universität wien

DISSERTATION

Titel der Dissertation

Strategies for computational noncoding RNA detection

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr. rer.nat.)

Verfasser:	Andreas Gruber
Dissertationsgebiet (lt. Studienblatt):	Molekulare Biologie
Betreuer:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Wien, November 2010

Contents

Abstract

Noncoding RNAs (ncRNAs) function directly at the level of transcripts without ever being translated into proteins. During the past few years it has become evident that ncRNAs are key players in many cellular processes. The set of actions is versatile, including transcriptional regulation, post-transcriptional regulation, chromatin modification or epigenetics. Genome-wide annotation and computational analysis of ncRNAs have met increased attention over the last years and RNA biology has become one of the primary research topics in modern molecular biology. Unlike protein coding genes, ncRNAs lack common statistically significant features, which makes the detection of novel ncRNAs a challenging task. In this thesis several computational strategies for noncoding RNA detection ranging from *de novo* detection to homology based methods are addressed. In particular, an improved version of the `RNAz` algorithm, an updated version of the `RNALfold` algorithm, and two homology search studies on the detection of new family members of 7SK RNA and sbRNAs are presented.

`RNAz` is a software package for the detection of conserved, thermodynamically stable RNA secondary structures. In this thesis an updated version of `RNAz` is presented. The use of a dinucleotide background model, a newly compiled training set, the ability to score structural RNA alignments and the use of Shannon entropy as a measure of sequence variation lead to an overall improved detection accuracy. When no or limited comparative genomics data is available the set of *de novo* detection methods for functional RNA structures becomes very sparse. For those cases, `RNALfoldz` an approach to quickly evaluate the set of local, thermodynamically stable structures in single genomic sequences has been developed. Efficient evaluation of thermodynamic stability is achieved by a modified support vector regression approach that significantly reduces execution time compared to former approaches.

Noncoding RNAs often evolve fast, retaining only a few sequence conserved elements. Conservation is, however, found at the level of secondary structures. This poses extreme challenges for RNA homology search methods. In this thesis, two studies on detection of new members belonging to the RNA families of 7SK RNA and sbRNAs are presented. By means of a computational ncRNA-specific promoter screen, 7SK genes are successfully identified in arthropod species, where experimental and computational studies previously failed to recover a candidate. The second study aims at the detailed characterization of the putative novel RNA family of sbRNAs. Using a set of several methods 240 new sbRNA genes are identified in nematode species. Detailed analysis of the structural features of sbRNAs shows that sbRNAs are not a novel RNA family, but are homologs of vertebrate Y RNAs.

Zusammenfassung

Noncoding RNAs (ncRNAs) sind RNA Moleküle, die ihre Funktion auf Ebene von Transkripten ausüben ohne jemals in Proteine übersetzt zu werden. In den vergangenen Jahren hat sich gezeigt, dass ncRNAs Hauptakteure in vielen zellulären Vorgängen sind, einschließlich Prozessen wie transkriptioneller und post-transkriptioneller Regulation, Chromatinmodifikation oder Epigenetik. Genomweite Annotation und bioinformatische Analysen von ncRNAs haben im letzten Jahrzehnt erhöhte Aufmerksamkeit erhalten und die RNA-Biologie ist zu einem der Hauptforschungsgebiete der modernen Molekularbiologie aufgestiegen. Im Gegensatz zu protein-kodierenden Genen weisen ncRNAs keine gemeinsamen statistisch signifikanten Eigenschaften auf. Dies macht das Auffinden von ncRNAs zu einer anspruchsvollen Aufgabe. In dieser Arbeit werden mehrere computergestützte Strategien zum effizienten Auffinden von ncRNAs präsentiert. Dabei werden sowohl Methoden zur *de novo* Erkennung als auch homologiebasierte Verfahren vorgestellt. Insbesondere beschäftigt sich diese Arbeit mit einer verbesserten Version des `RNAz` Algorithmus, einer aktualisierten Version des `RNALfold` Algorithmus und zwei Studien zur Homologiesuche von ncRNAs am Beispiel der beiden RNA Familien 7SK RNA und sbRNAs.

`RNAz` ist ein Softwarepaket zum Auffinden von konservierten, thermodynamisch stabilen RNA Sekundärstrukturen. Im Zuge dieser Arbeit wurde `RNAz` verbessert. Eine verbesserten Erkennungsgenauigkeit wird durch Verwendung eines dinucleotidbasierten Background-Modells, eines neu zusammengestelltes Training-Set, strukturellen RNA Alignments und der Shannon Entropie als Maß für Sequenzvariation erreicht. Wenn keine oder nur begrenzte Daten aus vergleichender Genomik vorliegen, gibt es nur wenige Methoden zur *de novo* Vorhersage funktioneller RNA Strukturen. Für solche Fälle, ist mit `RNALfoldz` ein Ansatz entwickelt worden, der es erlaubt lokale, thermodynamisch stabile Strukturen in einzelnen genomischen Sequenzen effizient zu finden. Die effiziente Berechnung der thermodynamischen Stabilität wird durch einen Modifikation der Support Vector Regression erzielt. Diesem neue Ansatz führt zu einer deutlichen Reduktion der Ausführungszeit im Vergleich zu früheren Methoden.

Noncoding RNAs evolvieren oft schnell und behalten häufig nur ein paar sequenzkonservierte Elemente. Da oft nur Sekundärstrukturelemente konserviert sind, stellt die RNA Homologiesuche eine extreme Herausforderung für Suchmethoden dar. In dieser Arbeit werden zwei Studien zur Homologiesuche von RNA Familien, nämlich für 7SK RNAs und sbRNAs, vorgestellt. Mit Hilfe eines computergestützten ncRNA-spezifischen Promoter-Screens gelang es 7SK Gene in der Gruppe der Arthropoden zu identifizieren, wo zuvor sowohl experimentelle als auch computergestützte Methoden gescheitert waren. Die zweite Studie behandelt die detaillierte Charakterisierung der vermeintlich neuen Familie der sbRNAs. Mit einer Reihe an verschieden Methoden gelang es 240 neue sbRNA Gene in Nematoden zu identifizieren. Eine detaillierte Analyse der strukturellen Merkmale von sbRNAs zeigte schliesslich, dass sbRNAs nicht eine neuartige RNA Familie sind, sondern homolog zur Familie der Y RNAs.

Acknowledgements

Research presented in this thesis was conducted at the Institute of Theoretical Chemistry, University of Vienna, and at the Professur für Bioinformatik, University of Leipzig, during the years 2007 - 2010. First and foremost I would like to thank my supervisor Ivo L. Hofacker for extraordinarily supporting my scientific career and giving me the freedom to explore every (scientific) idea that came to my mind. I am especially grateful for funding all of my conference and summer school attendances. Many of the ideas and findings in this thesis originated from discussion with and inspiration by Peter F. Stadler. *Mahalo*, Peter, for all the support and the good times we had over a couple of beers.

Special thanks go to my room mates in Vienna Ronny, Christionchus, and Caro, and my room mates in Leipzig Dom, Sven, and Manja. I know it hasn't been easy to do some proper work with me talking all the time. You all were my motivation to get up every day and push science forward. Thank you, Ronny, for 24h support in programming tasks. Thank you, Christionchus, for bringing my laptop back to life and all the other stuff nobody actually knows what it is all about. Thank you, Manja, for the many hours of (scientific) talk, late-night shows at the institute, and for letting me win in Tetris a few times. Thank you, Sven, for your great collaboration on `RNAz` and for providing your sofa for a couple of days. Thank you, Dom, for the many hours and nights of joy we had together. Although Berni was never officially a room mate of mine, he spent half the day in my office. Thank you, Berni, for being present whenever needed. A big thank you, of course, goes across the pond to Wash. `perl -e 'print "Thank you, Xtof!\n"'`. I also thank all the other colleagues at the TBI and in Leipzig, and the Google translate service for assisting in preparation of the German version of the abstract.

The last words go to *Tini*. Thank you for being at my side!

"Genießt jede Minute, die ihr spielend verbringt!"

Alexey Pajitnov, Inventor of Tetris

1

Introduction

In a visionary act Francis Crick was the first to define relations between DNA, RNA, and proteins, the main macromolecules found in a cell (Crick, 1958, 1970). He proclaimed the *central dogma of molecular biology* and since then this dogma has shaped the scientific community's view on the roles of these macromolecules. RNA has long been regarded as an intermediate to promote the flow of information from DNA to proteins. Over the last decade evidence has, however, mounted that RNA molecules have versatile functions inside a cell ranging from catalytic processes to complex patterns in gene regulation (Fedor and Williamson, 2005; Amaral et al., 2008; Sharp, 2009; Waters and Storz, 2009). Research on RNA molecules and their functions has now again become a primary research topic in molecular and computational biology.

1.1 The early steps in (computational) RNA biology

The ground-breaking work of Watson and Crick (1953) in describing the double helical structure of DNA was one of the first contributions that helped to establish the field of *molecular biology*. At that time scientists were working hard at putting together the pieces of the puzzle of life, not even knowing if they possessed all pieces. One of those pieces was the macromolecule RNA. Neither questions about the role of RNA in the cell nor if RNA could also form double helical structures could be clearly answered. First experiments on the structure of RNA (Rich and Davies, 1956; Felsenfeld et al., 1957) soon showed that RNA is also capable of forming helical structures and that RNA can adopt complex structures by intra-molecular base-pairings (Fresco et al., 1960). Continued work by Rich (1960) also revealed that DNA-RNA complexes can be formed. Francis Crick canalized all information available at that

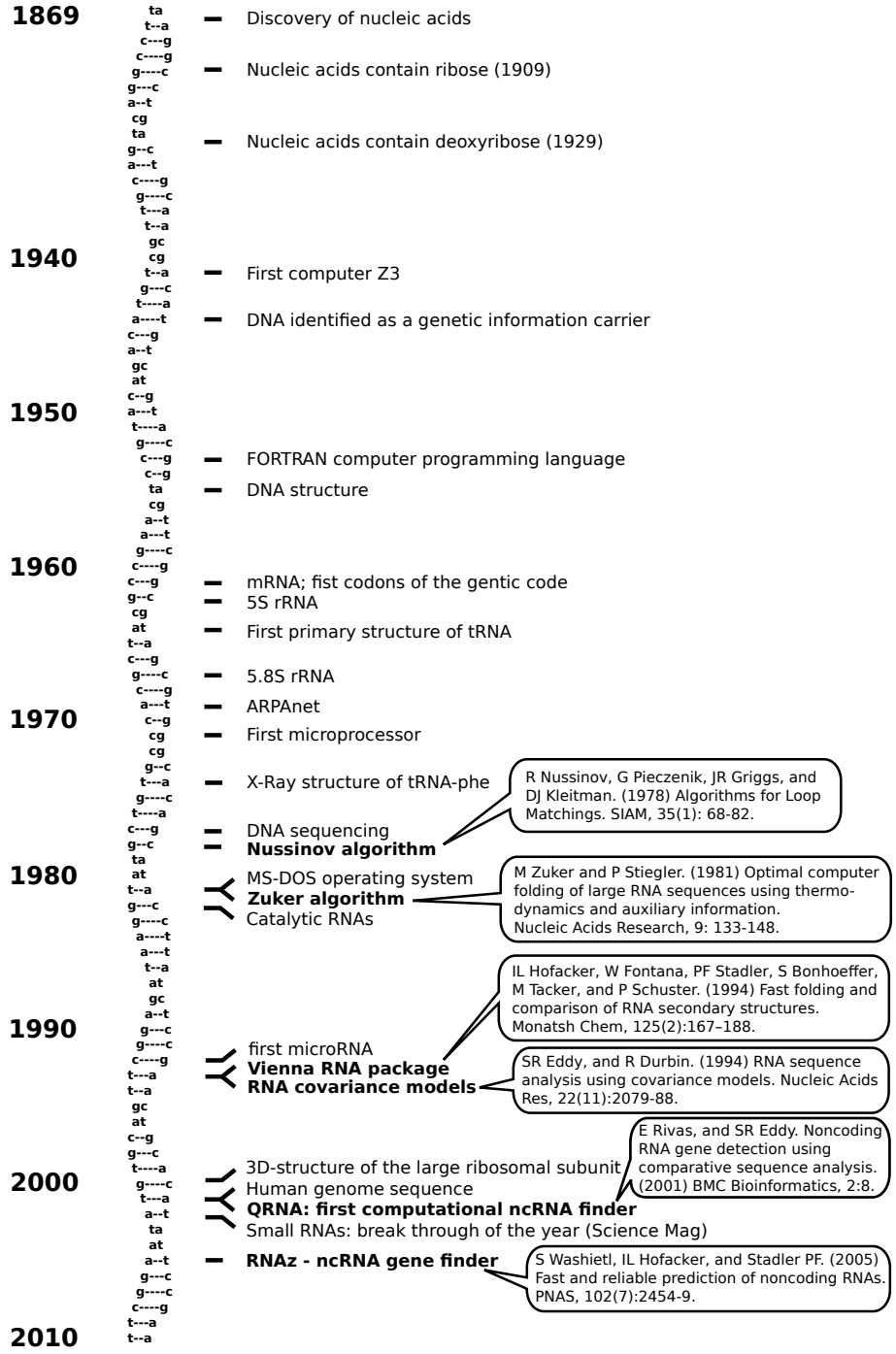


Figure 1.1. *Timeline depicting selected, major findings and inventions in computer science, molecular biology, and computational RNA biology.*

time and postulated the *central dogma of molecular biology* (Crick, 1970). The central dogma shaped the view of the roles of the biological macromolecules DNA, RNA, and proteins for the next decades. DNA is assigned the task of information storage, proteins are responsible for catalytic events in a cell, and RNA acts as a vehicle for information transfer from DNA to proteins. Figure 1.1 depicts a timeline of some selected, major findings and inventions in computational RNA biology and molecular biology discussed in this chapter.

It had been realized early on that RNA molecules can adopt complex conformations (Fresco et al., 1960), and scientist were striving to develop models that could be used to assess and quantify the network of base-pairing interactions. Based on experimental measurements a rudimentary set of energy parameters were available that allowed to calculate the free energy of a secondary structure (Tinoco et al., 1971). Nussinov and colleagues were the first to present efficient algorithms for the prediction of RNA secondary structures by either maximizing base-pairings (Nussinov et al., 1978) or minimizing the free energy (Nussinov and Jacobson, 1980). Building on the work of Nussinov, Zuker and Stiegler (1981) proposed an RNA folding algorithm that uses a more refined energy model that takes different loop types into account. A variant of the Zuker algorithm is also implemented in the *Vienna RNA package* (Hofacker et al., 1994), which is extensively used in this thesis. There are, however, also methods for structure prediction that do not consider a thermodynamical model. Gutell et al. (2002) reviewed the power of comparative sequence analysis for RNA structure prediction. In this work covariation analysis, which aims at detecting and quantifying exchanges between a set of base-pair types that covary with one another at a specific position, has been successfully applied to derive structure models for 16S and 23S ribosomal RNA molecules.

With the findings of the groups of Altman and Cech (Cech et al., 1981; Guerrier-Takada et al., 1983) that some RNA molecules have enzymatic activity a new era for RNA molecules was started. RNA molecules that can act as chemical catalysts were named *ribozymes*, short for *ribo*nucleic en*zymes*. The group led by Cech discovered that an intron within a pre-ribosomal RNA from *Tetrahymena thermophila* can catalyze its own cleavage (called self-splicing) to form the mature ribosomal RNA product, while the group of Altman showed that the active component of the RNase P particle responsible for cleavage of a phosphodiester bond to form the mature transfer RNA is actually an RNA molecule. The fact that RNA can both store information and act as an enzyme led to the *RNA world hypothesis* suggesting that RNA was the original molecule of life (Gilbert, 1986). This hypothesis has been recently fueled by the works of Lambert et al. (2010) and Barks et al. (2010). Both works report on advances in determining on how building blocks of nucleic acids might have emerged at abiotic conditions.

1.2 ncRNAs: the new hot topic

Besides the RNA classes of ribosomal RNAs (rRNA) and transfer RNAs (tRNAs), a functional description of which can be found already in any standard high school biology textbook, a lot of new noncoding RNA (ncRNA) classes have been discovered. Noncoding RNA or non-protein-coding RNA is a term that has become increasingly popular over the past years. In general, noncoding RNA describes RNA molecules that do not carry information for the translation to proteins, but rather exert their function in a cell as RNA molecules themselves. Recent research on noncoding RNA has to a great extent been fueled by findings of genome sequencing studies such as the ENCODE project (ENCODE Project Consortium, 2007). It has been shown that a significant fraction of the human genome is transcribed. Parts of the genome that had previously been considered as "Junk DNA" have since then moved into the focus of research. Moreover, genome sequencing studies and accompanying papers have shown that the number of protein-coding genes alone is not sufficient to explain the differences between simple and complex life forms.

The known spectrum of biological functions, ncRNAs are involved in, has considerably broadened over the past years. At this point, let us briefly review some of the most important RNA classes. Because of the ability of RNA molecules to form inter-molecular base-pairing interactions, many functional RNAs are involved in biological processes that involve other RNA molecules, e.g. tRNAs "read" base triples of mRNAs encoding information for amino acids, the group of small nuclear RNAs (snRNAs) is involved in splicing of mRNA (Valadkhan, 2010), or the class of small nucleolar RNAs (snoRNAs) guides chemical modifications (methylation and pseudouridylation) of ribosomal RNAs (Bachellerie et al., 2002). Transfer-messenger RNA (tmRNAs) have structural and functional properties of both a tRNA and a mRNA. They are able to rescue stalled transcriptional complexes and are involved in protein quality control by adding tags for proteolysis to ribosome-associated protein-fragments (Dulebohn et al., 2007). On the other hand, the family of Y RNAs is involved in RNA quality control (Stein et al., 2005) and has recently been shown to be required for initiation of DNA replication in human cells by a yet unknown mechanism (Gardiner et al., 2009). Other housekeeping RNAs are telomerase RNAs, which serve as template for elongating telomeres, 7SK RNA, which is involved in controlling eukaryotic gene expression by regulating the fraction of active RNA polymerase II molecules, or the RNA component of the signal recognition particle (SRP), which acts in promoting protein translocation across the endoplasmic reticulum membrane. The family of microRNAs with its first member discovered by Lee et al. (1993) has profoundly changed our view of gene regulation. The binding of microRNAs to complementary sequences in messenger RNA molecules eventually leads to silencing of the targeted gene.

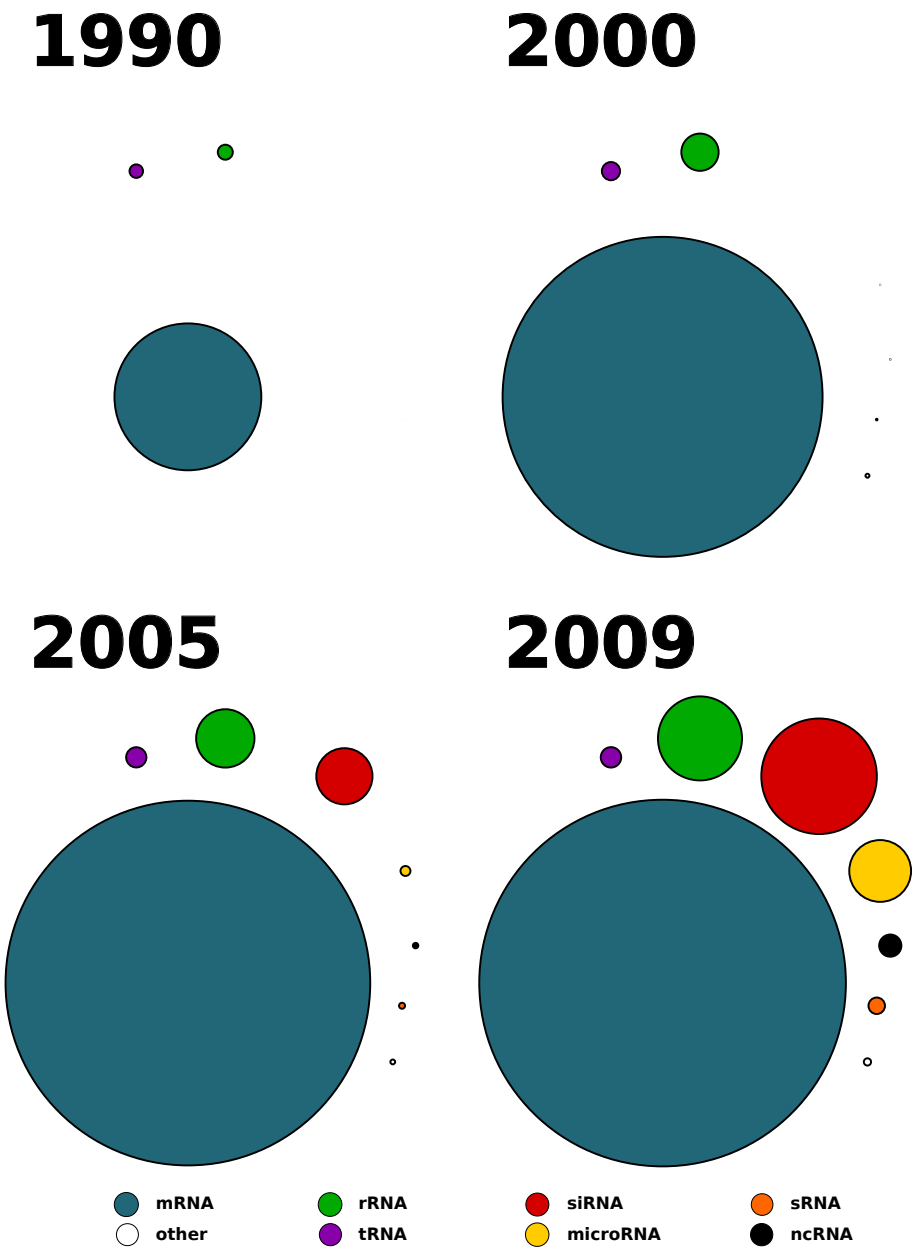


Figure 1.2. ***Comparative analysis of Pubmed indexed articles.*** *Circles are drawn proportionally to the number of publications matching the corresponding RNA class. The scientific community's research focus on RNA moved away from a protein synthesis centric view (mRNA, tRNA, rRNA) to a more diverse set of biological functions, where RNA molecules also account for various gene regulatory processes (miRNA, siRNA, sRNA).*

Current estimates start from at least 800 microRNAs in the human genome (Bentwich et al., 2005). MicroRNAs seem to be an eukaryotic innovation, but also in bacteria RNA molecules, often termed small RNAs (sRNA), have been found regulating synthesis of proteins by a multitude of different actions (Massé et al., 2003). While the RNA families discussed so far have been fairly well characterized in terms of function, structure and evolutionary history, there are still a lot of RNA molecules where little is known about their function. Long noncoding RNAs currently experience a hype not only in the RNA community. In a recent contribution Tsai et al. (2010) showed that long noncoding RNAs can regulate chromatin states and epigenetic inheritance and Huarte et al. (2010) identified a novel regulation mechanism of the p53 tumor suppressor gene. These recent findings clearly demonstrate that we are just beginning to understand the multitude of biological processes and pathways RNA molecules are involved in.

Figure 1.2 compiles results of a literature screen of Pubmed indexed articles. It clearly shows how the scientific community's view on RNA has changed over the last twenty years. In the years 1990 to 2000 research on RNA was dominated by RNAs involved in protein synthesis. mRNAs, tRNAs and ribosomal RNAs were the major RNA species investigated at that time. 2005 marks a trend of new RNA classes that shapes our current view (2009) of RNA. MicroRNAs, siRNAs (small interfering RNAs), and sRNAs exert their function not as catalytically active molecules or molecules in the protein synthesis pathways of a cell, but rather function in gene regulation at various stages. Today's role of RNA molecules is, hence, seen threefold: i) information transfer, ii) catalytic functions, and iii) gene regulation.

In this thesis we explore computational methods for the efficient detection of novel functional RNA structures and RNA genes. In detail, we address *de novo* detection of functional RNA secondary structures by means of improved versions of the `RNAz` (Washietl et al., 2005b) and `RNALfoldz` (Hofacker et al., 2004b) algorithms. We also present two studies on homology search and detailed evolutionary characterization of the RNA families of sbRNAs (Deng et al., 2006) and 7SK RNA (Zieve and Penman, 1976). In the following section we give a brief introduction into the topic of computational noncoding RNA detection and highlight current approaches.

1.3 Computational noncoding RNA detection

The emerging interest in noncoding RNAs has also led the scientific community to focus on the development of computational tools that are capable of detecting novel ncRNAs. First steps were done in the field of RNA homology search. tRNAs were among the first well characterized RNA molecules, and the well defined secondary structure and the internal promoter elements led to the development of a series of computational tools for the efficient detection of tRNA genes (Fichant and Burks, 1991; Pavesi et al., 1994; Lowe and Eddy, 1997). `tRNAscan-SE` (Lowe and Eddy, 1997) is still the state-of-the-art program for tRNA detection. There are, however, only few RNA families that are so abundant and of broad interest that specialized tools have been developed. Notable examples are tmRNAs (`BRUCE` - Laslett et al. (2002), `ARAGORN` - Laslett and Canback (2004)), RNase P (`Bcheck` - Yusuf et al. (2010)), snoRNAs (`snoScan` - Lowe and Eddy (1999), `snoGPS` - Schattner et al. (2004), `snoSeeker` - Yang et al. (2006), `snoReport` - Hertel et al. (2008), `Fisher` - Freyhult et al. (2008)) and microRNAs (`MiRscan` - Lim et al. (2003), `MiRseeker` - Lai et al. (2003), `ProMiR` - Nam et al. (2005), `FindMiRNA` - Adai et al. (2005), `RNAmicro` - Hertel and Stadler (2006)). These RNA classes have additional features that can be exploited to improve the detection sensitivity.

Model- or descriptor-based tools such as `Infernal` (Nawrocki et al., 2009) or `RNABOB` (Eddy, 1996) are the most widely used tools for ncRNA homology search nowadays. Their generic framework does not restrict the application of the method to a certain class of RNA molecules. In case of `Infernal` the input consists of a structure-annotated multiple sequence alignment, from which a covariance model (Eddy and Durbin, 1994) is generated. `RNABOB` takes as input a set of user defined sequence and structure patterns describing the structure of the RNA molecule of interest.

While in RNA homology search we now see a set of technically mature algorithms and tools, the establishment of tools for *de novo* detection of functional RNA structure is still in progress. The most obvious reason for that is that noncoding RNAs, unlike proteins, lack common statistically detectable signals. There are no clear start and stop signals and there is no equivalent for the codon bias in noncoding RNAs. The first idea that has been pursued is that RNA genes, since their functions heavily depend on their structures, should have structures that are more thermodynamically stable than expected by chance. This has, however, been a topic of controversial discussion (Le et al., 1990b,a; Seffens and Digby, 1999; Workman and Krogh, 1999; Rivas and Eddy, 2000; Clote et al., 2005; Freyhult et al., 2005). The current view is that certain RNA families do indeed show signatures of thermodynamic stability, but this concept cannot be applied to all functional RNA molecules. `NCRNASCAN` (Rivas and Eddy,

2000) was the first attempt to predict ncRNA genes based on thermodynamic stability. The clear message of this study was that thermodynamic stability alone cannot serve as a good signal for ncRNA genes in a genome-wide search. On the other hand, a series of tools such as `QRNA` (Rivas and Eddy, 2001), `ddbRNA` (di Bernardo et al., 2003), `MSARi` (Coventry et al., 2004), or `Evofold` (Pedersen et al., 2006) has been developed that try to exploit the network of compensatory (two base changes) and consistent mutations (one base change) often found in homologous RNA sequences (see Fig. 1.3). Approaches are quite different in their nature though.

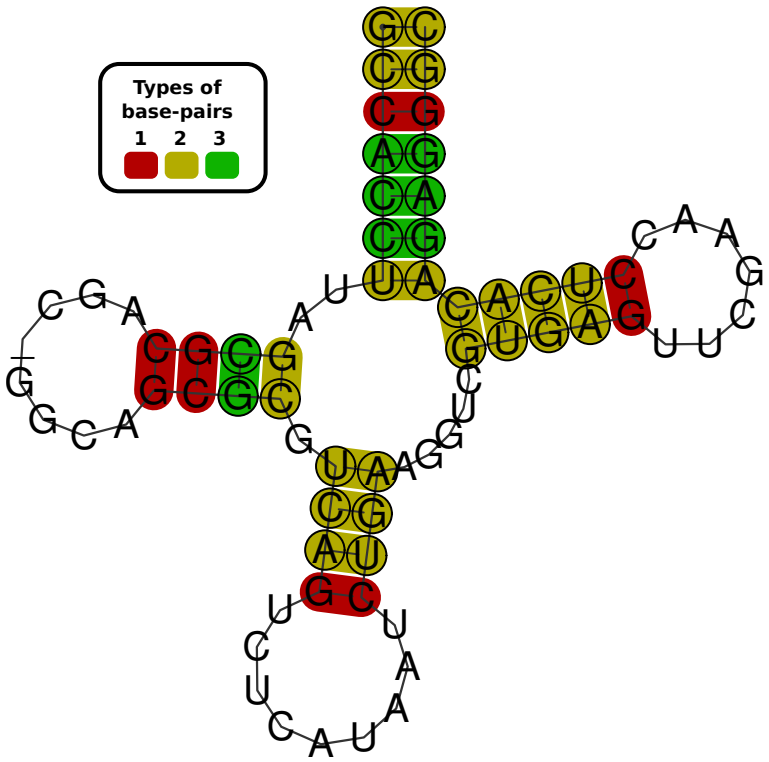


Figure 1.3. ***Consensus structure of human tRNA-met genes.*** *Base-pairs are colored according to the number of compensatory and consistent mutations supporting a base-pair at a particular position. Sequences were obtained from tRNAdb (Jühling et al., 2009). Consensus structure prediction was done with* `RNAalifold` *(Bernhart et al., 2008).*

`QRNA` operates on pairwise alignments and uses stochastic context free grammars to discriminate between three models (RNA, protein and a null hypothesis). `Evofold` is a follow-up and extends the model of `QRNA` to operate on multiple sequence alignments. `ddbRNA` and `MSARi` search for conserved stem structures and evaluate the significance on randomly shuffled versions of the input alignment and on a distribution mixture model, respectively. With

`Alifoldz` Washietl and Hofacker (2004) offered an approach that combines both thermodynamic stability assessment and evaluation of evolutionary conservation of RNA secondary structures. Again, significance was evaluated on randomly shuffled versions of the input alignment. In a later work (Gesell and Washietl, 2008) – `SISSIz` – this concept has been improved to consider a dinucleotide background model. `RNAz` (Washietl et al., 2005b) is currently the most widely used program for noncoding RNA detection. It is also based on the principles of thermodynamic stability and evolutionary conservation of secondary structures but uses a machine learning approach to allow fast and accurate detection on a genome-wide scale. Assessment of the evolutionary conservation is heavily influenced by the quality of the input alignments with regard to RNA secondary structures. With increasing sequence variation sequence-only based alignment programs often fail to do a good job though. Two contributions in this field `RSSVM` (Xu et al., 2009) and `Dynalign+SVM` (Uzilov et al., 2006) seize the power of structural alignments to increase detection sensitivity. So far, approaches discussed are general in their concept in a sense that there are no species-specific restrictions these algorithms can be applied to. In the field of bacterial small noncoding RNA detection several efforts have taken advantage of promoter and transcriptional terminator signals (Argaman et al., 2001; Chen et al., 2002; Livny et al., 2005; Yachie et al., 2006; Sridhar et al., 2010). There is also evidence that methods using base composition statistics can be successful in detecting noncoding RNAs (Carter et al., 2001; Schattner, 2002; Wang et al., 2006; Salari et al., 2009). In hyper-thermophilic organisms, especially, screens for GC-rich regions have effectively identified noncoding RNAs (Klein et al., 2002; Upadhyay et al., 2005; Larsson et al., 2008).

So far, approaches that infer novel ncRNAs from genomic sequences have been discussed. Many of the above cited works have subsequently confirmed that some of their predicted ncRNA candidates are indeed expressed and a transcript can be found. Searching for noncoding RNAs directly at the level of transcripts seems therefore a promising strategy, since there is already evidence that the sequence is expressed. A first contribution in this field was by MacIntosh et al. (2001). They screened collections of *Arabidopsis* expressed sequence tags (ESTs) for transcripts that do not show coding potential. Similar efforts (FANTOM Consortium, 2002; Numata et al., 2003; Liu et al., 2006) were done on data generated by the FANTOM project (The FANTOM Consortium, 2001), which aimed at the functional annotation of a full-length mouse cDNA collections. It has long been in question if noncoding RNAs can really be found in EST data, since many EST projects aimed at collecting mRNAs and various experimental filtering stages were applied prior to sequencing. There is, however, a series of contributions using quite different approaches, that report on successful identification of noncoding transcripts in EST collections (Tupy et al., 2005; Seemann et al., 2007; Xue

et al., 2008; Arrial et al., 2009). Two recent works (Jung et al., 2010; Langenberger et al., 2010) make use of next generation sequencing data and analyze the shapes of read patterns to find homologs of known ncRNA classes.

At this point let me state a few words on validation of predicted functional noncoding RNAs. Based on the assumption that expression is inherently linked to function, the standard protocol to verify a candidate is to show that there is a transcript that corresponds to the predicted locus either by RT-PCR or Northern blotting. This has been accepted by the scientific community for years. Studies like the ENCODE project (ENCODE Project Consortium, 2007) have shown that there is a good chance that any region in the genome is transcriptionally active at some time in some tissue. Hence, demonstrating that a predicted ncRNA is really functional merely based on transcriptional evidence will become less accepted in future times.

Although there have been major advances in the field of computational noncoding RNA detection it is still very competitive. Especially, the growing number of long noncoding RNAs poses new challenges for novel and improved algorithms.

1.4 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 gives a general overview of methods, algorithms, techniques and findings essential to this thesis. In particular, important chemical and biological aspects of RNA molecules, RNA folding algorithms, the concept of support vector machines, and tools for ncRNA detection are discussed. Chapter 3 introduces the particular problem statements discussed in this thesis. Chapters 4, 5, 6, and 7 are original research articles addressing problems of *de novo* detection and homology search of noncoding RNAs. In detail, Chapter 4 presents an improved version of the noncoding RNA gene finding algorithm `RNAz`. In Chapter 5, we describe a modified version of the `RNALfold` algorithm and its application and usefulness for genome-wide ncRNA detection. Chapter 6 presents a study that reports on the successful detection of 7SK RNA homologs in arthropod species. In Chapter 7, a study revealing new Y RNA family members in nematodes is presented. Finally, in Chapter 8, we conclude our work and discuss directions for future research.

2

Background

This chapter briefly discusses the basics essential to this thesis ranging from biological and chemical aspects of RNA to modern concepts of machine learning algorithms. In detail, we discuss RNA secondary structure prediction algorithms, the alignment problem for related sequences, the concept of support vector machines, and computational tools for both *de novo* detection and homology search of noncoding RNAs.

2.1 Chemistry and biology of RNA molecules

Like any macro-molecule ribonucleic acid (RNA) is composed of smaller building blocks. For RNA these building blocks are called nucleotides and consist of a nitrogenous hetero-cyclic base, a pentose sugar, and a phosphate group. Nucleotides are linked by phosphodiester bonds to form a polymer. The bases adenine (`A`) and guanine (`G`) belong to the group of purines and form a double ring, whereas cytosine (`C`) and uracil (`U`) are pyrimidine derivatives. Since the work of Watson and Crick, it is well known that nucleic acids can form complex base-pairing patterns via non-covalent hydrogen bonds between bases. In general, we distinguish between the Watson-Crick or canonical base-pairs (`A` · `U`, `U` · `A`, `C` · `G`, `G` · `C`), and the Wobble base-pairs (`G` · `U`, `U` · `G`). The hydrogen bonding patterns of RNA base-pair interactions are depicted in Fig. 2.1. All other base-pairing interactions are observed less frequently and are therefore referred to as non-standard base-pairs (Leontis and Westhof, 2001). Although non-standard base-pairs can account for a significant fraction of base-pairs in some RNA molecules and currently meet increased attention (Parisien and Major, 2008; Zhong et al., 2010), Watson-Crick and Wobble base-pairs are considered as the main driving forces in shaping the *secondary structure* of an RNA molecule. RNA secondary structure can be defined as the architecture

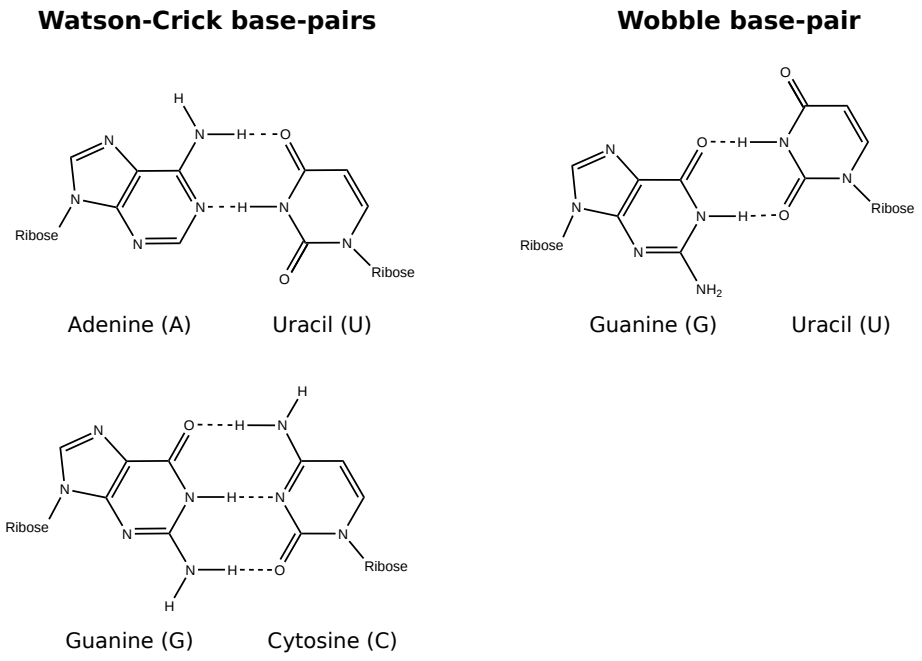


Figure 2.1. ***RNA base-pairing interactions.*** *A-U base-pairs and G-C base-pairs belong to the set of Watson-Crick or canonical base-pairs. G-U base-pairs are referred to as Wobble base-pairs.*

of helical regions and loops formed by intra-molecular base-pairings. Hydrogen bonding is responsible for selective pairing of bases, most of the energy gain from adopting a particular secondary structure formation comes, however, from π-π stacking of the aromatic systems of the bases (Petersheim and Turner, 1983). The folding of an RNA molecule can hence be seen as a hierarchical process, where the 3D structure (tertiary structure) is shaped to a large extent by secondary structural elements (Tinoco and Bustamante, 1999). This process is schematically depicted in Fig. 2.2 showing the primary (sequence), secondary, and tertiary structure of a tRNA molecule.

Double-stranded DNA usually adopts a helical form called B-helix. Double-stranded RNA (dsRNA) is not able to form a B-helix, since the additional hydroxyl group at the 2' position of ribose prevents adopting of this conformation. The usual form of dsRNA is the A-helix. This additional hydroxyl group compared to DNA makes RNA more catalytically active, see e.g. the cleavage reaction catalyzed by the hammerhead RNA (Scott et al., 1995). So far, when discussing secondary structure elements we were talking only about intra-molecular base-pairs. Pairing interactions are not restricted to be intra-molecular. Inter-molecular pairing is crucial in many biological processes. Consider e.g. the RNA primer in DNA replication,

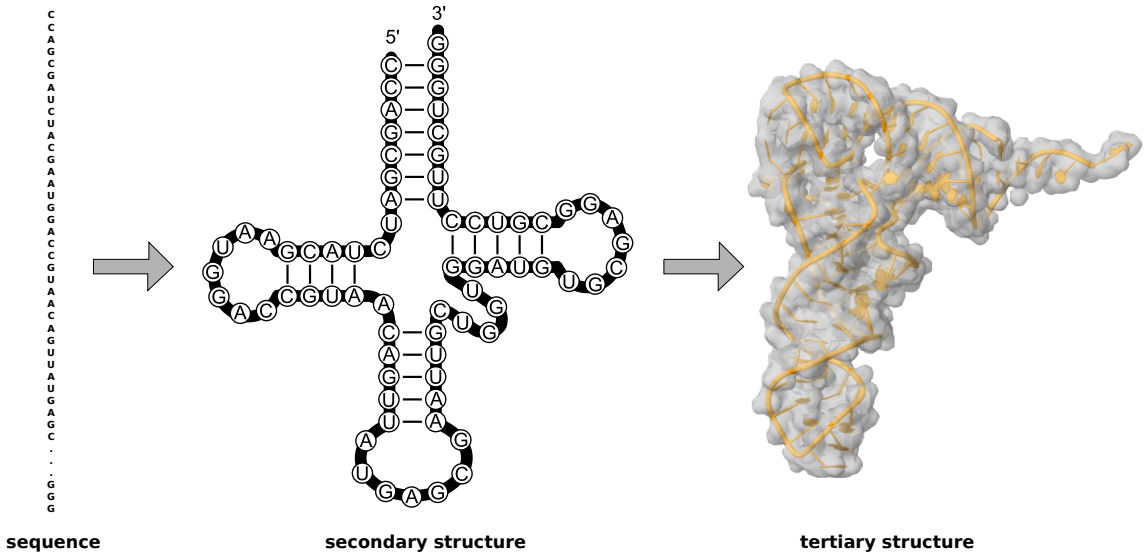


Figure 2.2. ***Folding hierarchies of an RNA molecule illustrated on a tRNA.*** *The formation of base-pairs between complementary regions forms a network of stems and loops, referred to as secondary structure. The secondary structure is energetically much more stable than the tertiary structure. The process of RNA folding can hence be seen as hierarchical in its nature.*

where we see a heterodimer of DNA and RNA or tRNA-mRNA interactions in translation. microRNAs and snoRNA also recognize their targets via base-pairing.

