche Ergebnisse, d. h. Vorhersagen ab einem RNAz-Score von 0,5, sind im Anhang in Abschnitt A.3.3 zu finden. Die Positionsangabe der identifizierten Strukturen beziehen sich auf die Position in der entsprechenden UTR-Sequenz. In der letzten Spalte der Tabellen sind die Spezies angegeben, welche bei der fRNA-Detektion im Zusammenhang mit dem fRNA-Kandidaten aus der jeweiligen Zeile ebenfalls eine potentielle fRNA aufwiesen.

5.4 Diskussion

In diesem Kapitel haben wir nach Sekundärstrukturelementen in den 5'- und 3'-UTRs ausgewählter mRNAs in *S. cerevisiae* gesucht, welche die Translation beeinflussen könnten. Die betrachteten Kandidaten wurden im Rahmen einer Kooperation mit der Abteilung Molekulare Mikrobiologie und Genetik (Prof. Braus), der Universität Göttingen identifiziert. Sie sind durch eine erhöhte Proteinmenge unter Aminosäuremangel aufgefallen. Da die Transkriptmenge von einigen dieser Gene gleichzeitig gleich blieb oder sogar sank, konnte auf eine post-transkriptionelle Regulation geschlossen werden. Motiviert von der Existenz regulatorischer Strukturele-

Gen	sys. Name	UTR-Länge	Start	Ende	Trefferspezies
ABF1	YKL112W	176	1	95	MIT_Spar, MIT_Smik, WashU_Sbay
ADE16	YLR028C	213	1	210	MIT_Spar
ADR1	YDR216W	1158	227	822	MIT_Spar
BMH2	YDR099W	124	61	120	MIT_Spar
CIN5	YOR028C	323	19	241	MIT_Spar
FKH2	YNL068C	288	26	255	MIT_Spar
GCV1	YDR019C	1200	491	570	MIT_Smik, WashU_Skud
			661	784	MIT_Smik
HAP1	YLR256W	300	16	208	MIT_Spar
MBP1	YDL056W	176	61	146	MIT_Spar
MET6	YER091C	70	1	70	MIT_Spar, MIT_Smik
MTG1	YMR097C	467	11	300	MIT_Spar, WashU_Skud
MXR1	YER042W	1155	231	315	MIT_Spar
			331	510	MIT_Spar
			916	965	MIT_Smik, WashU_Smik
NDD1	YOR372C	70	15	69	MIT_Spar
PCL5	YHR071W	362	4	62	MIT_Spar, MIT_Smik, WashU_Smik
			138	357	MIT_Spar, MIT_Smik
SKN7	YHR206W	292	11	109	MIT_Spar, MIT_Smik
			154	289	MIT_Spar
SOK2	YMR016C	281	86	154	MIT_Spar
STE12	YHR084W	323	76	163	MIT_Smik, WashU_Smik
			242	315	MIT_Spar
SWI5	YDR146C	177	35	162	MIT_Spar

Tab. 5.2: Kandidaten für fRNAs in der **5'-UTR** ausgewählter Gene in *S. cerevisiae*. Es sind nur Kandidaten mit einem RNAz-Score von mindestens 0,9 aufgelistet.

Gen	sys. Name	UTR-Länge	Start	Ende	Trefferspezies
ADE1	YAR015W	97	5	89	MIT_Spar
ADE8	YDR408C	210	47	168	MIT_Spar, MIT_Smik, WashU_Smik
ADR1	YDR216W	119	38	107	MIT_Spar
BNA1	YJR025C	171	27	130	MIT_Spar
CPR1	YDR155C	124	33	107	MIT_Spar
FLO8	YER109C	334	205	259	MIT_Spar
FPR1	YNL135C	209	82	136	MIT_Spar, WashU_Smik
HAP4	YKL109W	463	82	182	MIT_Spar, MIT_Smik, WashU_Smik
			201	300	MIT_Smik, WashU_Smik
HAP5	YOR358W	224	6	120	MIT_Spar
NDD1	YOR372C	266	202	259	MIT_Spar
STE12	YHR084W	307	177	296	MIT_Spar

Tab. 5.3: Kandidaten für fRNAs in der **3'-UTR** ausgewählter Gene in *S. cerevisiae*. Es sind nur Kandidaten mit einem RNAz-Score von mindestens 0,9 aufgelistet.

mente in 5'-UTR-Sequenzen in *S. cerevisiae* [41, 84], haben wir nach entsprechenden Strukturen in der 5'- und der 3'-UTR der auffälligen mRNAs gesucht. Das methodische Vorgehen ist in [81] beschrieben. Die Ergebnisse der bioinformatischen und experimentellen Untersuchung werden zur Veröffentlichung vorbereitet.

Um mögliche sequenzielle Ähnlichkeiten unter den 5'-UTRs bzw. den 3'-UTRs aufzuspüren, haben wir die Sequenzen mit BlastN untereinander verglichen. Der Vergleich blieb jedoch weitestgehend erfolglos (Abschnitt 5.3.1). In der Hoffnung, die UTR-Sequenzen könnten Ähnlichkeiten zu bekannten regulatorischen Strukturelementen aus der Datenbank `Rfam` aufweisen, haben wir jede UTR mit jeder fRNA-Familie verglichen. Der Vergleich erfolgte mit INFERNAL. Jedoch auch dieser Schritt brachte keine signifikanten Ergebnisse (Abschnitt 5.3.2), da alle Treffer mit niedrigen Scores bewertet wurden. Die am höchsten bewerteten Treffer mit einem INFERNAL-Score von mindestens 20 zeigten vorwiegend Ähnlichkeiten zu Familien von fRNA-kodierenden Genen (Tabelle A.9). Es ist zwar möglich, aber unwahrscheinlich, dass der gleiche DNA-Abschnitt sowohl für ein eigenständiges fRNA-Molekül als auch für ein Protein kodiert. Viel wahrscheinlicher ist hingegen das Vorkommen einer regulatorischen Struktur innerhalb der UTR einer mRNA, welche an der Translation der entsprechenden mRNA beteiligt ist.

Insgesamt haben wir in den 3'-UTRs zweier Gene Ähnlichkeiten zu regulatorischen Elementen nachgewiesen. Die Mitglieder dieser fRNA-Familien wurden bisher jedoch nur in 5'-UTR-Sequenzen gefunden [109, 89]. Die geringe Sequenzkomplexität (extrem geringer GC-Gehalt) und die niedrigen Scores, mit denen die Kandidaten bewertet wurden (Abbildung 5.3), sind deutliche Hinweise für ein zufälliges Auftreten der Ähnlichkeiten.

Da weder eine signifikante Sequenzähnlichkeit zwischen den UTRs, noch eine überzeugende Ähnlichkeit zu bekannten fRNAs zu finden war, haben wir die Strukturstabilität untersucht. Dabei gingen wir von der Annahme aus, dass regulatorische Strukturelemente eine stabile Struktur aufweisen, da die Struktur der entscheidende Regulationsfaktor ist.

Die Strukturstabilität wird als MFE angegeben und ist als absoluter Wert schwer zu interpretieren, da sie von verschiedenen Merkmalen, wie der Sequenzlänge und der Dinukleotidzusammensetzung abhängt. Um ein vergleichbares Kriterium zu erhalten, haben wir den z-Score der MFE berechnet (Abschnitt 5.2.3).

Mit Hilfe des z-Scores wird die Stabilität einer Struktur, im Vergleich zu der Stabilität von Strukturen zufälliger Sequenzen mit den gleichen Eigenschaften wie die betrachtete Sequenz, beurteilt. Ein negativer z-Score gibt an, dass die Struktur der betrachteten Sequenz im Schnitt stabiler ist, als die Strukturen der Vergleichssequenzen. Dennoch ist das nicht immer ein ausreichender Hinweis auf eine regulatorische Struktur. Einer Studie von Rivas *et al.* [86] zufolge, kann die Strukturstabilität erst für einen z-Score von weniger als -4, als signifikant angesehen werden.

Wir haben zuerst die UTR-Sequenzen in voller Länge gefaltet und jeweils den z-Score für die Gesamtstruktur berechnet (Abschnitt 5.2.3). Für keine der Strukturen hat der z-Score die Signifikanzschwelle von -4 unterschritten (Abschnitt 5.3.3). Ein Grund dafür könnte sein, dass sich eine regulatorische Struktur nicht über die gesamte UTR erstrecken muss, falls sie überhaupt existiert. Um das zu berücksichtigen, haben wir in einem zweiten Schritt zuerst Teilsequenzen der UTR identifiziert, die eine lokale MFE-Struktur ausbilden (Abschnitt 5.2.3) und für diese den z-Score berechnet. Mit der Strategie konnten wir in den UTR-Sequenzen von acht mRNAs eine auffallend stabile Struktur, mit einen z-Score von höchstens -4, identifizieren. Für die mRNAs von *ABF1*, *ADE16*, *ADR1*, *MXR1* und *PCL5* befand sich die Struktur in der 5'-UTR, für die mRNA von *HAP4*, *YHI9* und *SKN7* war sie in der 3'-UTR (siehe Anhang: Tabelle A.10, Tabelle A.11).

Bei der Suche nach regulatorischen Strukturen mit Hilfe des z-Scores, wird außer der Sequenz selbst, keine weitere Information benötigt. Leider weisen viele bekannte fRNAs, deren Funktion von ihrer Struktur abhängt, einen höheren als den empfohlenen z-Score von höchstens -4 auf und können mit diesem Ansatz nicht lokalisiert werden [86]. Daher ist zu befürchten, dass auch in unserer Suche mögliche fRNAs übersehen wurden.

Da neben der Strukturstabilität auch eine konservierte Struktur in verwandten Spezies ein Hinweis auf ein regulatorisches Element ist, haben wir die Strategie zur fRNA-Detektion aus Kapitel 2 auf die 5'- bzw. 3'-UTR-Sequenzen angewandt. Für diesen komparativen Ansatz sind geeignete Vergleichssequenzen notwendig. Es ist nicht möglich, die komparative fRNA-Detektion nur auf die Menge der 5'-UTR-Sequenzen, bzw. nur auf die Menge der 3'-UTR-Sequenzen anzuwenden, da die Sequenzen untereinander keine ausreichende Sequenzähnlichkeiten aufweisen (Abschnitt 5.3.1). Daher haben wir Sequenzen aus verwandten Spezies als Vergleich herangezogen (Abschnitt 5.1.2).

Insgesamt haben wir mit dem komparativen Ansatz eine deutlich höhere Anzahl an Kandidaten (siehe Tabelle 5.2 und Tabelle 5.3) lokalisiert, als mit dem z-Score-Ansatz. Bis auf zwei Ausnahmen, die Kandidaten in der 3'-UTR zu *YHI9* und *SKN7*, wurden alle Kandidaten aus der z-Score-Untersuchung ebenfalls mit dem komparativen Ansatz entdeckt. Eine Komponente des komparativen Ansatzes ist die Anwendung von RNAz zur Bewertung der Strukturstabilität und der Strukturkonservierung zwischen mehreren alignierten Sequenzen. Dabei geht unter anderem der z-Score der MFE der einzelnen Strukturen als ein Merkmal in die Bewertung ein.

Der z-Score ist nicht das einzige Entscheidungskriterium beim komparativen Ansatz. So können Sequenzen trotz eines niedrigeren z-Scores, als fRNA-kodierend eingestuft werden, wenn die Struktur in den verglichenen Sequenzen besonders stark konserviert ist (siehe Abbildung 1.7). Das erklärt die höhere Anzahl an Kandidaten, im Vergleich zur Anwendung des z-Scores als einziges Entscheidungskriterium. Ein signifikanter z-Score, wie das bei einem Wert von weniger als -4 der Fall ist, wirkt sich dennoch auf die Vorhersage mit RNAz aus. Daher ist es nicht verwunderlich, dass die meisten Kandidaten aus der z-Score-Untersuchung ebenfalls vom komparativen Ansatz als fRNAs eingestuft wurden.

Im Ausnahmefall von *YHI9* haben wir keine entsprechenden Sequenzähnlichkei-

ten in den verwandten Spezies entdeckt. Damit fehlt die Basis für eine komparative Vorhersage. Im Fall von *SKN7* wurde zwar eine Sequenzähnlichkeit in *S. paradoxus* lokalisiert. Die entsprechende Teilsequenz der 3'-UTR, umfasste jedoch nicht den gesamten Bereich, welcher in der z-Score-Untersuchung durch eine stabile Struktur aufgefallen ist. Somit war es nicht möglich, die entsprechende Struktur zu identifizieren.

Von den hier verwendeten Ansätzen, um regulatorische Strukturelemente aufzuspüren, brachte der komparative Ansatz die meisten interessanten Kandidaten. Ein Problem des Ansatzes ist seine Abhängigkeit von geeigneten Vergleichssequenzen. Sind diese nicht gegeben, kann keine Vorhersage erfolgen. Der z-Score der MFE hat diese Abhängigkeit nicht. Um jedoch signifikante Vorhersagen zu erhalten, muss ein strikter Schwellenwert gewählt werden. Viele bekannte fRNAs erreichen diesen Wert nicht und werden nicht erkannt. Daher empfiehlt sich der z-Score-Ansatz ergänzend zum komparativen Ansatz, aber nicht als einziges Vorhersagekriterium.

Wir haben in diesem Kapitel eine Liste regulatorischer Strukturelemente in der 5'-UTR oder der 3'-UTR der betrachteten Gene vorhersagen können. Ob diese Strukturen an der Regulation der Translation beteiligt sind oder eine andere Funktion ausüben, muss experimentell verifiziert werden.

6 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein Ansatz zur Detektion strukturbasierter funktioneller RNAs entwickelt und unter Verwendung bekannter Werkzeuge wie BlastN, ClustalW, RNAz und INFERNAL implementiert. Die Anwendung ist darauf ausgerichtet, fRNAs vor allem in langen Sequenzen, wie z. B. kompletten Genomsequenzen, vorherzusagen, kann aber ebenso auf eine Vorauswahl kurzer Sequenzen angewandt werden. Dabei werden zwei unterschiedliche Strategien kombiniert. Ein komparativer Ansatz, der ohne a priori Informationen über bekannte fRNA-Familien auskommt und ein Kovarianzmodell-basierter Ansatz, der Informationen über bekannte fRNA verwendet. Im ersten Ansatz werden Gruppen von Sequenzen anhand konservierter Sequenzinformationen in verwandten Spezies zusammengefasst und aligniert. Die Alignments werden auf Anzeichen konservierter und stabiler Sturkturen untersucht, die ein deutlicher Hinweis auf funktionelle RNAs sind. Im zweiten Ansatz wird nach Sequenz- und Strukturähnlichkeiten zu bekannten fRNA-Familien gesucht.

Es wird sowohl nach fRNA-kodierenden Genen, als auch nach strukturbasierten regulatorischen Elementen gesucht. Da der komparative Ansatz ohne a priori Informationen auskommt, können damit vor allem neue, bisher unbekannte fRNAs aufgespürt werden. In der komparativen Herangehensweise werden mehrere Sequenzen gleichzeitig auf das Vorkommen von fRNAs untersucht. Wir erhalten daher Gruppen von potentiellen fRNAs, die sowohl sequenzielle als auch strukturelle Gemeinsamkeiten aufweisen. Der Vergleich des Genomkontexts ähnlicher Kandidaten, kann erste Hinweise auf ihre Funktion liefern. Liegen solche Kandidaten z. B. alle in der Promotorregion orthologer Gene, so spricht das für eine regulatorische Funktion der potentiellen fRNAs während der Expression des nachfolgenden Gens. Informationen über die Funktion der fRNA-Kandidaten erhalten wir ebenfalls mit Hilfe des zweiten Ansatzes, welcher auf der Suche nach Ähnlichkeiten zu bekannten fRNA-Familien basiert.

Bei einem erfolgreichen Test des Ansatzes auf dem Musterorganismus *E. coli*,

konnte ein Großteil der bekannten fRNAs, sowie neue fRNA-Kandidaten vorhergesagt werden. Danach haben wir den Ansatz im Rahmen verschiedener Kooperationen auf weitere Datensätze angewandt. Im Fokus der Untersuchungen standen: *Bacillus amyloliquefaciens* FZB42 [16, 17], *Methanosarcina mazei* Go1, *Streptomyces coelicolor* A3(2), *Oceanobacillus iheyensis* HTE831, *Pyrococcus furiosus* DSM 3638 und *Rhizobium* sp. NGR234 (zur Veröffentlichung eingereicht [94]). In den Genomsequenzen dieser Organismen sowie derer zum Vergleich herangezogener Spezies konnte eine große Anzahl vielversprechender Kandidaten identifiziert werden.