In cells RNA is synthesized by the process known as transcription, where an RNA molecule is generated by an enzyme complex from a DNA or in the case of RNA viruses an RNA template. The enzyme that catalyzes the reaction is called RNA polymerase (pol). In eukaryotes there are several types of RNA polymerases, responsible for the transcription of different classes of RNA molecules. RNA polymerase II is responsible for transcribing protein coding genes (mRNA) and several snRNAs. RNA polymerases I and III have a more limited set of action. RNA polymerase I synthesizes a pre-rRNA 45S, which matures into 28S, 18S and 5.8S rRNAs, the core RNA components of the ribosome. The set of transcribed genes by RNA polymerase III is also limited. In particular, the RNA families of tRNAs, 5S rRNA, 7SK, U6 snRNA, U6atac snRNA, 7SK RNA, RNase P, RNase MRP and Y RNAs are known to be transcribed by pol III. The recruitment of a specific RNA polymerase is achieved by specific promoter elements. Figure 2.3 depicts promoter structures of RNA genes transcribed by RNA pol III. Type 1 (rRNA) and type 2 (tRNA) are internal promoter elements that reside within the RNA gene. Type 3 promoter elements are found upstream of the RNA gene, consisting of

a TATA box, a proximal sequence element (PSE) and an enhancer element named distal sequence element (DSE). The transcription termination signal is a stretch of a at least four `T` residues.

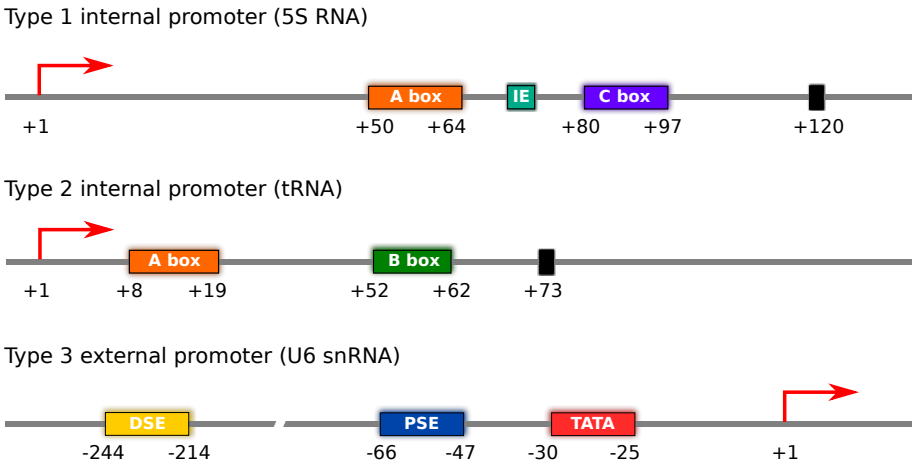


Figure 2.3. ***Promoter elements of RNA genes transcribed by RNA pol III.*** *Figure adapted from Cassimeris et al. (2010). Abbreviations: IE - internal element, PSE - proximal sequence element, DSE - distal sequence element. Black box marks the transcription terminator composed of four or more* `T` *residues.*

2.2 RNA secondary structure prediction

While in the previous section we loosely defined RNA secondary structure as the architecture of helical regions and loops formed by intra-molecular base-pairings, we aim to give a more formal definition here. Given a finite alphabet $\mathbb{A}_{RNA} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ we define the primary structure or sequence of an RNA molecule as a string $S = s_1 s_2 \ldots s_n$, where n is the number of nucleotides in the molecule and $s \in \mathbb{A}_{RNA}$. A base-pair between nucleotides s_i and s_j is denoted in the following by $i \cdot j$. An RNA secondary structure Y is then defined as the set of base-pairs $i \cdot j$ $(i < j)$ meeting following criteria:

(i) $i \cdot j \in \mathbb{B}$ where $\mathbb{B} = \{\mathtt{A} \cdot \mathtt{U}, \mathtt{U} \cdot \mathtt{A}, \mathtt{C} \cdot \mathtt{G}, \mathtt{G} \cdot \mathtt{C}, \mathtt{G} \cdot \mathtt{U}, \mathtt{U} \cdot \mathtt{G}\}$.

(ii) Two base-pairs $i \cdot j$ and $k \cdot l$ are either identical, or else $i \neq k$ and $j \neq l$.

(iii) Two base-pairs $i \cdot j$ and $k \cdot l$ with $i < k$ satisfy either $i < j < k < l$ (Fig. 2.4b) or $i < k < l < j$ (Fig. 2.4c).

(iv) For any base-pair $i \cdot j$: $|i - j| > 3$.

These conditions help us to reduce the folding space of RNA molecules that has to be considered to a well defined set of structures and to deduce algorithms that are capable of evaluating this folding space in reasonable time. Condition (i) restricts the set of base-pairs to Watson-Crick and Wobble base-pairs. This is not a strict condition. It is, however, motivated by the fact that most of the free energy of an RNA molecule is contributed by these base-pairs and that only for these standard base-pairs energy parameters have been reliably measured (Mathews et al., 1999, 2004). Once parameters for non-standard base-pairs will have been determined, the set of valid base-pairs has to be extended for sure. In fact, a recent statistically motivated approach for RNA folding reported on the successful prediction of non-standard pairing interactions (Parisien and Major, 2008). Condition (ii) is to ensure that each base can take part in at most one base-pair (Fig. 2.4a), while condition (iii) prohibits the formation of pseudo-knots (Fig. 2.4d). Base-triplets and pseudo-knots are, however, frequently observed in RNA molecules and form functionally important structural elements. Fig. 2.4e shows canonical base-pairs, base triples, and pseudo-knots of a tRNA molecule. A recent survey also demonstrated that these interactions are often evolutionarily conserved (Messmer et al., 2009). Despite the biological importance of the pseudo-knots, there are two good reasons to exclude them from the set of secondary structure motifs. First, there is no energy parameter set available accounting for all possible pseudo-knotted interaction types. And second, the computational prediction of pseudo-knotted structures without any restriction of the folding

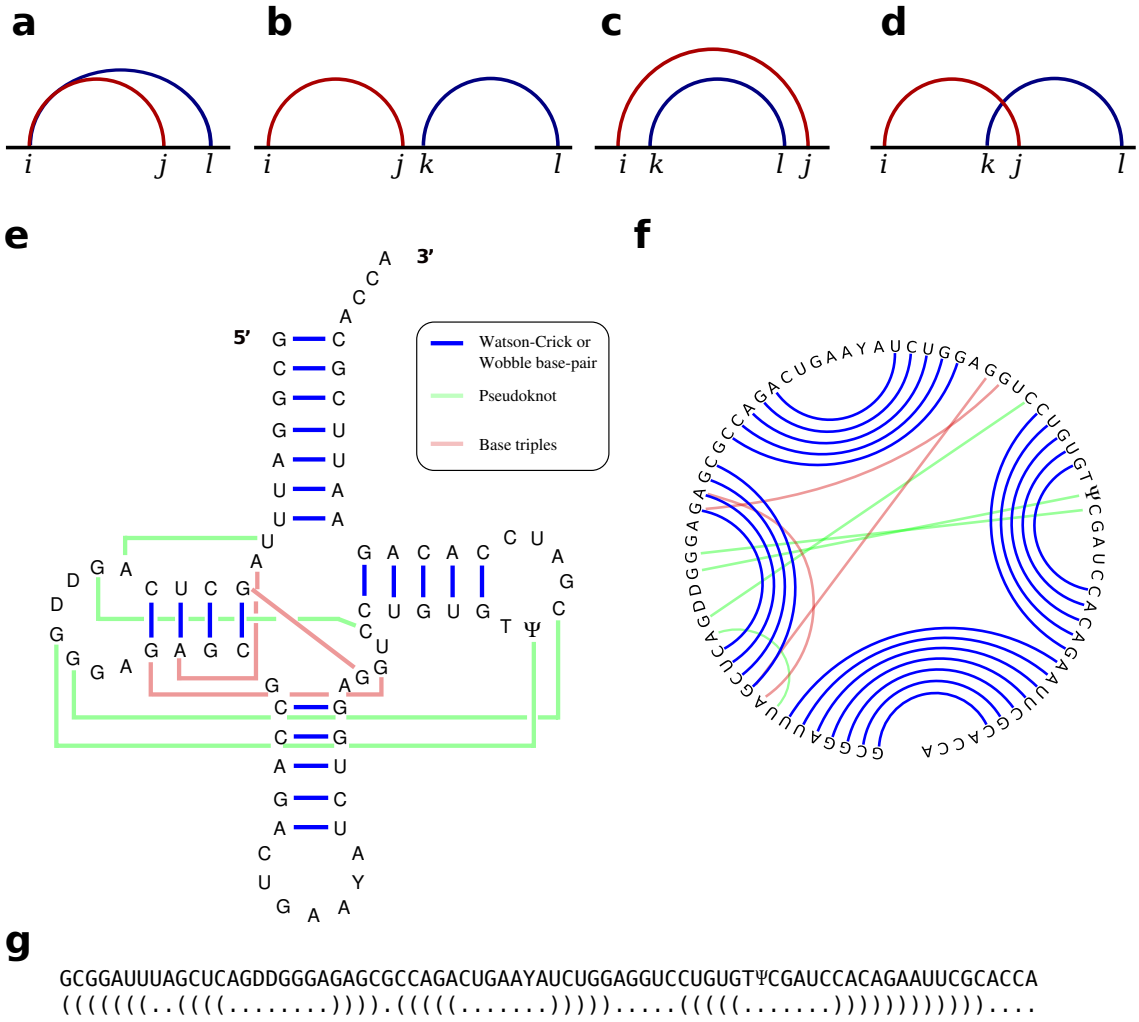


Figure 2.4. ***RNA secondary structure rules and visualizations.*** ***a-d*** *Visualization of the rule set of RNA secondary structures.* ***a****. Base triples (tertiary structure motif).* ***b****. Two adjacent base-pairs.* ***c****. Nested base-pairs.* ***d****. Pseudo-knot (tertiary structure motif).* ***e****. Visualization of the secondary structure of a tRNA molecule (with modified residues) indicating standard base-pairs in blue, and tertiary motifs such as pseudo-knots (green) and base triples (red). Non-secondary structural motifs are indicated by pale colors. Figure adapted from (Messmer et al., 2009)* ***f****. Circular Feynman representation of the tRNA structure from (e). Tertiary structure motifs are indicated in pale colors. By definition, a secondary structure is free of crossing edges.* ***g****. Sequence and secondary structure of the tRNA in the Vienna dot-bracket notation.*

space using energy-based models is NP-hard (Lyngsø and Pedersen, 2000). There exists, however, a series of practical approaches that either limit the search space to a certain set of pseudo-knots or are based on a heuristic, see e.g. Rivas and Eddy (1999), Reeder and Giegerich (2004) or Ren et al. (2005). To the knowledge of the author there is currently no computational approach for predicting base triples, which are in most cases non-standard base-pairs anyway. Due to these difficulties pseudo-knots and base triples are commonly referred to as tertiary structure motifs. The last condition (iv) prohibits sharp U-turns. This is argued with sterical hindrance by the RNA sugar-phosphate backbone. Any valid secondary structure can be represented as a string over the alphabet $\mathbb{A}_{Structure} = \{(,),.\}$. Characters "(" and ")" correspond to the 5' base and the 3' base in a base-pair, while "." denotes an unpaired residue (Fig. 2.4g). This dot-bracket notation is intuitive as it follows mathematical rules for setting parentheses, but is not well suited for easily recognizing the fold of an RNA molecule. Visualization of RNA structures is in many cases a key step in the analysis of a molecule's function. From a mathematical point of view an RNA secondary structure can be considered as an outer-planar graph (Fig. 2.4f). By shortening the edges of this graph to a fixed length, the commonly known representation of secondary structures is deduced. This can be achieved by special layout algorithms (Bruccoleri and Heinrich, 1988) or force-field like approaches (Wiese et al., 2005). We only want to briefly mention here that there exist many other visualization methods such as mountain plots, dot plots or ordered, rooted trees, each suited to highlight different aspects of RNA secondary structures (Hogeweg and Hesper, 1984; Shapiro, 1988; Fontana et al., 1993).

2.2.1 Secondary structure prediction algorithms

As briefly discussed in Chapter 1, first efforts to predict RNA secondary structures date back to 1978, when Ruth Nussinov and colleagues presented a dynamic programming algorithm for maximizing the number of base-pairs known as the maximum matching problem (Nussinov et al., 1978). Although the non-thermodynamic scoring model used by the algorithm is too simple to predict RNA secondary structures with adequate accuracy, the algorithmic principle applied is fundamental in its nature. Current state-of-the-art approaches still rely on the basic principle of this algorithm introduced more than 30 years ago. In the following we will shortly discuss Nussinov's solution to the maximum matching problem.

Let us assume an RNA sequence x with n nucleotides, then x_i denotes the i^{th} nucleotide in sequence x. As mentioned before when formally defining the concept of RNA secondary structure, only Watson-Crick and Wobble base-pairs are allowed, no pseudo-knots and no base triples are allowed to occur. For sake of simplicity, we skip the minimum distance of

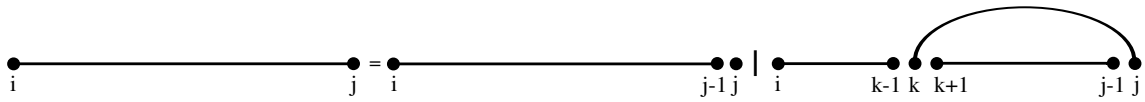


Figure 2.5. ***Decompositions used in the Nussinov algorithm.*** *There are only two ways a structure on sub-sequence $x[i...j]$ can be composed of. Either j is unpaired or j is paired to some nucleotide k splitting $x[i...j]$ into two smaller sub-sequence $x[i...k-1]$ and $x[k+1...j-1]$.*

three nucleotides in this setting. $x[i...j]$ denotes a sub-sequence of x from position i to j and $M_{i,j}$ the maximum number of base-pairs on the sub-sequence $x[i...j]$. The basic idea is that there are only two ways a structure on the sub-sequence $x[i...j]$ can be composed of (Fig. 2.5). Let us assume that we have already computed all the maximum matching scores on the interval $x[i...j-1]$. If we now add the nucleotide x_j one of the two scenarios has to match: i) j is unpaired, or ii) j is paired to some base k with $i \leq k < j$. The second case splits $x[i...j]$ into two smaller sub-sequences $x[i...k-1]$ and $x[k+1...j-1]$. Since we have already computed all matching scores on the interval $x[i...j-1]$, we know the scores of $M_{i,k-1}$ and $M_{k+1,j-1}$ and can now easily calculate the score $M_{i,j}$. The maximum matching on a sub-sequence $x[i...j]$ is hence given by following recursion:

$$M_{i,j} = \max \begin{cases} M_{i,j-1} \\ \max\limits_{\substack{i \leq k \leq j-1 \\ (k,j) \in \mathbb{B}}} M_{i,k-1} + M_{k+1,j-1} + 1 \end{cases} \tag{2.1}$$

While this approach yields the maximum number of base-pairs on a sequence, it does not instantaneously produce the secondary structure with those base-pairs. The list of base-pairs has to be retrieved via *backtracking*. This is simply done by inverting the algorithm using the calculated values of the forward recursion to reconstruct the optimal path (set of base-pairs) that give rise to the maximum matching score $M_{1,n}$. The Nussinov algorithm scales with $\mathcal{O}(n^3)$ in CPU time and $\mathcal{O}(n^2)$ in memory requirements.

Nussinov's solution for the maximum matching problem can easily be extended to use a toy thermodynamic model, where a pseudo-energy E is minimized:

$$E_{i,j} = \min \begin{cases} E_{i,j-1} \\ \min\limits_{\substack{i \leq k \leq j-1 \\ (k,j) \in \mathbb{B}}} E_{i,k-1} + E_{k+1,j-1} + \varepsilon(k,j) \end{cases} \tag{2.2}$$

ε is an energy scoring function yielding e.g. -3 for G · C and C · G pairs, -2 for A · U and U · A pairs, and -1 for G · U and U · G pairs. Current state-of-the-art approaches still use the basic structure of the algorithm discussed above, but apply a more sophisticated, thermodynamic scoring model, the so called loop-based energy model or nearest neighbor model. Any RNA secondary structure can uniquely be decomposed into a set of loops. A position k is called *immediately interior* to a base-pair $i \cdot j$, if $i < k < j$ and there is no other base-pair $p \cdot q$ such that $i < p < k < q < j$. Hence, any loop is uniquely determined by its closing base-pair $i \cdot j$. The exterior loop L_0 refers to all bases not enclosed by a base-pair. Loops are characterized by the number of unpaired bases and the number of base-pairs k, discriminating between *interior* and *closing* base-pairs. The term k-loop denotes a loop composed of $k-1$ interior base-pairs and one closing base-pair. A *hairpin loop* has only a closing base-pair and all bases between the base-pair are unpaired. The degree k is 1 in this case. Loops with a degree k of 2, are called *interior loops* or *internal loops*. *Bulged loops*, or *bulges* for short, and *stacked pairs* are special cases of interior loops. Bulges are asymmetric interior loops where only one side has unpaired bases and stacked pairs are two adjacent base-pairs containing no unpaired bases. Multiple stacked pairs give rise to stems or helical regions. *Multi-loops* are of degree 3 or more. Fig. 2.6 schematically shows all loop types. The k-loop decomposition is the basis for the energy model used by the programs of the *Vienna RNA package*.

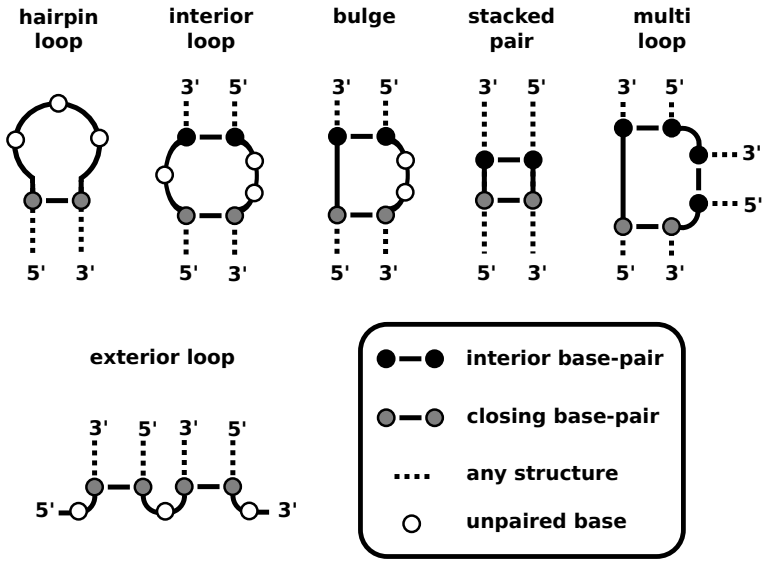
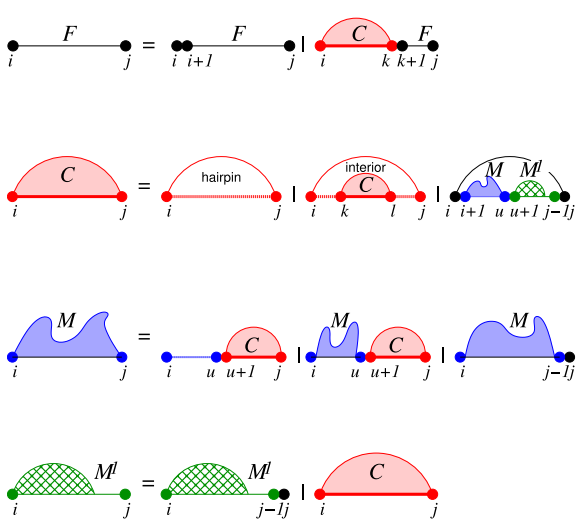


Figure 2.6. ***RNA secondary structure loop types.*** *In a k-loop decomposition a hairpin loop is of degree one. Interior loops, including the special cases of bulged loops and stacked pairs, are of degree two. Multi loops are of degree three or more. The exterior loop collects all bases not enclosed by a base-pair.*

At this point, let us briefly resume the objective of our proposed secondary structure prediction algorithm. Given an RNA sequence x and a set of energy parameters $\mathcal{M}$ we want an algorithm $\mathcal{A}(x, \mathcal{M})$ that returns one or more RNA secondary structures that can be adopted by the sequence x and, let's say, are of biological interest. Based on the assumption that RNA molecules tend to fold into a state of *minimum free energy* (MFE), it is hence a reasonable choice to ask for the minimum free energy structure given the energy parameter set M. It is important to note at this point that whenever we are talking about *free energy* here we actually refer to the *free energy change* ΔG that quantifies the difference in energy between the unfolded and the folded state. A folded RNA has a negative free energy change, and the lower it is the more stable the particular fold. Let ΔG be a function that quantifies the free energy of a structural element. The total free energy of a structure Y is then given by the sum of the individual contributions of all loops composing Y:

$$\begin{aligned} \Delta G(Y) \;=\; & \Delta G(\text{exterior loop}) \;+ \\ & \sum \Delta G(\text{stacked pairs}) \;+ \\ & \sum \Delta G(\text{interior loops}) \;+ \\ & \sum \Delta G(\text{bulged loops}) \;+ \\ & \sum \Delta G(\text{hairpin loops}) \;+ \\ & \sum \Delta G(\text{multi-loops}) \end{aligned} \tag{2.3}$$

Energy parameters can be derived experimentally from RNA oligomer unfolding experiments (Xia et al., 1998; Mathews et al., 1999, 2004) or inferred by statistical methods (Andronescu et al., 2007, 2010). The formation of helical regions (series of stacked pairs) is the dominant stabilizing factor, while all other loop types have, in general, destabilizing contributions. For stacked base-pairs and small hairpin loops one usually uses tabulated values, free energies for other loop types are calculated with models derived from polymer theory. The first efficient algorithm that utilizes such an energy model was proposed by Zuker and Stiegler (1981). It is again a dynamic programming solution, but the different loop types of the nearest neighbor model make it necessary to keep track which structural element yields the lowest free energy at a particular position. Instead of filling one two-dimensional matrix one now has to fill four matrices. Recursions as implemented in the Vienna RNA package (Hofacker et al., 1994) are depicted in Fig. 2.7. The matrix F stores the global free energy, while C, M, and M^1 hold values of particular sub-components. Once the forward recursion is completed, the minimum free energy is found at position $F_{1,n}$ and the corresponding secondary structure is again derived by backtracking. If the number of unpaired bases in an interior loop is restricted by a constant c (e.g., $c = 30$) CPU time requirements are still in $\mathcal{O}(n^3)$.



$$F_{i,j} = \min \begin{cases} F_{i+1,j}, \\ \min\limits_{i<k\leq j} C_{i,k} + F_{k+1,j} \end{cases}$$

$$C_{i,j} = \min \begin{cases} \mathcal{H}(i,j), \\ \min\limits_{i<k<l<j} C_{k,l} + \mathcal{I}(i,j;k,l), \\ \min\limits_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a \end{cases}$$

$$M_{i,j} = \min \begin{cases} \min\limits_{i<u<j} (u-i+1)c + C_{u+1,j} + b, \\ \min\limits_{i<u<j} M_{i,u} + C_{u+1,j} + b, \\ M_{i,j-1} + c \end{cases}$$

$$M^1_{i,j} = \min \begin{cases} M^1_{i,j-1} + c, \\ C_{i,j} + b \end{cases}$$

Figure 2.7. ***Vienna RNA package loop decompositions and recursions.*** *$F_{i,j}$ holds the free energy of the optimal sub-structure on the interval $[i...j]$. $C_{i,j}$ the free energy of the optimal sub-structure given the constraint that i and j form a base-pair. $M_{i,j}$ the free energy of the optimal sub-structure given the constraint that $i..j$ is part of a multi-loop and has at least one component. $M^1_{i,j}$ the free energy of the optimal sub-structure given the constraint that $[i...j]$ is part of a multi-loop and has exactly one component. $\mathcal{H}(k,l)$ denotes the energy contribution of a hairpin loop. $\mathcal{I}(k,l;p,q)$ denotes the energy contribution of an interior loop including stacked pairs and bulges. Energy contributions of multi-loops are calculated by an additive model $\mathcal{M} = a + b + c$, where a is the contribution of the closing pair, $b = b' \times B$ the contribution of helices with B being the number of interior base-pairs, and $c = c' \times \ell$ the contribution of unpaired bases with ℓ being the number of unpaired bases. Drawings are adapted from Bompfünewerer et al. (2008) and Hofacker and Stadler (2007) and recursions are due to Hofacker et al. (1994).*

At room temperature there is usually not a single stable fold Y an RNA molecule is trapped in, rather it is an ensemble of structures $\mathcal{Y}$ that the molecule will adopt. Statistics describing this ensemble are of particular interest. The *partition function* Z is the sum over all Boltzmann weighted structures, formally defined as

$$Z = \sum_{Y \in \mathcal{Y}} \exp(-\frac{1}{RT}\Delta G(Y)) \tag{2.4}$$

where R is the molar gas constant and T the absolute temperature in Kelvin. McCaskill (1990) proposed a dynamic programming algorithm that allows the efficient computation of the partition function of the ensemble of RNA structures. In principle, *min* operators in the MFE recursions are exchanged to sum operators $\sum$, and additions to multiplications. Once the value of the partition function is known, one can easily derive the probability P of a single structure Y

$$P(Y) = \frac{\exp(-\frac{1}{RT}\Delta G(Y))}{Z} \tag{2.5}$$

and subsequently the probability p of a certain base-pair $i \cdot j$

$$p_{i,j} = \sum_{Y \in \mathcal{Y}} P(Y)\ \delta_{i,j}(Y) \tag{2.6}$$

where the function δ is 1 if the particular base-pair $i \cdot j$ is found in Y and 0 otherwise.

2.2.2 Prediction of local minimum free energy structures

Many of the RNAs in today's known repertoire of functional, structured noncoding RNAs are rather short in size. MicroRNA precursors and tRNAs typically have a length below 100, and even the longest house-keeping structured RNAs such as 7SK RNA or RNaseP RNA are below 400 nucleotides. When searching for RNA structures in long genomic sequences, one does not want to globally fold the whole genomic sequence and then pick structural sub-motifs. One is rather interested in efficiently predicting local structural motifs with base-pairs that do not span over a maximal distance L. This is of particular interest, when start and end positions of putative ncRNA genes are not known *a priori*. An efficient solution for this task was presented by Hofacker et al. (2004b) and is implemented in `RNALfold` in the Vienna RNA package. Considering again the Nussinov algorithm at this point, it is sufficient to pose a limit when searching for k paired to j. The recursion when processing a sequence from the 5' end to the 3' end is then given by

$$E_{i,j}^{L} = \min \begin{cases} E_{i,j-1} \\ \min\limits_{\substack{j-L \le k \le j-1 \\ (k,j) \in \mathbb{B}}} E_{i,k-1}^{L} + E_{k+1,j-1}^{L} + \varepsilon(k,j) \end{cases} \tag{2.7}$$

The implementation of the `RNALfold` algorithm in the Vienna RNA package uses the full loop-based energy model. The overall memory consumption scales efficiently with $\mathcal{O}(n + L^2)$, where n is the length of the sequence. The computational complexity is $\mathcal{O}(n\,L^2)$. The output consists of a list of self-contained RNA secondary structures, corresponding to minimum free energies and positions in the sequence.

2.2.3 Consensus structure prediction of aligned sequences

As briefly discussed in Chapter 1, the use of a set of related RNA molecules is a powerful method to predict RNA secondary structures, see e.g. Gutell et al. (2002) and references therein. The basic assumption is that related RNA molecules with identical functions are expected to have identical or related structures. Structural elements will be conserved despite of sequence variation. Such patterns of sequence variation that preserve base-pair interactions, called *covariation*, give additional evidence that a predicted base-pair might indeed be correct. In the early days of RNA secondary structure prediction this was widely employed computational method to infer secondary structures (Woese et al., 1980). `RNAalifold` is a computational approach that combines thermodynamic folding with covariation analysis (Hofacker et al., 2002; Bernhart et al., 2008). `RNAalifold` predicts a consensus secondary structure common to the sequences in the input alignment. "Common" in this case means that a base-pair can be formed by at least 50% of the sequences in the alignment. To include covariation information the standard energy model is modified by introducing a (base-pair) conservation score $\gamma(i,j)$ that is added as a pseudo-energy. Given an alignment $\mathbb{D}$ with n sequences, the consensus energy $\beta_{ij}^{\mathbb{D}}$ of two columns i and j is given by

$$\beta_{ij}^{\mathbb{D}} = \frac{1}{n} \sum_{x \in \mathbb{D}} \varepsilon(x_i, x_j) - \phi_2 \gamma_{i,j}, \tag{2.8}$$

where γ is composed of two parts γ' and γ''. γ' measures the covariance contribution and γ'' adds a penalty if a base-pair cannot be formed by a sequence. γ' is calculated as follows

$$\gamma'(i,j) = \frac{1}{2} \sum_{\substack{x,y \in \mathbb{D} \\ x \neq y}} \begin{cases} \mathfrak{h}(x_i, y_i) + \mathfrak{h}(x_j, y_j) & \text{if } (x_i, x_j) \in \mathcal{B} \text{ and } (y_i, y_j) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

where i and j denote two columns in the alignment, $\mathfrak{h}$ is the Hamming distance, and $\mathcal{B}$ is the set of valid base-pairs. Using this pairwise comparison setting, compensatory and consistent mutations will be awarded 2 and 1, respectively, while total conservation is assigned 0. With γ' we do not capture inconsistent base-pairs. γ'' quantifies these counter-examples to structural conservation by the following score matrix:

$$\gamma''(i,j) = \sum_{x \in \mathbb{D}} \begin{cases} 0 & \text{if } (x_i, x_j) \in \mathcal{B} \\ 0.25 & \text{if } x_i \text{ and } x_j \text{ are gaps} \\ 1 & \text{otherwise} \end{cases} \tag{2.10}$$

Finally, γ is given by $\gamma = \gamma' + \phi_1 \gamma''$. `RNAalifold` uses the two scaling factors ϕ_1 and ϕ_2 to control the influence of the covariance score. ϕ_1 weights the contribution of counter examples, and ϕ_2 controls the overall contribution of the covariance score to the consensus energy. Lindgreen et al. (2006) demonstrated that the `RNAalifold` scoring schema, despite its simplicity, is well suited to quantify covariance. `RNAalifold` is employed by several noncoding RNA gene finders to measure structural conservation (Washietl and Hofacker, 2004; Gesell and Washietl, 2008; Washietl et al., 2005b). In the latest version of `RNAalifold` RIBOSUM matrices replace the Hamming distances in the covariance evaluation. While this modification results in increased prediction accuracy, it does not directly lead to better discrimination capability in terms of the Structure Conservation Index (see Section 2.5; Bernhart et al. (2008)).

2.3 Alignments: sequence vs. sequence/structure based

Comparison of two or more sequences, be they DNA, RNA or protein, is a crucial task in the analysis of biological sequences. The generation of sequence alignments is often the first step to later identify homologous positions or sites and infer an evolutionary history. Since comparative sequence analysis is of such fundamental nature it has been addressed already at the very beginnings of computational biology. Needleman and Wunsch (1970) proposed a solution for the *global alignment problem*, where both sequences are aligned along their entire length and an optimal solution is found. In the *local alignment problem* one aims for an optimal alignment of sub-sequences. An efficient solution to this task was presented by Smith and Waterman (1981). At this point, let us briefly discuss how to generate a global alignment of two sequences and what optimality in this case means. When aligning two sequences we have three basic options, which reflect evolutionary events: i) match, ii) mismatch, and iii) gaps. A *match* is given when two identical characters are compared, a *mismatch* otherwise. *Gaps* are introduced to model biological events such as insertions or deletions. Aligning two gap characters is always forbidden. Given a scoring system that defines appropriate scores for matches, mismatches, and gaps we can ask for an alignment with the highest score when maximizing similarity or with the lowest score when minimizing distances. These are equivalent procedures and will return the same result (Smith et al., 1981). The solution to the global alignment problem by Needleman and Wunsch (1970) is based on dynamic programming, and follows the same principles as the Nussinov algorithm, namely recursively deducing the optimal solution from optimal solutions of smaller parts. Given two sequences $\mathbf{x}$ and $\mathbf{y}$ with a sequence length of $|\mathbf{x}|$ and $|\mathbf{y}|$, respectively, a matrix Q is filled using the recursion below.

$$Q_{i,j} = \max \begin{cases} Q_{i-1,j} + \gamma, \\ Q_{i,j-1} + \gamma, \\ Q_{i-1,j-1} + \sigma(x_i, y_j) \end{cases} \tag{2.11}$$

Where σ is a scoring function for matches and mismatches, and γ is the score for a gap. In this setting the word penalty is often used for mismatches and gaps. $Q_{i,j}$ is the optimal score given that sub-sequence $x[1...i]$ is aligned to sub-sequence $y[1...j]$. Therefore the optimal score of the global alignment of the two sequences x and y is found in matrix Q at position $Q_{|\mathbf{x}|,|\mathbf{y}|}$. The actual sequence alignment is then retrieved via backtracking starting with the last column.

Alignments of multiple sequences can also be computed via dynamic programming. The complexity scales, however, exponentially with the number of sequences. For every-day-use, heuristics are applied to generate multiple sequence alignments. *Progressive alignment* introduced by Feng and Doolittle (1987) is one strategy to do so. Starting with the alignment of the two most closely related sequences the next closest sequence or sequence group is added. This process continues in an iterative manner. Positioning of gaps is adjusted in all sequences at each iteration. `CLUSTAL W` (Thompson et al., 1994; Larkin et al., 2007) is the most widely used approach implementing this strategy.

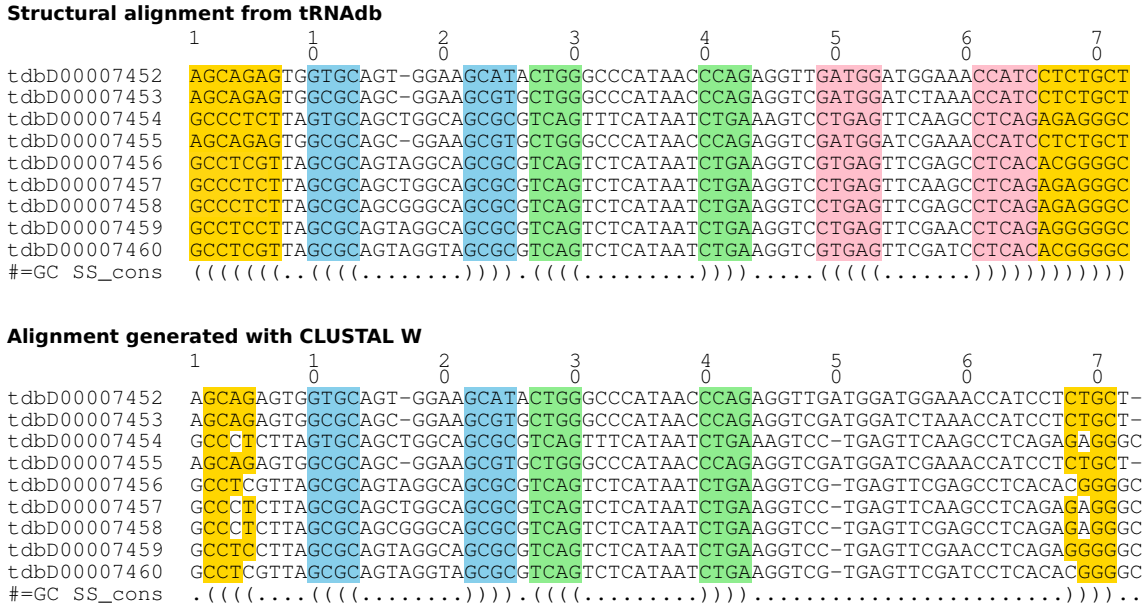


Figure 2.8. ***Structural alignment and CLUSTAL W generated alignment of human tRNA-met sequences.*** *While the D-loop stem and the Anticodon-loop stem are predicted in both cases, the TΨC-loop stem is only present in the structural alignment. Moreover, the closing acceptor stem is also only correctly predicted in the structural alignment. Consensus structures were computed with* `RNAalifold` *(Bernhart et al., 2008). Alignments were visualized using the* `RALEE-mode` *in* `emacs` *(Griffiths-Jones, 2005).*

The application range of purely sequence-based aligners is limited in computational RNA biology, since RNA molecules often evolve fast on sequence level and only retain their structural features. An alignment program that only considers sequence motifs and not shared structural components will hence fail to yield a good alignment in terms of RNA secondary structure. Figure 2.8 illustrates the difference in quality on a structural and a `CLUSTAL W` generated alignment for human tRNA genes. Until position 49 both alignments are identical. The gap character at position 50 introduced by `CLUSTAL W` to better fit the sequence patterns

in the remaining part disassembles the consensus structure. `RNAalifold` is not able to predict the last hairpin and the closing stem correctly then.

Sankoff (1985) proposed a dynamic programming solution for the problem of simultaneously aligning and folding of RNA sequences. Due its high computational cost of $\mathcal{O}(n^6)$, where n is the length of the two sequences to be aligned, it is not suited for practical use. Restricted versions of Sankoff's algorithm are implemented in `foldalign` (Havgaard et al., 2005), `dynalign` (Mathews and Turner, 2002), `PMcomp` (Hofacker et al., 2004a), or `LocARNA` (Will et al., 2007). At this point let us briefly discuss the underlying approach of `PMcomp` and `LocARNA` as reviewed in Bompfünewerer et al. (2008). In principle, the algorithms follow Sankoff's solution, but are split in two separate stages. First, base-pairing probability matrices are calculated using `RNAfold` for each sequence. Pairwise alignments of two sequences $\mathbf{x}$ and $\mathbf{y}$ are then generated by following recursions (Fig. 2.9), where $Q_{i,j,k,l}$ denotes the maximal score of an alignment of sub-sequences $\mathbf{x}[i...j]$ and $\mathbf{y}[k...l]$. The optimal score is found in matrix Q at position $Q_{1,|\mathbf{x}|,1,|\mathbf{y}|}$.

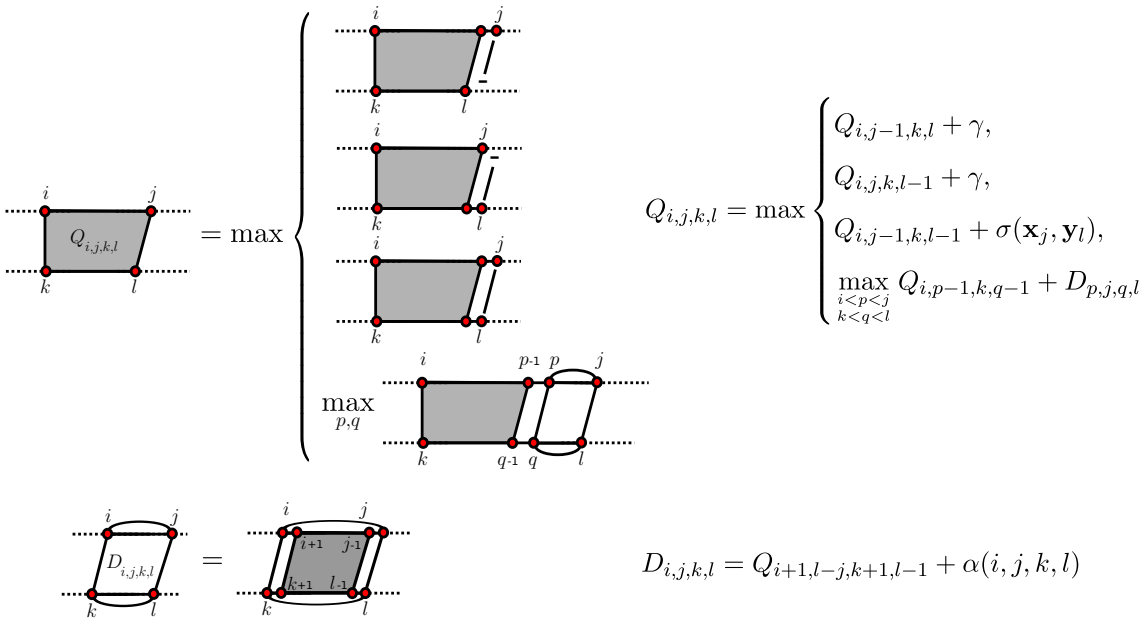


Figure 2.9. ***Recursion scheme used by*** *`LocARNA`. There are four distinct cases for calculating scores in matrix Q and a single decomposition for the entries in D. Figure adapted from Bompfünewerer et al. (2008).*

As seen before for plain sequence alignment (Eq. 2.11) γ and σ are scores or scoring functions for gaps and unpaired (mis)matches, respectively. At $D_{i,j,k,l}$ we find the optimal score of an

alignment of sub-sequences $\mathbf{x}[i...j]$ and $\mathbf{y}[k...l]$ given the condition that the base-pairs $i \cdot j$ and $k \cdot l$ are matched. α is a scoring function for base-pair matches using base-pair scores that are derived from the base-pairing probability matrices of the two individual sequences. `LocARNA` uses only base-pairs that have a base-pairing probability higher than a certain cut-off. This improves the time complexity to $\mathcal{O}(n^4)$ in CPU time, and $\mathcal{O}(n^2)$ in memory requirements. `LocARNA` is used in this thesis for generation of the structural alignment training set for `RNAz` (cf. Chapter 4). Moreover, we applied the `LocARNA-RNAclust` clustering pipeline to group ncRNA candidates based on sequence/structure similarity (cf. Chapter 7).

2.4 Machine learning using support vector machines

Machine learning is a discipline in computer science that deals with the development and design of algorithms to perform tasks commonly associated with artificial intelligence. A formal definition on how to define and evaluate "learning" in terms of computer programs was given by Mitchell (1997):

> *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

Given for example the task T of recognizing handwritten digits, P can be defined as the percentage of correctly classified digits and E as a database of handwritten digits. Actually, recognizing handwritten digits was among the first real-world problems where an efficient solution was found employing machine learning techniques (Matan et al., 1990). Over the years, many concepts and algorithms have emerged (Alpaydin, 2004; Bishop, 2006). Most prominent approaches include neural networks, kernel methods such as the *support vector machines* (SVMs), hidden Markov models, k-nearest neighbor algorithms, Gaussian mixture models, naïve Bayes classifiers, or decision trees. While artificial neural networks dominated the field in the beginning, support vector machines (Cortes and Vapnik, 1995) are currently the tools of the trade in many disciplines. Especially in computational biology, SVMs have been applied to a wide variety of problems, including protein-coding gene detection (Schweikert et al., 2009), protein sub-cellular localization prediction (Shi et al., 2007; Lei and Dai, 2005), protein fold recognition (Sun and Huang, 2006; Shamim et al., 2007), detection of translation initiation sites (Zien et al., 2000), splice site detection (Sonnenburg et al., 2002; Hiller et al., 2009), cancer tissue classification (Chu and Wang, 2005; Chiu et al., 2008), promoter prediction (Sonnenburg et al., 2006; Towsey et al., 2008), or microRNA gene prediction (Xue et al., 2005; Hertel and Stadler, 2006; Ng and Mishra, 2007; Xu et al., 2008; Li et al., 2010). Since SVM techniques are extensively used in this thesis, we briefly outline the theoretical principles of SVMs based on contributions by Bennett and Bredensteiner (2000) and Bennett and Campbell (2000) in the following.

2.4.1 Supervised learning and the conceptual idea of SVMs

In *supervised learning* the machine learning algorithm trains on input-output pairs and learns a decision function to map the input to the output. Depending on whether the output is discrete or continuous, problems are inferred as *classification* and *regression*, respectively.

Support vector machines are prototype examples of supervised learning algorithms and can be applied to both classification and regression problems.

Following the reasoning of Bennett and Campbell (2000), the success of support vector machines is largely attributed to these features: i) the SVM approach is systematic, reproducible, and properly motivated by statistical learning theory, and ii) since training involves optimizing of a convex cost function, the algorithm does not get trapped in local minima and the optimal solution can always be found. Moreover, ready-to-go software implementations like `libSVM` (Chang and Lin, 2001) or `SVMLight` (Joachims, 1999) have made the use of SVMs fairly easy. The conceptual idea behind this powerful machine learning technique is depicted in Fig. 2.10. Given a binary classification problem, a set consisting of data points from two classes that is not linearly separable in the input space is transformed via a function Φ into a higher dimensional feature space, where a linear separation is possible.

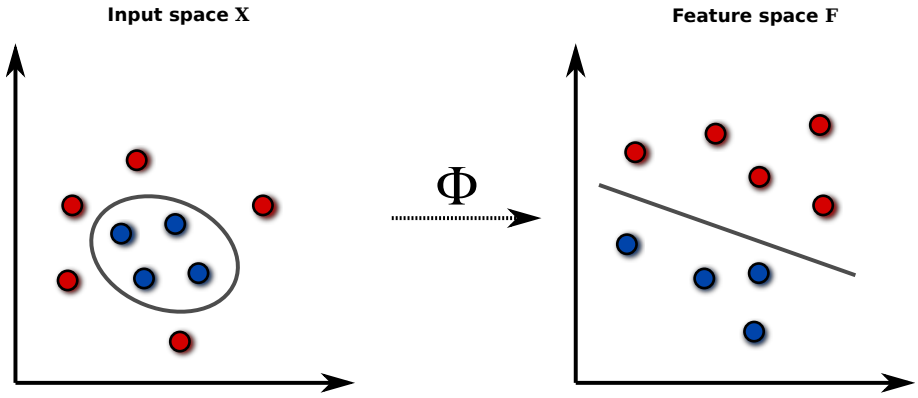


Figure 2.10. ***Conceptual idea of classification with SVMs.*** *Data points from two classes (red and blue) are mapped from the input space X to a higher dimensional feature space F via Φ. A maximum margin separating hyperplane is constructed in F, which yields a non-linear decision boundary in the input space. Figure adapted from Schölkopf and Smola (2002).*

2.4.2 Hyperplane classifiers and the kernel trick

Let us assume a binary classification problem and a set of ℓ training instances $\{\boldsymbol{x}_i, y_i\}$ with $i = 1, ..., \ell$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$. Each training instance is a vector of n features $\boldsymbol{x} = (x_1, ..., x_n)^\mathsf{T}$ and belongs either to the positive or the negative class. Based on the training instances we seek a classification function $f(\boldsymbol{x})$ such that $\boldsymbol{x}$ is assigned to the positive class if

$f(\boldsymbol{x}) \geq 0$, or to the negative class otherwise. In the simple case, where the data set is truly linearly separable, a hyperplane of the form

$$D(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b = \sum_{1=1}^{n} w_i x_i + b \tag{2.12}$$

with $\boldsymbol{w} \in \mathbb{R}^n$ and a bias term $b \in \mathbb{R}$ can be constructed that correctly classifies all instances. However, an infinite number of such hyperplanes exist. Fig. 2.11a shows two possible hyperplanes (dashed and solid lines), and certainly the question arises: which one provides the better classifier? Without any additional information, the hyperplane represented by the solid line seems to be the better choice, since it is likely to generalize better on future data. From a geometric point of view this hyperplane can be described as being "furthest" from both classes. In other words, the optimal hyperplane in terms of generalization ability adopts the maximal distance from any of the two sets.

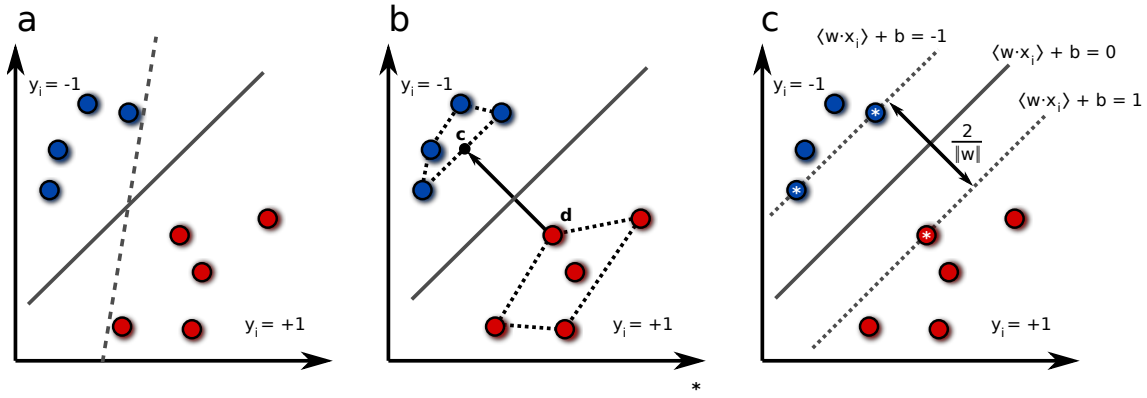


Figure 2.11. ***Construction of an optimal hyperplane in a binary classification problem.*** *(**a**) There exist an infinite number of linearly separating hyperplanes. The hyperplane represented by a solid line is not that sensitive to small perturbations in the training data as the hyperplane drawn with a dashed line. (**b**) Convex hulls of each class are indicated by dashed lines.* c *and* d *mark the two closest points of the two hulls. The optimally separating hyperplane bisects these two closest points. (**c**) Supporting hyperplanes are indicated by dashed lines. The optimally separating hyperplane (solid line) is defined as the hyperplane with the maximal margin with respect to the supporting hyperplanes.*

There are two possible ways to determine the optimally separating hyperplane. The methods are graphically outlined in detail in Fig. 2.11b and Fig. 2.11c. In the convex hull approach, a convex hull (convex set that contains all data points) is generated for each class. Next the two closest points of each hull c and d are found, which can be done efficiently via a quadratic

minimization problem (Bennett and Campbell, 2000). The optimal hyperplane bisects these two closest points. An alternative approach is to maximize the margin between two parallel supporting hyperplanes (Fig. 2.11c dashed lines). A hyperplane is called supporting if all points of a class are on one side. $\boldsymbol{w}$ and b are rescaled such that the points closest to the hyperplane satisfy $D(\boldsymbol{x}) = y(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) = 1$ and thus $D(x_i) = y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1$ is valid for all points $\boldsymbol{x}_i$ with $i = 1, ..., \ell$. When $D(\boldsymbol{x}) = y(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) = 0$ the separating hyperplane is in the middle of the two supporting hyperplanes. The distance from the separating hyperplane to the nearest training point is called *margin*. The optimal separating hyperplane with the maximal margin is obtained by following constrained quadratic minimization problem for $\boldsymbol{w}$ and b:

$$\begin{aligned} &\text{minimize } Q(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to } \; y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) - 1 \geq 0, \qquad \forall i = 1, ..., \ell \end{aligned} \tag{2.13}$$

Data points that satisfy the equality constraint are called *support vectors*. In Fig. 2.11c, data points marked with an asterisk indicate support vectors. The convex optimization problem of minimizing $Q(\boldsymbol{w}, b)$ could in theory be solved by quadratic programming techniques, but today's efficient SVM implementations rely on conversion into the equivalent *dual* problem, where the number of variables is the number of training data:

$$\begin{aligned} &\text{maximize } Q(\alpha) = \sum_{i=1}^{\ell} \alpha_i \; - \; \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \, \alpha_j \, y_i \, y_j \, \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j \\ &\text{subject to } \sum_{i=1}^{\ell} y_i \alpha_i = 0 \qquad \alpha_i \geq 0 \qquad \forall i = 1, ..., \ell \end{aligned} \tag{2.14}$$

Where α_i are non-negative Lagrange multipliers, arising from conversion into the dual problem. So far, we have considered a perfectly separable data set and deduced a classifier, which is commonly referred as *hard margin support vector machine*. In real-world applications the scenario of perfectly separable classes is rarely seen and a hard margin SVM will fail in the case of inseparable data. However, the concept of *soft margin support vector machines* allows errors to be made during the training process (Cortes and Vapnik, 1995). The strict constraints in Eq. 2.13 are relaxed and a margin parameter C is then used to control the trade-off between maximization of the margin and minimization of the classification error. If the data is not linearly separable a soft margin SVM will succeed in generating a classifier. Although the hyperplane has been optimally determined, the classifier may suffer from poor generalization ability. To handle non linearly separable data, SVM techniques generally apply a mapping of data from the *input space* to a higher dimensional *feature space* using a

nonlinear vector function Φ (cf. Fig. 2.10). The separating hyperplane is then given by

$$D(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) \rangle + b \tag{2.15}$$

Instead of using $\Phi(\boldsymbol{x})$ a *kernel* function $H(\boldsymbol{x}^\mathsf{T}, \boldsymbol{x})$ is used since this elegantly avoids treating the high dimensional feature space explicitly. This technique is commonly referred to as *kernel trick* (Aizerman et al., 1964). A kernel function must be continuous, symmetric, and most preferably it should have a positive (semi-)definite Gram matrix, since this guarantees that the optimization problem will be convex and the solution unique. Among the most widely used kernel functions is the radial basis function (RBF) kernel

$$H(\boldsymbol{x}^\mathsf{T}, \boldsymbol{x}) = \exp -\gamma \|\boldsymbol{x}^\mathsf{T} - \boldsymbol{x}\| \tag{2.16}$$

where γ is a positive parameter controlling the radius.

2.4.3 Support vector regression

The idea of finding a separating hyperplane in a binary classification task can be easily generalized to regression analysis. Training data is still in the form of input-output pairs $\{\boldsymbol{x}, y\}$ while now $y \in \mathbb{R}$. In linear regression usually a squared error function $E(r)$ based on the residual r is used. Such a quadratic error function is not an optimal choice in the case of support vector regression since it will generate no sparseness in the support vectors (Gunn, 1998). Instead a piecewise, linear function of the form

$$E(r) = \begin{cases} 0 & if\, |r| \leq \epsilon \\ |r| - \epsilon & otherwise \end{cases} \tag{2.17}$$

is used where ϵ is a small positive value. This simply means that in the evaluation of the error function errors that are smaller than ϵ are ignored.

2.5 Structured noncoding RNA detection with RNAz

RNAz (Washietl et al., 2005b) is a noncoding RNA gene finder that relies on signatures of thermodynamic stability and evolutionary conservation of RNA secondary structures. RNAz ncRNA screens have been conducted in several organisms including mammals (Washietl et al., 2005a, 2007; Seemann et al., 2007), fish (Rose et al., 2008), nematodes (Missal et al., 2006), arthropods (Rose et al., 2007), yeast (Washietl et al., 2005b; Steigele et al., 2007), plants (Song et al., 2009), bacteria (del Val et al., 2007; Sonnleitner et al., 2008; Pánek et al., 2008) or even metagenomics data (Shi et al., 2009). The RNAz algorithm or parts of it have also been used in a series of other programs (Song and Deng, 2010; Xu et al., 2009; Hertel and Stadler, 2006; Hertel et al., 2008; Reiche and Stadler, 2007). Besides the core program the RNAz package comes with a series of helper scripts that allow the user to easily set up a computational pipeline for genome-wide screens. There is also a RNAz web server that allows to easily conduct RNAz ncRNA screens online (Gruber et al., 2007). A schematic overview of the RNAz algorithm is given in Fig. 2.12.

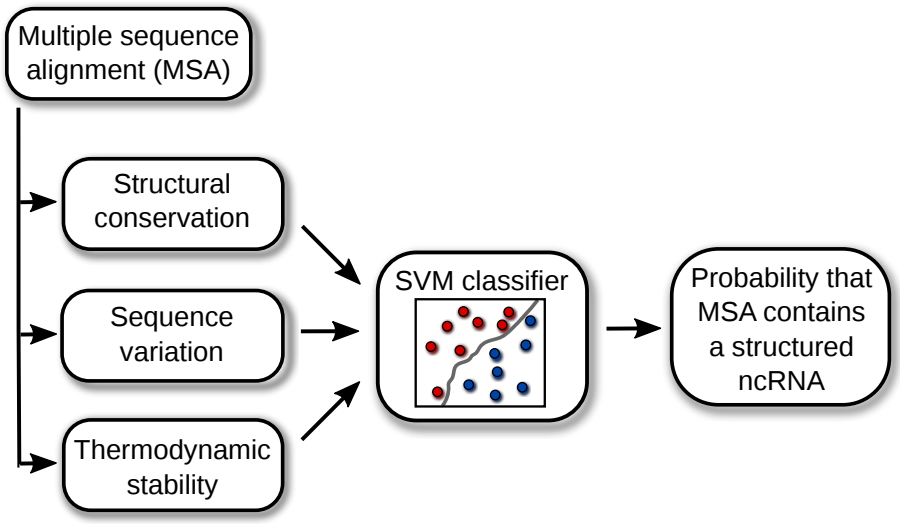


Figure 2.12. ***Overview of the RNAz algorithm.*** *A multiple sequence alignment serves as input. Structural conservation, sequence variation, and thermodynamic stability are measured and serve as input values for a SVM classifier. The output is the probability that the alignment shows signatures of evolutionary conserved, thermodynamic RNA structures.*

A multiple-sequence alignment serves as input and three properties, namely *thermodynamic stability*, *sequence variation*, and *structural conservation* are measured. Subsequently, these features serve as input for a binary SVM classifier that outputs the probability that the multi-

ple sequence alignment shows signatures of thermodynamically stable, evolutionary conserved RNA secondary structures. First, for each sequence in the alignment the thermodynamic stability is calculated in terms of a z-score of the form $z = (E - \mu)/\sigma$. E denotes the energy of the minimum free energy structure of the native sequence and μ and σ are the average energy and standard deviation of the energies of a set of shuffled sequences with the same base composition. The conventional approach of generating the set of shuffled sequences by explicitly shuffling and folding is far too expensive to apply it on a genome-wide scale in a reasonable amount of time. Instead, support vector regression models have been trained to estimate μ and σ, which can be done at a fraction of the time needed in the explicit approach. Finally, the averaged z-score of all sequences in the alignment is calculated.

For training of the regression models, Washietl and colleagues generated training sequences by sampling the sequence space with a regularly space gird. In particular, the G+C content, the A/(A+T) ratio, and the C/(C+G) ratio were all varied from 0.25 to 0.75 in steps of 0.05. The length of the sequences was varied from 50 to 400 in steps of 50 nt. In total a set of 10,648 sequences was used. For each of these sequences 1,000 randomized sequences were generated using the Fisher-Yates shuffling algorithm and subsequently folded with RNAfold.

The average pairwise sequence identity and the number of sequences in the alignment serve as input features describing the sequence variation. Measuring sequence variation is necessary in order to somehow normalize the structure conservation index (SCI), which is used to measure the structural conservation of the sequences. This is motivated by a series of reasonable assumptions that with increasing sequence variation standard alignment algorithms fail to yield good alignments in terms of correctly aligned RNA secondary structures (Gardner et al., 2005). Hence with more sequence variation, we are expecting lower SCI values. "Somehow" in this context relates to the fact that normalization is not done explicitly via a known model, but implicitly by the black box SVM classifier. The SCI is formally defined as $SCI = E_{consensus}/\langle E_{single} \rangle$, where $E_{consensus}$ is the energy of the consensus structure predicted by RNAalifold (Hofacker et al., 2002; Bernhart et al., 2008) and $\langle E_{single} \rangle$ is the average of the energies of the single sequences. The SCI is an refinement of the concept first introduced by Washietl and Hofacker (2004). There the RNAalifold consensus energy was interpreted by calculating a z-score on consensus energies of randomized alignments. In contrast, the consensus energy in the SCI is normalized by the average of the single sequence folding energies. If all sequences in the alignment are able to fold into the consensus structure, the consensus energy will be close to the average of single sequence folding energies, hence yielding a SCI close to 1. Due to the rewarding schema for consistent and compensatory mutations in terms of bonus energies implemented in RNAalifold, the SCI can even have values higher

than 1. Note, that structural conservation is measured in terms of energies and not by taking the secondary structures into account. In a subsequent work (Gruber et al., 2008a) to `RNAz` we showed that the SCI is the overall best measure to quantify structural conservation. The concept of measuring structural conservation in terms of energies rather than the conservation of single base-pairs is not easily intuitive, but becomes more comprehensible when considering the following. RNA secondary structure predictions are far from yielding perfect predictions (Gardner and Giegerich, 2004). Estimates of the prediction accuracy depending on the used data set and metric are roughly speaking between 45-70% (Doshi et al., 2004). Taking for example 100,751 tRNA sequences present in Rfam (Gardner et al., 2009) and folding them with `RNAfold`, only 85% will adopt the typical tRNA shape (`[[][][]]`) as their predicted minimum free energy structure. The moderate overall prediction accuracy is only one problem. Another one is the presence of degenerate folding states. In the forward recursion of any dynamic programming RNA folding algorithm the minimum free energy is computed first, and then a secondary structure with that energy is determined by back-tracking. There can, however, be multiple secondary structure with the same folding free energy, as shown below:

```
UCGUUCCUGGCCGCCGGACUGAAAGUGAGCGUAGAACUCCGAUGGGGGUCUUGAAGCAACUACCUUUGUGAUUCUUCUUG
.(((((((((...))))((.....)))))))..((.(((((....)))))))..((((.((.(((....))).))))))... -14.60
.(((((((((...))))((.....)))))))..(((.((((....)))))))..((((.((.(((....))).))))))... -14.60
(((.(((.((...))))).))).(((..((..((.(((((....)))))))....)).)))..................... -14.60
(((.(((.((...))))).))).(((..((..(((.((((....)))))))....)).)))..................... -14.60
```

The minimum free energy of the sequence given above as predicted with `RNAfold` is -14.60. There is not a single structure associated with that energy, rather it is a set of four structures with quite different structural elements. If one aims to compute the structural conservation of this sequence with some other sequences and uses some kind of base-pair based metric, the output is heavily influenced by which structure was actually backtracked by the RNA folding program. The use of folding free energies to calculate structural conservation elegantly circumvents such pitfalls and has been shown to be a powerful strategy (Gruber et al., 2008a; Okada et al., 2010).

2.6 Computational tools for ncRNA homology search

2.6.1 Sequence based tools: `BLAST`

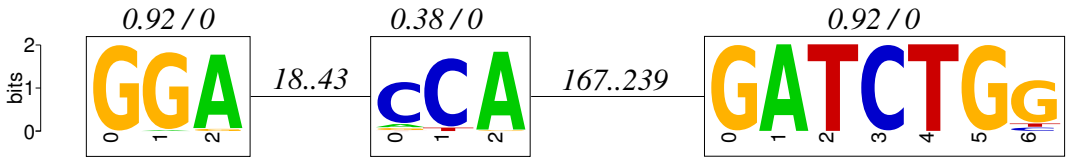
The running time of standard alignment algorithms makes it impractical to apply these algorithms for screening and searching of large collections of nucleotide or protein sequences. Instead, common day sequence analysis for finding a query sequence in a large data set relies on the use of heuristic methods. The `BLAST` software package (Altschul et al., 1990, 1997) and web resources (http://blast.ncbi.nlm.nih.gov/Blast.cgi) are the most commonly used tools for this purpose. `BLAST` is an acronym for Basic Local Alignment Search Tool. The dramatic speedup of `BLAST` compared to standard alignment algorithms is achieved by a reduction of the search space. Basically, the `BLAST` algorithm consists of two components: i) heuristic search, and ii) statistical evaluation of the computed solutions. In a first step, the sequence is split into a set of *words*. Words are sub-sequences of a defined length, typically 11 for nucleotide and 3 for protein sequences. Words that score above a certain threshold when aligned to the query sequence are selected as *seed* sequences and the database is subsequently screened with the seed sequences, i.e. determining all locations of all common words of the query and the target sequence. In a next step, hits are grouped and the algorithm tries to expand the ends by adding further alignment columns forming so called *high scoring segment pairs* (HSPs). At least two hits have to be found, that can be grouped, otherwise the search is unsuccessful. Finally, `BLAST` computes for each HSP a bit score and an E-value, which measures the number of alignments with an equal or better score that are estimated to occur by chance.

A `BLAST` search is usually the first step when searching for homologous RNA sequences. The web interface at NCBI allows easy and effortless searching of millions of nucleotide sequences by a simple copy-and-paste of the query sequence and hitting the run button. Although changing of the standard `BLAST` parameters may improve results, this search strategy is deemed to fail to retrieve homologs in distantly related species since RNA genes are known to show only weak sequence conservation. Nevertheless, `BLAST` can be used to collect an initial pool of homologous sequences.

2.6.2 Sequence based tools: `fragrep`

Because of the moderate sequence conservation of many ncRNA families or large insertions and deletions `BLAST` fails to recover distantly related RNA sequences in many cases. This does not necessarily mean that sequence based search methods are no use for noncoding

RNA detection. It is quite common that the researcher has identified sequence stretches based on conservation analysis or wet-lab experiments that seem to be more important than others for the biological function of the RNA molecule. Since `BLAST` is a fully automated software package, there is no easy way to incorporate this kind of expert knowledge into the algorithm. A software package that allows to construct user defined sequence queries is `fragrep` (Mosig et al., 2006, 2007a). In detail, it is a dynamic programming algorithm that allows to search for fragmented sequence patterns in long genomic sequences. Scanning speed depends on the complexity of the query patterns and the length of the genome, but run-times are usually in the range of a few minutes. `fragrep` has been successfully applied in a series of noncoding RNA detection approaches, e.g. vault RNAs (Stadler et al., 2009), 7SK snRNA (Gruber et al., 2008b), telomerase RNAs (Mosig et al., 2007a), and Y RNAs (Mosig et al., 2007b).




```
3 matrices
    0     0 M0:GGA 0.92 0
   18    43 M1:CCA 0.38 0
  167   239 M2:GATCTGG 0.92 0

M0:GGA              M1:CCA              M2:GATCTGG
   0    1   48        4    0   50        0   51    0    0    0    0    0
   0    0    0       43   48    0        0    0    0   51    0    0    4
  50   49    2        3    0    1       51    0    0    0    0   51   43
   0    0    0        1    3    0        0    0   51    0   51    0    4
#  G    G    A        C    C    A        G    A    T    C    T    G    G
```

Figure 2.13. ***Outline of `fragrep` query patterns.*** *The upper panel shows the search patterns as a combination of sequence logos. Numbers between the logos indicate minimal and maximal distances. The lower panel shows the corresponding `fragrep` input pattern consisting of header information and position weight matrices for each sub-sequence.*

While in its first version (Mosig et al., 2006) `fragrep` only allowed to define a query as a combination of sequence strings in IUPAC nomenclature, the newest version (Mosig et al., 2007a) uses position weight matrices derived from a multiple alignment of the sequences of interest. Fig. 2.13 shows a typical `fragrep` input pattern. Usually, the pattern is generated from user annotation of a multiple alignment, but the user is free to modify the generated pattern to his or her needs. Especially, the adjustment of minimal and maximal distances between two sequence patterns can help to recover distantly related homologous sequences.

2.6.3 Model based tools: RNABOB

Although there is no official publication describing the algorithm used in RNABOB (Eddy, 1996), RNABOB is a widely used tool to search for RNA structure and sequence motifs (Riccitelli and Lupták, 2010). It falls into the group of descriptor based search algorithms. The pioneer program in this field was RNAMOT (Gautheret et al., 1990), which serves also as basis for RNABOB. Recent contributions are RNAmotif (Macke et al., 2001) and RNAMST (Chang et al., 2006). Common to all these methods is the use of descriptors, which are basically a set of sequence or structure patterns describing the RNA molecule of interest. The easy, yet powerful descriptor syntax of RNABOB has for sure contributed to its wide acceptance. The upper panel of Fig. 2.14 depicts the consensus secondary structure of a toy RNA molecule and one potential RNABOB descriptor.

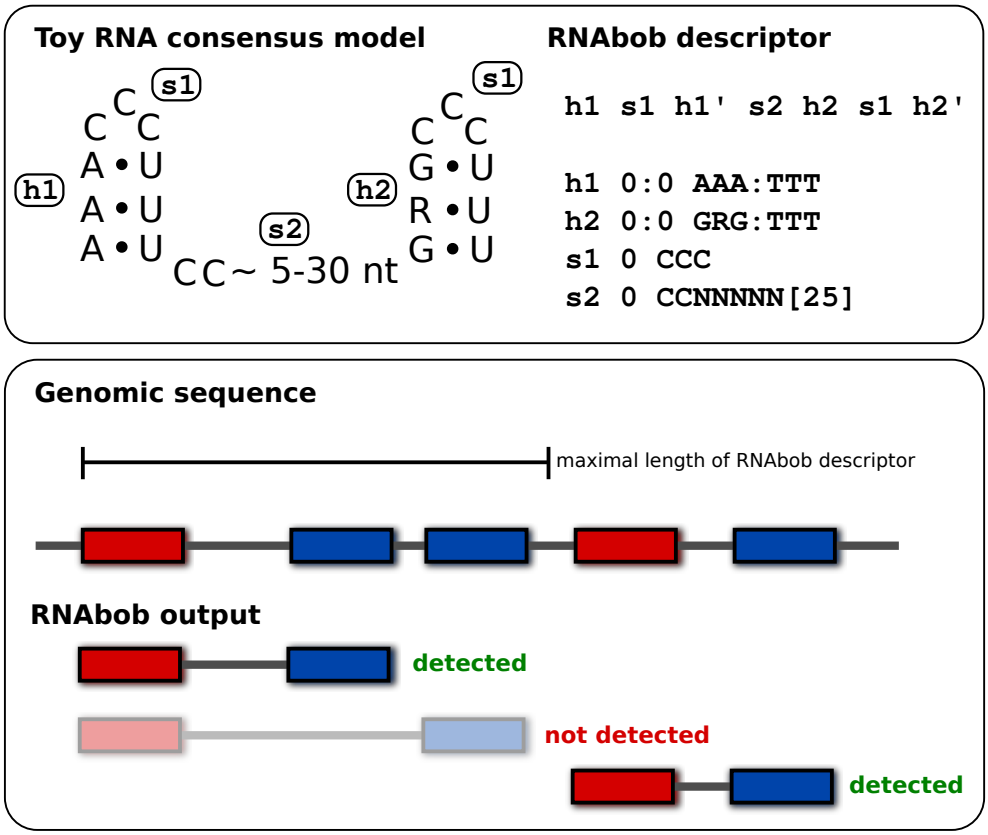


Figure 2.14. *RNABOB descriptors and pitfalls. The upper panel shows the secondary structure of toy RNA consensus model and one potential RNABOB descriptor. The lower panel depicts pitfalls in genome-wide search with RNABOB. Not all putative hits are detected, although they are in the maximal length span defined in the descriptor.*

"h" is used to describe helical regions, while "s" denotes single-stranded regions. The user also has to define the number of allowed sequence mutations and a sequence pattern for the region. The sequence pattern "CCNNNNN[25]", for example, resembles a sequence that starts with two C residues, followed by a minimum of five residues up to a maximum of 30 residues. When such length-variable regions are used one has to carefully evaluate the RNABOB output. One has to keep in mind that the more loosely defined a sequence/structure pattern is the more likely it will produce a match in the target sequence. RNABOB reports only the first matching hit, although there might be another (better) one within the maximal sequence spans defined in the descriptor. An *ad hoc* solution to this problem is to build multiple descriptor files and gradually increase the number of residues. Due to combinatorial explosion this becomes, however, intractable when multiple variable regions are defined.

3

This Thesis

In this thesis we address several computational strategies for noncoding RNA detection ranging from *de novo* detection to homology based methods. In the following, we will give a brief introduction into each topic. Moreover, we will discuss scientific works and developments that motivated and guided the studies in this thesis.

With the `RNAz` gene finding software package Washietl and colleagues have developed the currently most widely used, general noncoding RNA gene finder. One part of this thesis deals with the technical improvement of the `RNAz` algorithm. We have discussed the `RNAz` algorithm in great detail in Chapter 2, but let us at this point briefly summarize the approach. `RNAz` uses a machine learning approach to identify functional RNA secondary structures based on the z-score as a measure of thermodynamic stability and the SCI as a measure of structural conservation. There has been a serious debate whether the z-score can serve as a good discriminator of functional RNA secondary structures against randomized decoys at all (Le et al., 1990b,a; Seffens and Digby, 1999; Workman and Krogh, 1999; Rivas and Eddy, 2000; Clote et al., 2005; Freyhult et al., 2005). The current point of view is that there are RNA classes that do indeed show signatures of thermodynamic stability, and this effect is even more pronounced when a dinucleotide background model is considered. As discussed in Chapter 2, state-of-the-art RNA folding algorithms use a so called nearest neighbor energy model, i.e. a model where stacking energies of base-pairs are the main stabilizing force. These stacking energies are dependent on the immediate neighboring base-pairs, and hence the dinucleotide composition of a sequence is of great influence. Washietl et al. (2007) showed that this is an important factor one has to consider when applying `RNAz` on a genome-wide scale. Authors of this study reported that the distribution of z-scores for screened sequences was not centered around zero, but shifted slightly to -0.5. One could argue that there might be evolutionary

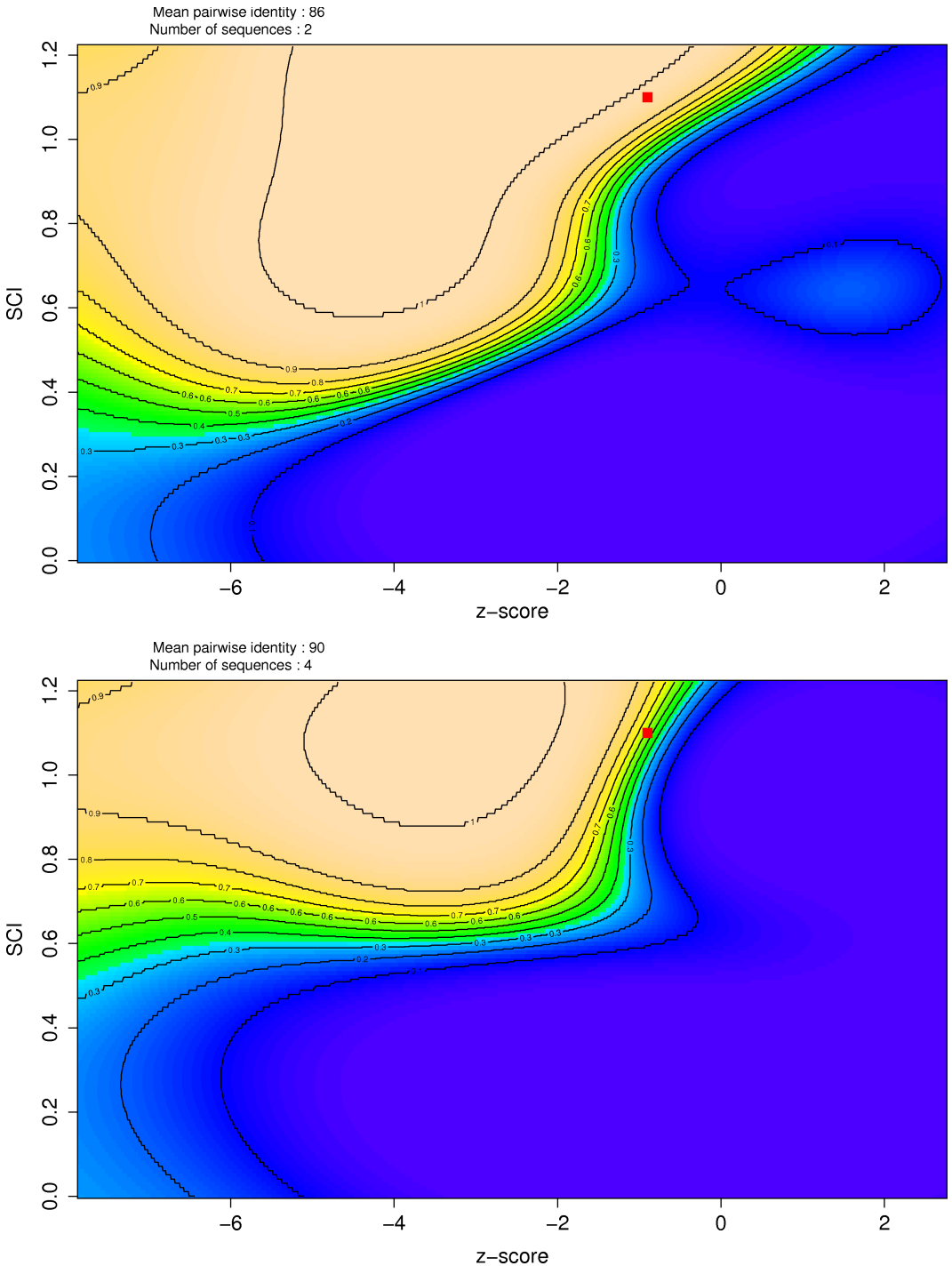


Figure 3.1. ***Visualizations of probability landscapes of the RNAz 1.0 classification SVM.*** *Probability landscapes are shown depending on the SCI and z-score given a fixed sequence variation (MPI and number of sequences). The upper panel depicts a two-way alignment with a MPI of 86%, while the lower panel depicts a four-way alignment with a MPI of 90%. The red square indicates a sample alignment.*

pressure towards increased thermodynamic stability. The fact that this shift vanishes when dinucleotide preserving shuffling is used suggests that this is more likely to be an artifact than of true biological significance. In its initial release `RNAz` used a support vector regression to estimate the z-score of a sequence only trained on mononucleotide shuffled sequences. The reason that a dinucleotide background model for regression (and also classification) was not considered right from the beginning is twofold. First, data generation and training of SVMs for the z-score regression considering dinucleotides is not as straightforward as it is for mononucleotides. For the estimation of the z-score by SVM regression, Washietl and colleagues generated approximately 10,000 synthetic sequences as training instances in a four-dimensional regularly spaced grid. There is, however, no simple and intuitive way of how to uniformly sample sequences from the 16-dimensional dinucleotide space, a space with higher order dependencies. Because of these higher order dependencies (dinucleotide composition controls mononucleotide composition), the grid-like approach for generating sequences is not applicable and even when using an Order-1 Markov model to simulate sequences one is in the need to be able to uniformly sample dinucleotide probability matrices. The second reason was that there was no method available to generate negative instance alignments given a dinucleotide background model. With the works of Anandam et al. (2009) and Gesell and Washietl (2008) two solutions were presented to overcome this lack of an appropriate negative instance set of alignments. Given an input alignment both approaches generate approximately dinucleotide preserving, randomized alignments, either by shuffling or simulation. The many studies conducted with `RNAz` and resulting user feed-back have also identified some other shortcomings of the original approach. Sequence variation in the input alignment is measured by two parameters, namely, number of sequences and mean pairwise identity (MPI). Based on the training data used, this implies an upper limit of six sequences in an alignment that can be processed. There is, however, a more severe implication of using the number of sequences as an input parameter to the classification SVM. The mixed use of a discrete and continuous parameters poses, in general, problems in many machine learning algorithms, and especially in the setting of `RNAz`. Let us assume we have an alignment with two sequences with a MPI of 86%, a SCI of 1.1 and a z-score of -1, as illustrated in the upper panel of Fig. 3.1. It will be assigned a probability of being a ncRNA close to 1 by the `RNAz` SVM classifier. If we generate a four-way alignment by simply copying and pasting the sequences in the alignment into the alignment, we obtain a four-way alignment with a MPI of 90%. SCI and z-score will remain unchanged. This new alignment will be assigned a much lower ncRNA-class probability around 0.6%. Although we did not add new information to the alignment, classification results differ. This effect can be attributed to two factors. On the one hand, it is the `RNAz` training data, which obviously shaped the probability landscapes of the SVM

classifier differently for pairwise and four-way alignments. On the other hand, measuring sequence variation with a combination of two parameters is not an optimal choice. A simple doubling of sequences, which does not add any additional information that has not already been present in the alignment, leads to a change in the parameters. In a previous study (Gruber et al., 2008a), we have shown that the Shannon entropy can be used as a qualified measure to capture sequence variation in the context of structural conservation evaluation. In the same study we also showed that the use of structural alignments can significantly boost the discriminative power of truly conserved structures against randomized decoys. An obvious conclusion is that a structural alignment program should be used instead of a plain sequence-only based aligner. As `RNAz` is only trained on `CLUSTAL W` generated alignments, the use of structurally aligned sequences will result in higher SCI values, mainly contributed by compensatory mutations. Overall higher SCI values compared to levels of sequences based alignments will subsequently lead to an increased amount of false positives. In order to use structural alignments the `RNAz` classification engine also needs to be trained on structural alignments. `RNAz` is a highly successful noncoding RNA gene finder, but taken together, there are still some shortcomings. In Chapter 4 we describe an improved implementation of the `RNAz` algorithm posing solutions to the above described issues.