Unser Ansatz zur fRNA-Detektion kann nicht nur auf komplette Genomsequenzen, sondern ebenfalls auf kleine Datensätze aus ausgewählten Sequenzen angewandt werden. Im Fall von *S. cerevisiae* waren wir an einer Auswahl Proteinkodierender Gene interessiert, deren mRNAs unter Aminosäuremangel (+3AT) eine post-transkriptionelle Regulation der Translation vermuten ließ. Da die untranslatierten Regionen einer mRNAs dafür bekannt sind, strukturbasierte regulatorische Elemente zu beherbergen, haben wir uns auf die Untersuchung der 5'- und 3'-UTRs dieser mRNAs konzentriert. In ungefähr 30 % der 5'-UTR-Sequenzen und in 22 % der 3'-UTR-Sequenzen konnten wir eine stabile und teilweise auch konservierte Struktur vorhersagen. Eine Zusammenfassung des methodischen Vorgehens wir in Kürze in Rachfall *et al.* [81] veröffentlicht. Die in dieser Arbeit vorgestellten Ergebnisse der fRNA-Vorhersage auf *S. cerevisiae* und die Ergebnisse der experimentellen Untersuchungen, die unsere Kooperationspartner durchführen, werden zur Veröffentlichung vorbereitet.

Die Anwendung zur fRNA-Vorhersage wird in Kürze über ein Webinterface öffentlich zugänglich gemacht. Damit wird ein Werkzeug bereitgestellt, das es ermöglicht, eine schnelle und umfangreiche fRNA-Vorhersage in kompletten Genomsequenzen durchzuführen, ohne das notwendige Expertenwissen der einzelnen verwendeten Programme vorauszusetzen.

A Anhang

A.1 Vergleich der Kandidaten

In der Tabelle A.1 und der Tabelle A.2 ist die Menge der vorhergesagten Kandidaten für alle untersuchten Datensätze zusammengefasst. Zuerst ist die Anzahl und der prozentuale Anteil der Kandidaten angegeben, die mit Hilfe des komparativen Ansatzes (RNAz-Score von mindestens 0,9) vorhergesagt wurden. Dann ist die Anzahl der Kandidaten angegeben, die mit Hilfe des Kovarianzmodell-basieren Ansatzes (INFERNAL-Score von mindestens 50) vorhergesagt wurden. In der letzten Spalte sind überlappende Kandidaten aus beiden Ansätzen angegeben.

	RNAz-0,9		INF.-50	RNAz-0,9 / INF.-50
	Anz.	%	Anz.	
E. coli	2465	10.28	65	61 / 62
S. flexneri	2786	10.95	63	57 / 61
S. enterica	945	2.31	51	41 / 41
E. coli	684	2.30	65	43 / 41
Y. pestis	568	1.66	36	18 / 18
K. pneumoniae	503	1.08	52	36 / 35
B. amyloliquefaciens	588	1,68	59	54 / 49
B. licheniformis	265	0,68	59	47 / 40
B. subtilis	635	1,69	62	56 / 51
B. anthracis	198	0,38	98	27 / 25
M. mazei	912	2,69	1	1 / 1
M. acetivorans	1654	3,95	25	3 / 2
M. barkeri	1491	4,38	3	3 / 2
S. coelicolor	966	1,44	20	17 / 14
S. avermitilis	916	1,15	14	15 / 12
T. fusca	27	0,15	11	2 / 2
O. iheyensis	206	1,06	66	6 / 6
B. licheniformis	250	0,63	59	39 / 35
B. subtilis	250	0,64	62	39 / 35
P. furiosus	99	0,67	9	6 / 6
P. abyssi	95	0,64	8	6 / 6
P. horikoshii	102	0,70	6	5 / 5

Tab. A.1: Übersicht der Kandidaten des komparativen und des Kovarianzmodell-basierten Ansatzes.

Genom	Replikon	RNAz-0,9		INF.-50	RNAz-0,9 / INF.-50
		Anz.	%	Anz.	
R. NGR234	chrNGR234	1018	2,98	12	11 / 10
	pNGR234a	75	1,87	0	- / -
	pNGR234b	284	1,42	5	3 / 3
A. tumefaciens	cchrC58	195	0,76	13	7 / 7
	lchrC58	94	0,54	6	3 / 3
	pTi	6	0,22	0	- / -
	pAt	13	0,25	0	- / -
R. etli	chrRetli42	297	0,71	14	8 / 8
	p42a	62	5,09	0	- / -
	p42b	6	0,46	1	0 / 0
	p42c	5	0,19	0	- / -
	p42d	77	2,98	1	1 / 1
	p42e	21	0,36	1	0 / 0
	p42f	21	0,38	0	- / -
S. medicae	chrSMED	1231	4,26	14	13 / 14
	pSMED01	373	2,72	3	3 / 3
	pSMED02	383	4,88	0	- / -
	pSMED03	71	5,20	1	1 / 1
S. meliloti	chrSM1021	1586	5,72	14	13 / 13
	pSymA	303	3,11	1	1 / 1
	pSymB	421	2,83	3	2 / 2

Tab. A.2: Übersicht der Kandidaten in *R.* NGR234 und Vergleichsorganismen. Alle angegebenen Sequenzen wurden zusammen untersucht.

A.2 Beziehungen zwischen Kandidaten des komparativen Ansatzes

In den folgenden Tabellen (Tabelle A.3 bis Tabelle A.7) stellen wir die Beziehungen zwischen den Kandidaten des komparativen Ansatzes (RNAz-Score von mindestens 0,9) in den verschiedenen Genomen dar. Dabei geben wir zuerst die Anzahl der sogenannten Eigentreffer an, d. h. derjenigen Kandidaten, die mit Hilfe des komparativen Ansatzes ausschließlich im Zusammenhang mit Sequenzen aus dem gleichen Genom als fRNA klassifiziert wurden. Danach vergleichen wir die Genome paarweise. Für die Kandidaten in den Genomen in jeder Zeile wurde bestimmt, wieviele davon im Zusammenhang mit Sequenzen der Genome, die in den Spalten angegeben sind, als potentielle fRNAs eingestuft werden. Wird ein Genom mit sich selbst verglichen, so ist die Gesamtanzahl der Kandidaten für dieses Genom angegeben. Diese Zahl ist fett hervorgehoben. In der letzten Spalte gegeben wir an, wieviele der Kandidaten des Genoms in der jeweiligen Zeile in Beziehung zu mindestens einem Kandidaten aus jedem der anderen Genome stehen.

Die Tabelle für den ESS- und den EYK-Datensatz ist in Abschnitt 3.2.1 zu finden.

	Eigentreffer	*B. amyloliquefaciens*	*B. licheniformis*	*B. subtilis*	*B. anthracis*	Alle
B. amyloliquefaciens	23	**588**	157	550	14	12
B. licheniformis	35	158	**265**	205	18	14
B. subtilis	33	556	191	**635**	16	13
B. anthracis	174	18	18	17	**198**	12

Tab. A.3: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes in *B. amyloliquefaciens*, *B. licheniformis*, *B. subtilis* und *B. anthracis*.

	Eigentreffer	*M. mazei*	*M. acetivorans*	*M. barkeri*	Alle
M. mazei	128	**912**	674	355	245
M. acetivorans	611	717	**1654**	592	266
M. barkeri	717	396	637	**1491**	259

Tab. A.4: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes in *M. mazei*, *M. acetivorans* und *M. barkeri*.

	Eigentreffer	*S. coelicolor*	*S. avermitilis*	*T. fusca*	Alle
S. coelicolor	152	**966**	814	2	2
S. avermitilis	96	819	**916**	4	3
T. fusca	24	2	3	**27**	2

Tab. A.5: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes in *S. coelicolor*, *S. avermitilis* und *T. fusca*.

	Eigentreffer	*O. iheyensis*	*B. licheniformis*	*B. subtilis*	Alle
O. iheyensis	196	**206**	10	9	9
B. licheniformis	37	10	**250**	212	9
B. subtilis	48	9	202	**250**	9

Tab. A.6: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes in *O. iheyensis*, *B. licheniformis* und *B. subtilis*.

	Eigentreffer	*P. furiosus*	*P. abyssi*	*P. horikoshii*	Alle
P. furiosus	30	**99**	37	52	20
P. abyssi	10	36	**95**	70	21
P. horikoshii	3	48	71	**102**	20

Tab. A.7: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes in *P. furiosus*, *P. abyssi* und *P. horikoshii*.