`RNAz` uses comparative genomics data to infer putative noncoding RNAs. There are, however, many scenarios where the use of comparative data is applicable only in a limited way. Concerning *de novo* detection of functional RNA secondary structures the set of available methods is very sparse in these cases. There are specialized methods to screen for ncRNA in AT-rich genomes, but generally applicable methods other than simple sequence-based statistics to detect at least a set of putatively biologically interesting regions are not available. The aim to obtain such a set of putative biologically interesting regions is of particular interest, as intersection with transcriptomics sequencing data or promoter regions can yield novel insights. `RNALfold` is a general approach for predicting locally stable, self-contained RNA secondary structures in long genomic sequences (Hofacker et al., 2004b). A graphical summary of the `RNALfold` output for an *E. coli* tRNA cluster is shown in Fig. 3.2. In this experiment, `RNALfold` was called with a maximum base-pair span of 120 nt. For each of the seven tRNAs a predicted structure is found in the `RNALfold` output that almost perfectly matches the sequence boundaries of the corresponding tRNA. Since `RNALfold` predicts all locally stable structures, the whole tRNA structure and also structural sub-elements are found in the output. The huge amount of predicted structures limits the usefulness of the `RNALfold` approach. For the complete *E. coli* genome a total of 1,387,136 structures is obtained, which approximately corresponds to a new structure predicted every third nucleotide. Hence, an efficient filtering strategy is needed to reduce the set of candidate sequences to a reasonable

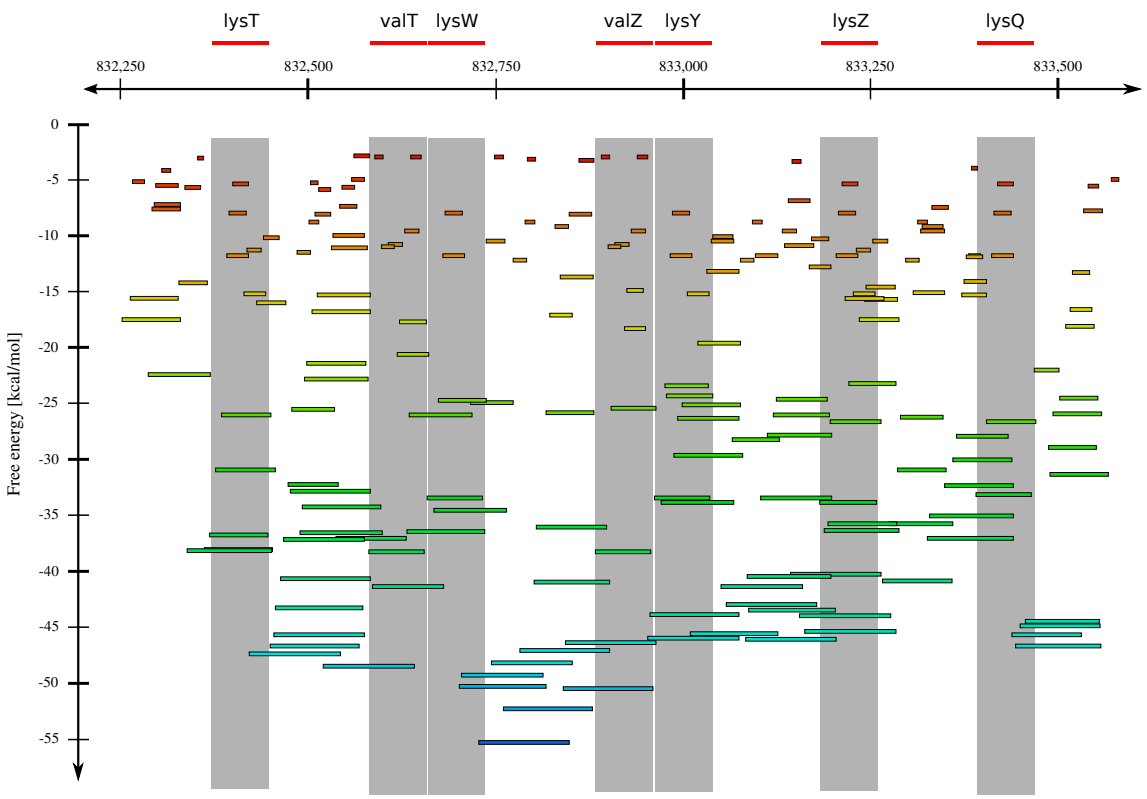


Figure 3.2. ***Visualization of the `RNALfold` output for a sequence region in the E. coli genome.*** *The region contains a cluster of seven tRNA genes indicated by red bars and gray boxes. Structures are colored by their minimum free energy.*

amount. To ask for a function that takes an RNA sequence and/or structure as input and returns a value of biological significance is out of reach. A feature or value that we can ask for is the thermodynamical stability as previously employed by `RNAz`. In Chapter 5 of this thesis, we report on the successful approach of integrating a z-score based filtering strategy into the `RNALfold` algorithm and demonstrate the applicability of the approach to obtain a set of putatively biologically interesting regions in the setting of a genome-wide screen.

Two studies in this thesis are devoted to homology search for new members of an already known RNA family. In particular, we report on the successful detection of new members of the 7SK snRNA family (Zieve and Penman, 1976; Wassarman and Steitz, 1991) and in a second work we characterize in detail a putatively novel ncRNA family named sbRNAs (Deng et al., 2006). As pointed out in a recent review by Menzel et al. (2009), RNA homology search is a non-trivial task that requires expert knowledge and manual curating of steps at multiple stages. Despite recent advances in this field, there is no plug-and-play software package that

allows effortless detection of new family members as one might be used to when doing protein homology search.

7SK snRNA is a highly abundant noncoding RNA in human cells (Zieve and Penman, 1976) that regulates the RNA polymerase II transcription elongation process (Yang et al., 2001; Nguyen et al., 2001). In detail, 7SK snRNA and HEXIM proteins associate in a reversible process with the elongation factor P-TEFb (Li et al., 2005), which leads to inhibition of P-TEFb activity. By this mechanism, gene expression is controlled by regulating the fraction of RNA polymerase II molecules that generate full-length mRNAs (Price, 2000). 7SK snRNA is highly conserved in mammals with only a few known homologs in lower vertebrates (Gürsoy et al., 2000). In human 7SK snRNA has been shown to be transcribed by polymerase III using an external promoter (Murphy et al., 1987). In a previous study we successfully identified 7SK homologs in lower deuterostomia and lophotrochozoans, but failed to recover any plausible candidate in ecdysozoans (Gruber et al., 2008b). We used a combination of `fragrep` search patterns focusing on major structural and functional elements to identify candidate sequences. Manual evaluation of structure conservation and comparison of upstream promoter elements to other snRNAs were then used to detect homologous sequences. We observed that only two structural elements, namely the 5' and the 3' stem structures, are conserved. The remaining part of the molecule shows both low sequence and low structural conservation and large indel regions. The lack of identification of putatively homologous sequences in arthropods or nematodes can, in principle, be of two reasons. First, a 7SK/HEXIM control mechanism does not exist in these species, or second, 7SK molecules in these species are structurally so diverged that they cannot be detected with the methods used so far. The P-TEFb complex has been originally identified in Drosophila (Marshall and Price, 1995), and due to its importance in cell cycle control homologs of P-TEFb are easily identified in all eukaryotes. However, no components of the 7SK/HEXIM snRNP have been experimentally identified in Drosophila so far. Hanyu-Nakamura et al. (2008) reported on an alternative regulatory pathway of P-TEFb by the protein *Pgc* in Drosophila primordial germ cells. Authors of this study argue that such a mechanism may also apply to somatic cells, and that this explains the apparent lack of 7SK in insects. On the other hand, our previous study on 7SK homologs showed that 7SK RNA evolves rapidly with only few well conserved elements. It seemed therefore likely that a homologous gene in insects might have been missed. In Chapter 6 we present a study that successfully identified 7SK RNA homologs in arthropod species. Since a purely structure based homology search was not successful in the previous contribution, we employed a computational promoter screen to identify putative ncRNAs that are transcribed by polymerase III. Subsequent structural characterization of candidates and refined search patterns then identified 7SK RNA homologs in a broad range of arthropod species.

There has been a large number of noncoding RNA screens in various species, both experimentally and/or computationally. While in many studies a large number of putative novel ncRNAs is reported, a functional characterization is often missing. Even the assignment to already known RNA families is a non-trivial task. Homology search is an appropriate strategy towards a functional characterization of molecules, because it helps to identify conserved elements which are in turn important for the function of the molecule. Deng et al. (2006) report on a genome-wide identification of noncoding RNA transcripts in *C. elegans*. Among new snoRNAs and microRNAs Deng and colleagues described in their work a putative novel RNA family with at least seven members in *C. elegans* and a few homologs in *C. briggsae*. Based on the predicted secondary structure, which consists of a stem structure interspersed by a small bulge, the family was termed sbRNAs. In a series of contributions by the group of Runsheng Chen (He et al., 2006, 2007; Li et al., 2008; Aftab et al., 2008) expression profiles of sbRNAs among other ncRNAs were studied, but no clue about the function of sbRNAs could be drawn from that data. In Chapter 7 we report on RNA homology search for sbRNAs in nematodes and, moreover, show that sbRNAs do not constitute a novel RNA class, but instead are members of the Y RNA family.

4

RNAz 2.0: improved noncoding RNA detection

Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF (2010) **RNAz 2.0: improved noncoding RNA detection.** In Proceedings of the Pacific Symposium on Biocomputing, volume 15, pages 69-79. Hawaii, USA.
DOI: 10.1142/9789814295291_0009

Authors' contributions: ILH, SW and PFS initiated the study and guided in study design. ARG developed the dinucelotide z-score regression. ARG and SF generated the new training and test data sets. ARG retrained RNAz on new data. ARG and SF evaluated the performance of RNAz. ARG and SF wrote the manuscript.

Pacific Symposium on Biocomputing 15:69-79(2010)

RNAZ 2.0: IMPROVED NONCODING RNA DETECTION

ANDREAS R. GRUBER[1,2], SVEN FINDEIß[1], STEFAN WASHIETL[2,3],
IVO L. HOFACKER[2] AND PETER F. STADLER[1,2]

[1] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig*
Härtelstrasse 16-18, D-04107 Leipzig, Germany

[2] *Institute for Theoretical Chemistry, University of Vienna*
Währingerstrasse 17, A-1090 Wien, Austria.

[3] *European Molecular Biology Laboratory – European Bioinformatics Institute*
Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

`RNAz` is a widely used software package for *de novo* detection of structured noncoding RNAs in comparative genomics data. Four years of experience have not only demonstrated the applicability of the approach, but also helped us to identify limitations of the current implementation. `RNAz 2.0` provides significant improvements in two respects: (1) The accuracy is increased by the systematic use of dinucleotide models. (2) Technical limitations of the previous version, such as the inability to handle alignments with more than six sequences, are overcome by increased training data and the usage of an entropy measure to represent sequence similarities. `RNAz 2.0` shows a significantly lower false discovery rate on a dinucleotide background model than the previous version. Separate models for structural alignments provide an additional way to increase the predictive power. `RNAz` is open source software and can be obtained free of charge at: `http://www.tbi.univie.ac.at/~wash/RNAz/`

Keywords: RNA structure; noncoding RNA; structure conservation; comparative genomics; gene prediction

1. Introduction

Noncoding RNAs (ncRNAs) are transcripts that are not translated to proteins but function directly on the RNA level. During the past few years it has become evident that such "RNA genes" are more common than previously thought. MicroRNAs, for instance, have profoundly changed our view of gene regulation, and several completely new classes of ncRNAs were discovered recently.[1] They have been found to be involved in such diverse processes as transcriptional regulation,[2–4] post-transcriptional regulation,[5] chromatin modification and epigenetics,[6,7] and development.[8] Non-coding RNAs thus are key players in cellular regulation, a realization that has also moved the computational analysis and the annotation of ncRNAs at genome-wide scales into the focus of attention.

With the rapidly increasing availability of genomic sequence data, the *de novo* prediction of ncRNAs is of particular interest. While protein gene prediction is a classical problem in computational biology and has been studied for more than 15 years, RNA gene prediction is still in its infancy. Nevertheless, significant progress has been made regarding the prediction of "structured ncRNAs". This class of ncRNAs is characterized by evolutionary conserved secondary structures which appear to be important for their function. Most of the well-characterized ncRNAs belong to this class. Leading software tools developed for *de novo* RNA gene finding therefore use evolutionary conservation of functional secondary structures as the main signal to detect these ncRNAs.[9–13]

`RNAz` also detects structural ncRNAs by means of a comparative approach. In addition to measuring evolutionary conservation, however, it also explicitly evaluates the thermodynamic stability of the secondary structure.[14] A support vector machine (SVM) is then used to evaluate both criteria. `RNAz 1.0` has been used successfully to map structural ncRNAs in a wide variety of genomes.[15–20] A large number of these predictions have also been verified experimentally.[21–23] Moreover, the generic approach and many algorithmic details developed for `RNAz 1.0` have been re-used, extended, and adapted to other problems in the field of RNA gene-finding.[11,24–30]

The wide-spread use of `RNAz 1.0` also helped to identify some of its limitations and to point our directions

for improvements. In this contribution, we describe a major update of the RNAz program. It is based on the results of two follow-up studies,[31,32] on our experiences gained during many real-life applications, in particular the ENCODE pilot project,[33,34] and last but not least, on the user feedback we received over the past four years.

One major improvement is that RNAz 2.0 now allows to calculate thermodynamic stability scores based on a dinucleotide background model. It has been noted early-on that folding algorithms utilizing stacking energies of adjacent base-pairs in their energy model are sensitive to the dinucleotide content.[35] In the context of genome-wide ncRNA predictions, this effect can lead to an increased number of false positive calls as pointed out several times.[32,33,36] The new dinucleotide model in RNAz 2.0 now avoids this source of potential false positives and increases the accuracy of the program.

Another major limitation of RNAz 1.0 was the fact that only alignments with at most six sequences could be scored. This rather arbitrary restriction was the result of the limited amount of comparative data sets that were available at the time. During the past few years, however, comparative data sets have grown massively and therefore we adapted the algorithm to allow flexible analysis of alignments of any size.

2. Methods

2.1. *Overview of the RNAz algorithm*

RNAz predicts functional RNA structures on two independent criteria: (i) thermodynamic stability and (ii) structural conservation.

A common way to express thermodynamic stability is in terms of a z-score. This is simply the number of standard deviations by which the minimum free energy (MFE) deviates from the mean MFE of a set of randomized sequences with the same length and base composition. A negative z-score thus indicates that a sequence is more stable than expected by chance. As this procedure involves energy evaluation of a large set of random sequences it is not applicable for large-scale genomic screens. RNAz instead uses support vector regression (SVR) to estimate the mean and the standard deviation based on the nucleotide composition of a sequence.

RNAz evaluates evolutionary conservation of RNA structures in terms of the structure conservation index (SCI). A consensus secondary structure is predicted using the RNAalifold algorithm,[42] which is an extension of standard minimum free energy folding algorithms with the constraint that all sequences have to fold into a common structure. Compensatory mutations, i.e. mutations that preserve a certain base pair, yield bonus energies, while inconsistent mutations add penalty energies. RNAz measures structural conservation by calculating the ratio of the consensus folding energy to the unconstrained folding energies of the single sequences.

Both criteria are combined by another support vector machine model that classifies the input alignment as "structural RNA" or "other". A graphical overview of the RNAz algorithm is depicted in Fig. 1. In the following, we describe independent refinements of these steps that improve the overall prediction accuracy of the RNAz approach.

2.2. *z-score regression for dinucleotide shuffled sequences*

As in RNAz 1.0, we use support vector regression to compute z-scores for folding energies because the direct approach via repeated shuffling and folding is too costly for genome-wide applications.

In order to efficiently train the regression engine of RNAz 2.0, we used the following grid-like procedure: We first generated synthetic sequences of length 50 with G+C content, A/(A+U) ratio, and C/(C+G) ratio ranging from 0.20 to 0.80 in steps of 0.05. For each of these start sequences we then generated 500,000 mononucleotide shuffled sequences and discarded those sequences where the relative difference between the observed dinucleotide frequency and the expected frequency exceeded the threshold of 1.5. Evaluation on human ENCODE sequences showed that only a small fraction of approximately 1% of the sequences have a higher value and it was hence considered to be a reasonable threshold. Sequences of length 100, 150 and

Pacific Symposium on Biocomputing 15:69-79(2010)

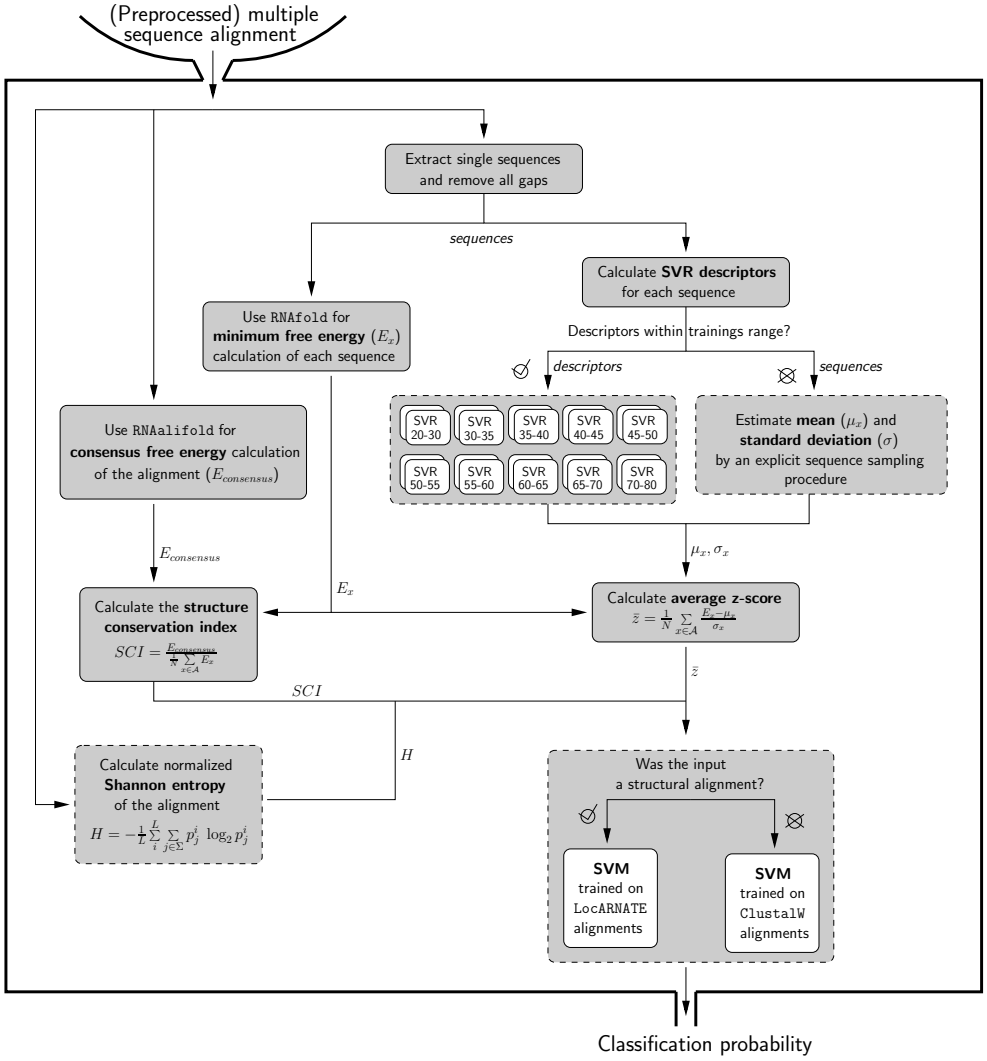


Fig. 1. Outline of the RNAz 2.0 work-flow and algorithm. In a first step large genomic multiple alignments are processed using rnazWindow.pl into smaller alignments. This filtering procedure involves several steps: (i) overlapping windows given a fixed window and step size are created, (ii) sequences that contain too many gaps are removed and (iii) from the remaining sequences only those sequences are kept that meet a predefined average pairwise identity threshold. The resulting alignments are then separately processed by RNAz. First, structure and energy predictions are performed for both the single sequences and the alignment. These results can be immediately combined to calculate the SCI as the measure of the evolutionary conservation of the RNA sequences in the alignment. In a second step, the mean free energy and the standard deviation used for the calculation of the z-score are estimated. For this purpose descriptors based on the nucleotide composition (G+C content, A/(A+U) ratio, C/(C+G) ratio, all 16 dinucleotide frequencies and the length of the sequence) are calculated for each sequence. If descriptors are within the training boundaries they are passed to the corresponding support vector regression (SVR) based on the G+C content. Otherwise, the mean and the standard deviation are evaluated explicitly by folding of 1,000 randomized sequences with the same dinucleotide composition. In a final step the average z-score of the sequences, the SCI and the normalized Shannon entropy of the alignment are passed to the classification SVM, which returns a probability estimate that the given alignment harbours thermodynamically stable and/or evolutionary conserved RNA secondary structures. Parts that are highlighted in dashed boxes are new or modified components of RNAz algorithm. RNAfold and RNAalifold are part of the Vienna RNA Package. Numbers in the SVR boxes indicate the G+C content the particular SVR is trained on. For a detailed explanation of the formulas we refer to section 2.3.

200 where then generated by concatenating the initial set of sequences 2 to 4 times. This initial set can be generated very quickly and served as the basis for the selection of a much smaller, approximately evenly

spaced, training set with representative dinucleotide frequencies. A sequence from the initial set was only added to the representative training set if the Euclidean distance of the dinucleotide frequencies to any sequence already present in the representative set was above a certain threshold (0.075 for a G+C content of 0.20 and 0.80, 0.100 for a G+C content of 0.25, 0.30, 0.70 and 0.75, and 0.125 for the remaining range). For the final training set we also added sequences of length 75, 125 and 175, which were generated as described above, resulting in a total of 1,155,737 training instances.

For each of these instances, we generated 1,000 randomized sequences by the Altschul-Erikson algorithm[37] with the same dinucleotide composition and used `RNAfold`[38] with parameter `-d2` to evaluate their folding free energy. More than 1 million training instances are by far too many to be used in SVM training procedures in reasonable time. For this reason we split the training instances into smaller subsets according to their G+C content. In total we have 10 subsets with at most 150,000 training instances. We used the SVM library `LIBSVM` to train regression models for the mean and the standard deviation for each of the ten subsets. As input features we used the G+C content, the A/(A+U) ratio, the C/(C+G) ratio, all 16 dinucleotide frequencies and the length of the sequence scaled to the interval [0,1]. The regression for estimating the mean free energy was trained to learn energy per nucleotide, while the standard deviation was not scaled. We chose the ν variant of regression and a radial basis function kernel. The standard grid search approach was used to find optimal combinations for SVM parameters. Regression accuracy was monitored on an independent test set compiled from randomly selected sequences of variable length from 50 to 200 nt from the human ENCODE regions. The average number of support vectors for the mean and the standard deviation regression models are 8,763 and 8,607, respectively.

2.3. *Training data generation and training of the SVM classifier*

Training and test sets are based on the data available in the `Rfam` 9.1 database.[39] 93 RNA families were selected based on their signals for thermodynamic stability and structural conservation. The `RNAz 2.0` training set covers a broad range of different RNA families including major classes such as tRNAs, snoRNAs, microRNAs, riboswitches, and bacterial regulatory RNAs.

For each RNA family, a set of alignments with varying numbers of sequences and average pairwise identities was generated using the following strategy: `Rfam` full alignments were used if they contained less than 300 sequences, otherwise we used the seed alignments. For our purpose the use of at most 300 sequences proofed well to generate a set of alignments over the desired range of average pairwise identities. `Rfam` alignments were utilized only as a source to retrieve family members of a particular ncRNA class and only extracted, ungapped RNA sequences were used for subsequent analyses.

First, `Rfam` alignments were filtered to remove nearly identical sequences, so that the training alignments contained sequences with at most 98% identity. The sequences were then re-aligned using `ClustalW`. For each of these ncRNA family alignments we then proceeded as follows: for each number of sequences from 2 to 15 we generated at most 10 alignments with a randomly chosen average pairwise identity between 50 and 98% and with a maximum relative difference in sequence lengths of 65% using `rnazWindow.pl` which is part of the `RNAz` analysis pipeline.[44]

To ensure that this set of positive training examples contained only instances with good structural conservation signals we filtered alignments by using tree editing distances between the structures of the sequences in the alignment as a quality measure of structural conservation. Ordered, rooted trees can be deduced from the dot-bracket notation of RNA secondary structures. Tree editing defines a metric in the space of trees by a set of operations (deletions, insertion and relabeling of nodes) and hence can be used to calculate distances between RNA secondary structures.[31] For each alignment we extracted sequences, removed gaps and calculated the averaged pairwise tree editing distance using `RNAdistance` with options `-d2 -Dh` to enable dangling ends and to use the HIT representation for RNA secondary structures. We repeated this for a set of 100 randomized alignments and calculated an empirical p-value as a measure of structural conservation. Alignments with a p-value higher than 0.05 were removed from the training set. Alignments retained after this filtering procedure were realigned with `ClustalW` with standard options for

Pacific Symposium on Biocomputing 15:69-79(2010)

application to sequence-based alignments.

For the generation of structural alignments for the training set we chose to use `LocARNATE`,[40] which is a structural alignment program based on the Sankoff algorithm for the simultaneous solution of the RNA folding and the alignment problem. `LocARNATE` uses `RNAfold` for structure predictions and hence the same energy parameters as `RNAz` does. `LocARNATE` was called with options `--no-seq --no-struc` to generate global, structural alignments.

Negative instances of the training set were generated by shuffling using `multiperm`[41] v. 0.9.3 if the normalized Shannon entropy of the alignment[31] was less than 0.50. Otherwise, alignments were simulated using `SISSIz`[32] to ensure full randomization for the more diverse alignments where shuffling can become inefficient. The final training set was composed of 10,538 alignments for each the positive and the negative class.

The `RNAz 2.0` SVM classifier uses three features to detect structured noncoding RNAs: (i) the average minimum free energy z-score $\bar{z}$ estimated from a dinucleotide shuffled background, (ii) the SCI and (iii) the normalized Shannon entropy H of the alignment as a measure for the content of evolutionary information.

Consider an alignment $\mathcal{A}$ consisting of N sequences. Let E_x denote the minimum free energy of sequence x, and let μ_x and σ_x be the mean and standard deviation, respectively, of the folding energies of a large number of random sequences of the same length and same dinucleotide composition as x. The averaged z-score of the alignment $\mathcal{A}$ is defined as

$$\bar{z} = \frac{1}{N} \sum_{x \in \mathcal{A}} \frac{E_x - \mu_x}{\sigma_x}$$

The SCI of an alignment is given as the fraction of the consensus folding free energy ($E_{consensus}$) to the average of the folding free energies of the single sequences:

$$SCI = \frac{E_{consensus}}{\frac{1}{N} \sum\limits_{x \in \mathcal{A}} E_x}$$

The normalized Shannon entropy H of an alignment $\mathcal{A}$ of RNA sequences over the alphabet $\Sigma = \{$A, C, G, U, -$\}$ is defined as the sum of the Shannon entropies of the individual columns divided by the length of the alignment denoted by L:

$$H = -\frac{1}{L} \sum_{i}^{L} \sum_{\alpha \in \Sigma} p_\alpha^i \ \log_2 \ p_\alpha^i$$

The probability p_α^i is approximated by the observed frequency of character α in alignment column i (normalized by the number N of sequences in the alignment). All features were scaled to a range of [-1,1]. Standard grid search combined with a 10-fold cross validation was applied to find optimized SVM parameters. Among the models with the best cross-validation accuracy (top 20) we chose the model that showed best performance on an independent test set created the same way as the training set. The output of the final classification SVM is a probability estimate that the input alignment contains thermodynamically stable and/or structurally conserved RNA sequences.

A second, independent, SVM classifier was trained on sequence/structure-based alignments generated by `LocARNATE` using the same procedure.

3. Results

3.1. *Dinucleotide based z-scores*

To estimate the mean and standard deviation of folding energies for mononucleotide shuffled sequences it is feasible to sample uniformly simply by varying variables describing the four mononucleotide frequencies and the length of the sequence on a grid. This approach cannot, however, be extended that easily for dinucleotide shuffled sequences. One has to consider the much larger space of dinucleotide compositions that is occupied by

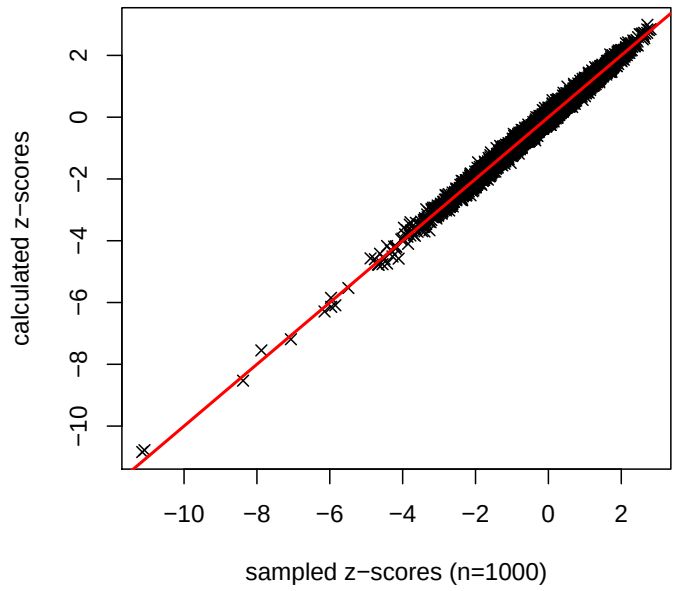


Fig. 2. z-scores calculated by support vector regression in comparison with z-scores determined from 1,000 random samples preserving dinucleotide frequencies for 10,000 randomly drawn sequences from the human ENCODE regions. Correlation of z-scores is 0.996 and the mean absolute error is 0.076.

sequences of practical interest. In this work we use a grid-like approach, where we first apply uniform sampling to cover the mononucleotide space and then choose, for each data point in the grid, a representative set of sequences that covers the dinucleotide space for that particular base composition. However, this procedure still gave more than one million training instances. The training data was split into different ranges of the G+C content to guarantee efficient training and fast prediction. This comes at the price of increased memory consumption but keeps the number of support vectors comparable to the approach used in `RNAz 1.0`. Accuracy of the z-score regression for dinucleotide shuffled sequences was evaluated on 10,000 randomly chosen sequences of variable length from 50 to 200 nt from the human ENCODE regions[34] (Fig. 2) and genomic sequences of *D. melanogaster* and *E. coli*. The mean absolute error (MAE) and the correlation (R) of z-scores calculated by SVM regression compared to z-scores determined from 1,000 random samples is 0.0748 and 0.996, respectively ($n = 30,000$; genomic sequence from ENCODE regions, *D. melanogaster*, and *E. coli*). Comparisons of z-scores determined from 1,000 dinucleotide shuffled sequences to 100 dinucleotide shuffled sequences (MAE= 0.107, $R = 0.992$) and to 1,000 mononucleotide shuffled samples (MAE= 0.420, $R = 0.916$) clearly demonstrate that our method is a suitable approach for fast and efficient estimation of dinucleotide controlled z-scores. `RNAz 1.0` also showed restrictions on the base composition because of the training range of the SVR. This limitation is now overcome by explicit generation of shuffled sequences once the base composition of a sequence is out of the training range. Since boundaries have been chosen broadly (e.g. G+C content from 20 to 80%) this will only apply in a small minority of cases.

3.2. *New training sets and improved classification model*

Since the postulation of the SCI, it has been a major point of criticism that the SCI evaluates structural conservation on the energy level rather than on the RNA structures themselves. However, in previous study[31] it has been shown that the SCI is on average the most powerful method and that it is only outperformed by

Pacific Symposium on Biocomputing 15:69-79(2010)

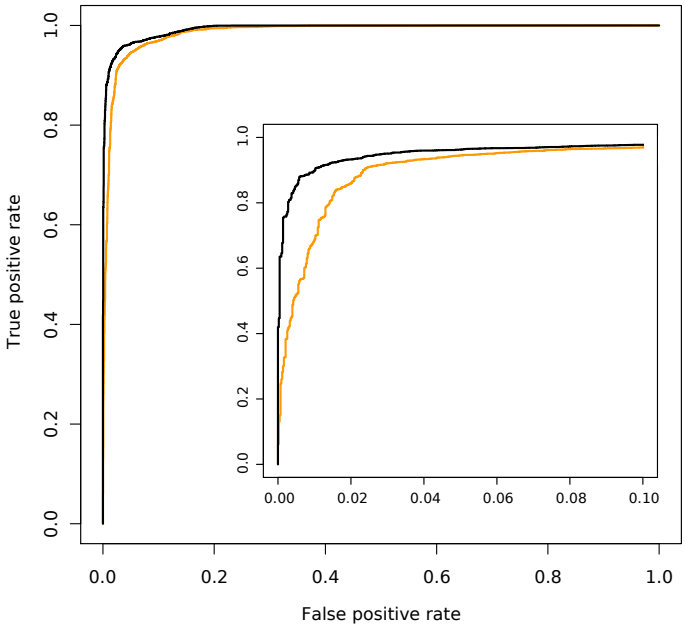


Fig. 3. Accuracy of `RNAz 2.0` classification (black) vs. `RNAz 1.0` classification (orange) on a previously published data set for the evaluation of noncoding RNA gene finders.[32] The positive instance data set consists of 4,303 alignments of structural RNA families (5S ribosomal RNA, U2 spliceosomal RNA, tRNA, Hammerhead ribozyme, U3 snoRNA, U5 spliceosomal RNA, Group II catalytic intron, and Mir-10 microRNA) with two to six sequences per alignment. The negative instance data set consists of 4,303 alignments taken from random genomic location, which resemble approximately the same dinucleotide composition and conservation degree as the positive set. The inset shows the region of high specificity were `RNAz 2.0` clearly outperforms the old version.

other approaches in the high sequence identity range. Attempts to use other conservation measure methods than the SCI, however, failed to give results of comparable quality (data not shown).

To use the SCI for efficient classification one has to take into account the average pairwise identity and the number of sequences as well. Due to the lack of comparative data at the time of training of the initial `RNAz` algorithm limits on these two descriptors were rather arbitrarily chosen. In this work we generated a new training set covering a broader range of RNA families and evaluate sequence variation in terms of the normalized Shannon entropy which has been shown to combine both sequence variation and the number of sequences into one measure.[31] This does not only result in dimensionality reduction of the final classification model, but also overcomes the need to set an upper boundary to the number of sequences in an alignment.

The new `RNAz 2.0` algorithm now uses the average z-score of the sequences in the alignment based on a dinucleotide background model, the SCI and the normalized Shannon entropy as features in the final classification model. To evaluate the predictive power of `RNAz 2.0` we chose a test set used in a previous study.[32] This test set is especially well suited as it contains randomly chosen genomic regions from vertebrate alignments as negative controls. The background dinucleotide content in vertebrate genomes is known to be the main reason for false positive calls in `RNAz 1.0`.[33] Although both versions perform well on this test set, `RNAz 2.0` clearly outperforms version `1.0` in the high specificity range (Fig. 3). For example, at the generally used 0.01 false-positive cutoff, `RNAz 2.0` shows 0.899 sensitivity compared to 0.688 in the old version.

It is a well known fact that sequence-based alignment methods fail to give high quality alignments regarding RNA secondary structures in low average pairwise identity ranges. By using structural alignments one can expect an improvement in discrimination capability of the SCI for alignments with low sequence

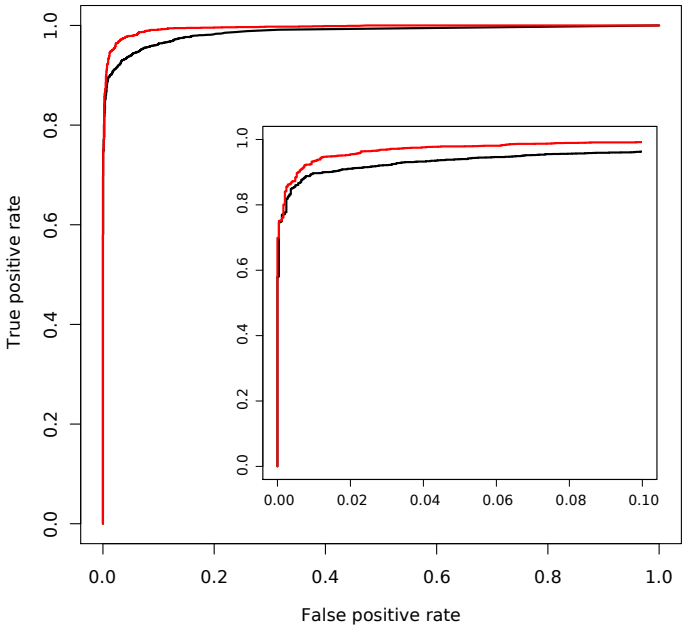


Fig. 4. ROC curves for the `RNAz 2.0` prediction accuracy on sequence-based alignments (black) vs. structural alignments (red). A significant improve of the overall predictive power of `RNAz 2.0` is achieved by use of structural alignments. The test set is composed of 2,455 alignments of various ncRNA families with an average pairwise identity between 30 and 70%, as well as a negative set consisting of 2,455 alignments derived by randomization of reference alignments with `multiperm` or `SISSIz` as described in section 2.2. Sequence-based alignments were generated with `ClustalW`, while structural alignments were generated with `LocARNATE`.

similarity.[11] Therefore, we trained a separate SVM decision model based on sequence/structure alignments, similar to the approach used in `RSSVM`.[30] Structural alignments were generated using `LocARNATE`, a multiple alignment variant of `LocARNA`.[46] As depicted in Fig. 4 structural alignments improve the overall predictive power of `RNAz`.

Recent studies (e.g. Washietl *et al.*[33]) have shown that `RNAz` suffers from a high false discovery rate (FDR). We therefore evaluated the performance of both versions for the human ENCODE regions. 17-way MAF alignments based on the human genome assembly hg.17 were downloaded from the UCSC genome browser. In total we screened 193,634 MAF alignments derived by pre-filtering with `rnazWindow.pl` with standard options (window length is 120 nt, step size is 40 nt, average pairwise identity the resulting alignment is optimized to is 80%, and at most six sequences are allowed). Both reading directions were considered in our analysis. A dinucleotide background model was generated with `SISSIz`[32] and all hits detected by `RNAz` on this data set were considered to be false positives. Results are summarized in Tab. 1. While `RNAz 1.0` shows a very high FDR of around 80%, the FDR of `RNAz 2.0` is much lower being around 54% for high confident hits (classification probability > 0.9). It seems noteworthy, that in a previous study[33] the FDR for `RNAz 1.0` on ENCODE data was estimated to be around 50%. This estimate was based on a rather simplistic *ad hoc* method to correct for the dinucleotide bias. The new results are based on the more accurate `SISSIz` null model and demonstrate that `RNAz 1.0` is even more affected by the dinucleotide bias than previously assumed. The new version, however, reduces this source of false positives significantly.

To investigate a potential G+C bias of `RNAz` that was observed for version `1.0`,[33] we also trained a classification model that included the G+C content as fourth feature. This additional feature, however, had little impact on the predictions. In particular, the distribution of the G+C content of the positive predictions

Pacific Symposium on Biocomputing 15:69-79(2010)

Table 1. Comparison of the false discovery rate (FDR) based on ENCODE regions and a dinucleotide background model for low (P > 0.5) and high (P > 0.9) confidence hits. A hit corresponds to a single alignment derived from pre-filtering of ENCODE MAF alignments with `rnazWindow.pl`.

	`RNAz 1.0`		`RNAz 2.0`	
	# low conf.	# high conf.	# low conf.	# high conf.
ENCODE regions	17,814	6,854	6,880	2,259
background	14,489	5,596	4,090	1,219
estimated FDR	81%	82%	59%	54%

remained nearly unchanged (data not shown). This suggests that the elevated G+C content of `RNAz` hits is not an artificial bias, but rather reflects the G+C content of true functional RNAs. Consistent with this observation, the G+C bias of structured RNAs has been used successfully for *de novo* prediction of RNA genes.[43] Preliminary analysis of the ENCODE data showed that the effect is smaller for `RNAz 2.0` than in the earlier version.

3.3. *Computational speed*

The performance of `RNAz 2.0` in comparison to `RNAz 1.0` was benchmarked on 50,000 randomly chosen MAF alignments from the ENCODE data set. Alignment length was 120 nucleotides and alignments contained at most six sequences. Experiments were conducted on an Intel Xeon 2.40GHz CPU. For each alignment both reading directions were examined, resulting in a total of 100,000 alignments that had to be scored. The execution time required by `RNAz 1.0` was 202 min, `RNAz 2.0` with explicit shuffling switched off was 252 min and `RNAz 2.0` using explicit shuffling was 1,230 min. Although explicit shuffling had to be used for only 1% of the sequences (5,524 out of 549,210), it comes with an tremendous overhead increasing the run time of `RNAz 2.0` almost 5-fold. We extracted those alignments where explicit shuffling was used and compared the classification probability to the one derived from calling `RNAz` with option `--no-shuffle` to avoid explicit shuffling. For the vast majority of cases (96%) the change in classification probability was less than 1%. For this data set the maximal observed difference was 0.21. In general, we observed larger differences in the range from 0.2 to 0.8 than in the regions close to 0 or 1.

With option `--no-shuffle`, `RNAz 2.0` has an execution time that is increased by about 25% compared to `RNAz 1.0`.

4. Future directions

In this work we present a major update of the `RNAz` algorithm. Evaluation of thermodynamic stability has been improved by considering a dinucleotide background model. This directly translates into a significantly lower false discovery rate. In addition to the dinucleotide z-score, the overall prediction accuracy is improved by a combination of the use of a new training set and the normalized Shannon entropy as a measure of sequence variation. Furthermore, the updated version is not any more restricted to limitations concerning the base composition or number of sequences in the input alignment.

The generation of structural alignments is computationally expensive but we showed that they can improve the `RNAz` classification power. This is true in particular for alignments of low average pairwise identity. Given that the overall computational complexity of `LocARNATE` is $O(n^4)$, the routine use of structural alignments on a genome-wide scale is still out of reach, at least when off-the shelf hardware is used. In general, it has to be questioned if ncRNA gene finding would benefit from realigning genomic alignments available to date with a structural aligner. These alignments have been generated by means of sequence-only based methods and therefore are not likely to contain homologous RNA sequences that evolve fast on nucleotide level but retain structural conservation. A feasible strategy, however, is the pre-selection of

syntenic regions based on better-conserved flanking regions.[13] Such an approach could be employed for the detection of conserved local structures in the untranslated regions of protein-coding mRNAs, where orthology is established based on similarities of the much better conserved coding sequences. The re-scoring of positively scored hits of a sequence-based `RNAz` screen after re-aligning them with a structural aligner may help to increase the overall accuracy, in particular for relatively poorly conserved alignment slices. One could also use `RNALfold`[45] augmented with the z-score prediction engine of `RNAz` to screen for loci that show signature of increased thermodynamic stability then re-evaluate these loci using structural alignments with `RNAz 2.0` to also account for structural conservation.