		R. NGR234			*A. tumefaciens*				*R. etli*							*S. medicae*				*S. meliloti*		
	Eigentreffer	chrNGR234	pNGR234a	pNGR234b	cchrC58	lchrC58	pTi	pAt	chrRetli42	p42a	p42b	p42c	p42d	p42e	p42f	chrSMED	pSMED01	pSMED02	pSMED03	chrSM1021	pSymA	pSymB
chrNGR234	73	**1018**	49	278	58	21	0	0	122	9	2	3	1	10	10	586	91	17	11	768	28	139
pNGR234a	13	24	**75**	29	0	0	1	0	2	9	2	0	7	1	2	7	3	11	12	12	8	1
pNGR234b	27	137	36	**284**	12	10	0	1	20	9	3	1	7	5	4	50	52	15	16	63	36	58
cchrC58	6	63	0	15	**195**	114	0	2	91	0	0	0	0	7	0	41	3	0	1	68	1	20
lchrC58	8	14	0	9	63	**94**	0	3	22	3	1	1	1	3	0	11	2	4	3	18	1	12
pTi	0	0	1	0	0	0	**6**	1	0	2	0	0	1	0	0	0	0	1	0	0	0	0
pAt	4	0	0	1	2	2	1	**13**	0	0	0	0	2	0	1	0	1	3	2	1	2	1
chrRetli42	65	113	4	28	82	23	0	0	**297**	12	6	10	16	28	28	64	9	1	1	95	2	12
p42a	9	5	8	7	0	3	2	0	8	**62**	1	1	20	1	1	8	2	14	6	6	3	1
p42b	0	2	2	3	0	1	0	0	2	1	**6**	0	3	1	2	1	2	1	2	1	1	2
p42c	0	1	0	1	0	1	0	0	3	1	0	**5**	0	0	0	0	0	0	0	0	0	0
p42d	24	1	6	7	0	2	1	2	7	20	3	0	**77**	2	8	5	1	12	9	3	8	3
p42e	4	4	1	4	1	1	0	0	8	1	1	0	2	**21**	2	4	2	1	1	9	1	2
p42f	0	4	3	3	0	0	0	1	7	1	2	0	8	3	**21**	1	2	4	1	1	4	1
chrSMED	19	478	5	57	42	15	0	0	60	8	1	0	3	7	1	**1231**	63	26	11	1123	30	54
pSMED01	3	30	2	50	3	2	0	1	5	2	2	0	1	2	2	38	**373**	27	6	42	18	306
pSMED02	76	8	13	15	0	4	1	4	1	21	1	0	15	2	2	26	36	**383**	43	32	207	23
pSMED03	6	7	11	12	1	2	0	2	1	5	2	0	7	1	1	9	8	33	**71**	10	20	7
chrSM1021	63	744	22	158	72	26	0	1	101	7	1	0	3	24	1	1200	108	51	19	**1586**	67	262
pSymA	15	11	9	34	2	1	0	3	1	6	1	0	7	1	2	25	20	208	29	42	**303**	24
pSymB	5	61	1	60	11	10	0	1	9	1	2	0	3	2	1	42	309	22	8	88	21	**421**

Tab. A.8: Zusammenhang zwischen den Kandidaten des komparativen Ansatzes aller Replikons in *R.* NGR234, *A. tumefaciens*, *R. etli*, *S. medicae* und *S. meliloti*.

A.3 Zusätzliche Ergebnisse aus der Untersuchung der 5'-/3'-UTRs in S. cerevisiae

A.3.1 Ergebnisse des Vergleichs mit `Rfam`

In der Tabelle A.9 sind die Ergebnisse des INFERNAL-Vergleichs der 5'- und 3'-UTR-Sequenzen mit `Rfam` zusammengefasst. Es sind ausschließlich Treffer mit einem Score von mindestens 20 angegeben. Neben den Positionsinformationen, d. h. der Startposition S und der Endposition E, in der entsprechenden Sequenz werden der INFERNAL-Score, der Strang des Kandidaten und die entsprechende fRNA-Familie, zu der eine Ähnlichkeit gefunden wurde, angegeben. In der letzten Spalte steht der Typ der fRNA-Familie.

	Gen	sys. Name	S	E	INF.-Score	Strang	fRNA-Familie	Typ
5'-UTR	*PCL5*	YHR071W	276	362	23,34	+	*snoR20*	Gen
			299	361	20,55	+	*SNORD63*	Gen
			298	355	20,43	+	*snoMe28S-Cm3227*	Gen
3'-UTR	*STE12*	YHR084W	152	235	22,50	+	*SNORD79*	Gen
			136	201	20,17	-	*RyhB*	Gen
	CIN5	YOR028C	135	221	20,50	-	*snoR9_plant*	Gen
	NDD1	YOR372C	35	236	20,52	-	*SraC_RyeA*	Gen
	YHI9	YHR029C	18	69	22,39	+	*mir-1*	Gen
			18	69	20,98	+	*ctRNA_pND324*	Gen
	HAP4	**YKL109W**	143	284	25,42	+	*SNORA4*	Gen
			177	240	24,32	+	*SNORD78*	Gen
			149	249	23,88	+	*snopsi18S-1854*	Gen
			149	210	21,84	+	*SNORD77*	Gen
			135	226	21,17	-	*SNORA40*	Gen
			144	207	20,97	+	*ctRNA_pT181*	Gen
			144	**277**	**20,66**	**+**	***L20_leader***	**cis-El.**
			217	287	20,00	+	*ctRNA_pND324*	Gen
	COF1	YLL050C	43	118	21,69	-	*snoZ195*	Gen
	STB1	YNL309W	40	111	22,03	+	*snoR20*	Gen
	SKN7	**YHR206W**	221	266	23,47	-	*mir-1*	Gen
			216	**257**	**21,53**	**-**	***RbcL_stabil***	**cis-El.**

Tab. A.9: Ähnlichkeiten ausgewählter UTRs in *S. cerevisiae* zu bekannten fRNA-Familien in `Rfam`.

A.3.2 z-Score für Teilstrukturen der 5'-/3'-UTRs

In der Tabelle A.10 und der Tabelle A.11 ist der z-Score (Gleichung 1.1) für Teilsequenzen der 5'- und 3'-UTRs mit einer lokal optimalen MFE-Struktur angegeben. Es sind nur Teilsequenzen mit einem z-Score von maximal -3 aufgelistet.

Neben dem Gennamen, der Startposition und Länge der betrachteten Teilsequenz, ist außerdem der MFE-Wert aus der Struktur-Vorhersage angegeben. Jede Sequenz wurde jeweils 100-mal mit Dishuffle (Abschnitt 5.2.3) permutiert. Für die permutierten Sequenzen ist der Mittelwert und die Standardabweichung der MFEs angegeben. Die letzte Spalte gibt den aus den drei vorhergehenden Werten resultierenden z-Score an.

Gen	sys. Name	S	E	MFE	MFE-Mw.	MFE-Std.	z-Score
ABF1	**YKL112W**	**14**	**93**	**-23,30**	**-12,63**	**2,41**	**-4,43**
		17	90	-21,10	-12,35	2,47	-3,54
ADE16	**YLR028C**	**110**	**210**	**-18,40**	**-6,77**	**2,18**	**-5,34**
		115	213	-12,80	-5,58	2,05	-3,52
ADR1	**YDR216W**	207	68	-17,30	-10,46	2,11	-3,24
		285	259	-37,20	-27,60	2,80	-3,43
		583	**681**	**-26,90**	**-13,11**	**2,70**	**-5,11**
		591	**669**	**-23,70**	**-11,16**	**2,64**	**-4,75**
		722	813	-23,30	-15,02	2,70	-3,06
ARO8	YGL202W	32	86	-9,50	-3,59	1,53	-3,87
GCV1	YDR019C	689	782	-19,20	-10,48	2,34	-3,73
HAP4	YKL109W	177	255	-13,65	-7,47	1,96	-3,15
HOR2	YER062C	50	127	-18,59	-10,35	2,38	-3,47
LYS12	YIL094C	44	96	-12,39	-6,20	1,99	-3,11
MTG1	YMR097C	131	232	-27,30	-16,08	3,17	-3,54
		134	229	-25,70	-14,23	3,01	-3,81
MXR1	**YER042W**	239	302	-13,00	-5,97	2,23	-3,16
		373	**464**	**-28,80**	**-17,87**	**2,62**	**-4,17**
		400	474	-22,50	-13,07	2,79	-3,38
PCL5	**YHR071W**	**2**	**57**	**-20,00**	**-9,62**	**2,47**	**-4,19**
		258	315	-9,30	-3,44	1,64	-3,58
SWI5	YDR146C	35	87	-12,70	-4,98	2,22	-3,48

Tab. A.10: z-Score für Teilsequenzen der 5'-UTR in *S. cerevisiae*.