An open question, not covered by this work, is how to address the growing number of species in genomic alignments. The use of the normalized Shannon entropy helped us to remove the upper limit on the number of sequences in the alignment. Preliminary analysis of `RNAz 2.0` on multiz 44-way, 28-way and 17-way alignments shows, however, that the simple use of more sequences does not necessarily correlated with improved classification power. To a large extent the increased conservation signal is counteracted by increasing levels of alignment errors. Structural variation of the ncRNAs themselves also poses technical challenges. To-date, an algorithm that addresses both possible misalignments and structural variation is still missing.

RNA secondary structure prediction is sensitive to the exact ends of the input sequence. The use of arbitrarily determined alignment windows of fixed width thus introduces noise. This issue will be alleviated in a forthcoming update of `RNAz` that addresses the pre-processing of long genomic alignments. Here, the sliding window approach will be replaced by the systematic use of `RNALalifold`,[47] an algorithm that computes locally stable consensus RNA secondary structures. These are then used to extract alignments of self-contained (sub)structures for `RNAz` scoring.

`RNAz 2.0` was trained on two particular alignment methods, `ClustalW` for sequence-based alignments and `LocARNATE` for structure-based alignments. As `RNAz` uses a machine learning approach, we have to expect some influence of the alignment algorithm since the features passed to the SVM implicitly also incorporate properties of the alignment algorithms themselves. It may thus become necessary to either re-align the input data or to train decision models for alternative alignment methods.

Supplementary material

An Electronic Supplement located at `www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-026/` compiles a supplemental figure and data sets used in this work.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network III" and "noncoding RNA II", the University of Vienna and the German Research Foundation (grants STA 850/7-1 under the auspices of SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes").

References

1. P. P. Amaral *et al.*, *Science* **319**, 5871 (2008).
2. T. Kuwabara *et al.*, *Cell* **116**, 6 (2004).
3. J. Feng *et al.*, *Genes Dev* **20**, 11 (2006).
4. C. A. Espinoza *et al.*, *RNA* **13**, 4 (2007).
5. A. Pagano *et al.*, *PLoS Genet* **3**, 2 (2007).
6. A. Wutz, *Trends Genet* **23**, 9 (2007).
7. J. L. Rinn *et al.*, *Cell* **129**, 7 (2007).
8. P. P. Amaral and J. S. Mattick, *Mamm Genome* **19**, 7-8 (2008).
9. E. Rivas and S. R. Eddy, *BMC Bioinformatics* **2**, (2001).
10. J. S. Pedersen *et al.*, *PLoS Comput Biol* **2**, 4 (2006).
11. A. V. Uzilov *et al.*, *BMC Bioinformatics* **7**, (2006).
12. Z. Yao *et al.*, *Bioinformatics* **22**, 4 (2006).

Pacific Symposium on Biocomputing 15:69-79(2010)

13. E. Torarinsson *et al.*, *Genome Res* **16**, 7 (2006).
14. S. Washietl *et al.*, *Proc Natl Acad Sci USA* **102**, 7 (2005).
15. S. Washietl *et al.*, *Nat Biotechnol* **23**, 11 (2005).
16. K. Missal *et al.*, *Bioinformatics* **21**, (2005).
17. K. Missal *et al.*, *J Exp Zoolog B Mol Dev Evol* **306**, 4 (2006).
18. D. Rose *et al.*, *BMC Genomics* **8**, (2007).
19. D. Rose *et al.*, *J Bioinform Comput Biol* **6**, 6 (2008).
20. A. M. McGuire and J. E. Galagan, *PLoS One* **7**, 3 (2008).
21. C. Weile *et al.*, *BMC Genomics* **8**, (2007).
22. T. Mourier *et al.*, *Genome Res* **18**, 2 (2008).
23. C. del Val *et al.*, *Mol Microbiol* **66**, 5 (2007).
24. J. Hertel *et al.*, *Bioinformatics* **24**, 2 (2008).
25. J. Hertel *et al.*, *Bioinformatics* **22**, 14 (2006).
26. K. Reiche *et al.*, *Algorithms Mol Biol* **2**, (2007).
27. P. P. Gardner *et al.*, *Nucleic Acids Res* **33**, 8 (2005).
28. P. W. Hsu *et al.*, *Nucleic Acids Res* **33**, (2006).
29. T. Sandmann and S. M. Cohen, *PLoS ONE* **2**, 11 (2007).
30. X. Xu *et al.*, *PLoS Comput. Biol.*, **5**, (2009).
31. A. R. Gruber *et al.*, *BMC Bioinformatics* **9**, (2008).
32. T. Gesell and S. Washietl, *BMC Bioinformatics* **9**, (2008).
33. S. Washietl *et al.*, *Genome Res* **17**, 6 (2007).
34. ENCODE Project Consortium, *Nature* **447**, 7146 (2007).
35. C. Workman and A. Krogh, *Nucleic Acids Res* **27**, 24 (1999).
36. T.Babak *et al.*, *BMC Bioinformatics* **8**, (2007).
37. S. F. Altschul and B. Erickson, *Mol Biol Evol.* **2**, 6 (1985).
38. I. L. Hofacker *et al.*, *Monatsh. Chem.* **125**, (1994).
39. P. P. Gardner *et al.*, *Nucleic Acids Res.* **37**, (2008).
40. W. Otto *et al.*, *Proceedings of the German Conference on Bioinformatics* **P-136**, (2008).
41. P. Anandam *et al.*, *Bioinformatics* **25**, (2009).
42. I. L. Hofacker *et al.*, *J.Mol.Biol.* **319**, (2002).
43. M. M. Meyer *et al.*, *BMC Genomics* **10**, (2009).
44. S. Washietl, *Methods Mol Biol.* **395**, (2007).
45. I. L. Hofacker *et al.*, *Bioinformatics* **20**, (2004).
46. S. Will *et al.*, *PLoS Comput. Biol.*, **3**, (2007).
47. A. F. Bompfünewerer Consortium, *J Exp Zoolog B Mol Dev Evol.* **308**, (2007).

5

RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures

Gruber AR, Bernhart SH, Zhou Y, Hofacker IL (2010) **RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures.** In Proceedings of the German Conference on Bioinformatics (GCB'2010), volume P-173 of Lecture Notes in Informatics (LNI), pages 11-21. Gesellschaft für Informatik (GI). ISBN 978-3-88579-267-3.

Authors' contributions: ILH initiated the study and guided in study design. ARG developed the z-score regression. SHB implemented the new approach in `RNALfold`. ARG performed all analyses and wrote the manuscript. YZ participated in a preliminary study that led to this study.

RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures

Andreas R. Gruber[1], Stephan H. Bernhart[1],
You Zhou[1,2], and Ivo L. Hofacker[1]
[1] *Institute for Theoretical Chemistry*
University of Vienna, Währingerstraße 17, 1090 Wien, Austria
[2] College of Computer Science and Technology
Jilin University, Changchun 130012, China
{agruber, berni, ivo}@tbi.univie.ac.at, zyou@jlu.edu.cn

Abstract: The search for local RNA secondary structures and the annotation of unusually stable folding regions in genomic sequences are two well motivated bioinformatic problems. In this contribution we introduce `RNALfoldz` an efficient solution two tackle both tasks. It is an extension of the `RNALfold` algorithm augmented by support vector regression for efficient calculation of a structure's thermodynamic stability. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

1 Introduction

Over the past decade noncoding RNAs (ncRNAs) have risen from a shadowy existence to one of the primary research topics in modern molecular biology. Today computational RNA biology faces challenges in the ever growing amount of sequencing data. Efficient computational tools are needed to turn these data into information. In this context, the search for locally stable RNA secondary structures in large sequences is a well motivated bioinformatic problem that has drawn considerable attention in the community. `RNALfold` [HPS04] has been the first in a series of tools that offered an efficient solution to this task. Instead of a straight-forward, but costly sliding window approach a dynamic programming recursion has been formulated that predicts all stable, local RNA structures in $\mathcal{O}(N \times L^2)$, where L is the maximum base-pair span and N the length of the sequence. Since its publication, the `RNALfold` algorithm has inspired a lot of work in this field, see e.g. `Rnall` by Wan *et al.* [WLX06] or `RNAslider` by Horesh *et al.* [HWL$^+$09]. All contributions so far in this field focused on improving the computational complexity of the algorithm, but none of the approaches has ever been used to unravel results of biological significance. In particular, *de novo* detection of functional RNA structures has been addressed, but application on a genome-wide scale with a low false discovery rate seems still out of reach. Even on the moderately sized genome of *E. coli* (4.6 Mb) one is drowning in hundreds of thousands of local structures. Unlike in the well established field of protein coding gene detection where one can exploit signals like codon usage, functional

RNA secondary structures, in general, do not show strong characteristics that make them easily distinguishable from random decoys. Successful approaches for ncRNA detection operating solely on a single sequence [HHS08, JWW$^+$07] are limited to specific RNA classes, where some outstanding characteristics can be harnessed. There is no master plan for the detection of functional RNA structures, but one would certainly want to limit the `RNALfold` output to a reasonable amount. So far, only the minimum free energy (MFE) of the locally stable secondary structures, which is intrinsically computed by the algorithm, has been considered as potential discriminator to limit the number of secondary structures. As demonstrated clearly by Freyhult and colleagues [FGM05] the MFE is roughly a function of the length of the sequence and the G+C content. Even normalizing the MFE by length of the sequence does not serve as a good discriminator between shuffled or coding sequences and functional RNA structures. A strategy that does work, however, is to compare the native MFE E of the RNA molecule to the MFEs of a set of shuffled sequences of same length and base composition [LM89]. This way we can evaluate the thermodynamic stability of the secondary structure. A common statistical quantity in this context is the z-score, which is calculated as follows

$$z = \frac{E - \mu}{\sigma}$$

where μ and σ are the average and the standard deviation of the energies of the set of shuffled sequences. The more negative the z-score the more thermodynamically stable is the structure. Efficient estimation of a sequence's z-score has been a profound problem already addressed in the very beginnings of computational RNA biology. A first strategy to avoid explicit shuffling and folding was based on table look-ups of linear regression coefficients [CLS$^+$90]. Clote and colleagues [CFKK05] introduced the concept of the asymptotic z-score, where the efficient calculation is also solved via table look-ups. The current state-of-the art approach for fast and efficient estimation of the z-score is to use support vector regression [WHS05].

The study by Clote and colleagues and a follow up to Chen *et al.* (1990) [LLM02] also report on the effort to predict thermodynamically stable structures using a sliding window approach. In this contribution we present `RNALfoldz` an algorithm that combines local RNA secondary structure prediction and the efficient search for thermodynamically stable structures. `RNALfoldz` is an extension of the `RNALfold` algorithm augmented by support vector regression for efficient calculation of a sequence's z-score. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

2 Methods

2.1 Fast estimation of the z-score using support vector regression

For the efficient estimation of the z-score we follow the strategy first introduced by Washietl *et al.* [WHS05]. Instead of explicit generation and folding of shuffled sequences in order to

determine the average free energy and the corresponding standard deviation support vector regression (SVR) models are trained to estimate both values. As described in detail in the previous work, we used a regularly spaced grid to sample sequences for the training set. Synthetic sequences ranged from 50 to 400 nt in steps of 50 nt. The G+C content, A/(A+T) ratio and C/(C+G) ratio were, however, extended to a broader spectrum, now ranging from 0.20 to 0.80 in steps of 0.05. A total of 17,576 sequences were used for training. For each sequence of the training set 1,000 randomized sequences were generated using the Fisher-Yates shuffle algorithm, and subsequently folded with `RNAfold` with dangling ends option `-d2` [HFS+94]. SVR models for the average free energy and standard deviation were trained using the `LIBSVM` package (`www.csie.ntu.edu.tw/~cjlin/libsvm`). While in the previous work input features and the dependent variables were normalized to a mean of zero and a standard deviation of one, we apply here a different normalization strategy that leads to a significantly lower number of support vectors for the final models. For the regression of the average free energy model the dependent variable is normalized by the length of the sequence, while for the standard deviation it is the square root of the sequence length. The length still remains in the set of input features and is scaled from 0 to 1. Other features remain unchanged. We used a RBF kernel, and optimized values for the SVM parameters were determined using standard protocols for this purpose. Final regression models were selected by balancing two criteria: (i) mean absolute error (MAE) on a test set of 5,000 randomly drawn sequences of arbitrary length (50-400) from the human genome, and (ii) complexity of the model (number of support vectors) , which translates to following procedure: from the top 10% of regression models in terms of MAE we selected the one that had the lowest number of support vectors. For the average free energy regression we selected a model with a MAE of 0.453 and 1,088 support vectors, and for the standard deviation regression a model with a MAE of 0.027 and 2,252 support vectors. To validate our approach we finally compared z-scores derived from the SVR to traditionally sampled z-scores on a set of 1,000 randomly drawn sequences from the human genome. The correlation coefficient (R) is 0.9981 and the MAE is 0.072. This is in fair agreement to results obtained when comparing two sets of sampled z-scores (R: 0.9986, MAE: 0.054, number of shuffled sequences = 1,000).

2.2 Adaption of the `RNALfold` algorithm

RNALfold computes locally stable structures of long RNA molecules. It uses a Zuker type secondary structure prediction algorithm [ZS81] and restricts the maximum base pair span to L bases to keep the structures local. The sequence is processed from the 3' (the sequence length n) to the 5' end. In order to keep the number of back trace operations low and the output at moderate size, we want to avoid backtracing structures that differ only by unpaired regions. Furthermore, only the longest helices possible are of interest. To achieve this, a structure starting at base i is only traced back if the total energy $F(i,n)$ is smaller than that of its 3' neighbor $F(i+1,n)$ while its 5' neighbor has the same energy: $F(i-1,n) = F(i,n) < F(i+1,n)$. The local minimum structure is found by identifying the pairing partner j of i so that $C(i,j) + F(j+1,n) = F(i,n)$, i.e. the minimum energy

from i to n is decomposed into the local minimum part i, j and the rest of the molecule. Here, $C(i, j)$ stands for the energy of a structural feature enclosed by the base pair i, j. As a result of this, the output of `RNALfold` contains components, i.e. structures that are enclosed by a base pair, only. Before we actually start the trace back, we evaluate two new criteria: (1) the sequence of the structure traced back has to be within the training parameters of the SVR, and (2) the z-score of the energy of this structure has to be lower than a predefined bound. Criterion (1) is simply imposed by the training boundaries of the SVMs. Boundaries have, however, been chosen carefully to cover a broad range of today's known spectrum of functional RNA structures. 99.79% of the sequences in the `Rfam` v. 10 full data set match the base composition requirements of the SVR and 90% of `Rfam` RNA families are in within the sequence length restrictions.

In order to get the exact sequence composition that is needed for the SVR evaluations, the 3' end of the structure (j) has to be computed first. This is done by a first, short backtracing step, where the decomposition $F(i, n) = C(i, j) + F(j + 1, n)$ is used to find j. Subsequently, the average free energy given the base composition of the sequence $s(i, j)$ is computed by calling the corresponding SVR model. The SVR model for the standard deviation has approximately twice the number of support vectors as the average free energy model. To minimize calls of this model, first the minimal standard deviation for the particular sequence length is looked up. We can then, using the free energy of $C(i, j)$, calculate a lower bound of the z-score. Only if this lower bound is below the minimal required z-score, the support vector regression for the standard deviation is called to calculate the actual z-score. If the z-score then still meets the minimal z-score criterion, the structure is fully traced back and printed out.

3 Results

The concept of fast and efficient estimation of the z-score by support vector regression was first introduced by Washietl *et al.* [WHS05], and implemented in the noncoding RNA gene finder `RNAz`. The speed up of this approach compared to explicit shuffling and folding, which is usually done on 1,000 replicas, is tremendous, at minimum a factor of 1,000. Moreover, computing time is invariant to the length of the sequence, while RNA folding is of complexity of $\mathcal{O}(N^3)$. When considering the z-score as evaluation criterion in the `RNALfold` algorithm, calculation of the z-score becomes a time consuming factor, as in a worst case scenario it has to be done almost for every nucleotide of the sequence. It is therefore a crucial concern to use support vector models that do not only have good accuracy, but also a moderate number of support vectors (SVs). In this work we extended the z-score support vector regression to cover a broader range of the sequence spectrum, but managed at the same time to build models that have significantly less SVs than the models used by `RNAz`. This was accomplished by normalizing the dependent variables of the regression, i. e. the average free energy and the standard deviation, by the sequence length. The dependent variables do not strictly linearly correlate with the sequence length and so we have to keep the sequence length as an input feature. Nevertheless, redundant points are created in the training set, which eventually leads to a smaller space to be trained. For

the average free energy model and the standard deviation model we were able to achieve a 6.3 and a 2.7 fold reduction, respectively, in the number of SVs compared to the RNAz equivalents.

3.1 Evaluation of RNALfoldz predicition accuracy

For the task of predicting local RNA secondary structures one would particularly be interested in following criteria: (i) to which extent can functional ncRNAs be discovered, (ii) how well do the molecule's predicted boundaries match to the real coordinates, and (iii) is there any significant difference between native, biological sequences and random decoys. To address these questions, we applied RNALfoldz to the genome of *E. coli* (Accession number: CP000948). A maximum base-pair span L of 120 nt and a z-score cutoff of -2 was used. Setting the cutoff at -2 is for sure restrictive, but it should still cover a large fraction of the ncRNA repertoire. Both strands were considered. A total of 202,126 structures have been obtained. In comparison, the regular RNALfold returned a total of 1,387,136 structures, 824, 000 of which have a length $\geq$ 50 nt. The RNALfoldz output (a true subset of the RNALfold output) is only a forth of the regular RNALfold output.

The *E. coli* genome Genbank file lists 119 ncRNAs with a maximum length of 120 nt in its current annotation. To investigate the extent annotated ncRNAs are covered in the RNALfoldz output, we define for a RNALfold/RNALfoldz prediction to be counted as hit a minimal coverage of 90% of the ncRNA sequence. Giving this setup a total of 106 and 89 ncRNAs can be found in the RNALfold and RNALfoldz output, respectively. Detailed results for each RNA gene are shown in an online supplementary table. With a z-score cutoff of -2, 17 ncRNAs that were found by RNALfold are not in output set of RNALfoldz. The detection success is directly proportional to the reduction rate of the RNALfold output. Modulating the z-score cutoff affects both quantities (Fig. 1). The failure to detect the 13 ncRNAs that were missed by both RNALfold and RNALfoldz results from the fact that the RNALfold algorithm predicts only self-contained RNA structures. For example, the two ncRNA genes *rprA* and *ryeE* that have only low covering RNALfoldz hits, have indeed multi-component structures at the MFE level (abstract shape notation [GVR04]: `[][][][]`, `[][][]`). In these cases RNALfoldz will rather produce multiple hits than one single hit covering the whole ncRNA. Overall, our findings confirm that most *E. coli* small ncRNAs are indeed more thermodynamically stable than expected by chance and that the RNALfoldz algorithm is able to detect these structures efficiently.

We further investigated how precisely the RNALfoldz predictions map to the coordinates of the annotated ncRNAs. This is a legitimate issue, but one has to keep in mind that functional RNAs adopt their structure at the transcription level, while in this experiment we used the genomic sequence to detect these structures. So it might easily happen that the RNA is predicted in a bigger structural context than its actual size. The underlying dynamic programming algorithm is the same for RNALfold and RNALfoldz, and hence results discussed here do hold for both versions. In this work we define *noise* as the fraction of the RNALfoldz hit that does not overlap with the annotated ncRNA. In total, 34 out of

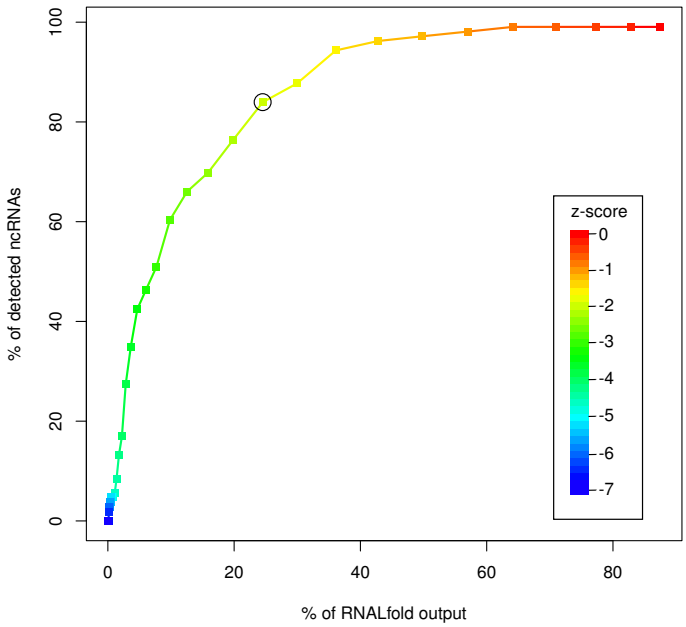


Figure 1: Non-coding RNA detection success vs. reduction of the `RNALfold` output. Given a z-score cutoff of 0 only one structure prediction is missed in the `RNALfoldz` output. With a z-score cutoff of -2 (circle) we see a four-fold reduction of the output, while at the same time covering 84% of the correct `RNALfold` ncRNA predictions.

the 89 predictions have less than 10% noise. Averaged over all hits ($\geq$ 90% coverage) we see *noise* of 18%. Using a classic sliding window approach with a length of 120 nt, one would expect more than 33% noise for a window containing a tRNA sequence (average length of tRNAs in E. coli: 78 nt). In the `RNALfoldz` output we find that 29 out of 73 tRNA predictions have less than 10% noise.

Finally, we address the significance of the predictions when compared to randomized controls. Therefore, we performed the same experiment on randomized sequences generated by (i) mononucleotide shuffling, (ii) simulation with an order-0 Markov model (mononucleotide frequencies) , and (iii) simulation with an order-1 Markov model (dinucleotide frequencies). Shuffling and simulations were done with `shuffle` from Sean Eddy's squid library using default parameters. A detailed comparison of the results of these four experiments is shown in Fig. 2. In general, we observe a tendency to more stable structures in the native sequence than in any randomized sequence. Structures with a z-score $\leq$ -8 are profoundly enriched in the native sequence, which might point to biological relevance of these structures. These are, however, extremes and most ncRNAs will have z-score values in a much higher range.

The value -2 for the z-score cutoff in this experiment was chosen arbitrarily. Moving to an even lower value than -2 will reduce the false discovery rate, but at the same time limit the number of ncRNAs that show such high thermodynamic stability. Using the `RNALfoldz` output from the experiment with randomized sequences (order-1 Markov model), we can calculate an empirical precision or positive predictive value (PPV), which is simply the

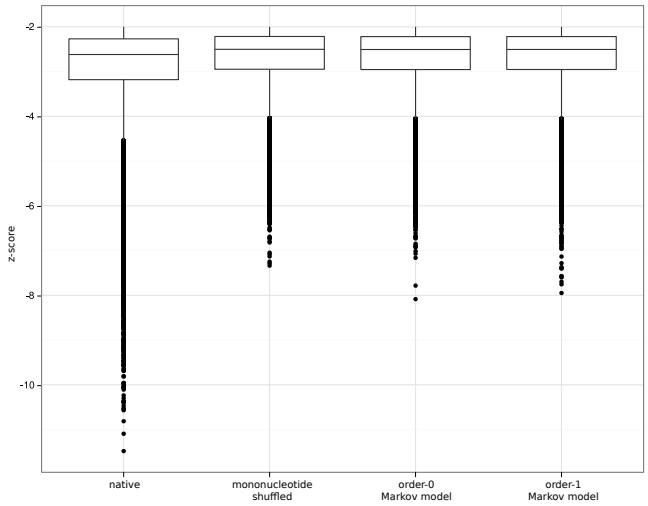


Figure 2: Comparison of the distribution of stable secondary structures from the native *E. coli* genome and randomized controls. The native *E. coli* sequence has a strong tendency to more stable local secondary structures. `RNALfoldz` predictions with a z-score below -8 are exclusively found in the native sequence.

proportion of true positives against all positive results. Assuming that thermodynamic stability is inherently linked to biologically function, we declare any `RNALfoldz` prediction with a z-score below a certain threshold from the native sequence and the randomized sequence as true positive and as false positive, respectively. Using then a PPV of 0.75, which corresponds to 25% estimated false positives, and, hence, a deduced z-score cutoff of -3.86 we can find 25 of the 106 annotated ncRNAs that are detectable with the `RNALfold` algorithm, while reducing the `RNALfoldz` to 21,715 predictions (3% of the `RNALfold` output). We further investigated if we can determine more specific z-score cutoffs when the `RNALfoldz` output is partitioned into different structural classes. This is motivated by the reasonable assumption that, for example, a short stable hairpin is more likely formed by chance than a stable, structurally more complex, multi-branched molecule. Hence, one would set different z-score cutoffs for different structural classes. To investigate this claim we map the MFE structures to the corresponding abstract RNA shape at the highest abstraction level. At this abstraction level only the helix nesting pattern is considered. As an example, the well-known cloverleaf structure of tRNA molecules is then simply represented as `[[][][]]`. The six major structural classes are shown in Tab. 1. We further partition structures according to their length into two classes *short* ($\leq$ 85 nt) and *long*.

Fig. 3 shows structure class specific precision values in dependency of the z-score, for those three classes that show the most deviation from the population precision. Using now class-specific z-score values when filtering the `RNALfoldz` output we can raise our prediction count from 25 to 38 ncRNAs, while keeping the expected false-positive rate fixed at 25%. The total number of `RNALfoldz` predictions increases slightly to 23,225.

Table 1: Major structural classes in the *E. coli* genome

frequency	abstract shape	length class	figure code	class specific z-score cutoff (PPV 0.75)
27%	`[[][]]`	long		-3.60
26%	`[[][]]`	short	SC2	-4.14
21%	`[]`	short	SC3	-4.16
7%	`[[][[][]]]`	long		-3.80
7%	`[[[][]][]]`	long		-3.74
4%	`[]`	long	SC6	-3.35
8%	rest			-3.35

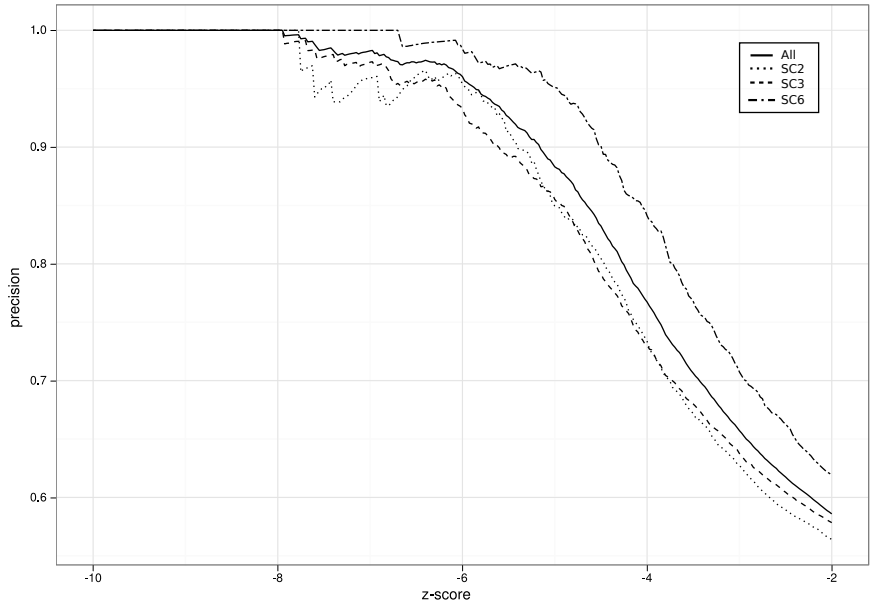


Figure 3: Precision values of different structural classes by the z-score. The solid line represents the whole `RNALfoldz` output.

3.2 Timing

The overall complexity $\mathcal{O}(N \times L^2)$ of the core algorithm does not change, the z-score calculation just adds a constant factor. We benchmarked both implementations on an Intel Quad Core2 CPU with 2.40 GHz. Detailed results are shown in Tab. 2.

At a maximal base-pair span of 120 nt `RNALfold` is able to scan at a speed of approx. 250 kb/min. At the same settings and with a minimal z-score cutoff of -2 scanning speed drops to 153 kb/min for `RNALfoldz`. The performance clearly depends on the number of times the support vector regression has to be called. When moving to a lower z-score cutoff of -4 the scanning speed increases to 207 kb/min.

Table 2: Timing results [sec] for RNALfold and RNALfoldz.

L	RNALfold	RNALfoldz		
		z-score $\leq$ -2	z-score $\leq$ -3	z-score $\leq$ -4
120	1,123	1,842	1,477	1,359
240	2,629	3,922	3,321	3,105

4 Discussion

In this work we have presented an extension of the RNALfold algorithm to predict thermodynamically stable, local RNA secondary structures. Using fast support vector regression models to calculate the z-score this approach comes with only a minor overhead in execution time compared to the former version, while yielding at the same time a much lower number of relevant structures. We have demonstrated that already with a z-score cutoff of -2, approx. 80% of the annotated *E. coli* small ncRNAs can be found in the RNALfoldz output. Comparison to randomized genome sequences showed that the native *E. coli* genome sequence has a strong bias to more stable secondary structures. This shift is, however, not significant enough to qualify RNALfoldz as a stand-alone RNA gene finder with an acceptable false discovery rate. We see the role of RNALfoldz mainly as a first filtering step in a cascade of computational ncRNA detection steps. In particular, the intersection of data from high throughput sequencing, promoter and transcription termination signals (see e.g. [SNS$^+$10]) or G+C content on AT rich genomes with RNALfoldz hits could be of value.

In this contribution, we assume that thermodynamic stability is inherently coupled to biological function. This is certainly true to some extent, but there are also a lot of RNA classes where stability is not a major issue for function, e.g. C/D box snoRNAs or ncRNAs that form interaction with other RNAs. It is therefore highly unlikely that these RNA classes will show up in the RNALfoldz output. In this context, RNALfoldz can, however, be used to define a set of highly stable loci which can then be excluded from further analysis.

It has been noted early on that thermodynamic stability alone is not a sufficient discriminator to distinguish ncRNAs from random sequences [RE00]. This is the main reason why most ncRNA gene finders rely solely on signals from evolutionary conservation of RNA secondary structures, or use thermodynamic stability only as an additional feature. These methods are clearly limited by the comparative genomics data available. Investigation of species that are distantly related to any species sequenced so far, or species specific RNA genes are, hence, out of scope for these methods. The RNALfoldz algorithm presented in this work will not be a magic tool suddenly shedding light on these dark areas. The search for extraordinarily stable structures, however, can help to give first clues to putatively functional RNA secondary structure elements, where other methods fail.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network III" and "regulatory non-coding RNAs". `RNALfoldz` is part of the Vienna RNA package and can be obtained free of charge at: `http://www.tbi.univie.ac.at/~ivo/RNA/`. An electronic supplement located at `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/RNALfoldz` compiles training and test data.

References

[CFKK05] P Clote, F Ferré, 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.

[CLS+90] J H Chen, S Y Le, B Shapiro, K M Currey, and J V Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Comput Appl Biosci*, 6(1):7–18, 1990.

[FGM05] E Freyhult, P P Gardner, and V Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241–241, 2005.

[GVR04] R Giegerich, B Voss, and M Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res*, 32(16):4843–4851, 2004.

[HFS+94] I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[HHS08] J Hertel, I L Hofacker, and P F Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2008.

[HPS04] 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.

[HWL+09] Y Horesh, Y Wexler, I Lebenthal, M Ziv-Ukelson, and R Unger. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics*, 10:76–76, 2009.

[JWW+07] P Jiang, H Wu, W Wang, W Ma, X Sun, and Z Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*, 35(Web Server issue):339–344, 2007.

[LLM02] S Y Le, W M Liu, and J V Maizel. A data mining approach to discover unusual folding regions in genome sequences. *Knowledge-Based Systems*, 15(4):243 – 250, 2002.

[LM89] S Y Le and J V Maizel. A method for assessing the statistical significance of RNA folding. *J Theor Biol*, 138(4):495–510, 1989.

[RE00] 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.

[SNS+10] J Sridhar, S R Narmada, R Sabarinathan, H Y Ou, Z Deng, K Sekar, Z A Rafi, and K Rajakumar. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*, 5(8), 2010.

[WHS05] S Washietl, I L Hofacker, and P F Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, 2005.

[WLX06] X F Wan, G Lin, and D Xu. Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. *J Bioinform Comput Biol*, 4(5):1015–1031, 2006.

[ZS81] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.

6

Arthropod 7SK RNA

Gruber AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF (2008) **Arthropod 7SK RNA.** Molecular Biology and Evolution, volume 25, issue 9, pages 1923-30.
DOI:10.1093/molbev/msn140

Authors' contributions: ARG initiated the study, performed the promoter screen, and did the structural analysis. ILH and PFS guided in study design. ARG, ILH, AM, and PFS wrote the manuscript. AM assisted in application of fragrep and helped to discover distant homologs. CK and WH performed the biological characterization.

Arthropod 7SK RNA

Andreas R. Gruber[*,a], Carsten Kilgus[b,h], Axel Mosig[b,c], Ivo L. Hofacker[a], Wolfgang Hennig[b,h], Peter F. Stadler[d,e,f,a,g,*],

[a] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[b] *Department of Combinatorics and Geometry (DCG), CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences (SIBS), Shanghai, China*

[c] *Max Planck Insitute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*

[d] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[e] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[f] *Fraunhofer Institut für Zelltherapie und Immunologie — IZI Perlickstrasse 1, D-04103 Leipzig, Germany*

[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[h] *Institut für Genetik, Johannes Gutenberg-Universität Mainz, J.-J.-Becherweg 32, 55099 Mainz,Germany*

[*] corresponding author: Tel.: +43 5277 52731 Fax: +43 4277 52793

e-mail: agruber@tbi.univie.ac.at

Abstract

The 7SK small nuclear RNA (snRNA) is a key player in the regulation of polymerase (pol) II transcription. The 7SK RNA was long believed to be specific to vertebrates where it is highly conserved. Homologs in basal deuterostomes and a few lophotrochozoan species were only recently reported. On longer timescales 7SK evolves rapidly with only few conserved sequence and structure motifs. Previous attempts to identify the Drosophila homolog thus have remained unsuccessful despite considerable efforts. Here we report on the discovery of arthropod 7SK RNAs using a novel search strategy based on pol III promoters, as well as the subsequent verification of its expression. Our results demonstrate that a 7SK snRNA featuring two highly structured conserved domains was present already in the bilaterian ancestor.

Key words: 7SK snRNA, homology search, non-coding RNA

Research Article *30 May 2008*

1 Introduction

The 7SK small nuclear RNA (snRNA) is one of the most highly abundant non-coding RNAs (ncRNAs) in vertebrate cells. Due to its abundance it has been known since the 1960s. Its function as a transcriptional regulator, however, has only recently been discovered: 7SK mediates the inhibition of transcription elongation factor P-TEFb, a critical regulator of RNA polymerase (pol) II transcription which stimulates the elongation phase (Nguyen et al., 2001; Yang et al., 2001; Michels et al., 2004; Blazek et al., 2005; Egloff et al., 2006; Peterlin and Price, 297-305; Krueger et al., 2008). In addition, 7SK RNA suppresses the deaminase activity of APOBEC3C and sequesters this enzyme in the nucleolus (He et al., 2006).

The pol III transcript with a length of about 330nt (Krüger and Benecke, 1987; Murphy et al., 1987) is highly conserved in vertebrates (Gürsoy et al., 2000). In contrast to the nearly perfect sequence conservation in jawed vertebrates, the 7SK RNA from the lamprey *Lampetra fluviatilis* differs in more than 30% of its nucleotide positions from its mammalian counterpart (Gürsoy et al., 2000). Based on several unsuccessful attempts to clone 7SK homologs, the molecule has long been believed to be vertebrate specific. In a recent contribution (Gruber et al., 2008), however, we reported on the computational detection and experimental verification of 7SK sequences from several basal deuterostomes as well as a few Lophotrochozoa. Direct experimental evidence is available for the hagfish *Myxine glutinosa*, the lancet *Branchiostoma lanceolatum*, and the snail *Helix pomatia*. In contrast, neither experimental cloning procedures nor computational homology search revealed a plausible 7SK candidate in *Drosophila melanogaster* or any other sequenced genome of an ecdysozoan.

In this contribution, we report on the computational discovery of the 7SK snRNA homologue in Drosphilidae and other arthropod genomes, on its bioinformatical characterization, and its subsequent verification in *Drosophila melanogaster*.

2 Materials and Methods

2.1 Sequence Data

Genomic sequences were downloaded from ENSEMBL (version 48, `www.ensembl.org`), the Joint Genome Institute (`www.jgi.doe.gov`), and the Broad Institute (`www.broad.mit.edu`) Web sites. Details on the assemblies used here are listed in the Electronic Supplement. 100nt upstream regions of the annotated U6 (CR31379, CR32867, CR31539) and U6atac (CR32989) RNAs were retrieved from FlyBase (`www.flybase.org`). Previously described 7SK sequences and their alignment were taken from Gruber et al. (2008).

2.2 Homology Search

From the 100nt upstream regions of the *Drosophila melanogaster* U6 and U6atac snRNAs, we generated a multiple sequence alignment using `MAFFT` (Katoh et al., 2002). Guided by previous findings (Mount et al., 2007; Hernandez Jr et al., 2007), we selected the search pattern such that it contained the conserved promoter region, two conserved Thymidine residues to guarantee distinguishability from pol II recognized PSE elements and the TATA-box. Then we scanned the *D. melanogaster* genome using `fragrep` (version 2) (Mosig et al., 2007) in position weight matrix mode. The sequence conservation pattern downstream of the resulting hits was visually inspected in the UC Santa Cruz (UCSC) genome browser (`genome.ucsc.edu`). Neoptera species were searched iteratively using the `blast` front-end at the FlyBase Web site, using previously identified hits as additional queries. In addition, we searched GenBank using NCBI's Web interface (`www.ncbi.nlm.nih.gov/blast`). Sean Eddy's `rnabob` (`selab.janelia.org/software.html`) was used for pattern-based RNA structure searches.

2.3 Sequence-Structure Alignments

Initial alignments were generated using `ClustalW` (Thompson et al., 1994), `dialign2` (Morgenstern, 1999; Morgenstern et al., 2006), and `MAFFT` (Katoh et al., 2002). Initial structure annotation was produced using `RNAalifold` (Hofacker et al., 2002). This information was used as the basis for a semimanual alignment edited in `emacs` using the `ralee` mode (Griffiths-Jones, 2005). The 5' and 3' domains were re-aligned using `locarna` (Will et al., 2007).

2.4 Northern Blot

Total RNA was isolated from *D. melanogaster* (Canton S) flies according to Chomczynski and Sacchi (1987). For Northern blots, 15 μg of total RNA were separated in 2% agarose-formaldehyde gels and blotted onto Hybond-N membrane (Roche) according to Sambrook et al. (2001). A DIG-labeled probe of 344nt of the 7SK RNA was obtained by amplification of the respective fragment on genomic DNA of *D. melanogaster* (Canton S) with the primers `CGATATTCAGGTAACTGCATCTG` (positions 35 to 58 in the predicted transcript) and `CGAAAATCCGAAGCTAAGCTACT` (positions 356 to 379) and the PCR DIG labeling mix (Roche, catalog number 11636090910). Hybridization was carried out in 5x standard saline citrate, 0.1%N-lauryl-sarcosine, 1% milk powder, 0.02% SDS at 65° C overnight. The membranes were washed with 0.1 M Tris/HCl (pH 7.5), 0.15 M NaCl, 0.3% Tween 20. The same buffer with additional 1% milk powder was used for the blocking. For detection, we used the alkaline phophatase-conjugated anti-DIG-antibody (Roche, catalog number # 11093657910) in a dilution of 1:7500 in the same buffer at room temperature for 2h. For detection, 7ml AP-buffer (0.1M Tris/HCl, 0.1M NaCl, 5mM MgCl2, pH 9.5) was freshly mixed with 14μl NBT (Nitro-Blue Tetrazolium Chloride; 100mg/ml) and 21μl BCIP (5-Bromo-4-Chloro-3 p-Toluidine Salt; 50mg/ml). The substracte reaction was stopped when a signal appeared (after 20 to 30 min) by adding ddH_2O to decrease the pH-value.

3 Results

3.1 Initial Search

Because direct homology search had failed previously, we employed a different strategy. The snRNAs, including the 7SK snRNA, exhibit a characteristic promoter structure (Hernandez, 2001) that is fairly well conserved in evolution. The spliceosomal snRNAs had recently been studied in great detail in Drosophilidae (Mount et al., 2007; Hernandez Jr et al., 2007), and their promoter sequence motifs are known in detail for most of the 12 sequenced drosophilid fly species. The 7SK snRNA has a canonical pol III type 3 promoter in vertebrates, see Bannister et al. (2007) and the references therein. We thus derived a search pattern for canonical type 3 pol III promoters, Fig. 1, using a region of 100nt upstream of the U6 and U6atac snRNAs as template.

The pattern was used to search the *D. melanogaster* genome. In addition to recovering the U6 and U6atac snRNAs, we uncovered 4 hits, summarized in Tab. 1. One of them belongs to a known snoRNA previously described in Huang et al. (2005). Pol III regulated expression of snoRNAs has not been described in Drosophilidae so far. The observation is not unexpected, however,

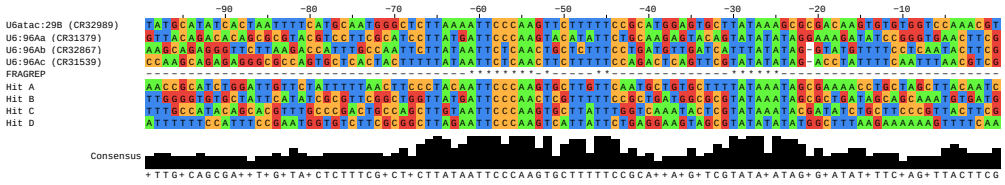


Fig. 1. Upstream regions of known U6 and U6atac genes and the four candidate sequences listed in Tab. 1. The `fragrep` pattern used to scan the genome of *D. melanogaster* is indicated in the middle.

Table 1
Characterization of conserved loci with putative U6-like snRNA promoter motifs. Evidence for evolutionarily conserved secondary structure is taken from a recent `RNAz`-based survey. The numbers refer to the loci listed in the Supplemental Material of Rose et al. (2007). Evidence from ChIP-on-chip data for binding of TRF1 and BRF refers to the loci listed in the Supplemental Material of Isogai et al. (2006).

	Location	`RNAz`	pol-III	Note	Ref.
A	*3L*:7 632 840-7 632 900(+)	—	—	CR34703 C/D snoRNA Me18S-A1806-RA	(Huang et al., 2005)
B	*3R*:3 300 270-3 300 900(-)	1077	TRF1 (1582) BRF (1580)	CR33925 smnRNA:331-RA	(Yuan et al., 2003)
C	*3R*:19 555,800-19 556 250(-)	7371, 7372	BRF (9494) BRF (9495)	CR33682 smnRNA:342 RNAse MRP	(Yuan et al., 2003; Woodhams et al., 2007; Piccinelli et al., 2005)
D	*X*:21 308 600-21 308 750(+)	—	—	—	—

because pol III transcription of snoRNAs has been observed previously in *Saccharomyces cerevisiae* (Moqtaderi and Struhl, 2004). The candidate located on the *X* chromosome shows no direct evidence for pol III transcription in the study (Isogai et al., 2006).

Two candidates on chromosome *3R* overlap small nonmessenger RNAs cloned in an experimental survey of small RNAs in *D. melanogaster* (Yuan et al., 2003). While both Woodhams et al. (2007) and Piccinelli et al. (2005) list candidate **C** as RNAse MRP, no further annotation is available for candidate **B**. A comparison with a recent computational survey of structure conserved ncRNAs in flies shows that both loci have been detected by `RNAz` (Rose et al., 2007). Furthermore, there is direct evidence that these regions are transcribed by pol III. Isogai et al. (2006) showed that unlike in most other eukaryotes,

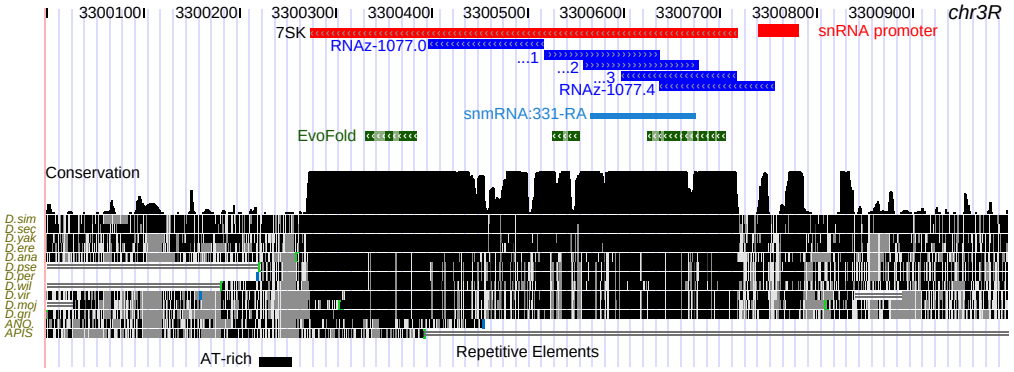


Fig. 2. Detailed genomic view of the 7SK candidate **B** at *3R*(3.3M). Adapted from a USCS Genome Browser view. The upper bar indicates the predicted 7SK transcript along with the snRNA promoter element. The predicted transcript overlaps the conserved secondary structures reported by `RNAz` and `evofold` as well as the fragment cloned in Yuan et al. (2003).

TRF1/BRF binding appears responsible for the initiation of all classes of pol III transcription and they have mapped TRF1 and BRF binding sites in the respective sites.

3.2 Homology Search

Candidate **B**, located on chromosome *3R* at 3.3M, Fig. 2, shows strong evidence for pol III transcription, strong evidence for an evolutionarily well conserved secondary structure, and a characteristic T-rich region indicative of a pol III terminator. With an overall length of about 450nt, the conserved sequence is only slightly longer than the previously known 7SK snRNAs (Gruber et al., 2008). Note however, that the ends of the transcripts cannot be predicted accurately. In *D. melanogaster*, an AT-rich low-complexity region is located immediately downstream of the annotated conserved region, which could be (partially) transcribed. The human 7SK, for instance, shows some variability in the exact position of its 3' end, which consists of a short U-rich tail of length 5-7. In addition, a fraction of the human transcripts are adenylated posttranscriptionally (Sinha et al., 1998). For the bioinformatic analysis, we defined the 3' end of the arthropod candidate sequence before the low complexity region.

The high level of sequence conservation in Drosophilidae promoted us to search for homologs in additional arthropod genomes. In Neoptera species, these could easily be retrieved by iterative `blast` searches. As `blast` failed to recover a homologue in *Ixodes scapularis*, we constructed a `fragrep` pattern

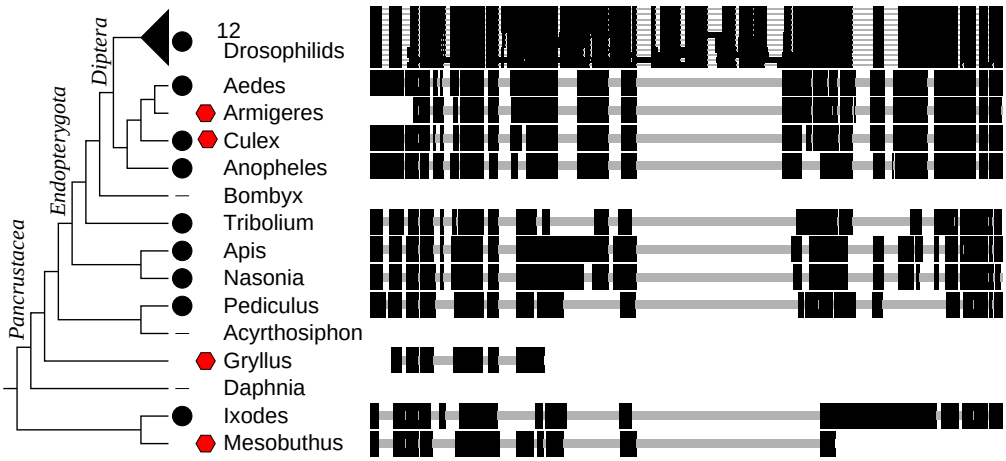


Fig. 3. Phylogenetic distribution of 7SK candidate sequences. A bullet indicates a match in the genomic sequence, the hexagons for *Armigeres*, *Culex*, *Gryllus* and *Mesobuthus* refer to partial ESTs. For complete genomic sequences a sketch of the alignment structure highlights the large insertion domains in Pancrustacea and in Drosophilidae in particular. Aligned blocks are shown in black, gray bars indicate gaps in the alignment, missing sequence data adjacent to EST regions appear white. The underlying tree is composed from the genome-wide near intron positions (Krauss et al., 2008), a phylogeny of mosquitoes (Harbach and Kitching, 1998) and two recent studies of arthropod phylogeny (Cameron et al., 2004; Kjer, 2004), for the relationships outside the Endopterygota.

from already identified arthropod sequences (see Electronic Supplement [1]).

We recovered candidate sequences from most of the available arthropod genomes, with the notable exceptions of the lepidopteran *Bombyx mori* and the aphid *Acyrthosiphon pisum*, and the crustacean *Daphnia pulex* see Fig. 3 and Electronic Supplement. In these cases, it is plausible to assume that no candidate was found due the quality of the current draft assemblies, although we cannot rule out that the sequence is too derived to be recognizable by our search methods.

In addition to genomic DNA, we also searched the NCBI EST database using all the genomic hits as `blast` queries. This resulted in some evidence for expression of the 7SK candidates beyond the fragments reported in Yuan et al. (2003).

A `blast` search of the NCBI NR and EST collections revealed additional evidence for transcription of this locus in several species, namely *Culex pipiens* (multiple ESTs from an unpublished EST project), *Armigeres subalbatus* (a single EST from the ref. Aliota et al. (2007)), *Gryllus bimaculatus* (a single un-

[1] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-008/`

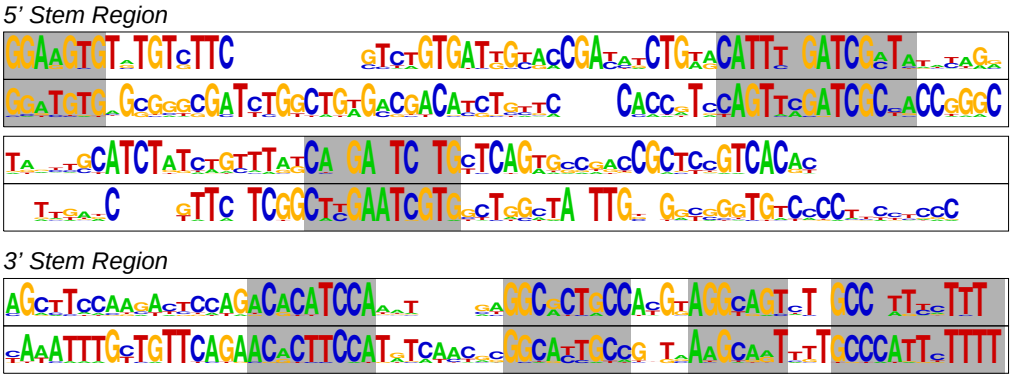


Fig. 4. Sequence Logos (Schneider and Stephens, 1990) representing an alignment of the 7SK consensus sequences reported in Gruber et al. (2008) *(upper sequences)* and the consensus of the arthropod candidates *(lower sequences)*. The logos were computed using `aln2pattern` (Mosig et al., 2007) separately for the two groups of sequences. Common well-conserved motifs are shaded.

published cDNA), *Mesobuthus gibbosus* (five ESTs from an unpublished EST project). GenBank accession numbers are listed in the Electronic Supplement.

A multiple sequence alignment (Electronic Supplement) shows that the candidate sequences have two well conserved domains located at the 5' and the 3' termini, whereas the intermediate portion appears to evolve rapidly and contains large insertions and deletions, see also Fig. 3. Overall, this organization conforms the observations for the known 7SK sequences (Gruber et al., 2008): the highest sequence conservation among the known 7SK snRNAs is also observed in the 5' and 3' hairpin regions.

Fig. 4 demonstrates substantial similarities between the 7SK snRNAs reported in (Gruber et al., 2008) and the candidate sequences discovered in this contribution. The domains with similar sequences are located in a similar structural context, see below.

3.3 Structural Characterization

We therefore constructed a structural consensus model of the arthropod sequences and compared this with the structural models derived in Gruber et al. (2008). Two distinct secondary structure elements are highly conserved throughout vertebrates (Egloff et al., 2006): a 5'-terminal hairpin structure that binds both HEXIM1 and P-TEFb, and a 3'-terminal hairpin that interacts with P-TEFb only. A recent study (Krueger et al., 2008) revealed that 7SK snRNA is stably associated with LARP7, a close relative of La, which is associated with many nascent pol III transcripts, including 7SK snRNA

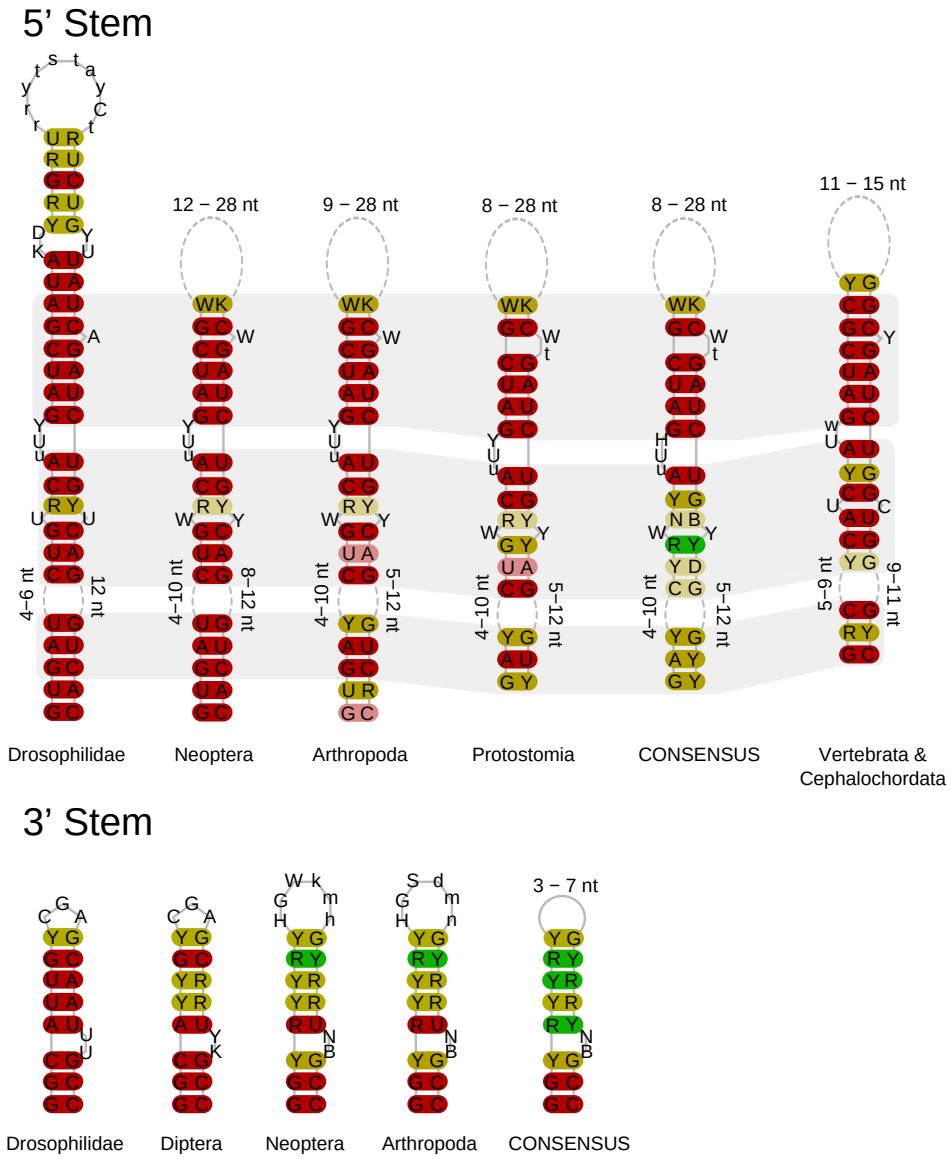


Fig. 5. Comparison of structural motifs of 7SK snRNAs. Conserved nucleotides in stems are shown in red, while ocher (green) indicates two (three) different supporting compensatory mutations. Pale colors indicate that a base-pair cannot be formed by all the sequences. Lower case letters imply a deletion in some sequences. The variable-size regions close to the hairpin loops, which have no clear consensus folds, are drawn as dashed ellipses. Correspondences of helices are highlighted by a gray background.

(Hogg and Collins, 2007). It is unknown, however, how LARP7 binds to 7SK. Interestingly, LARP7 has a well-known homologue in *Drosophila melanogaster* (Krueger et al., 2008), namely *mxc* (*multi sexcomb reduced*), a member of the

polycomb group regulating gene expression during development (Rajasekhar and Begemann, 2007).

Structural alignments of identified candidate sequences based on previously published 7SK sequence data (Gruber et al., 2008) were generated for both the 5' region and the terminal 3' region. Independent models were generated for the 5' regions of Drosophilidae, Neoptera, and all Arthropoda, respectively. Using also the previously published sequence data on 7SK (Gruber et al., 2008), we furthermore constructed combined models for Arthropoda+Lophotrochozoa, and Vertebrata+Cephalochordata. Their combination was then used to suggest a consensus model.

Overall, the secondary structure of the 5'-stem region of arthropods is quite similar to its vertebrate and lophotrochozoan counterpart. While the lower part of the stem-loop structure is very similar in all know sequences, the closing hairpin loop varies considerably in size and base composition. In drosophilid flies, this stem is extended by a helical element consisting of five base-pairs (supported by several compensatory mutations), while otherwise the terminal loop consists of 8-15 nt. The hairpin loop is closed by a stem that is highly conserved in both sequence and structure. This stacked region is only interspersed by a positionally conserved bulge loop. The outer part of this stem comprises the **GAUC-GAUC** motif enclosed by positionally conserved bulge loops. The functional importance of this motif is discussed in detail in Egloff et al. (2006). The consensus model shows that there exists only a structural, not a strong sequence constraint on the other elements of the 5'-stem region.

Both helices in the 3'-stem region are supported by many compensatory mutations. The position of the bulge loop as well as the position of the hairpin loop are highly conserved. While both Vertebrata and Lophotrochozoa show a sequence constraint in the hairpin loop, this does not seem to be the case in arthropods. For Diptera, the hairpin loop is reduced to a minimal size of three nt. Based on the structure model for Deuterostomia and Lophotrochozoa suggested in Gruber et al. (2008) and the arthropod model derived here, we suggest a universal structural model of the 3' terminal stem.

The sequence similarities, Fig. 4, the very similar structural organization of both the 5' and the 3' conserved domains, Fig. 5, and the fact the *Drosophila* loci have the typical organization of a pol III transcript with a type 3 pol III promoter demonstrate beyond reasonable doubt that the *3R*(3.3M) locus **B** indeed harbors a 7SK homologue.

The conserved elements in Figs. 4 and 5 can in principle be used to construct sequence or sequence/structure patterns for further homology searches. Attempts to find a 7SK homologue in the shotgun traces of the *Daphnia pulex* genome remained unsuccessful, however, with both `fragrep` and `rnabob`.

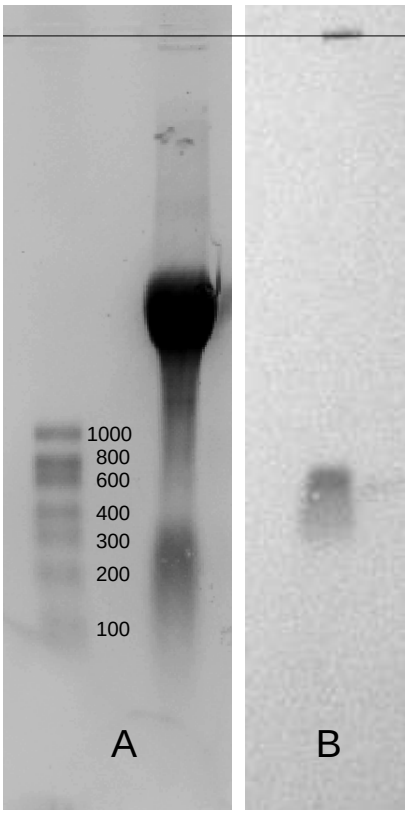


Fig. 6. (A) Electrophoretic separation of total RNA from *D. melanogaster* adults in a 2% agarose gel (ethidium bromide stained), (B) Northern blot hybridized with the 7SK DNA probe of length 344bp.

3.4 Expression in Drosophila melanogaster

In order to verify the expression of the 7SK locus, for which a previous study had already reported a partial transcript (Yuan et al., 2003), we performed a standard Northern blot experiment. We used a 344bp probe located between position 35 and 379 within the 445nt long predicted 7SK gene.

The DIG-labelled PCR fragment was hybridized to a blot of total RNA from flies, separated on an agarose gel. For the detection of the hybrids we used alkaline phosphatase-labelled anti-DIG antibody for the reaction with NBT/BCIP as substrate which yields a purple reaction product.

Fig. 6 shows the electrophoretic separation of the total RNA and the Northern blot, which resulted in a clear single band. Comparison between the marker in the gel and the blot shows that detected transcript appears somewhat larger than the predicted 7SK gene.

4 Discussion

Homology search for non-coding RNAs has turned out to be a surprisingly hard problem in bioinformatics. Standard methods of homology search often fail due to large variations in sequence length and oftentimes extremely poor sequence conservation, see, e.g. (Mosig et al., 2007; Gruber et al., 2008; Xie et al., 2008) for recent examples. Indeed, the arthropod 7SK RNAs reported in this contribution were not discovered by straightforward search but rather by an indirect strategy that uses the typical promoter structure of 7SKs (Bannister et al., 2007), experimental evidence for pol III transcripts in *Drosophila melanogaster* (Isogai et al., 2006), sequence conservation (*Drosophila* 12 Genomes Consortium, 2007) and *de novo* prediction of evolutionarily conserved RNA secondary structure (Rose et al., 2007). Once the representative sequences in Drosophilidae were found, conventional `blast`-based searches revealed additional homologs, which could then be used as starting-point for pattern-based searches that resulted in 7SK sequences spanning most of the arthropod tree.

A detailed analysis of sequence motifs and the construction of RNA secondary models based on a combination of thermodynamic folding and sequence covariation demonstrates that our candidate sequences share key features, namely the the 5' and 3' stem regions, with deuterostome and lophotrochozoan 7SK RNAs, demonstrating that we have indeed found the 7SK snRNA.

A search of EST and cDNA data revealed evidence for transcription of the 7SK locus in several species across the Arthropoda. We furthermore performed a Northern blot to verify the 7SK in Drosophilidae directly. The resulting transcript is somewhat longer than expected. There is, however, an AT-rich repetitive region immediately downstream of the 7SK RNA which may be at least partially transcribed. Human 7SK ends are known to be heterogeneous (Sinha et al., 1998). Furthermore, an extension of pol-III transcripts beyond a putative 3'end inferred from homology search was for instance observed in mouse and rat vault RNAs (compared to most other mammalian vault RNAs) (Vilalta et al., 1994; Kickhoefer et al., 2003). The smear observed in the Northern blot below the major signal might indicate the presence of a series of smaller transcripts due to earlier termination.

Our results demonstrate that a 7SK snRNA featuring two highly structured conserved domains was present already in the bilaterian ancestor. This suggests that also the function of the 7SK snRNA is evolutionary conserved despite a recent report that the inhibition of P-TEFb by the peptide Pgc is RNAse insensitive in primordial germ cells (Hanyu-Nakamura et al., 2008). The hypothesis of functional conservation is further supported by the observation that all major protein components of the human 7SK snRNP (P-TEFb, HEXIM, and LARP7) have homologs in *D. melanogaster* (P-TEFb, CG3508,

and *mxc*, respectively). More generally, the presumably ancient origin of 7SK snRNA and the ubiquitous role of 6S RNA as transcriptional regulator in bacteria (Barrick et al., 2005) suggests that the recently uncovered variety of non-coding RNAs regulating the transcriptional machinery (Goodrich and Kugel, 2006; Barrandon et al., 2008) may also be evolutionary ancient (Lu et al., 2008).

Supplemental Information

An Electronic Supplement located at `http:www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-008/` compiled sequence data, primers, alignments in machine-readable form, and `fragrep2` search patterns.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network II" and "non coding RNA", as well as by the Priority Program *SPP 1258: Sensory and regulatory RNAs in Prokaryotes* of the Deutsche Forschungsgemeinschaft (DFG), and the PICB. PFS thanks the CAS-MPG Partner Institute for Computational Biology in Shanghai for its hospitality in spring 2008, where much of this work was performed.

References

Aliota, M. T., J. F. Fuchs, G. F. Mayhew, C. C. Chen, and B. M. Christensen. 2007. Mosquito transcriptome changes and filarial worm resistance in *Armigeres subalbatus*. BMC Genomics **8**:463.

Bannister, S. C., T. G. Wise, D. M. Cahill, and T. J. Doran. 2007. Comparison of the chicken 7SK and U6 RNA polymerase III promoters for hairpin RNA expression. BMC Biotech **7**:79.

Barrandon, C., B. Spiluttini, and O. Bensaude. 2008. Non-coding RNAs regulating the transcriptional machinery. Biol. Cell **100**:83–95.

Barrick, J. E., N. Sudarsan, Z. Weinberg, W. L. Ruzzo, and R. R. Breaker. 2005. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. RNA **11**:774–784.

Blazek, D., M. Barboric, J. Kohoutek, I. Oven, and B. M. Peterlin. 2005. Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb. Nucleic Acids Res. **33**:7000–7010.

Cameron, S. L., K. B. Miller, C. A. D'Haese, M. F. Whiting, and S. C. Barker. 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Arthropoda). Cladistics **20**:534–557.

Chomczynski, P., and N. Sacchi. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem **162**:156–159.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature **450**:203–218.

Egloff, S., E. Van Herreweghe, and T. Kiss. 2006. Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. Mol. Cell. Biol. **26**:630–642.

Goodrich, J. A., and J. F. Kugel. 2006. Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol **7**:612–616.

Griffiths-Jones, S. 2005. `RALEE`—RNA ALignment editor in `Emacs`. Bioinformatics **21**:257–259.

Gruber, A. R., D. Koper-Emde, M. Marz, H. Tafer, S. Bernhart, G. Obernosterer, A. Mosig, I. L. Hofacker, P. F. Stadler, and B.-J. Benecke. 2008. Invertebrate 7SK snRNAs. J. Mol. Evol. **107-115**:66.

Gürsoy, H.-C., D. Koper, and B.-J. Benecke. 2000. The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). J. Mol. Evol. **50**:456–464.

Hanyu-Nakamura, K., H. Sonobe-Nojima, A. Tanigawa, P. Lasko, and A. Nakamura. 2008. *Drosophila* Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. Nature **451**:730–733.

Harbach, R. E., and I. J. Kitching. 1998. Phylogeny and classification of the Culicidae (Diptera). Syst. Entomol. **23**:327–370.

He, W. J., R. Chen, Z. Yang, and Q. Zhou. 2006. Regulation of two key nuclear enzymatic activities by the 7SK small nuclear RNA. Cold Spring Harb Symp Quant Biol. **71**:301–311.

Hernandez, N. 2001. Small Nuclear RNA Genes: a Model System to Study Fundamental Mechanisms of Transcription. J. Biol. Chem. **276**:26733–26736.

Hernandez Jr, G., F. Valafar, and W. E. Stumph. 2007. Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity. Nucleic Acids Res. **35**:21–34.

Hofacker, I. L., M. Fekete, and P. F. Stadler. 2002. Secondary Structure Prediction for Aligned RNA Sequences. J. Mol. Biol. **319**:1059–1066.

Hogg, J. R., and K. Collins. 2007. RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. RNA **13**:868–880.

Huang, Z. P., H. Zhou, H. L. He, C. L. Chen, D. Liang, and L. H. Qu. 2005. Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. RNA **11**:1303–1316.

Isogai, Y., S. Takada, R. Tjian, and S. Keles. 2006. Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. EMBO J .

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**:3059–3066.

Kickhoefer, V. A., N. Emre, A. G. Stephen, M. J. Poderycki, and L. H. Rome. 2003. Identification of conserved vault RNA expression elements and a non-expressed mouse vault RNA gene. Gene **309**:65–70.

Kjer, K. M. 2004. Aligned 18S and Insect Phylogeny. Syst. Biol. **53**:506–514.

Krauss, V., C. Thümmler, F. Georgi, J. Lehmann, P. F. Stadler, and C. Eisenhardt. 2008. Near intron positions are reliable phylogenetic markers: An application to Holometabolous Insects. Mol. Biol. Evol. Doi:10.1093/molbev/msn013.

Krueger, B. J., C. Jeronimo, B. B. Roy, et al. 2008. LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated. Nucleic Acids Res. .

Krüger, W., and B. J. Benecke. 1987. Structural and functional analysis of a human 7 S K RNA gene. J. Mol. Biol. **195**:31–41.

Lu, J., Y. Shen, Q. Wu, S. Kumar, B. He, S. Shi, R. W. Carthew, S. M. Wang, and C. I. Wu. 2008. The birth and death of microRNA genes in *Drosophila*. Nat Genet **40**:351–355.

Michels, A. A., Q. Fraldi, A. Li, T. E. Adamson, et al. 2004. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. **23**:2608–2619.

Moqtaderi, Z., and K. Struhl. 2004. Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes. Mol. Cell Biol. **24**:4118–4127.

Morgenstern, B. 1999. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**:211–218.

Morgenstern, B., S. J. Prohaska, D. Pohler, and P. F. Stadler. 2006. Multiple sequence alignment with user-defined anchor points. Algo. Mol. Biol. **1**:6.

Mosig, A., J. L. Chen, and P. F. Stadler. 2007. Homology Search with Fragmented Nucleic Acid Sequence Patterns. In: Giancarlo, R., and S. Hannenhalli, editors, Algorithms in Bioinformatics (WABI 2007), volume 4645 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Verlag, 335–345.

Mount, S. M., V. Gotea, C.-F. Lin, K. Hernandez, and W. Makałowski. 2007. Spliceosomal small nuclear RNA genes in 11 insect genomes. RNA **13**:5–14.

Murphy, S., C. Di Liegro, and M. Melli. 1987. The *in vitro* transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter. Cell **51**:81–87.

Nguyen, V. T., T. Kiss, A. A. Michels, and O. Bensaude. 2001. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature **414**:322–325.

Peterlin, B. M., and D. H. Price. 297-305. Controlling the elongation phase of transcription with P-TEFb. Mol. Cell. **2006**:23.

Piccinelli, P., M. A. Rosenblad, and T. Samuelsson. 2005. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res **33**:4485–4495.

Rajasekhar, V. K., and M. Begemann. 2007. Concise review: roles of polycomb group proteins in development and disease: a stem cell perspective. Stem Cells **25**:2498–2510.

Rose, D. R., J. Hackermüller, S. Washietl, S. Findeiß, K. Reiche, J. Hertel, P. F. Stadler, and S. J. Prohaska. 2007. Computational RNomics of Drosophilids. BMC Genomics **8**:406.

Sambrook, J., D. Russell, and J. Sambrook. 2001. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press.

Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res **18**:6097–6100.

Sinha, K. M., J. Gu, Y. Chen, and R. Reddy. 1998. Adenylation of small RNAs in human cells: Development of a cell-free system for accurate a adenylation on the 3'-end of human signal recognition particle RNA. J. Biol. Chem. **273**:6853–6859.

Thompson, J. D., D. G. Higgs, and T. J. Gibson. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. Nucl. Acids Res. **22**:4673–4680.

Vilalta, A., V. A. Kickhoefer, L. H. Rome, and D. L. Johnson. 1994. The rat vault RNA gene contains a unique RNA polymerase III promoter composed of both external and internal elements that function synergistically. J Biol Chem. **269**:29752–29759.

Will, S., K. Missal, I. L. Hofacker, P. F. Stadler, and R. Backofen. 2007. Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. PLoS Comp. Biol. **3**:e65.

Woodhams, M. D., P. F. Stadler, D. Penny, and L. J. Collins. 2007. RNase MRP and the RNA processing cascade in the eukaryotic ancestor. BMC Evol Biol **7 Suppl 1**.

Xie, M., A. Mosig, X. Qi, Y. Li, P. F. Stadler, and J. J.-L. Chen. 2008. Size Variation and Structural Conservation of Vertebrate Telomerase RNA. J. Biol. Chem. **283**:2049–2059.

Yang, Z., Q. Zhu, K. Luo, and Q. Zhou. 2001. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature **414**:317–322.

Yuan, G., C. Klämbt, J. P. Bachellerie, J. Brosius, and A. Hüttenhofer. 2003. RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. Nucleic Acids Res. **31**:2495–2507.

7

Nematode sbRNAs: homologs of vertebrate Y RNAs

Boria I*, Gruber AR*, Tanzer A*, Bernhart SH, Lorenz R, Mueller MM, Hofacker IL, Stadler PF (2010) **Nematode sbRNAs: homologs of vertebrate Y RNAs.** Journal of Molecular Evolution, volume 70, issue 4, pages 346-58.
DOI: 10.1007/s00239-010-9332-4
* These authors contributed equally to this work.

Authors' contributions: PFS initiated the study. ILH and PFS guided in study design. ARG, AT and PFS wrote the manuscript. IB, ARG, TA, SH, LR performed the initial sequence based searches. ARG developed and performed the promoter screen. ARG and IB performed homology screens with `RNABOB`. ARG developed and implemented the scoring schema for `RNABOB`. ARG and IB structurally characterized sbRNAs. ARG found the link to Y RNAs. AT described the evolutionary history of sbRNAs. MM assisted in biological characterization.

J. Mol. Evol. manuscript No.
(will be inserted by the editor)

Ilenia Boria*, Andreas R. Gruber*, Andrea Tanzer*, Stephan H. Bernhart, Ronny Lorenz, Michael M. Mueller, Ivo L. Hofacker, Peter F. Stadler

Nematode sbRNAs: homologs of vertebrate Y RNAs

March 10, 2010

Abstract Stem-bulge RNAs (sbRNAs) are a group of small, functionally yet uncharacterized noncoding RNAs first described in *C. elegans*, with a few homologous sequences postulated in *C. briggsae*. In this study we report on a comprehensive survey of this ncRNA family in the phylum Nematoda. Employing homology search strategies based on both sequence and secondary structure models and a computational promoter screen we identified a total of 240 new sbRNA homologs. For the majority of these loci we identified both promoter regions and transcription termination signals characteristic for pol-III transcripts. Sequence and structure comparison with known RNA families revealed that sbRNAs are homologs of vertebrate Y RNAs. Most of the sbRNAs show the characteristic Ro protein binding motif, and contain a region highly similar to a functionally required motif for DNA replication previously thought to be unique to vertebrate Y RNAs. The single Y RNA that was previously described in *C. elegans*, however, does not show this motif, and in general bears the hallmarks of a highly derived family member.

Keywords
sbRNA, nematodes, Y RNA, homology search, noncoding RNA

I. Boria
Department of Medical Sciences and Interdisciplinary Research Centre for Autoimmune Diseases, Università del Piemonte Orientale, via Solaroli 17, I-28100 Novara, Italy.

I. Boria, A.R. Gruber, A. Tanzer, S.H. Bernhart, R. Lorenz, I.L. Hofacker, P.F. Stadler
Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.
E-mail: {ilenia,agruber,at,berni,ronny,ivo}@tbi.univie.ac.at

A.R. Gruber, A. Tanzer, P.F. Stadler
Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

M.M. Mueller
Department of Chromosome Biology, Max F. Perutz Laboratories, University of Vienna, A-1030 Vienna, Austria.

P.F. Stadler
Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
Fraunhofer Institut für Zelltherapie und Immunologie (IZI), Perlickstraße 1, D-04103 Leipzig, Germany.
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.
E-mail: stadler@bioinf.uni-leipzig.de
*These authors contributed equally.

Introduction

Stem-bulge RNAs (sbRNAs) were discovered in the nematode *C. elegans* three years ago in a systematic screen of a ncRNA-specific full-length cDNA library by Deng et al. (2006). This initial study identified 9 distinct members of the family. In a subsequent contribution, Aftab et al. (2008) annotated three additional experimentally verified ncRNAs as sbRNAs. These seed sequences are listed in Tab. 1. They share two conserved internal motifs at the 5'- and 3'-end of the molecules, respectively. Computational predictions showed that these regions are able to form a long stem interrupted by a small bulge. The term "stem-bulge RNA" was coined because of this feature (Deng et al. 2006). A BLAST-based comparison with the *C. briggsae* genome revealed eleven putative homologs (Deng et al. 2006), providing further support for the stem-structure and indicating that the loop regions evolve rapidly.

The sbRNAs in *C. elegans* as well as their *C. briggsae* homologs show a common promoter structure consisting of a proximal sequence element B (PSE B) and a TATA-box (Deng et al. 2006). This type of pol-III promoter is closely related to that of snRNAs (Hernandez 2001), from which it differs by the lack of the conserved PSE A box in the proximal element, see Fig. 1 top. In a subsequent, detailed analysis of the sbRNA promoter, Li et al. (2008) showed that in contrast to the other promoters analyzed, transcription – albeit reduced by 30 to

Table 1 Seed set of sbRNAs.
All twelve sbRNAs are found in the ncRNA set identified by Deng et al. (2006). Ref. **b** indicates that they were first annotated as sbRNA by Aftab et al. (2008). The sequences marked **c** were also reported in Zemann et al. (2006). RNAi experiments were conducted for sequences marked **d** (Kamath et al. 2003) and **e** (Sönnichsen et al. 2005). A Y RNA homolog computationally predicted by Perreault et al. (2007) is marked by **f**. Column *L* denotes the length.

Name	Wormbase	Acc.No.	*L*	Refs.
CeN71	F08G2.13	AY948635	74	**c**
CeN72	–	AY948636	98	
CeN73-1	–	AY948637	133	
CeN73-2	–	AY948638	131	
CeN74-1	M163.13	AY948639	79	**c**
CeN74-2	M163.12	AY948640	77	**c**
CeN75	–	AY948593	70	
CeN76	W01D2.8	AY948641	77	
CeN77	fragmented	AY948602	69	
CeN135	F08G2.12	AM286261	67	**b**,**d**
CeN133	C15H11.12	AM286259	95	**b**,**d**, **e**
CeN134	F35E12.11	AM286260	119	**b**,**f**

50% – was detectable when only one of the two parts of the promoter (either PSE B or TATA-box) was present. Taken together with the fact that sbRNAs are uncapped and terminate with a poly-U stretch, these observations leave little doubt that sbRNAs are transcribed by RNA polymerase III.

Most sbRNAs are differentially expressed in developmental stages. The highest levels of expression have been found in mature adult worms, dauer larvae and especially worms after heat shock (Deng et al. 2006). In an unrelated study focusing on the snoRNAs complement of *C. elegans*, Zemann et al. (2006) confirmed two of Deng's sbRNAs.

For two sbRNAs (CeN135 and CeN133), along with almost 20,000 other genes, knock-down experiments were performed (Kamath et al. 2003). No phenotype was reported for these two knock-downs. CeN133 was also knocked down in a study by Sönnichsen et al. (2005), again with no visible phenotype. Considering latest results on the efficiency of RNAi on ncRNAs (Ploner et al. 2009) it has to be questioned if sbRNA expression levels were sufficiently decreased to see an effect. Furthermore, functionally required motifs may reside in the highly structurally conserved stem common to all sbRNAs. It is plausible, therefore, that other sbRNAs may functionally compensate for the reduced levels of a particular paralog.

A first attempt to gain insight into the putative biological functions of sbRNAs is reported by Aftab et al. (2008). Some sbRNAs showed increased levels of expression after depletion of the protein components of the snoRNPs. A detailed understanding of these findings is still missing and, up to now, biological functions and processes the sbRNAs are involved in remain to be uncovered.

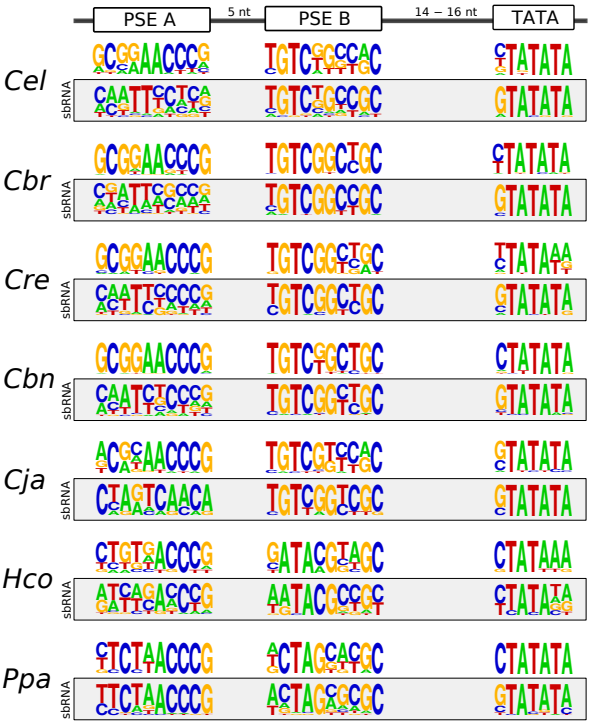


Fig. 1 Comparison of promoter elements of sbRNAs to other pol-III transcripts. The upper row for each species shows sequence logos (Crooks et al. 2004) of the promoter motifs for other pol-III transcripts (U6 snRNA, RNase P, RNase MRP, tRNA-SeC, Y RNAs), while the lower row denotes the corresponding elements for sbRNAs. High similarity is observed for the PSE B and the TATA-box for all species, while high similarity for PSE A is only observed for *H. contortus* and *P. pacificus*. Similarity was measured using the averaged Kullback-Leibler divergence of position frequency matrices of the corresponding motifs, see e.g. Aerts et al. (2003). A value of 0.20 and below can be considered to indicate high similarity. Abbreviations: Cel - *C. elegans*, Cbr - *C. briggsae*, Cre - *C. remanei*, Cbn - *C. brenneri*, Cja - *C. japonica*, Hco - *H. contortus*, Ppa - *P. pacificus*.

In this contribution we report on a comprehensive homology search for sbRNAs in the phylum Nematoda, and on an in depth analysis of the large gene family uncovered by this survey. We show that, unexpectedly, sbRNAs are homologs of Y RNAs.