Gen	sys. Name	S	E	MFE	MFE-Mw.	MFE-Std.	z-Score
BNA1	YJR025C	42	133	-28,20	-16,79	3,06	-3,73
BMH2	YDR099W	73	136	-13,21	-6,96	2,07	-3,02
CIN5	YOR028C	110	209	-24,13	-14,15	2,73	-3,66
FLO8	YER109C	271	334	-19,40	-11,36	2,54	-3,16
FPR1	YNL135C	64	146	-16,65	-9,35	2,42	-3,01
HAP4	**YKL109W**	**230**	**280**	**-13,40**	**-5,73**	**1,90**	**-4,03**
YHI9	**YHR029C**	**11**	**76**	**-23,40**	**-12,71**	**1,96**	**-5,44**
		16	69	-18,30	-10,06	2,14	-3,85
SKN7	**YHR206W**	**199**	**277**	**-14,40**	**-5,39**	**2,10**	**-4,30**
		208	277	-14,15	-6,40	1,98	-3,91

Tab. A.11: z-Score für Teilsequenzen der 3'-UTR in *S. cerevisiae*.

A.3.3 Komparativer Ansatz mit RNAz-Schwellenwert von 0,5

Ergänzend zu den Ergebnissen in Abschnitt 5.3.4, sind hier alle positiven Vorhersagen des komparativen Ansatzes, d. h. alle Vorhersagen mit einem RNAz-Score von mindestens 0,5 aufgelistet. Im Gegensatz dazu wurden in Abschnitt 5.3.4 nur die signifikantesten Ergebnisse (RNAz-Score $\geq$ 0,9) angegeben. Somit sind alle Kandidaten aus Abschnitt 5.3.4 in den hier vorgestellten Ergebnissen enthalten.

In der Tabelle A.12 und der Tabelle A.13 wird unter anderem ein Score für jeden Kandidaten angegeben. Ein Kandidat besteht meistens aus mehreren, sich teilweise überlappenden Vorhersagen, die mit unterschiedlichen Scores bewertet wurden. Aber nur Vorhersagen mit einem Score, der über einer vorgegebenen Toleranzgrenze liegt, werden zu einem Kandidaten zusammengefasst. In Abschnitt 5.3.4 lag diese bei 0,9; in den hier vorgestellten Ergebnissen liegt sie bei 0,5. Damit ist eine untere Grenze für den Score bekannt. Da in den Kandidaten teilweise deutlich bessere Vorhersagen enthalten sind, geben wir den Score der besten in diesem Kandidaten enthaltenen Vorhersage an. Es ist jedoch zu berücksichtigen, dass eventuell nur ein Teil des Kandidaten mit diesem Score bewertet wurde.

Gen	sys. Name	UTR-Länge	S	E	max. Score	Trefferspezies
ABF1	YKL112W	176	1	105	0.99	MIT_Smik, WashU_Sbay, MIT_Spar
ADE12	YNL220W	111	1	75	0.52	MIT_Spar
ADE16	YLR028C	213	1	210	1.00	MIT_Spar, WashU_Skud
ADR1	YDR216W	1158	227	878	1.00	MIT_Spar
			884	1096	0.63	MIT_Spar
BMH2	YDR099W	124	56	120	0.98	MIT_Spar
CIN5	YOR028C	323	19	241	1.00	MIT_Spar
FKH2	YNL068C	288	21	260	1.00	MIT_Spar
FLO8	YER109C	175	58	121	0.66	MIT_Spar, MIT_Smik, WashU_Smik
GCV1	YDR019C	1200	401	460	0.76	MIT_Smik, WashU_Skud
			491	570	1.00	MIT_Smik WashU_Skud
			631	796	0.96	MIT_Smik WashU_Skud
HAP1	YLR256W	300	6	289	0.98	MIT_Spar
HAP4	YKL109W	337	185	247	0.81	MIT_Spar
HSF1	YGL073W	101	18	95	0.77	MIT_Spar
ILV5	YLR355C	91	16	75	0.72	MIT_Spar
MBP1	YDL056W	176	61	151	0.91	MIT_Spar
MET6	YER091C	70	1	70	0.99	MIT_Spar MIT_Smik
MTG1	YMR097C	467	11	345	1.00	MIT_Spar, WashU_Skud, WashU_Sbay, MIT_Sbay, WashU_Smik, MIT_Smik
MXR1	YER042W	1155	226	540	0.98	MIT_Spar, MIT_Smik
			576	735	0.85	MIT_Spar
			871	980	0.93	WashU_Smik, MIT_Smik, MIT_Spar
NDD1	YOR372C	70	10	69	0.99	MIT_Spar
PCL5	YHR071W	362	4	62	1.00	MIT_Spar, WashU_Smik, MIT_Smik
			103	357	1.00	MIT_Smik, MIT_Spar, WashU_Skud
PGK1	YCR012W	82	1	70	0.83	WashU_Smik, MIT_Spar, MIT_Smik
RPS26B	YER131W	439	192	266	0.89	MIT_Spar
SAM4	YPL273W	138	49	103	0.65	MIT_Spar
SKN7	YHR206W	292	6	113	1.00	MIT_Spar, MIT_Smik
			125	289	0.99	MIT_Spar
SOK2	YMR016C	281	76	160	1.00	MIT_Spar
			176	222	0.57	MIT_Spar
STE12	YHR084W	323	44	163	1.00	MIT_Spar, WashU_Smik, MIT_Smik
			237	320	0.90	MIT_Spar
SWI5	YDR146C	177	30	162	1.00	MIT_Spar

Tab. A.12: fRNA-Kandidaten mit RNAz-Score $\geq$ 0,5 in 5'-UTR in *S. cerevisiae*.

Gen	sys. Name	UTR-Länge	S	E	max. Score	Trefferspezies
ADE1	YAR015W	97	5	89	1.00	MIT_Spar
ADE8	YDR408C	210	47	168	0.97	MIT_Spar MIT_Smik WashU_Smik
ADR1	YDR216W	119	38	107	1.00	MIT_Spar
BMH2	YDR099W	175	79	172	0.88	MIT_Smik MIT_Spar
BNA1	YJR025C	171	12	155	1.00	MIT_Spar MIT_Smik
COF1	YLL050C	130	17	101	0.86	MIT_Smik WashU_Smik
CPR1	YDR155C	124	28	107	1.00	MIT_Spar
FLO8	YER109C	334	200	269	0.98	MIT_Spar
FPR1	YNL135C	209	77	136	1.00	MIT_Spar WashU_Smik
HAP4	YKL109W	463	70	187	1.00	MIT_Spar MIT_Sbay MIT_Smik WashU_Smik
			196	330	1.00	WashU_Smik MIT_Smik MIT_Spar
HAP5	YOR358W	224	1	210	1.00	MIT_Spar MIT_Smik
ILV5	YLR355C	138	51	95	0.67	MIT_Spar MIT_Smik WashU_Smik
NDD1	YOR372C	266	184	265	0.98	MIT_Smik MIT_Spar
PGK1	YCR012W	76	1	61	0.79	MIT_Sbay WashU_Sbay MIT_Spar
RHR2	YIL053W	156	21	94	0.83	MIT_Spar
STE12	YHR084W	307	78	142	0.55	MIT_Spar
			168	306	1.00	MIT_Spar

Tab. A.13: fRNA-Kandidaten mit RNAz-Score $\geq$ 0,5 in 3'-UTR in *S. cerevisiae*.

Literaturverzeichnis

[1] Jmol: an open-source Java viewer for chemical structures in 3D. URL http://www.jmol.org/.

[2] SGD project: Saccharomyces Genome Database, 2007. URL http://www.yeastgenome.org/.

[3] S. Altschul and B. Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, 2(6):526–538, 1985. URL http://mbe.oxfordjournals.org/cgi/content/abstract/2/6/526.

[4] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, 25(17):3389–3402, 1997. (doi:10.1093/nar/25.17.3389).

[5] V. Ambros and X. Chen. The regulation of genes and genomes by small RNAs. *Development*, 134:1635–1641, 2007. (doi:10.1242/dev.002006).

[6] I. Axmann, P. Kensche, J. Vogel, S. Kohl, H. Herzel, and W. Hess. Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biology*, 6(9):R73, 2005. ISSN 1465-6906. (doi:10.1186/gb-2005-6-9-r73).

[7] T. Babak, B. Blencowe, and T. Hughes. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, 8(1):33, 2007. ISSN 1471-2105. (doi:10.1186/1471-2105-8-33).

[8] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucl. Acids Res.*, 34:D16–20, 2006. (doi:10.1093/nar/gkj157).

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28(1):235–242, 2000. (doi:10.1093/nar/28.1.235).

[10] Y. Bessho, R. Shibata, S.-i. Sekine, K. Murayama, K. Higashijima, C. Hori-Takemoto, M. Shirouzu, S. Kuramitsu, and S. Yokoyama. Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA. *Proceedings of the National Academy of Sciences*, 104(20):8293–8298, 2007. (doi:10.1073/pnas.0700402104).

[11] F. R. Blattner, I. Plunkett, Guy, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, 1997. (doi:10.1126/science.277.5331.1453).

[12] W. J. Broughton, S. Jabbouri, and X. Perret. Keys to Symbiotic Harmony. *J. Bacteriol.*, 182(20):5641–5652, 2000. (doi:10.1128/JB.182.20.5641-5652.2000).

[13] D. Capela, F. Barloy-Hubler, J. Gouzy, G. Bothe, F. Ampe, J. Batut, P. Boistard, A. Becker, M. Boutry, E. Cadieu, S. Dréano, S. Gloux, T. Godrie, A. Goffeau, D. Kahn, E. Kiss, V. Lelaure, D. Masuy, T. Pohl, D. Portetelle, A. Pühler, B. Purnelle, U. Ramsperger, C. Renard, P. Thébault, M. Vandenbol, S. Weidner, and F. Galibert. Analysis of the chromosome sequence of the legume symbiont Sinorhizobium meliloti strain 1021. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9877–9882, 2001. (doi:10.1073/pnas.161294398).

[14] R. J. Carter, I. Dubchak, and S. R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.*, 29(19):3928–3938, 2001. (doi:10.1093/nar/29.19.3928).

[15] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[16] X. H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W. R. Hess, O. Reva, H. Junge, B. Voigt,

P. R. Jungblut, J. Vater, R. Süssmuth, H. Liesegang, A. Strittmatter, G. Gottschalk, and R. Borriss. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium Bacillus amyloliquefaciens FZB42. *Nature Biotechnology*, 25:1007–1014, September 2007. (doi:10.1038/nbt1325).

[17] X. H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W. R. Hess, O. Reva, H. Junge, B. Voigt, P. R. Jungblut, J. Vater, R. Süssmuth, H. Liesegang, A. Strittmatter, G. Gottschalk, and R. Borriss. Genomanalyse eines phytostimulatorischen Bacillus-Stammes. *GenomXPress 3.07*, 3.07:11–13, September 2007.

[18] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting. *Science*, 301(5629):71–76, 2003. (doi:10.1126/science.1084337).

[19] P. CLOTE, F. FERRE, E. KRANAKIS, and D. KRIZANC. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, 2005. (doi:10.1261/rna.7220505).

[20] J. Cocke and J. T. Schwartz. *Programming languages and their compilers: Preliminary notes*. Courant Institute of Mathematical Sciences of New York University, New York, 1970.

[21] G. N. Cohen, V. Barbe, D. Flament, M. Galperin, R. Heilig, O. Lecompte, O. Poch, D. Prieur, J. Quérellou, R. Ripp, J.-C. Thierry, J. V. der Oost, J. Weissenbach, Y. Zivanovic, and P. Forterre. An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. *Molecular Microbiology*, 47(6):1495–1512, 2003. URL `http://dx.doi.org/10.1046/j.1365-2958.2003.03381.x`.

[22] A. Coventry, D. J. Kleitman, and B. Berger. msari : Multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33): 12102–12107, 2004. (doi:10.1073/pnas.0404193101).

[23] P. P. Dennis and A. Omer. Small non-coding RNAs in Archaea. *Current Opinion in Microbiology*, 8:685–694, 12 2005.

[24] U. Deppenmeier, A. Johann, T. Hartsch, R. Merkl, R. Schmitz, R. Martinez-Arias, A. Henne, A. Wiezer, S. Bäumer, C. Jacobi, T. Brüggemann, H. Lienard, A. Christmann, M. Bömeke, S. Steckel, A. Bhattacharyya, A. Lykidis, R. Overbeek, H. Klenk, R. Gunsalus, H. Fritz, and G. Gottschalk. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol.*, 4(4):453–461, Juli 2002.

[25] D. di Bernardo, T. Down, and T. Hubbard. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19(13):1606–1611, 2003. (doi:10.1093/bioinformatics/btg229).

[26] M. E. Dinger, T. R. Mercer, and J. S. Mattick. RNAs as extracellular signaling molecules. *J Mol Endocrinol*, 40(4):151–159, 2008. (doi:10.1677/JME-07-0160).

[27] E. A. Doherty and J. A. Doudna. RIBOZYME STRUCTURES AND MECHANISMS. *Annual Review of Biophysics and Biomolecular Structure*, 30 (1):457–475, 2001. (doi:10.1146/annurev.biophys.30.1.457).

[28] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[29] L. Duret, C. Chureau, S. Samain, J. Weissenbach, and P. Avner. The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. *Science*, 312(5780):1653–1655, 2006. (doi:10.1126/science.1126316).

[30] S. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998. (doi:10.1093/bioinformatics/14.9.755).

[31] S. Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3(1):18, 2002. ISSN 1471-2105. (doi:10.1186/1471-2105-3-18).

[32] S. Eddy. *INFERNAL User's Guide (Version 0.81)*. HHMI Janelia Farm, May 2007. URL http://infernal.janelia.org/.

[33] S. Eddy. SQUID. URL http://selab.janelia.org/software.html.

[34] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22(11):2079–2088, 1994. (doi:10.1093/nar/22.11.2079).

[35] EMBL-EBI. URL http://www3.ebi.ac.uk/Services/WebFeat/.

[36] C. Freiberg, R. Fellay, A. Bairoch, W. J. Broughton, A. Rosenthal, and X. Perret. Molecular basis of symbiosis between Rhizobium and legumes. *Nature*, 387(6631):394–401, May 1997. (doi:10.1038/387394a0).

[37] E. K. Freyhult, J. P. Bollback, and P. P. Gardner. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research*, 17(1):117–125, 2007. (doi:10.1101/gr.5890907).

[38] J. E. Galagan, C. Nusbaum, A. Roy, M. G. Endrizzi, P. Macdonald, W. FitzHugh, S. Calvo, R. Engels, S. Smirnov, D. Atnoor, A. Brown, N. Allen, J. Naylor, N. Stange-Thomann, K. DeArellano, R. Johnson, L. Linton, P. McEwan, K. McKernan, J. Talamas, A. Tirrell, W. Ye, A. Zimmer, R. D. Barber, I. Cann, D. E. Graham, D. A. Grahame, A. M. Guss, R. Hedderich, C. Ingram-Smith, H. C. Kuettner, J. A. Krzycki, J. A. Leigh, W. Li, J. Liu, B. Mukhopadhyay, J. N. Reeve, K. Smith, T. A. Springer, L. A. Umayam, O. White, R. H. White, E. C. de Macario, J. G. Ferry, K. F. Jarrell, H. Jing, A. J. Macario, I. Paulsen, M. Pritchett, K. R. Sowers, R. V. Swanson, S. H. Zinder, E. Lander, W. W. Metcalf, and B. Birren. The Genome of M. acetivorans Reveals Extensive Metabolic and Physiological Diversity. *Genome Res.*, 12(4):532–542, 2002. (doi:10.1101/gr.223902).

[39] P. P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8): 2433–2439, 2005. (doi:10.1093/nar/gki541).

[40] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman.

Rfam: updates to the RNA families database. *Nucl. Acids Res.*, page gkn766, 2008. (doi:10.1093/nar/gkn766).

[41] W. V. Gilbert, K. Zhou, T. K. Butler, and J. A. Doudna. Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. *Science*, 317(5842):1224–1227, 2007. (doi:10.1126/science.1144467).

[42] V. González, R. I. Santamaría, P. Bustos, I. Hernández-González, A. Medrano-Soto, G. Moreno-Hagelsieb, S. C. Janga, M. A. Ramírez, V. Jiménez-Jacinto, J. Collado-Vides, and G. Dávila. The partitioned Rhizobium etli genome: Genetic and metabolic redundancy in seven interacting replicons. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10): 3834–3839, 2006. (doi:10.1073/pnas.0508502103).

[43] S. Graf, D. Strothmann, S. Kurtz, and G. Steger. HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns. *Nucl. Acids Res.*, 29(1):196–198, 2001. (doi:10.1093/nar/29.1.196).

[44] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucl. Acids Res.*, 31(1):439–441, 2003. (doi:10.1093/nar/gkg006).