Materials and Methods

Sequence Data

Genomic sequences of nematode species were downloaded from Wormbase (WS198, www.wormbase.org), the Sanger

Institute (www.sanger.ac.uk), TraceDB (www.ncbi.nlm.nih.gov/pub/TraceDB), the Sophia-Antipolis Institute (meloidogyne.toulouse.inra.fr) (Abad et al. 2008), the *M. hapla* Genome Sequencing Group (www.hapla.org). Details on the assemblies used here are listed in the Electronic Supplement. The phylogenetic relations of the investigated species are depicted in Fig. 2.

Sequence-Based Homology Search

Starting from an initial set of experimentally verified sbRNAs, listed in Tab. 1, we performed a blastn search with default parameters against the available genome assemblies of nematode species. Due to the high sequence variation in the central loop region, this initial step recovered only a few full length sbRNAs in other species. Blastn hits that showed a query coverage of at least 50% were extended by flanking sequence and manually compared to known sbRNAs in a structural alignment. In addition, we extracted putative sbRNA sequences from the multiz 6-way alignments of nematode species available at the UCSC Genome browser (genome.ucsc.edu) for known *C. elegans* sbRNA loci.

Homology Search with Promoter Elements

We applied a computational promoter search using the characteristic promoter elements of sbRNAs (PSE B and TATA-box) in species of the genus *Caenorhabditis*, in *P. pacificus* and in *H. contortus*. In the first step, we extracted regions 200 nt upstream of RNase P, RNase MRP, U6 snRNAs, and Selenocysteine tRNAs. These noncoding RNAs are known to utilize very similar PSE B and TATA-Box promoter elements. For *C. elegans* the sequences for RNase P, RNase MRP, and Selenocysteine tRNA could easily be retrieved from annotated Wormbase entries (rpr-1, mrpr-1, K11H12.t1) or, in case of U6, snRNAs from the literature (Dávila López et al. 2008; Marz et al. 2008). Simple blastn searches were sufficient to identify their orthologs in other nematode species. We then created multiple sequence alignments of the upstream regions using Jalview (Waterhouse et al. 2009) for each species, marked blocks corresponding to the PSE B and the TATA-box and generated a FRAGREP (Mosig et al. 2007b) search pattern. The FRAGREP search resulted in approx. 1,200 hits in *C. remanei* and more moderate numbers for the other nematodes. For each hit we searched the 300 nt of genomic DNA downstream of the putative promoter regions for a possible terminator consisting of a consecutive run of at least four T residues. The region ranging from 20 nt downstream of the TATA-box to the putative terminator was extracted for further analysis.

We then applied sequence-structure based clustering using the LocARNA-RNAclust pipeline (Will et al. 2007; Kaczkowski et al. 2009) to these putative transcripts. Default parameters were used for both LocARNA and RNAclust. Clusters were visually examined for sequence-structure similarity to already identified sbRNAs using the RNAsoupViewer (www.bioinf.uni-leipzig.de/pages/40/software.html).

This approach offers two major advantages over purely sequence-based or (structure) model-based searches, where only the ncRNA itself is used as query: (i) since promoter elements that are shared with other ncRNA classes are used for initial filtering of the genomic data, knowledge on the variability of the sequence and/or structure of the query ncRNA is irrelevant at this stage. Instead, a search using the query ncRNA is only performed on the small set of putative transcripts. Thus, more sensitive but also computationally much more expensive tools can be used in this second step; (ii) the canonical promoter structure lends additional credibility to the candidates. The feasibility of this strategy was recently demonstrated for identifying 7SK snRNAs of arthropods (Gruber et al. 2008).

Model-Based Homology Search

Multiple sequence alignments of the seed sequences and the hits of both the sequence-based homology search and the promoter screen were constructed. In a first analysis, sbRNAs were manually grouped into clusters based on length and sequence identity and aligned. RNAalifold (Hofacker et al. 2002; Bernhart et al. 2008) predictions for each group were then used as starting point for deriving a consensus structure for the well-conserved parts. These initial alignments were then refined manually and combined to a global alignment in the emacs text editor, making use of the RALEE mode (Griffiths-Jones 2005), which explicitly handles secondary structure annotation.

These structure-annotated alignments were then used to deduce a non-stringent sequence/structure model (available in the Electronic Supplement), which was then employed to screen the nematode genomes with RNABOB (selab.janelia.org/software.html) with default parameters. The resulting initial candidates were filtered using a modified position weight matrix scoring in which base-pairs are treated like individual letters:

Let $\mathcal{A} = \{A, C, G, T\}$ be the nucleotide alphabet. Then $\mathcal{B} = \{AA, AC, AG, AT, ..., TT\}$ is the alphabet of all standard and non-standard base pairs. The modified equation for the information vector I at position i in the approach of Kel et al. (2003) is

$$I(i) = \sum_{b \in \mathcal{A} \text{ or } \mathcal{B}} f_{i,b} \ln(k(b)\, f_{i,b}) \qquad (1)$$

where i is now either an unpaired nucleotide or a base pair, and $k(b) = 4$ if $b \in \mathcal{A}$ and $k(b) = 16$ if $b \in \mathcal{B}$. We implemented a Perl script that takes the RNABOB output and position weight matrices derived from the

structural alignment as input and outputs RNABOB hits augmented by a matrix similarity score (mSS) as defined by Kel et al. (2003). Hits with a mSS > 0.65 were then compared manually to previously identified sbRNAs. Recognizable homologs were incorporated into the sequence-structure alignment.

Identification of Promoter Elements

For the five species of the genus *Caenorhabditis*, *P. pacificus*, and *H. contortus* we were able to collect a sufficient number of upstream regions of ncRNAs that share at least partially the same promoter elements as sbRNAs. We created separate position weight matrices (PWMs) for the PSE A and the PSE B for each species as well as a general TATA-box PWM and used the approach by Kel et al. (2003) to score corresponding elements in the upstream sequences of our sbRNA candidates. Sequence motifs corresponding to PSE A were only classified as reliable if their score was higher than 0.75 and if they were exactly located 5 nt upstream of a PSE B. Alignments and PWMs are available in the Electronic Supplement.

Identification of Syntenic Regions

The UCSC Genome Browser provides gene annotations for all *Caenorhabditis* genomes used in this study. The advantage of this resource is that *C. elegans* genes were mapped using `tblastn` to other *Caenorhabditis* proteins so that the gene identifiers are available across all genomes. Wormbase, on the other hand, uses different gene identifiers for the individual species and does not supply a read-to-use homology table. In order to construct local synteny maps between *Caenorhabditis* genomes, we first used a simple `blastn` search to map our sbRNA sequence to the genomes version provided by the sequence repository at UCSC, which are older than the genome assemblies for the other analyses used here. We then extracted gene annotations within ±40 kb of each sbRNA location. In the next step, sbRNAs and adjacent genes were compared between all genomes. If sbRNAs in different genomes are located in the vicinity of genes with identical annotation, we consider these locations syntenic.

All genomes used here, except for *C. elegans* and *C. briggsae*, have not been assembled to the level of chromosomes. Thus, sbRNAs and adjacent protein coding genes might resided on different contigs making it difficult to identify both upstream and downstream markers. As a consequence, our strategy for detecting sbRNAs in syntenic regions requires at least one homologous protein within ±40 kb flanking a sbRNA.

Using this approach, we found that only two *C. elegans* sbRNA clusters, namely those on chromosome III and chromosome X have syntenically conserved locations in other *Caenorhabditis* species. These two clusters where then in detail examined using the synteny resources available at wormbase.org.

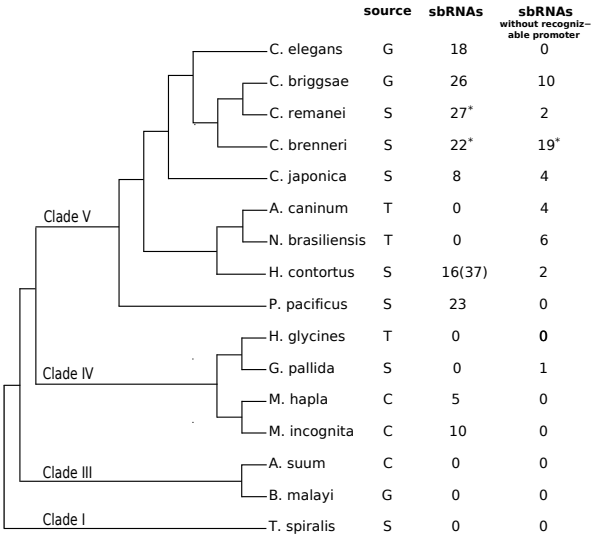


Fig. 2 Phylogenetic distribution of the 240 identified sbRNA homologs. Hits are divided into sbRNAs with confirmed promoter regions, and those hits that did not yield any significant homology to known ncRNA promoters. The species phylogeny is represented as a cladogram with arbitrary branch lengths, combining the Caenorhabditis species phylogeny by Sudhaus and Kiontke (2007) with the phylogeny of phylum Nematoda by Blaxter et al. (1998) and Mitreva et al. (2005) and the work from Chilton et al. (2006). *Accounting for allelic variants (Barrière et al. 2009), the number of sbRNAs in *C. remanei* is reduced to 26, while in *C. brenneri* 19 copies with intact promoter and 15 without are genomically distinct. The column "source" denotes the assembly status of the genomic DNA sequences (T: Traces, C: contigs, S: supercontigs, G: chromosomal level). For *H. contortus* we found a hit with 37 adjacent copies. For the list of sbRNA with verified promoter regions this hit was just counted once.

Results

Homology Searches

Starting from the seed sequences, both the analysis of the multiz alignments and an iterative BLAST search resulted only in a moderate number of additional homologs in the *Caenorhabditis* species and a few hits in *P. pacificus*, and failed to give any plausible candidate in other nematodes. In a second approach to identify new sbRNAs, we took advantage of the well characterised promoter elements of known sbRNAs (Li et al. 2008) and performed a computational promoter screen. sbRNAs found to that point were used to construct a promiscuous search pattern for RNABOB, whose results were filtered further using a PWM-based method to de-

tect faint sequence similarities as described in detail in the Methods section.

After manual inspection of the search results, we retained a list of 240 sbRNAs distributed over the nematode clade V (Strongylida, Diplogasterida, and Rhabditida) and clade IV (Tylenchida, Cephalobina, and Panagrolaimida), summarized in Fig. 2. It was recently shown that a considerable fraction of the genome assemblies of *C. remanei* and *C. brenneri* represents two alleles rather than distinct genomic loci (Barrière et al. 2009). In *C. brenneri* 14 sbRNAs that assembled to separate contigs show extensive sequence similarities (> 80% identity) within 1,000 nt examined flanking regions. Six out of these 14 show nearly perfect sequence conservation in the 3' flanking region, while the 5' flanking region does not. For these cases it is likely that we see an assembly artifact instead of an allelic variant. In *C. remanei* we find two sbRNAs that are located on separate contigs and show extensive sequence identity in the flanking regions. High identity is, however, only observed in the 5' flanking region suggesting that it might again be an assembly artifact. We conclude that 8 of our 240 sbRNA sequences are duplicates.

In particular, we report a total of 18 sbRNAs genes in the *C. elegans* genome, all having confirmed promoter elements. In the other species we also list a significant number of sbRNAs that do not show significant matches to known ncRNA promoter elements. One of the hits we identified in *H. contortus* has several (37) adjacent copies on one contig. We cannot exclude that this might be an assembly artifact and therefore we count this hit just once in the list of sbRNAs with promoter elements. Our survey failed to retrieve homologs in the genomes of *A. suum*, *B. malayi* and *T. spiralis* and in the shotgun trace sequences of *Heterodera glycines*.

Analysis of Upstream Regions

For *C. elegans* the core promoter of sbRNAs has been shown to consist only of a PSE B and a TATA-box (Li et al. 2008), while other polymerase III transcripts including the previously described Y RNA (Van Horn et al. 1995) have an additional conserved element located 5 nt upstream of the PSE B, called PSE A (Thomas et al. 1990; Missal et al. 2006). In other species of the phylum nematoda, studies of snRNA promoters of this type (pol-III type 3) have not been conducted so far. For all species except *C. elegans*, we identified corresponding promoter elements by sequence and positional conservation.

A detailed analysis of the upstream regions of sbRNAs with position weight matrices used in the computational promoter screen revealed that the shortened core promoter characteristic for sbRNAs in *C. elegans* can only be found in the genus *Caenorhabditis*. Upstream sequences of sbRNAs in *P. pacificus* and *H. contortus* show the presence of both a PSE A and a PSE B. A detailed representation of the core promoter for these species is shown in Fig. 1 together with corresponding elements of other putative pol-III transcripts. For *A. caninum*, *N. brasiliensis*, *G. pallida*, *M. hapla*, and *M. incognita* we were not able to find a sufficient number of high-confidence homologs of other pol-III transcripts to build reliable species-specific position weight matrices (PWMs) or to determine the exact position of PSEs and the TATA-box. In these cases upstream regions were visually compared for stretches of homologous regions. Results of promoter analysis are summarized in Fig. 2.

Secondary Structure

In order to derive a consensus secondary structure, we used the subset of those 155 (out of 240) sbRNA homologs that exhibit clearly recognizable pol-III promoters to avoid contamination by possible pseudogenes. The structural alignment was constructed manually. Due to high sequence variation in the central loop this region remained unaligned and was investigated separately.

The combination of thermodynamic structure predictions and phylogenetic analysis revealed several conserved structural elements, summarized in Fig. 3. Nematode sbRNAs exhibit three conserved stem structures:

S1 Stem S1 consists of at least four conserved base-pairs. It is extended at the outer end in most of the sequences. The closing inner AU pair of stem S1 is absolutely conserved in all sequences.

S2 Stem S2 is composed of three base-pairs only, and the majority of sequences shows two GU wobble-pairs. From a thermodynamic point of view this is a rather weak stem, but supporting evidence is given by compensatory mutations.

S3 Stem S3 is composed of nine base-pairs. The outer part of S3 shows many compensatory mutations, suggesting that the ability to form this double stranded region is more important than the actual sequence. Stem S3 closes with three conserved GC pairs, preceded by a conserved UA pair. Only 13 sequences, all from *H. contortus*, show an AU pair at this position.

B Stems S1 and S2 are separated by a conserved single bulged cytosine.

I Stems S2 and S3 are separated by a small internal loop. Although some related sbRNAs show conservation of some nucleotide positions, it does not seem to be a general sequence motif for the entire set of sbRNAs there.

H The central loop enclosed by the stem starts with the conserved sequence motif UUAUC. Detailed analysis of this motif showed that it is in general not involved in a structural context. For short sbRNAs, the entire central region is generally unstructured, forming a single hairpin loop. The longer sbRNA homologs tend to form short structural elements that appear conserved within subgroups.

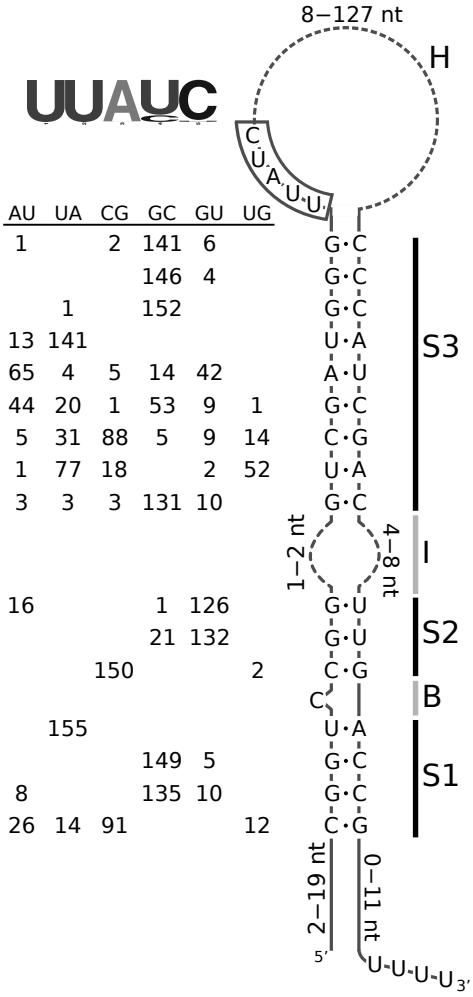


Fig. 3 Secondary structure model of sbRNAs derived from 155 sbRNAs with verified promoter regions. The table on the left shows the absolute counts of canonical and wobble base-pairs observed at a given position. The schematic drawing of the structure displays the most frequent base-pair. The sequence logo shows the frequencies of nucleotides for the UUAUC motif, which immediately follows the conserved stem. In only two out of 240 sbRNAs we observed one or two additional G residues inserted between the stem and this motif.

T At the 3' end we generally observe a stretch of at least four U residues, which are believed to function as transcription termination signals. For most sbRNAs further poly U/T stretches, which may serve as alternative termination signals (Gunnery et al. 1999; Guffanti et al. 2004) can be observed downstream of their genomic location.

sbRNAs are Y RNAs

Comparison with other RNA families revealed that nematode sbRNAs show substantial similarities in both sequence and secondary structure to vertebrate Y RNAs (see Mosig et al. (2007a) and Perreault et al. (2007) for Y RNA structure). The sbRNA CeN134 was reported as a possible Y RNA in the kingdom-wide survey for Y RNA homologs by Perreault et al. (2007). The connection of Y RNAs and sbRNAs was not commented on, and other sbRNA family members in *C. elegans* were not recognized, however. Fig. 4 summarizes a detailed comparison of the Nematode sbRNA consensus with the analysis of vertebrate Y RNAs by Mosig et al. (2007a) and the orthologs of the previously reported *C. elegans* Y RNAs from the genus *Caenorhabditis*. The latter were found using GotohScan (Hertel et al. 2009) starting from the experimentally known *C. elegans* CeY sequence (Van Horn et al. 1995).

All three structures share not only the overall organization but also several sequence features. In particular the inner part of stem S3, the two outer pairs of stem S2, the conserved cytidine bulge B, and the inner pairs of stem S1 are the same. These regions largely coincide with the most conserved ones within each of the three groups.

In mammals, stem S1, the bulged cytidine (B), and stem S2 have been shown to be required for Ro binding (Green et al. 1998; Stein et al. 2005), and thus for the formation of the Ro RNP particles, which are involved in RNA quality control. These features are well conserved between Y RNAs (vertebrates and nematodes) and sbRNAs (Fig. 4B). This strongly suggests that sbRNAs contain a functional Ro binding site.

Recently, it has been shown that Y RNAs are also required for chromosomal DNA replication in human cell nuclei (Christov et al. 2006; 2008). The primary motif for this function resides at the 3' end of stem S3 and consists of a stretch of three base-pairs (denoted by red stars in Fig. 4A) (Gardiner et al. 2009). In particular the UA base-pair turned out to be crucial for Y RNA functionality in DNA replication. Indeed, *C. elegans* CeY and a Y RNA homolog from *D. radiodurans* (Chen et al. 2007), both lacking this feature, were not able to compensate for vertebrate Y RNAs in DNA replication. All sbRNAs with the exception of 13 *H. contortus* sequences also show the conserved UA base-pair at this position.

Overall, nematode sbRNAs show more similarities with vertebrate Y RNAs than the previously reported *Caenorhabditis* Y RNAs. In addition to unambiguous structure homology in the helical regions, the conserved loop motif UUAUC is also present in the paralogous vertebrate subfamilies Y1 and Y3.

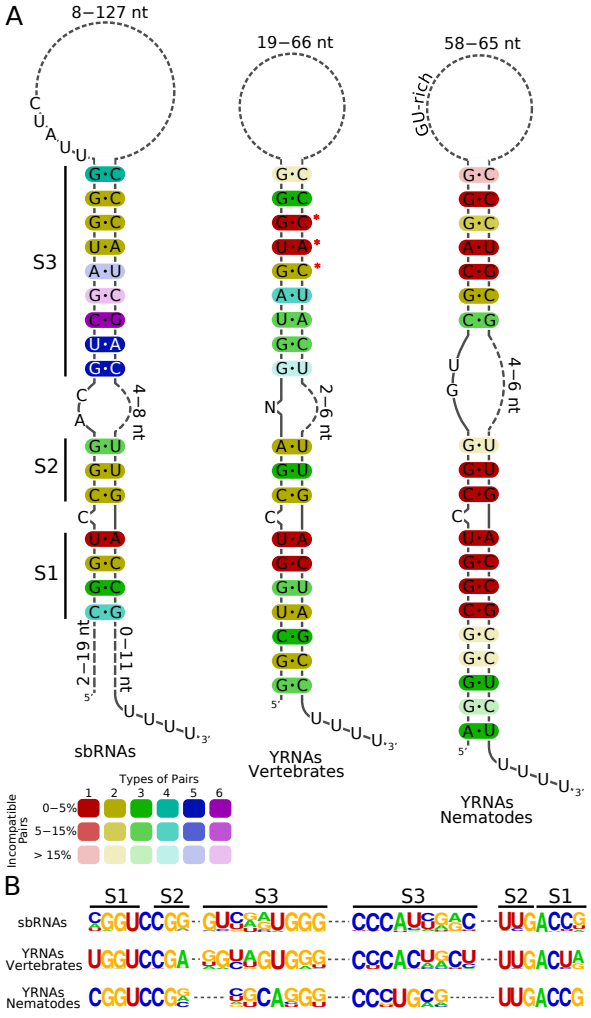


Fig. 4 **A** Comparison of secondary structures for nematode sbRNAs, vertebrate Y RNAs and the previously described Y RNA family in the genus *Caenorhabditis*. Red stars denote the region identified by Gardiner et al. (2009) to be crucial for the function of Y RNA in DNA replication. **B** Sequence logos for helical regions S1, S2, and S3.

Evolutionary History of sbRNAs

In *C. elegans* we uncovered six new sbRNA homologs (Tab. 2) in addition to the twelve previously described sbRNAs. All six are supported by promoter elements. Three hits have already been assigned a Wormbase ID, and for two of these there is evidence of transcription from a previously conducted study by Zemann et al. (2006). The same study annotated Cel7 as a C/D box snoRNA. This sequence yields a negative snoRNA classification by snoReport (Hertel et al. 2008) and can be unambiguously recognized as a sbRNA homolog based on both sequence and secondary structure.

Table 2 Newly identified sbRNA homologs in *C. elegans*. Hits marked with * are also reported by Zemann et al. (2006).

Name	Location	Other names	L
Cel1	intergenic	W01D2.7, Ce150*	81
Cel2	intergenic	–	85
Cel3	intronic	–	155
Cel5	intergenic	–	121
Cel6	intergenic	M163.15	83
Cel7	intergenic	M163.14, Ce94*	98

Due to the rapid evolution of the relatively short sbRNA sequences it is impossible to derive a reliable gene phylogeny based on sequence information alone. We therefore follow the strategy introduced for microRNA clusters by Tanzer and Stadler (2004). Furthermore, we systematically included synteny information. Syntenic clusters were identified in the genus *Caenorhabditis* based on their flanking protein coding genes (see Methods for details). Surprisingly, syntenic conservation can be established only for two of the five clusters: those located on *C. elegans* chr. III and chr. X. For the other clusters, only the sequence information could be used.

Standard phylogenetic methods are not applicable because the loop-part of the sbRNAs cannot be reliably aligned, while at the same time the better conserved stems barely contain phylogenetic information. We therefore used a z-score approach (Tanzer and Stadler 2004; 2006). In brief, the significance of pairwise alignments is evaluated by comparing the score with the score distribution of of pairwise alignments of shuffled input sequences. The resulting z-scores serve as similarity measure that can be used to construct hierarchical clustering. While this approach of course does not reconstruct an accurate phylogeny, it is capable of identifying clusters with statistically significant mutual similarities (Tanzer and Stadler 2006). The clustering not only identifies sbRNAs as unambiguous homologs of Y RNAs, it also confirms the observation that nematode sbRNAs are more similar to vertebrate Y RNAs than to the previously described nematode Y RNAs.

In vertebrates, Y RNAs show features required for both their known functions in DNA replication and binding to Ro. Their nematode homologs apparently underwent subfunctionalization so that sbRNAs and Y RNAs contain different features, Fig. 4. The exact time point of the divergence of sbRNAs and the CeY lineage cannot be determined with any certainty. While the z-score clustering points at an early divergence, CeY homologs were detectable within the genus *Caenorhabditis* only, suggesting a late duplication. Within *Caenorhabditis*, we observe a rapid radiation of divergent sbRNAs, supporting the hypothesis of a late divergence of CeY and sbRNAs.

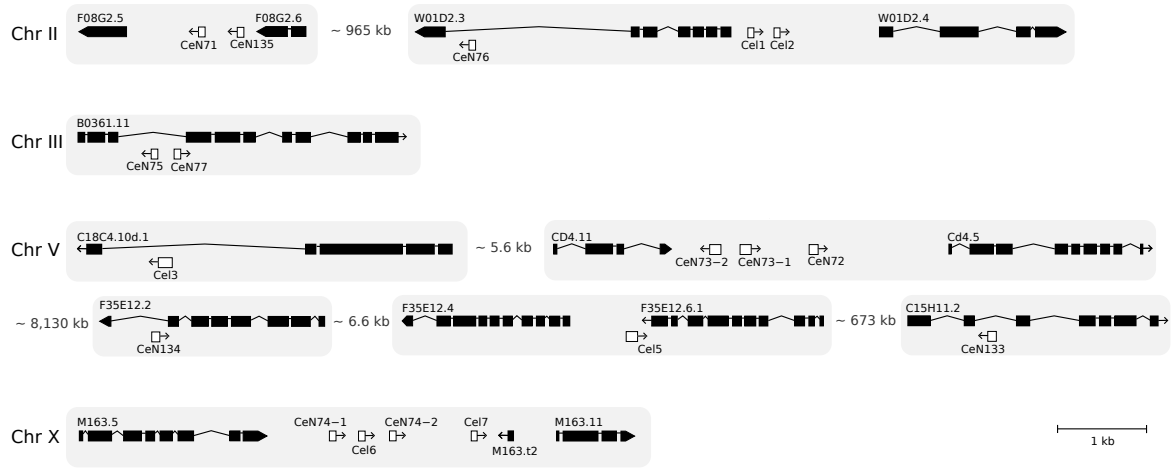


Fig. 5 Schematic drawing of the organization of the five sbRNA clusters in *C. elegans*. Each line represents a sbRNA cluster. White boxes denote sbRNAs, annotated Wormbase genes (release WS205) flanking sbRNA loci are shown in black.

The 18 *C. elegans* sbRNAs identified to-date are organized in five clusters, Fig. 5. Each cluster consists of multiple copies of one sbRNA subfamily. Thus, clusters seem to have arisen by local tandem duplications of one ancestral sbRNA. The mechanism by which sbRNAs were multiplied remains unknown. Nevertheless, we find evidence that not only single genes, but also groups of several sbRNAs might be affected by a single duplication event.

The sbRNA cluster on chromosome X. The chr. X cluster can be found with syntenic regions in all five *Caenorhabditis* species, Fig. 6 and Supplemental Fig. S2A. The cluster apparently derives from a single sbRNA, with *C. japonica* representing the ancestral state. The first duplication gave rise to two distinctive sbRNA families (A and B). In *C. elegans*, A was lost and B was copied 2 times. After the divergence of *C. elegans* and *C. briggsae*, B was duplicated leading to a cluster comprising three sbRNAs: A, B1 and B2, as found in *C. brenneri*. Clusters in both *C. briggsae* and *C. remanei* contain two copies of sbRNAs of family B2 suggesting a duplication prior to the speciation event. However, phylogenetic analysis rather suggest individual duplications in both species.

The sbRNA cluster on chromosome III. Both sequence similarity and cluster organisation indicate that the chr. III cluster has undergone different complex fates in each species, comprising multiple local duplication and deletion events (Fig. 6 and Supplemental Fig. S2A). Unlike the cluster on chromosome X, were single genes were effected, here two genes in tail-to-tail orientation seem to form a unit which is propagated. The two genes both contain their own PSEB and PSEA elements and thus do not seem to rely on promoter sharing.

In *C. elegans* the cluster is composed of one such unit (CeN75/CeN77) reflecting the ancestral state. Duplication of the ancestral 75/77 pair resulted in tandem copies 75A/77A and 75B/77B after the speciation event leading to *C. elegans*.

In *C. brenneri*, one of the two copies (75B/77B) was deleted and the other one (75A/77A) duplicated leading to Cbn29/Cbn30 and Cbn25/Cbn26. Thus, the cluster in *C. brenneri* consist only of members of families 75A and 77A. In addition, we find two more copies of 75A (Cbn28 and Cbn27), which most likely result from duplications of the adjacent Cbn29 (family 75A). In an alternative scenario, the whole unit of Cbn29/Cbn30 (77A/75A) was duplicated and each copy of 77B was subsequently lost. Such a scenario, however would be more costly than individual duplications and thus appears less probable. Cbn31, which is also present at this locus, shows some homology to the other members of the cluster. However, neither phylogenetic analysis nor sequence motifs in the loop regions allowed an unambiguous assignment to any of the the two families.

Members of both the 75A/77A and 75B/77B families are present in *C. remanei*. As in *C. brenneri*, we find an individual duplication of 75A (Cre12). The unit of 75B/77B was duplicated once such that in *C. remanei* there is one copy of 75A/77A (Cre10/Cre11), two copies of 75B/77B (Cre8/Cre9 and Cre14/Cre13) and another cape of 75A (Cre12). Interestingly, in *C. remanei* this locus seem to have undergone extensive genomic rearrangement. The exon structure of the surrounding gene (B0361.11) was altered, such that in *C. remanei* the sbRNA cluster resides in intron 2 instead of intron 3 (see location in *C. elegans*, Fig. 5).

In an alternative scenario, the duplication of the ancestral 75/77 pair took place after the speciation of *C.*

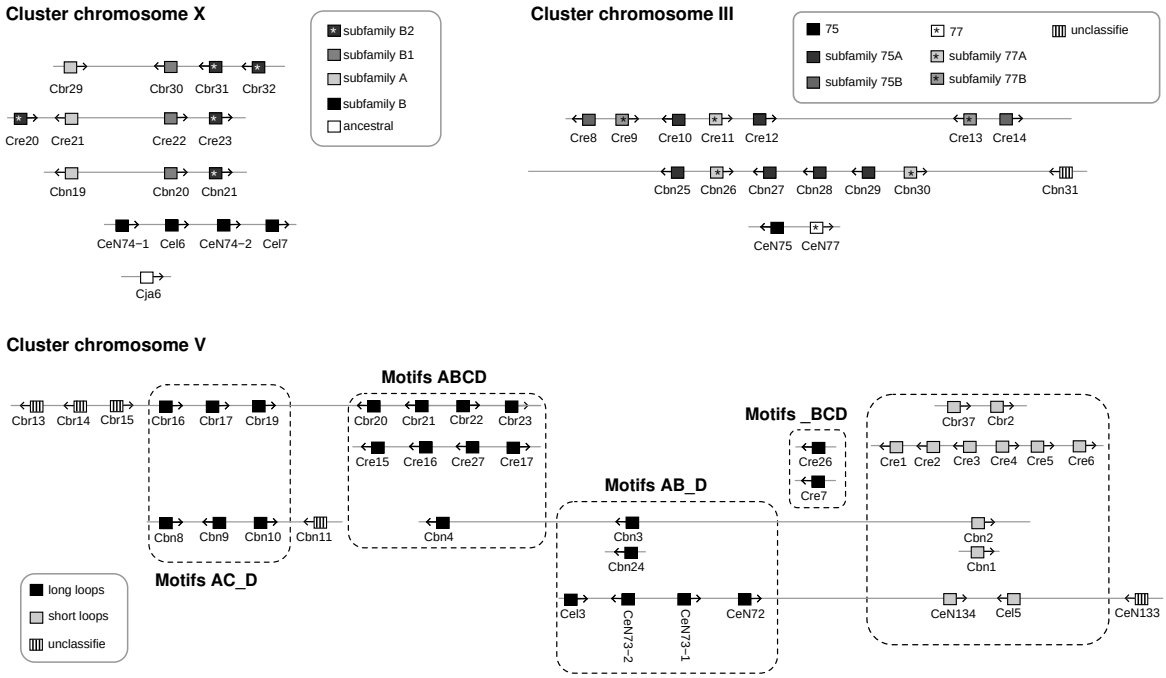


Fig. 6 Schematic drawing of the genomic organization of the sbRNA clusters on chromosome III, chromosome X, and chromosome V of *C. elegans* and their homologs. Based on phylogenetic analysis, conserved motifs and position within a cluster, individual genes were group into different subfamilies (shown as different shades of gray). Clusters are shaped by duplications of single genes as well as units of sbRNAs followed by deletions of individual genes. The cluster on chromosome V consist of two sbRNA families of different loop sizes (white boxes mark shorter ones, black the longer ones). The shorter ones date back to *H. contortus* (data not shown), whereas the longer ones appear in *Caenorhabditis*. Besides the structure and sequence motifs common to all sbRNAs, both families of this cluster reveal no homology in the heavily structured loops and therefore do not seem to have arisen by gene duplication. Gene duplications of the "long" sbRNAs coincided with duplications of substructures in the multiloop. A, B, C, D refers to these substructures. The loop region of CeN72 is too degenerated to assign this gene to either of the two groups based on sequence similarity. Due to its close vicinity to CeN73-1, Fig. 5, we grouped it with the long ones. For details see text and Supplemental Fig. S1 and Fig. S2. The figure shows the organisation of sbRNA clusters only and does not reflect genomic distances. Arrows indicate sbRNA orientation: plus strand (→) and minus strand (←). sbRNA which could not be assigned unambiguously to a subfamily are labelled as "unclassified". Abbreviations: Ce – *C. elegans*, Cbr – *C. briggsae*, Cre – *C. remanei*, Cbn – *C. brenneri*, Cja – *C. japonica*

brenneri. However, motifs in the loop region of all 75A family members in both *C. brenneri* and *C. remanei* are highly conservation and thus support the scenario outlined above.

Corresponding sbRNAs in *C. briggsae* seem to have been lost, since the corresponding intron is just 60 nt in size. *C. japonica* has a normal sized intron of 2,000 nt as seen in other species, but no sbRNA signatures have been detected there.

Two sbRNA clusters on chromosome V. The clusters on *C. elegans* chromosome V, Fig. 6, are distinct from all other sbRNAs discussed so far because their loop regions are both much longer than those of other sbRNAs and heavily structured. The clusters belong to two distinct sbRNA subfamilies of different length. Members of the shorter ones, white boxes in Fig. 6, are present in *C. japonica*, *C. elegans*, *C. brenneri*, *C. remanei*, and *C. briggsae* were also found in *H. contortus* (Electronic Supplement). The longer paralogs, indicated by filled boxes in Fig. 6, appear in *Caenorhabditis* only. Both families represented here seem to be ancestral to (or at least as old as) the family comprising the majority of sbRNAs. Further support for their evolutionary age comes from the presence of at least one of these families in *H. contortus*. As in the chr. III and chr. X clusters, there are multiple duplications and deletions of individual genes.

Taking a closer look at the loop regions of the individual genes showed that several gene duplications coincided with changes of the organization of the loop regions, i.e., regional duplications and deletions of substructures (see Supplemental Fig. S1). Thus, we grouped members of the cluster into four subfamilies based on their loop motifs (Fig. 6C). Sequence/structure alignments revealed that each of these subfamilies contains at least three stems in the loop region with hairpin A being the best conserved

one. In addition, each *Caenorhabditis* species seem to have undergone individual duplications of a subfamily. Based on both phylogenetic analysis and structure information, we deduced the following evolutionary scenario:

The ancestral copy of the long sbRNAs probably consisted of three hairpins (A_CD). This first round of gene duplications results in subfamilies AB_D and ABCD, where hairpin B seems to have arisen by a duplication of the upstream hairpin A (Fig. 6D). In particular, the loop motifs are almost identical, suggesting that they have arisen in the ancestral sbRNA by the duplication of an already existing secondary structure element. In a subsequent duplication of subfamily ABCD hairpin C was deleted leading to subfamily AB_D.

In *C. elegans* A_CD was lost again, AB_D was copied 2 times and hairpin C of ABCD degraded (Ce3). *C. brenneri* still shows all three ancestral subfamilies, but again underwent individual gene duplications. After the speciation of *C. brenneri*, subfamily AB_D was lost, such that in *C. briggsae* we find only members of ABCD and A_CD. In *C. remanei*, the whole cluster was heavily remodelled. A_CD was deleted and a duplicate of ABCD lost hairpin A.

Our analysis suggests that at least loop regions of these sbRNAs contain functional motifs, possibly establishing interactions with binding partners such as proteins or RNAs. In particular, the high conservation of motifs in hairpin A and B (CTTG) is striking. Most sbRNAs here have at least one stem in the loop region of this type. Hairpins 3 and 4, in contrast, seem to be more flexible and may be responsible for gene specific functions.

Reconstructing such complex patterns of gene duplications strongly depends on the genome information available. Data from additional *Caenorhabditis* as well as fully assembled genomes would be required to disentangle the apparently complex history of this cluster with any certainty. Thus, additional data and improved assemblies of the *Caenorhabditis* genomes will help to resolve the ambiguities in the scenario described above and may favour a slightly different reconstruction of the details evolutionary history in particular of these "non-syntenic" sbRNA clusters.

The sbRNA cluster on chromosome II. The cluster on *C. elegans* chromosome II consists of very short sbRNAs. The loop motif does not exceed 20 nt in length and seems to be unstructured. Due to these short loop motifs the evolutionary history of this sbRNA cluster could not be resolved unambiguously.

Discussion

Deng et al. (2006) annotated sbRNAs as a novel RNA family because of their unique promoter structure and the lack of obvious sequence homology with other known RNA families. Our analysis of the patterns of sequence and structure conservation established that sbRNAs are homologs of Y RNAs. We identified sbRNA homologs in species of nematode clades IV and V by a combination of several search strategies. While homology search based solely on sequence failed to identify many of the sbRNAs, the computational promoter screen and the searches with secondary structure models were successful in a broader range of species. We show here that a screen for characteristic promoter elements can substantially improve both sensitivity and specificity of RNA homology searches. This strategy, however, requires prior knowledge of promoter or other regulatory DNA elements. The construction of the promoter search patterns itself requires a collection of known RNA genes that are under the control of similar promoters. Due to the lack of a comprehensive ncRNA annotation for most invertebrate genomes, this amounts again to a homology search problem – although for better conserved ncRNAs. So far, promoter-based approaches have been employed systematically only for pol-III type 3 promoters (Gruber et al. 2008; Pagano et al. 2007), which are associated with a quite limited set of ncRNA families. In a recent contribution, some of us reported on the identification of the 7SK snRNA homologs in arthropods (Gruber et al. 2008) using a similar approach. In that study, the small number of initial hits allowed a manual analysis. Here, we had to use a a less stringent search because of the variability in the promoter structure itself. The deviant pol-III promoter structure of the sbRNAs described by Deng et al. (2006) turned out to be restricted to the genus *Caenorhabditis*. As a consequence, a large number initial candidates has to be a evaluated. This task could be mastered only by computational methods such as sequence/structure-based clustering (Will et al. 2007). This approach is computationally expensive, but has the benefit that one is not limited to structure or sequence constraints that have to been known from the beginning. As a third strategy we applied model-based RNA homology search combining sequence and structure information gathered in the two previous steps. Instead of focusing on specificity, we opted for a non-stringent RNABOB model and used a PWM-based approach for subsequent filtering. In total we end up with 240 loci across the currently available genome data of Chromadorea that we identified as sbRNAs with very high confidence. Accounting for the allelic variants included in some genomes, this number reduced to 231 distinct sbRNA genes.

We have been unable to find unambiguous sbRNA/Y RNA genes in basal nematodes. This does not come as a surprise. *B. malayi*, *T. spiralis*, and *A. suum* are separated by large evolutionary distances from their closest relatives with sequenced genomes. Signals of sequence homology are therefore faint for the short sequences in question. In the case of the Chromadorea we could start from several experimentally validated sequences in *C. elegans* to retrieve a large number of homologs from closely

related species. It is these data that allowed a detailed study of the sequence and structure constraints of sbRNAs. These models, in turn, were necessary to recognize the homology of sbRNAs with the previously described Y RNA of *C. elegans* and with the vertebrate Y RNAs. The information contained in these models, however, does not provide sufficient specificity to retrieve homologs from distantly related genomes with acceptable confidence. This also explains the surprising fact that the descriptor-based survey for Y RNAs by Perreault et al. (2007) hit one of the sbRNAs with borderline significance but failed to recognize most other family members.

The number of sbRNAs detected in this study varies significantly between species. For the two syntenically conserved sbRNA clusters we showed in detail that they exhibit a complex evolutionary history resulting in very different sbRNA complements even in fairly closely related species. The syntenically non-conserved clusters provide further evidence for the rapid evolution of the sbRNA complement. Strictly speaking, we cannot rule out that there are additional, highly-derived, members of the sbRNA/Y RNA family. The group of sequences identified here, however, shows coherent features and we did not detect ambiguous borderline-cases. Sampling biases, e.g. due to incomplete genome assemblies, thus might affect the exact sbRNAs counts, such technical artifacts can by no means account for the large differences between closely related species within the genus *Caenorhabditis*. Most likely, therefore, the striking differences observed in other clades, also reflects evolutionary variation rather than computational limitations.

Recent results on the function of mammalian Y RNAs suggest that they have at least two distinct modes of action. On the one hand, they are part of the Ro-RNA particle which is involved RNA quality control (Stein et al. 2005). On the other hand, they are essential for chromosomal DNA replication (Christov et al. 2006).