[45] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33:121–124, 2005. (doi:10.1093/nar/gki081).

[46] D. H. Haft, J. D. Selengut, L. M. Brinkac, N. Zafar, and O. White. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, 21(3):293–306, 2005. (doi:10.1093/bioinformatics/bti015).

[47] I. Hofacker, M. Fekete, and P. Stadler. Secondary Structure Prediction for Aligned RNA Sequences. *Journal of Molecular Biology*, 319:1059–1066, 2002. (doi:10.1016/S0022-2836(02)00308-X).

[48] I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31 (13):3429–3431, 2003. (doi:10.1093/nar/gkg599).

[49] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chem.*, 125:167–188, 1994.

[50] I. L. Hofacker, B. Priwitzer, and P. F. Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2): 186–190, 2004. (doi:10.1093/bioinformatics/btg388).

[51] N. Jarrous and R. Reiner. Human RNase P: a tRNA-processing enzyme and transcription factor. *Nucl. Acids Res.*, 35(11):3519–3524, 2007. (doi:10.1093/nar/gkm071).

[52] L. Jovine, S. Djordjevic, and D. Rhodes. The crystal structure of yeast phenylalanine tRNA at 2.0 å resolution: cleavage by Mg2+ in 15-year old crystals. *Journal of Molecular Biology*, 301(2):401–414, 2000. URL `http://www.sciencedirect.com/science/article/B6WK7-45F5166-8D/1/9d1b59ead339fa652f4add6d254`.

[53] T. Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. Scientific report afcrl-65-758, Air Force Cambridge Research Lab, Bedford (Massachussetts), 1965.

[54] Y. Kawarabayasi, M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S.-i. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K.-i. Aoki, T. Yoshizawa, Y. Nakamura, F. T. Robb, K. Horikoshi, Y. Masuchi, H. Shizuya, and H. Kikuchi. Complete Sequence and Gene Organization of the Genome of a Hyper-thermophilic Archaebacterium, Pyrococcus horikoshii OT3. *DNA Res*, 5(2):55–76, 1998. (doi:10.1093/dnares/5.2.55).

[55] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003. ISSN 0028-0836. (doi:10.1038/nature01644).

[56] R. Klein and S. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(1):44, 2003. ISSN 1471-2105. (doi:10.1186/1471-2105-4-44).

[57] R. J. Klein, Z. Misulovin, and S. R. Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7542–7547, 2002. (doi:10.1073/pnas.112063799).

[58] F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S.-K. Choi, J.-J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M.-F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S.-M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauel, C. Medigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S.-H. Park, V. Parro, T. M. Pohl, D. Portetelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Roche, M. Rose, Y. Sadaie, T. Sato, E. Scanlan, S. Schleich, R. Schroeter, F. Scoffone, J. Sekiguchi, A. Sekowska, S. J. Seror, P. Serror, B.-S. Shin, B. Soldo, A. Sorokin, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Tognoni, V. Tosato, S. Uchiyama, M. Vandenbol, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, E. Wedler, H. Wedler, T. Weitzenegger, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H.-F. Yoshikawa, E. Zumstein, H. Yoshikawa, and A. Danchin. The complete

genome sequence of the Gram-positive bacterium Bacillus subtilis. *Nature*, 390(6657):249–256, Nov. 1997. ISSN 0028-0836. (doi:10.1038/36786).

[59] E. Lai, P. Tomancak, R. Williams, and G. Rubin. Computational identification of Drosophila microRNA genes. *Genome Biology*, 4(7):R42, 2003. ISSN 1465-6906. (doi:10.1186/gb-2003-4-7-r42).

[60] D. Laslett, B. Canback, and S. Andersson. BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucl. Acids Res.*, 30(15):3449–3453, 2002. (doi:10.1093/nar/gkf459).

[61] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. The microRNAs of Caenorhabditis elegans. *Genes Dev.*, 17(8):991–1008, 2003. (doi:10.1101/gad.1074403).

[62] S.-L. Lin, J. D. Miller, and S.-Y. Ying. Intronic MicroRNA (miRNA). *Journal of Biomedicine and Biotechnology*, 2006(26818), 2006. (doi:10.1155/JBB/2006/26818).

[63] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao, and R. Chen. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, 33:Database issue:D112–D115, Januar 2005. (doi:10.1093/nar/gki041).

[64] T. Lowe and S. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25(5):955–964, 1997. (doi:10.1093/nar/25.5.955).

[65] A. Lykidis, K. Mavromatis, N. Ivanova, I. Anderson, M. Land, G. DiBartolo, M. Martinez, A. Lapidus, S. Lucas, A. Copeland, P. Richardson, D. B. Wilson, and N. Kyrpides. Genome Sequence and Analysis of the Soil Cellulolytic Actinomycete Thermobifida fusca YX. *J. Bacteriol.*, 189(6):2477–2486, 2007. (doi:10.1128/JB.01899-06).

[66] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucl. Acids Res.*, 29(22):4724–4735, 2001. (doi:10.1093/nar/29.22.4724).

[67] D. L. Maeder, R. B. Weiss, D. M. Dunn, J. L. Cherry, J. M. Gonzalez, J. DiRuggiero, and F. T. Robb. Divergence of the Hyperthermophilic Archaea Pyrococcus furiosus and P. horikoshii Inferred From Complete Genomic Sequences. *Genetics*, 152(4):1299–1305, 1999. URL http://www.genetics.org/cgi/content/abstract/152/4/1299.

[68] D. L. Maeder, I. Anderson, T. S. Brettin, D. C. Bruce, P. Gilna, C. S. Han, A. Lapidus, W. W. Metcalf, E. Saunders, R. Tapia, and K. R. Sowers. The Methanosarcina barkeri Genome: Comparative Analysis with Methanosarcina acetivorans and Methanosarcina mazei Reveals Extensive Rearrangement within Methanosarcinal Genomes. *J. Bacteriol.*, 188(22):7922–7931, 2006. (doi:10.1128/JB.00810-06).

[69] D. Mathews, J. Sabina, M. Zuker, and D. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.

[70] J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*, 25(10):930–939, 2003. (doi:10.1002/bies.10332).

[71] J. S. Mattick. RNA regulation: a new genetics? *Nat Rev Genet*, 5(4):316–323, Apr. 2004. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg1321.

[72] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Human Molecular Genetics*, 15:17–29, 2006. (doi:10.1093/hmg/ddl046).

[73] R. K. Montange and R. T. Batey. Riboswitches: Emerging Themes in RNA Structure and Function. *Annual Review of Biophysics*, 37(1):117–133, 2008. (doi:10.1146/annurev.biophys.37.032807.130000).

[74] E. P. Nawrocki and S. R. Eddy. Query-Dependent Banding (QDB) for Faster RNA Similarity Searches. *PLoS Comput Biol*, 3(3):e56, 2007. (doi:10.1371/journal.pcbi.0030056).

[75] E. Nudler and A. S. Mironov. The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences*, 29(1):11–17, 1 2004. (doi:10.1016/j.tibs.2003.11.004).

[76] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978. (doi:10.1137/0135006).

[77] S. Ömura, H. Ikeda, J. Ishikawa, A. Hanamoto, C. Takahashi, M. Shinose, Y. Takahashi, H. Horikawa, H. Nakazawa, T. Osonoe, H. Kikuchi, T. Shiba, Y. Sakaki, and M. Hattori. Genome sequence of an industrial microorganism Streptomyces avermitilis : Deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21):12215–12220, 2001. (doi:10.1073/pnas.211433198).

[78] K. C. Pang, S. Stephen, M. E. Dinger, P. G. Engstrom, B. Lenhard, and J. S. Mattick. RNAdb 2.0-an expanded database of mammalian non-coding RNAs. *Nucl. Acids Res.*, 35:D178–182, 2007. (doi:10.1093/nar/gkl926).

[79] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, 1988. URL `http://www.pnas.org/content/85/8/2444.abstract`.

[80] K. V. Prasanth and D. L. Spector. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.*, 21(1):11–42, 2007. (doi:10.1101/gad.1484207).

[81] N. Rachfall, I. Heinemeyer, and O. Valerius. 5'-TRUE: Die wahre Translation? *BIOspektrum*, 2, 2009.

[82] D. A. Rasko, M. R. Altherr, C. S. Han, and J. Ravel. Genomics of the Bacillus cereus group of organisms. *FEMS Microbiology Reviews*, 29(2):303–329, Apr. 2005. URL `http://www.sciencedirect.com/science/article/B6T37-4FBW3PG-1/2/6be1202c10fcb686ed4493f14fcd8efd`.

[83] M. Redenbach, H. Kieser, D. Denapaite, A. Eichner, J. Cullum, H. Kinashi, and D. Hopwood. A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb Streptomyces coelicolor A3(2) chromosome. *Molecular Microbiology*, 21(1):77–96, Juli 1996.

[84] M. Ringnér and M. Krogh. Folding Free Energies of 5'-UTRs Impact Post-Transcriptional Regulation on a Genomic Scale in Yeast. *PLoS Comput Biol*, 1(7):e72, Dez 2005. (doi:10.1371/journal.pcbi.0010072).

[85] E. Rivas and S. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001. ISSN 1471-2105. (doi:10.1186/1471-2105-2-8).

[86] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7): 583–605, 2000. (doi:10.1093/bioinformatics/16.7.583).

[87] K. E. Rudd. EcoGene: a genome sequence database for Escherichia coli K-12. *Nucl. Acids Res.*, 28(1):60–64, 2000. (doi:10.1093/nar/28.1.60).

[88] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000. (doi:10.1093/bioinformatics/16.10.944).

[89] M. L. Salvador, L. Suay, I. L. Anthonisen, and U. Klein. Changes in the 5'-untranslated region of the rbcL gene accelerate transcript degradation more than 50-fold in the chloroplast of Chlamydomonas reinhardtii. *Current Genetics*, 45(3):176–182, März 2004. (doi:10.1007/s00294-003-0470-8).

[90] D. Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985. ISSN 00361399. URL `http://www.jstor.org/stable/2101630`.

[91] D. G. Sashital and S. E. Butcher. Flipping Off the Riboswitch: RNA Structures That Control Gene Expression. *ACS Chemical Biology*, 1(6):341–345, 2006. ISSN 1554-8929. URL `http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/cb6002465`.

[92] M. Schallmey, A. Singh, and O. P. Ward. Developments in the use of Bacillus species for industrial production. *Can. J. Microbiol.*, 50(1):1–17, 2004. (doi:10.1139/w03-076).

[93] P. Schattner. Searching for RNA genes using base-composition statistics. *Nucl. Acids Res.*, 30(9):2076–2082, 2002. (doi:10.1093/nar/30.9.2076).

[94] C. Schmeisser, H. Liesegang, D. Krysciak, N. Bakkou, A. Le Quéré, A. Wollherr, I. Heinemeyer, B. Morgenstern, A. Pommerening-Röser, M. Flores, R. Palacios, S. Brenner, G. Gottschalk, R. A. Schmitz, W. J. Broughton, X. Perret, A. W. Strittmatter, and W. R. Streit. Rhizobium sp. NGR234 possesses a remarkable number of secretion systems. *Applied and Environmental Microbiology*, 2009.

[95] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, U.K., 2000.

[96] W. R. Streit, R. A. Schmitz, X. Perret, C. Staehelin, W. J. Deakin, C. Raasch, H. Liesegang, and W. J. Broughton. An Evolutionary Hot Spot: the pNGR234b Replicon of Rhizobium sp. Strain NGR234. *J. Bacteriol.*, 186(2): 535–542, 2004. (doi:10.1128/JB.186.2.535-542.2004).

[97] H. Takami, Y. Takaki, and I. Uchiyama. Genome sequence of Oceanobacillus iheyensis isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. *Nucl. Acids Res.*, 30(18):3927–3935, 2002. (doi:10.1093/nar/gkf526).

[98] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22(22):4673–4680, 1994. (doi:10.1093/nar/22.22.4673).

[99] B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology*, 15(3):342–348, 6 2005. (doi:10.1016/j.sbi.2005.05.003).

[100] B. Veith, C. Herzberg, S. Steckel, J. Feesche, K. H. Maurer, P. Ehrenreich, S. Baeumer, A. Henne, H. Liesegang, R. Merkl, A. Ehrenreich, and G. Gottschalk. The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J. Mol. Microbiol. Biotechnol.*, 7: 204–211, 2004.

[101] E. Wagner, S. Altuvia, and R. P. *Antisense RNAs in bacteria and their genetic elements*, volume 46 of *Advances in genetics*. Academic Press, 2002.

[102] S. Washietl and I. L. Hofacker. Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics. *Journal of Molecular Biology*, 342(1):19–30, Sept. 2004. (doi:10.1016/j.jmb.2004.07.018).

[103] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *PNAS*, 102(7):2454–2459, 2005. (doi:10.1073/pnas.0409169102).

[104] A. Wilm, I. Mainz, and G. Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms for Molecular Biology*, 1(1):19, 2006. (doi:10.1186/1748-7188-1-19).

[105] D. W. Wood, J. C. Setubal, R. Kaul, D. E. Monks, J. P. Kitajima, V. K. Okura, Y. Zhou, L. Chen, G. E. Wood, J. Almeida, Nalvo F., L. Woo, Y. Chen, I. T. Paulsen, J. A. Eisen, P. D. Karp, S. Bovee, Donald, P. Chapman, J. Clendenning, G. Deatherage, W. Gillet, C. Grant, T. Kutyavin, R. Levy, M.-J. Li, E. McClelland, A. Palmieri, C. Raymond, G. Rouse, C. Saenphimmachak, Z. Wu, P. Romero, D. Gordon, S. Zhang, H. Yoo, Y. Tao, P. Biddle, M. Jung, W. Krespan, M. Perry, B. Gordon-Kamm, L. Liao, S. Kim, C. Hendrick, Z.-Y. Zhao, M. Dolan, F. Chumley, S. V. Tingey, J.-F. Tomb, M. P. Gordon, M. V. Olson, and E. W. Nester. The Genome of the Natural Genetic Engineer Agrobacterium tumefaciens C58. *Science*, 294(5550):2317–2323, 2001. (doi:10.1126/science.1066804).

[106] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.*, 27(24):4816–4822, 1999. (doi:10.1093/nar/27.24.4816).

[107] D. H. Younger. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208, 1967.

[108] P. D. Zamore and B. Haley. Ribo-genome: The Big World of Small RNAs. *Science*, 309:1519–1524, 2005. (doi:10.1126/science.1111444).

[109] J. M. Zengel and L. Lindahl. Diverse Mechanisms for Regulating Ribosomal Protein Synthesis in Escherichia coli. volume Volume 47, pages 331–370. Academic Press, 1994. (doi:doi:10.1016/S0079-6603(08)60256-1).

[110] A. Zhang, K. M. Wassarman, C. Rosenow, B. C. Tjaden, G. Storz, and S. Gottesman. Global analysis of small RNA and mRNA targets of Hfq. *Molecular Microbiology*, 50(4):1111–1124, 2003. (doi:10.1046/j.1365-2958.2003.03734.x).

[111] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989. (doi:10.1126/science.2468181).

[112] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31(13):3406–3415, 2003. (doi:10.1093/nar/gkg595).

[113] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9(1):133–148, 1981. (doi:10.1093/nar/9.1.133).

Lebenslauf

Name:	Isabelle Heinemeyer (geb. Schneider)
Geburtsdatum:	10. Juli 1977
Geburtsort:	Prudnik (Polen)
Nationalität:	deutsch

Abschlüsse:

11/2003:	**Diplom**, Note: Gut, Thema der Diplomarbeit: „Die Faktorisierungsmethode für das inverse Dirichlet-Problem für eine lokal gestörte Halbebene“
10/1999:	**Vordiplom**, Note: Gut
06/1997:	**Abitur**, Note: Gut

Schule und Studium:

seit 01/2004:	Doktorandin in der Abteilung für Bioinformatik, Georg-August-Universität Göttingen
09/1997 - 11/2003:	Studium der Mathematik mit Nebenfach Volkswirtschaftslehre, Georg-August-Universität Göttingen
02/2000 - 08/2000:	Auslandssemester, Université Victor Segalen Bordeaux 2 (Frankreich)
1994 - 1997:	Gymnasium, Salzgitter Bad
1990 - 1994:	Albert-Schloenbach-Realschule, Salzgitter Bad
1984 - 1990:	Grundschule, Riegersdorf (Polen) und Salzgitter Bad (Deutschland)

Studienbegleitende Tätigkeiten:

seit 05/2007	Wissenschaftliche Mitarbeiterin, Abteilung für Bioinformatik, Universität Göttingen, Tätigkeit: Forschung, Programmierung und Lehre
01/2004 - 04/2004	Wissenschaftliche Hilfskraft, Abteilung für Bioinformatik, Universität Göttingen, Tätigkeit: Programmierung
10/2002 - 09/2003	Studentische Hilfskraft am Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Tätigkeit: Lehre