Despite the fact that sbRNAs form a large and diverse family of ncRNAs, only a single representative, the most derived CeY RNA (encoded by the *yrn-1* gene) was found to bind the *C. elegans* Ro60 ortholog ROP-1 *in vivo* (Van Horn et al. 1995). The same study also reported that human Y RNAs are not bound by the ceROP-1 protein *in vitro*, whereas the CeY RNA is bound by human Ro60 even more efficiently than the human Y3 and Y4 RNAs. Van Horn et al. (1995) also noted that the human Ro60 protein significantly differs from its *C. elegans* ortholog. Of the 28 residues of the *Xenopus laevis* Ro60 protein that are in contact with the Y RNA (Stein et al. 2005), only 11 are conserved in frog and worm, while 27 are shared between human and frog. Of the 14 amino acids in contact with mis-folded RNAs, on the other hand, almost all that are conserved between frog and human are also conserved in the worm. We found here that the other nematode sbRNAs are more similar to human Y RNAs than to ceY, in particular in terms of their secondary structure. Taken together, this suggests that sbRNAs (except ceY) in fact do not bind to ROP-1 at all. In this context, the ill-defined role of *rop-1* in *C. elegans* dauer larvae formation is of interest, which suggests alternative binding partners of ROP-1. The *Caenorhabditis* sbRNAs, however, conserve a motif that was recently demonstrated by Gardiner et al. (2009) to be essential for the function of vertebrate Y RNAs in DNA replication.

It is very tempting, therefore, to speculate about an involvement of sbRNAs in nematode chromosomal DNA replication. Our unpublished data of a *C. elegans yrn-1* deletion, furthermore, indicate that the ceY RNA — in contrast to human Y RNAs — is not essential for chromosomal DNA replication. The available information suggests that the sbRNA family has undergone subfunctionalization that separated the RNA responsible for the Ro-related function (ceY) from a much larger family of sbRNAs responsible for the replication-associated functionality. If this is true, then reports (Kamath et al. 2003; Sönnichsen et al. 2005) that depletion of some individual sbRNAs does not cause a phenotype detectable in high throughput studies are not surprising. For the hypothetical role of sbRNAs in DNA replication it is plausible to speculate that either not all sbRNAs might be involved in nematode DNA replication or, alternatively, that different sbRNAs might substitute for each other similar to vertebrate Y RNAs (Gardiner et al. 2009; Christov et al. 2006). If this is indeed the case, research by reverse genetics will not be easy given that the sbRNA family comprises at least 18 paralogs in *C. elegans*.

All sbRNAs, including the previously described ceY RNA, are subject to strong evolutionary pressure on the conserved stem structure. The central loop, on the other hand, seems to evolve rapidly since conserved motifs in the central loop are only recognizable in closely related species. This extreme variability poses the question if these loop motifs are of biological relevance at all. Hogg and Collins (2008) suggest that the loop regions of Y RNAs might specify substrate specificities, although there is not direct evidence for this hypothesis. Without a clearly defined and experimentally supported functional role for sbRNA, one could only speculate about the reasons and implications of species-specific differences.

Supplemental Information

An Electronic Supplement located at http:www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-020/ compiles a list of detected sbRNAs, sequence data and alignments in machine-readable form.

Acknowledgments

This research originated from an RNA Bioinformatics course at the University of Vienna in the fall semester

2008. Subsequently, it was then funded in part by the Austrian GEN-AU projects "Bioinformatics Integration Network III" and "Noncoding RNA", the AMS Vienna and the DFG under the auspices of the SPPs 1174 "Deep Metazoan Phylogeny" and 1258 "Sensory and Regulatory RNAs" and the AMS Vienna.

References

Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EG, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, Caillaud MC, Coutinho PM, Dasilva C, De Luca F, Deau F, Esquibet M, Flutre T, Goldstone JV, Hamamouch N, Hewezi T, Jaillon O, Jubin C, Leonetti P, Magliano M, Maier TR, Markov GV, McVeigh P, Pesole G, Poulain J, Robinson-Rechavi M, Sallet E, Ségurens B, Steinbach D, Tytgat T, Ugarte E, van Ghelder C, Veronico P, Baum TJ, Blaxter M, Bleve-Zacheo T, Davis EL, Ewbank JJ, Favery B, Grenier E, Henrissat B, Jones JT, Laudet V, Maule AG, Quesneville H, Rosso MN, Schiex T, Smant G, Weissenbach J, Wincker P (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat. Biotechnol. 26:909–915

Aftab MN, He H, Skogerbø G, Chen R (2008) Microarray analysis of ncRNA expression patterns in *Caenorhabditis elegans* after RNAi against snoRNA associated proteins. BMC Genomics 9:278–278

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B (2003) Computational detection of cis-regulatory modules. Bioinformatics 19:5–14

Barrière A, Yang SP, Pekarek E, Thomas CG, Haag ES, Ruvinsky I (2009) Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res. 19:470–80

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474–474

Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, Vida JT, Thomas WK (1998) A molecular evolutionary framework for the phylum Nematoda. Nature 392:71–75

Chen X, Wurtmann EJ, Van Batavia J, Zybailov B, Washburn MP, Wolin SL (2007) An ortholog of the Ro autoantigen functions in 23s rRNA maturation in *D. radiodurans*. Genes Dev. 21:1328–1339

Chilton NB, Huby-Chilton F, Gasser RB, Beveridge I (2006) The evolutionary origins of nematodes within the order Strongylida are related to predilection sites within hosts. Mol. Phylogenet. Evol. 40:118-128

Christov CP, Gardiner TJ, Szüts D, Krude T (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol. Cell. Biol. 26:6993–7004

Christov CP, Trivier E, Krude T (2008) Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. Br. J. Cancer 98:981–988

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res. 14:1188–1190

Dávila López M, Rosenblad MA, Samuelsson T (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res. 36:3001–3010

Deng W, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, He H, Cai L, Sun H, Liu C, Li B, Bai B, Wang J, Jia D, Sun S, He H, Cui Y, Wang Y, Bu D, Chen R (2006) Organization of the *Caenorhabditis elegans* small noncoding transcriptome: genomic features, biogenesis, and expression. Genome Res. 16:20–29

Gardiner TJ, Christov CP, Langley AR, Krude T (2009) A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication. RNA 15:1375–85

Green CD, Long KS, Shi H, Wolin SL (1998) Binding of the 60-kDa Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. RNA 4:750–765

Griffiths-Jones S (2005) RALEE–RNA alignment editor in emacs. Bioinformatics 21:257–259

Gruber AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF (2008) Arthropod 7SK RNA. Mol. Biol. Evol. 25:1923–1930

Guffanti E, Corradini R, Ottonello S, Dieci G (2004) Functional dissection of RNA polymerase III termination using a peptide nucleic acid as a transcriptional roadblock. J. Biol. Chem. 279:20708–20716

Gunnery S, Ma Y, Mathews MB (1999) Termination sequence requirements vary among genes transcribed by RNA polymerase III. J. Mol. Biol. 286:745–757

Hernandez N (2001) Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J. Biol. Chem. 276:26733–26736

Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF (2009) Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. Nucleic Acids Res. 37:1602–1615

Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 24:158–164

Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 319:1059–1066

Hogg RJ, Collins K (2008) Structured non-coding RNAs and the RNP Renaissance. Curr. Op. Chem. Biol. 12:684–689

Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J (2009) Structural profiles of miRNA families from pairwise clustering. Bioinfor-

matics 25:291–294

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421:231–237

Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31:3576–3579

Li T, He H, Wang Y, Zheng H, Skogerbø G, Chen R (2008) In vivo analysis of *Caenorhabditis elegans* noncoding RNA promoter motifs. BMC Mol. Biol. 9:71–71

Marz M, Kirsten T, Stadler PF (2008) Evolution of spliceosomal snRNA genes in metazoan animals. J. Mol. Evol. 67:594–607

Missal K, Zhu X, Rose D, Deng W, Skogerbø G, Chen R, Stadler PF (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J. Exp. Zoolog. B Mol. Dev. Evol. 306:379–392

Mitreva M, Blaxter ML, Bird DM, McCarter JP (2005) Comparative genomics of nematodes. Trends Genet. 21:573–581

Mosig A, Guofeng M, Stadler BM, Stadler PF (2007a) Evolution of the vertebrate Y RNA cluster. Theory Biosci. 126:9–14

Mosig A, Chen JL, Stadler PF (2007b) Homology search with fragmented nucleic acid sequence patterns. In R. Giancarlo and S. Hannenhalli (Eds.), Algorithms in Bioinformatics (WABI 2007), Volume 4645 of Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 335–345. Springer Verlag.

Pagano A, Castelnuovo M, Tortelli F, Ferrari R, Dieci G, Cancedda R (2007) New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. PLoS Genet. 3:e1

Ploner A, Ploner C, Lukasser M, Niederegger H, Hüttenhofer A (2009) Methodological obstacles in knocking down small noncoding RNAs. RNA 15:1797–804

Perreault J, Perreault JP, Boire G (2007) Ro-associated Y RNAs in metazoans: evolution and diversification. Mol. Biol. Evol. 24:1678–1689

Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, Hewitson M, Holz C, Khan M, Lazik S, Martin C, Nitzsche B, Ruer M, Stamford J, Winzi M, Heinkel R, Röder M, Finell J, Häntsch H, Jones SJ, Jones M, Piano F, Gunsalus KC, Oegema K, Gönczy P, Coulson A, Hyman AA, Echeverri CJ (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature 434:462–469

Stein AJ, Fuchs G, Fu C, Wolin SL, Reinisch KM (2005) Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. Cell 121:529–539

Sudhaus W, Kiontke K (2007) Comparison of the cryptic nematode species *Caenorhabditis brenneri sp.n.* and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the *Caenorhabditis Elegans* group. Zootaxa 1456:45–62

Tanzer A, Stadler PF (2004) Molecular evolution of a microRNA cluster. J. Mol. Biol. 339:327–335

Tanzer A, Stadler PF (2006) Evolution of MicroRNAs. In: Ying, SY (ed) MicroRNA Protocols. Humana Press, Totowa, NJ, pp. 335–350

Thomas J, Lea K, Zucker-Aprison E, Blumenthal T (1990) The spliceosomal snRNAs of *Caenorhabditis elegans*. Nucleic Acids Res. 18:2633–2642

Van Horn DJ, Eisenberg D, O'Brien CA, Wolin SL (1995) *Caenorhabditis elegans* embryos contain only one major species of Ro RNP. RNA 1:293–303

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput. Biol. 3:e65

Zemann A, op de Bekke A, Kiefmann M, Brosius J, Schmitz J (2006) Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res. 34:2676–2685

8

Discussion

In this thesis we have explored several strategies for computational noncoding RNA detection ranging from *de novo* detection approaches like `RNAz` and `RNALfoldz` to homology search problems for the RNA families of 7SK RNA and sbRNAs. *De novo* detection of functional RNA structures or even RNA genes is an ill-defined problem. There are no statistically significant features common to all functional structural elements. Signals that mark the boundaries of an RNA gene like start and stop codons as in the case of open reading frames are missing. Moreover, there is no clear definition of what defines functional in terms of RNA structures. There is no doubt that independent RNA genes such as tRNAs or microRNAs are functional, but there are often a multitude of secondary structures in mRNAs that are considered as functional, see e.g. the iron responsive element (Hentze et al., 2004) or the SECIS element (Lambert et al., 2002). The set of actions of such elements is broad. They can serve for example as recognition elements for proteins or can control the access to other elements. The global assessment of such functional structures is still out of reach, but first steps in this direction have been recently made by Kertesz et al. (2010), who conducted genome-wide RNA structure probing in yeast. Recently, some progress has also been made in functional characterization of long noncoding RNAs (Tsai et al., 2010). Long noncoding RNAs, or often also termed long intergenic noncoding RNAs, can span several thousand nucleotides. The question if structural elements, or more precisely to which extent structural elements, are required for the function of these RNAs still remains unanswered.

So far, *de novo* detection approaches for functional RNAs from genomic sequence are limited to the set of structured RNAs, where the structure of the RNA is essential for its function. `RNAz` is a leading software package in this field. In detail, `RNAz` aims at the detection of thermodynamically stable, conserved RNA secondary structures. In Chapter 4 of this thesis

we have introduced an improved version of RNAz, which now operates using a dinucleotide background model. A prerequisite to this was the development of computational tools (Gesell and Washietl, 2008; Anandam et al., 2009) that allowed to generate randomized alignments preserving the dinucleotide composition of the input alignment. The task of establishing a support vector regression (SVR) for the z-score estimation that is based on dinucleotide-preserving shuffled sequences was the most challenging one. For the first version of RNAz it was sufficient for training of the SVR to generate a set of approximately 10,000 sequences to uniformly cover the sequence space of interest. When moving to the dinucleotide space it is not intuitively clear how to draw a uniformly distributed sample of sequence in that case. We have developed a strategy that is best described as "coverage-oriented sampling". Briefly recapping the approach, we first generate a set of sequences that uniformly cover the mononucleotide space of interest, and in a second step we draw a representative set of sequences covering the dinucleotide space of a sequence with a particular mononucleotide composition. This strategy generated a training set of several hundred thousand sequences. Efficient training and prediction was then mastered by splitting data by the G+C content to smaller data sets. The presented dinucleotide regression is comparable in speed and accuracy with the original mononucleotide approach. We also applied a series of modifications that affect the overall classification capability and usage of RNAz. The use of the Shannon entropy as a measure of sequence variation instead of the mean pairwise identity and the number of sequences, helped to get rid of artifacts in the SVM classifier probability landscape (cf. Fig. 3.1) and eliminated the upper limit on the number of sequences. We also compiled a new training set and extended the RNAz approach to consider structural alignments. In a recent work by Salari et al. (2009), it was argued that RNAz makes use of conserved structure databases covering only a small portion of the genome. In fact, RNAz is a machine learning approach that is based on training data and, hence, restricted in generalizing by the examples in the training data. Features used by RNAz to infer putative functional RNA structures are general in a sense that they are not restricted to a particular RNA family and allow to discover a broad range of functional structures as demonstrated by a multitude of successful studies. With RNAz 2.0 we have presented a completely refurbished version of the RNAz algorithm addressing all major shortcomings of the previous version. Next tasks that aim at improving the prediction accuracy of RNAz will focus on preprocessing strategies of alignments before they are scored with RNAz. In particular, the classic sliding window approach used to scan long genomic alignments will be addressed. RNALalifold, a scanning version of RNAalifold, can be used to better split longer alignments into smaller ones based on local structure predictions.To the knowledge of the author, the new version of RNAz has so far been applied to genome-wide screens in nematodes and bacteria of the Order Aquificales.

All general *de novo* computational RNA gene finding methods rely on signatures of evolutionary conservation of RNA secondary structures. When only limited or no comparative genomics data is available, the set of computational methods becomes very sparse. In bacteria for example, some strategies using promoter and termination signals or screens for elevated G+C content haven been proposed for ncRNA detection, but a general approach is missing. Indeed, options are limited when no comparative data is available, but one strategy that can be applied is to ask for unusually stable folding regions, regions that are more thermodynamically stable than expected by chance. With `RNALfoldz` (cf. Chapter 5) we have introduced a program that implements this strategy. `RNALfold` (Hofacker et al., 2004b) is a local, minimum free energy (mfe) folding algorithm that can be used to predict local RNA secondary structures in long genomic sequences. Filtering of structures is needed to reduce the amount of structures to a reasonably sized set. `RNALfoldz` is an extension of `RNALfold` combining local mfe prediction with evaluation of thermodynamic stability. We modified the `RNAz` support vector regression approach to yield support vector models of a equal accuracy, but with a reduced number of support vectors. The number of support vectors inherently determines the execution time of the z-score regression. Since in a worst case scenario the z-score regression has to be called once for every position in the genomic sequence, it is of crucial concern to have a fast approach for the z-score evaluation. Current work in progress is focusing on building regression models by multiple linear regression instead of support vector regression. If this strategy shows comparable results in terms of accuracy compared to the support vector approach, the time spent for z-score evaluation could be reduced to a fraction of the amount currently spent. We have also explored a strategy to control the empirical false discovery rate of `RNALfoldz` using abstract shapes. If one simply opts for the most stable structures, the set will certainly consist to a large extent of small hairpin structures. Grouping by the sequences' abstract shapes poses an elegant way to retain a structurally diverse set of the most stable structures found in the screen. So far, we have applied `RNALfoldz` in a prototype screen in *E. coli* and in a still unpublished study on ncRNA detection in *Aquificales* (Nickel et al., 2010). In the later work we intersected `RNALfoldz` predicted structures with next generation sequencing transcriptomics data to infer novel ncRNAs. The specificity of `RNALfoldz` is too low to function as a standalone ncRNA gene finder, but intersection of `RNALfoldz` predicted structures with transcriptomics data or promoter/terminator signals can give predictions additional confidence.

De novo detection is often used to get a first guess on putative novel functional RNA structures. Functional annotation of those predicted structures whether experimentally or computationally is still an open problem. In Chapter 6 we presented a study on homology search for 7SK RNA members in arthropod species. The *Drosophila melanogaster* 7SK RNA gene

had been previously seen both in computational screens (Rose et al., 2007; Stark et al., 2007) as well as a experimental screen (Yuan et al., 2003), but none of these studies was able to provide a functional annotation. The main problem in such cases is that RNAs are often too diverged in sequence and structure to be recognized as members of a particular RNA family by standard methods. Indeed, the Drosophila 7SK RNA is so diverged, that also a previously conducted homology search on 7SK RNA (Gruber et al., 2008b) failed to recover a candidate in arthropods. In Gruber et al. (2008b), we used the polymerase III type 3 promoter structure of 7SK RNA to verify hits found by homology search. The typical 7SK promoter consists of a TATA-like box, a proximal sequence element (PSE) and often a distal enhancer element. Based on the assumption that this promoter structure is conserved across species, we conducted a computational promoter screen to find ncRNAs that are expressed by a pol III type 3 promoter. Detailed inspection of candidate RNA genes then revealed that one candidate located on chromosome arm 3R is a 7SK RNA homolog. Homology search starting with this initial hit then identified 7SK RNA genes in a broad range of arthropod species. We also tried to find homologs in nematodes, but our search remained unsuccessful. In a recent study (Marz et al., 2009) on the evolution of 7SK RNA and its protein binding partners presented a 7SK RNA homolog in the nematode *C. elegans*. This 7SK RNA gene was, however, later shown to be a snoRNA involved in rRNA processing (Hokii et al., 2010). A computational promoter screen is a powerful method to first identify a set small of candidates that can then be analyzed in detail. Large-scale application of this strategy is, however, limited. Such well defined promoter structures as in the case of 7SK RNA are limited to a certain set of housekeeping ncRNAs. Moreover, one needs to identify a set of ncRNAs first to build promoter search patterns. Pagano et al. (2007) performed a screen using a similar strategy on the human genome, but results are difficult to interpret. Experimental verification of candidate genes showed that the promoter region was often found located within the transcript. Pol III type 3 promoters are, however, known to be external promoters found upstream of the RNA gene.

Structural analysis of detected 7SK snRNA genes highlighted conserved sequence and structure motifs that are of functional importance, e.g. the conserved motif GAUC-GAUC in the 5' stem (Egloff et al., 2006; Lebars et al., 2010). Identification of such conserved motifs is often a starting point for functional characterization. In Chapter 7, we presented a study that aimed at the detailed characterization of the putatively novel RNA family of sbRNAs (Deng et al., 2006). We applied various homology search strategies ranging from sequence based `blastn` searches, computational promoter screens with sequence-structure clustering, to descriptor based search with `RNABOB`. We chose to use a non-stringent `RNABOB` descriptor, which resulted in several thousands of predictions. In order to rank these predictions we

developed a scoring schema to efficiently filter the `RNABOB` output. High scoring candidates were then manually examined. With the established secondary structure model of sbRNAs, we were able to find homologies to other RNA families. In particular, sbRNAs turned out to be homologous to vertebrate Y RNAs. Y RNAs are involved in quality control of ncRNAs (Stein et al., 2005). The Ro60-Y RNA complex recognizes specific misfolded RNAs, which are then tagged for destruction. Y RNAs and sbRNAs share the Ro protein binding motif. The group of Torsten Krude recently identified that Y RNAs are required for the initiation step of DNA replication in human (Gardiner et al., 2009; Krude et al., 2009). In a very recent contribution they showed that this function is not dependent on the Ro particle formation (Langley et al., 2010), suggesting that Y RNAs do indeed have two separate modes of action in a cell. The sequence-structure motif that has been identified to be crucial for the function in DNA replication initiation, is also conserved in sbRNAs, leaving no doubt that sbRNAs are true homologs of Y RNAs. There have been two independent homology search studies on Y RNAs (Perreault et al., 2007; Mosig et al., 2007b), but none of these studies identified the huge (240!) group of sbRNAs in nematodes. This is, however, a general problem in RNA homology search. RNA structure models used by `INFERNAL` or `RNABOB` are based on so far identified sequence data. This limited view often results in failure to capture the whole structural spectrum a particular RNA family has adopted over the millions of years of evolution. The Y RNA binding protein Ro can be easily identified in other species by PSI-blast searches. Even a bacterial homolog is known (Ramesh et al., 2007). Y RNAs so far have only been identified in vertebrates and nematodes. In a still unpublished study on Y RNA homology search in the set of species, where no Y RNA homolog is known so far, we successfully identified Y RNAs in lower deuterostomes such as *S. purpuraturs*, *B. floridae*, and *S. saccoglossus*. However, not a single candidate could be discovered in arthropods.

In summary, this thesis addressed methods for *de novo* detection of functional RNA structures as well as strategies for the task of ncRNA homology search. With `RNAz 2.0` and `RNALfoldz` we have presented two computational strategies that can be readily used for *de novo* detection of ncRNAs. Moreover, with two case studies on ncRNA homology search we have not only introduced novel ways for RNA homology search, but also expanded the set of known ncRNA genes.

List of Figures

Bibliography

A Adai, C Johnson, S Mlotshwa, S Archer-Evans, V Manocha, V Vance, and V Sundaresan. Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res*, 15(1):78–91, 2005.

M N Aftab, H He, G Skogerbo, and R Chen. Microarray analysis of ncRNA expression patterns in *Caenorhabditis elegans* after RNAi against snoRNA associated proteins. *BMC Genomics*, 9:278–278, 2008.

A Aizerman, E M Braverman, and L I Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

E Alpaydin. *Introduction to machine learning*. Adaptive computation and machine learning. MIT Press, 2004.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

P P Amaral, M E Dinger, T R Mercer, and J S Mattick. The eukaryotic genome as an RNA machine. *Science*, 319(5871):1787–1789, 2008.

P Anandam, E Torarinsson, and W L Ruzzo. Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics*, 25(5):668–669, 2009.

M Andronescu, A Condon, H H Hoos, D H Mathews, and K P Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):19–28, 2007.

M Andronescu, A Condon, H H Hoos, D H Mathews, and K P Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 2010. In press.

L Argaman, R Hershberg, J Vogel, G Bejerano, E G Wagner, H Margalit, and S Altuvia. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol*, 11(12):941–950, 2001.

R T Arrial, R C Togawa, and M d e M Brigido. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics*, 10:239–239, 2009.

J P Bachellerie, J Cavaillé, and A Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84(8): 775–790, 2002.

H L Barks, R Buckley, G A Grieves, E Di Mauro, N V Hud, and T M Orlando. Guanine, adenine, and hypoxanthine production in UV-irradiated formamide solutions: relaxation of the requirements for prebiotic purine nucleobase formation. *Chembiochem*, 11(9):1240–1243, 2010.

K P Bennett and E J Bredensteiner. Duality and geometry in SVM classifiers. In *Proceedings of the 17th International Conference on Machine Learning*, pages 57–64. Morgan Kaufmann, 2000.

K P Bennett and C Campbell. Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.*, 2(2):1–13, 2000.

I Bentwich, A Avniel, Y Karov, R Aharonov, S Gilad, O Barad, A Barzilai, P Einat, U Einav, E Meiri, E Sharon, Y Spector, and Z Bentwich. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 37(7):766–770, 2005.

S H Bernhart, I L Hofacker, S Will, A R Gruber, and P F Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474–474, 2008.

C M Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006.

A F Bompfünewerer, R Backofen, S H Bernhart, J Hertel, I L Hofacker, P F Stadler, and S Will. Variations on RNA folding and alignment: lessons from Benasque. *J Math Biol*, 56(1-2):129–144, 2008.

R E Bruccoleri and G Heinrich. An improved algorithm for nucleic acid secondary structure display. *Comput Appl Biosci*, 4(1):167–173, 1988.

R J Carter, I Dubchak, and S R Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*, 29(19):3928–3938, 2001.

L Cassimeris, G Plopper, and V R Lingappa. *Lewin's Cells*. Jones and Bartlett Publishers, 2010.

T R Cech, A J Zaug, and P J Grabowski. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27(3):487–496, 1981.

C-C Chang and C-J Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

T H Chang, H D Huang, T N Chuang, D M Shien, and J T Horng. RNAMST: efficient and flexible approach for identifying RNA structural homologs. *Nucleic Acids Res*, 34(Web Server issue):423–428, 2006.

S Chen, E A Lesnik, T A Hall, R Sampath, R H Griffey, D J Ecker, and L B Blyn. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 65(2-3): 157–177, 2002.

S H Chiu, C C Chen, and T H Lin. Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artif Intell Med*, 44(3): 221–231, 2008.

F Chu and L Wang. Applications of support vector machines to cancer classification with microarray data. *Int J Neural Syst*, 15(6):475–484, 2005.

P Clote, F Ferré, 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.

C Cortes and V Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

A Coventry, D J Kleitman, and B Berger. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc Natl Acad Sci U S A*, 101(33):12102–12107, 2004.

F H Crick. The biological replication of macromolecules. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

F H Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

C del Val, E Rivas, O Torres-Quesada, N Toro, and J I Jiménez-Zurdo. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. *Mol Microbiol*, 66(5):1080–1091, 2007.

W Deng, X Zhu, G Skogerbo, Y Zhao, Z Fu, Y Wang, H He, L Cai, H Sun, C Liu, B Li, B Bai, J Wang, D Jia, S Sun, H He, Y Cui, Y Wang, D Bu, and R Chen. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res*, 16(1): 20–29, 2006.

D di Bernardo, T Down, and T Hubbard. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19(13):1606–1611, 2003.

K J Doshi, J J Cannone, C W Cobaugh, and R R Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105–105, 2004.

D Dulebohn, J Choy, T Sundermeier, N Okan, and A W Karzai. Trans-translation: the tmRNA-mediated surveillance mechanism for ribosome rescue, directed protein degradation, and nonstop mRNA decay. *Biochemistry*, 46(16):4681–4693, 2007.

S R Eddy. RNABOB: a program to search for RNA secondary structure motifs in sequence databases. Software available at ftp://selab.janelia.org/pub/software/rnabob, 1996.

S R Eddy and R Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res*, 22(11): 2079–2088, 1994.

S Egloff, E Van Herreweghe, and T Kiss. Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. *Mol Cell Biol*, 26:630–42, 2006.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

FANTOM Consortium. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–73, 2002.

M J Fedor and J R Williamson. The catalytic diversity of RNAs. *Nat Rev Mol Cell Biol*, 6(5):399–412, 2005.

G Felsenfeld, D R Davies, and A Rich. Formation of a three-stranded polynucleotide molecule. *J Am Chem Soc*, 79(8):2023–2024, 1957.

D F Feng and R F Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360, 1987.

G A Fichant and C Burks. Identifying potential tRNA genes in genomic DNA sequences. *J Mol Biol*, 220(3):659–671, 1991.

W Fontana, D A Konings, P F Stadler, and P Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9):1389–1404, 1993.

J R Fresco, B M Alberts, and P Doty. Some molecular details of the secondary structure of ribonucleic acid. *Nature*, 188:98–101, 1960.

E Freyhult, S Edvardsson, I Tamas, V Moulton, and A M Poole. Fisher: a program for the detection of H/ACA snoRNAs using MFE secondary structure prediction and comparative genomics - assessment and update. *BMC Res Notes*, 1:49–49, 2008.

E Freyhult, P P Gardner, and V Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241–241, 2005.

T J Gardiner, C P Christov, A R Langley, and T Krude. A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication. *RNA*, 15(7):1375–1385, 2009.

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. *Nucleic Acids Res*, 37(Database issue):136–140, 2009.

P P Gardner and R Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140–140, 2004.

P P Gardner, A Wilm, and S Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–2439, 2005.

D Gautheret, F Major, and R Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput Appl Biosci*, 6(4):325–331, 1990.

T Gesell and S Washietl. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics*, 9:248–248, 2008.

W Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618, 1986.

S Griffiths-Jones. RALEE–RNA alignment editor in emacs. *Bioinformatics*, 21(2):257–259, 2005.

A R Gruber, S H Bernhart, I L Hofacker, and S Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9:122–122, 2008a.

A R Gruber, D Koper-Emde, M Marz, H Tafer, S Bernhart, G Obernosterer, A Mosig, I L Hofacker, P F Stadler, and B J Benecke. Invertebrate 7SK snRNAs. *J Mol Evol*, 66(2):107–115, 2008b.

A R Gruber, R Neuböck, I L Hofacker, and S Washietl. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res*, 35(Web Server issue):335–338, 2007.

C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

S R Gunn. Support vector machines for classification and regression. Technical report, University of Southampton, 1998.

H C Gürsoy, D Koper, and B J Benecke. The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). *J Mol Evol*, 50(5):456–464, 2000.

R R Gutell, J C Lee, and J J Cannone. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, 12(3):301–310, 2002.

K Hanyu-Nakamura, H Sonobe-Nojima, A Tanigawa, P Lasko, and A Nakamura. Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature*, 451(7179): 730–733, 2008.

J H Havgaard, R B Lyngso, G D Stormo, and J Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–1824, 2005.

H He, L Cai, G Skogerbo, W Deng, T Liu, X Zhu, Y Wang, D Jia, Z Zhang, Y Tao, H Zeng, M N Aftab, Y Cui, G Liu, and R Chen. Profiling *Caenorhabditis elegans* non-coding RNA expression with a combined microarray. *Nucleic Acids Res*, 34(10):2976–2983, 2006.

H He, J Wang, T Liu, X S Liu, T Li, Y Wang, Z Qian, H Zheng, X Zhu, T Wu, B Shi, W Deng, W Zhou, G Skogerbo, and R Chen. Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res*, 17(10):1471–1477, 2007.

M W Hentze, M U Muckenthaler, and N C Andrews. Balancing acts: molecular control of mammalian iron metabolism. *Cell*, 117(3):285–297, 2004.

J Hertel, I L Hofacker, and P F Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2008.

J Hertel and P F Stadler. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):197–202, 2006.

M Hiller, S Findeiss, S Lein, M Marz, C Nickel, D Rose, C Schulz, R Backofen, S J Prohaska, G Reuter, and P F Stadler. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res*, 19(7):1289–1300, 2009.

I L Hofacker, S H Bernhart, and P F Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, 2004a.

I L Hofacker, M Fekete, and P F Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, 2002.

I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

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, 2004b.

I L Hofacker and P F Stadler. RNA secondary structures. In T Lengauer, editor, *Bioinformatics: From Genomes to Therapies*, volume 1, pages 439–489. Wiley-VCH, Weinheim, Germany, 2007.

P Hogeweg and B Hesper. Energy directed folding of RNA sequences. *Nucleic Acids Res*, 12(1 Pt 1): 67–74, 1984.

Y Hokii, Y Sasano, M Sato, H Sakamoto, K Sakata, R Shingai, A Taneda, S Oka, H Himeno, A Muto, T Fujiwara, and C Ushida. A small nucleolar RNA functions in rRNA processing in *Caenorhabditis elegans*. *Nucleic Acids Res*, 38(17):5909–5918, 2010.

M Huarte, M Guttman, D Feldser, M Garber, M J Koziol, D Kenzelmann-Broz, A M Khalil, O Zuk, I Amit, M Rabani, L D Attardi, A Regev, E S Lander, T Jacks, and J L Rinn. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3): 409–419, 2010.

T Joachims. Making large scale SVM learning practical. *Advances in Kernel Methods Support Vector Learning*, pages 169–184, 1999.

F Jühling, M Mörl, R K Hartmann, M Sprinzl, P F Stadler, and J Pütz. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, 37(Database issue):159–162, 2009.

C H Jung, M A Hansen, I V Makunin, D J Korbie, and J S Mattick. Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC Genomics*, 11:77–77, 2010.

M Kertesz, Y Wan, E Mazor, J L Rinn, R C Nutter, H Y Chang, and E Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, 2010.

R J Klein, Z Misulovin, and S R Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A*, 99(11):7542–7547, 2002.

T Krude, C P Christov, O Hyrien, and K Marheineke. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci*, 122:2836–45, 2009.

E C Lai, P Tomancak, R W Williams, and G M Rubin. Computational identification of drosophila microRNA genes. *Genome Biol*, 4(7), 2003.

A Lambert, A Lescure, and D Gautheret. A survey of metazoan selenocysteine insertion sequences. *Biochimie*, 84(9):953–959, 2002.

J B Lambert, S A Gurusamy-Thangavelu, and K Ma. The silicate-mediated formose reaction: bottom-up synthesis of sugar silicates. *Science*, 327(5968):984–986, 2010.

D Langenberger, C I Bermudez-Santana, P F Stadler, and S Hoffmann. Identification and classification of small RNAs in transcriptome sequence data. *Pac Symp Biocomput*, pages 80–87, 2010.

A R Langley, H Chambers, C P Christov, and T Krude. Ribonucleoprotein particles containing non-coding Y RNAs, Ro60, La and nucleolin are not required for Y RNA function in DNA replication. *PLoS One*, 5(10), 2010.

M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

P Larsson, A Hinas, D H Ardell, L A Kirsebom, A Virtanen, and F Söderbom. *De novo* search for non-coding RNA genes in the AT-rich genome of *Dictyostelium discoideum*: performance of Markov-dependent genome feature scoring. *Genome Res*, 18(6):888–899, 2008.

D Laslett and B Canback. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*, 32(1):11–16, 2004.

D Laslett, B Canback, and S Andersson. BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res*, 30(15):3449–3453, 2002.

S Y Le, J H Chen, and J Maizel. Efficient searches for unusual folding regions in RNA sequences. *Structure and methods: Human Genome Initiative and DNA recombination*, 1:127–136, 1990a.

S Y Le, M H Malim, B R Cullen, and J V Maizel. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res*, 18(6): 1613–1623, 1990b.

I Lebars, D Martinez-Zapien, A Durand, J Coutant, B Kieffer, and A C Dock-Bregeon. HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements. *Nucleic Acids Res*, 2010. In press.

R C Lee, R L Feinbaum, and V Ambros. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

Z Lei and Y Dai. An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, 6:291–291, 2005.

N B Leontis and E Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7 (4):499–512, 2001.

Q Li, J P Price, S A Byers, D Cheng, J Peng, and D H Price. Analysis of the large inactive P-TEFb complex indicates that it contains one 7SK molecule, a dimer of HEXIM1 or HEXIM2, and two P-TEFb molecules containing Cdk9 phosphorylated at threonine 186. *J Biol Chem*, 280(31): 28819–28826, 2005.

S C Li, W C Chan, L Y Hu, C H Lai, C N Hsu, and W C Lin. Identification of homologous microRNAs in 56 animal genomes. *Genomics*, 96(1):1–9, 2010.

T Li, H He, Y Wang, H Zheng, G Skogerbo, and R Chen. In vivo analysis of *Caenorhabditis elegans* noncoding RNA promoter motifs. *BMC Mol Biol*, 9:71–71, 2008.

L P Lim, M E Glasner, S Yekta, C B Burge, and D P Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540–1540, 2003.

S Lindgreen, P P Gardner, and A Krogh. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, 22(24):2988–2995, 2006.

J Liu, J Gough, and B Rost. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet*, 2(4), 2006.

J Livny, M A Fogel, B M Davis, and M K Waldor. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res*, 33(13):4096–4105, 2005.

T M Lowe and S R Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–964, 1997.

T M Lowe and S R Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283(5405):1168–1171, 1999.

R B Lyngsø and C N Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7:409–27, 2000.

G C MacIntosh, C Wilkerson, and P J Green. Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol*, 127:765–76, 2001.

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. *Nucleic Acids Res*, 29(22):4724–4735, 2001.

N F Marshall and D H Price. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem*, 270(21):12335–12338, 1995.

M Marz, A Donath, N Verstraete, V T Nguyen, P F Stadler, and O Bensaude. Evolution of 7SK RNA and its protein partners in metazoa. *Mol Biol Evol*, 26(12):2821–2830, 2009.

E Massé, N Majdalani, and S Gottesman. Regulatory roles for small RNAs in bacteria. *Curr Opin Microbiol*, 6(2):120–124, 2003.

O Matan, R K Kiang, C E Stenard, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, L D Jackel, and Y Le Cur. Handwritten character recognition using neural network architectures. In *Proceedings of the 4th United States Postal Service Advanced Technology Conference*, volume 2, 1990.

D H Mathews, M D Disney, J L Childs, S J Schroeder, M Zuker, and D H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, 101(19):7287–7292, 2004.

D H Mathews, J Sabina, M Zuker, and H Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J Mol Biol*, 288:911–940, 1999.

D H Mathews and D H Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, 317(2):191–203, 2002.

J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

P Menzel, J Gorodkin, and P F Stadler. The tedious task of finding homologous noncoding RNA genes. *RNA*, 15(12):2075–2082, 2009.

M Messmer, J Pütz, T Suzuki, T Suzuki, C Sauter, M Sissler, and F Catherine. Tertiary network in mammalian mitochondrial tRNAAsp revealed by solution probing and phylogeny. *Nucleic Acids Res*, 37:6881–95, 2009.

K Missal, X Zhu, D Rose, W Deng, G Skogerbo, R Chen, and P F Stadler. Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J Exp Zool B Mol Dev Evol*, 306(4):379–392, 2006.

T M Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

A Mosig, J Chen, and P F Stadler. Homology search with fragmented nucleic acid sequence patterns. In Raffaele Giancarlo and Sridhar Hannenhalli, editors, *Algorithms in Bioinformatics*, volume 4645 of *Lecture Notes in Computer Science*, pages 335–345. Springer Berlin / Heidelberg, 2007a.

A Mosig, M Guofeng, B M Stadler, and P F Stadler. Evolution of the vertebrate Y RNA cluster. *Theory Biosci*, 126(1):9–14, 2007b.

A Mosig, K Sameith, and P Stadler. Fragrep: an efficient search tool for fragmented patterns in genomic sequences. *Genomics Proteomics Bioinformatics*, 4(1):56–60, 2006.

S Murphy, C Di Liegro, and M Melli. The in vitro transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter. *Cell*, 51(1):81–87, 1987.

J W Nam, K R Shin, J Han, Y Lee, V N Kim, and B T Zhang. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11):3570–3581, 2005.

E P Nawrocki, D L Kolbe, and S R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970.

K L Ng and S K Mishra. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–1330, 2007.

V T Nguyen, T Kiss, A A Michels, and O Bensaude. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*, 414(6861):322–325, 2001.

A Nickel, M Lechner, B Beckmann, C Sharma, A R Gruber, J Vogel, and R K Hartmann. Genomewide comparison and novel ncRNAs in Aquificales. In preparation, 2010.

K Numata, A Kanai, R Saito, S Kondo, J Adachi, L G Wilming, D A Hume, Y Hayashizaki, M Tomita, RIKEN GER Group, and GSL Members. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res*, 13(6B):1301–1306, 2003.

R Nussinov and A B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, 77(11):6309–6313, 1980.

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.

Y Okada, K Sato, and Y Sakakibara. Improvement of structure conservation index with centroid estimators. *Pac Symp Biocomput*, pages 88–97, 2010.

A Pagano, M Castelnuovo, F Tortelli, R Ferrari, G Dieci, and R Cancedda. New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. *PLoS Genet*, 3(2), 2007.

J Pánek, J Bobek, K Mikulík, M Basler, and J Vohradsk. Biocomputational prediction of small non-coding RNAs in Streptomyces. *BMC Genomics*, 9:217–217, 2008.

M Parisien and F Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–5, 2008.

A Pavesi, F Conterio, A Bolchi, G Dieci, and S Ottonello. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res*, 22(7):1247–1256, 1994.

J S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, E S Lander, J Kent, W Miller, and D Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, 2(4), 2006.

J Perreault, J P Perreault, and G Boire. Ro-associated Y RNAs in metazoans: evolution and diversification. *Mol Biol Evol*, 24(8):1678–1689, 2007.

M Petersheim and D H Turner. Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry*, 22:256–63, 1983.

D H Price. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol*, 20(8):2629–2634, 2000.

A Ramesh, C G Savva, A Holzenburg, and J C Sacchettini. Crystal structure of Rsr, an ortholog of the antigenic Ro protein, links conformational flexibility to RNA binding activity. *J Biol Chem*, 282:14960–7, 2007.

J Reeder and R Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104, 2004.

K Reiche and P F Stadler. RNAstrand: reading direction of structured RNAs in multiple sequence alignments. *Algorithms Mol Biol*, 2:6–6, 2007.

J Ren, B Rastegari, A Condon, and H H Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11:1494–504, 2005.

N J Riccitelli and A Lupták. Computational discovery of folded RNA domains in genomes and in vitro selected libraries. *Methods*, 52(2):133–140, 2010.

A Rich. A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids. *Proc Natl Acad Sci U S A*, 46(8):1044–1053, 1960.

A Rich and D R Davies. A new two stranded helical structure: polyadenylic acid and polyuridylic acid. *J Am Chem Soc*, 78(14):3548–3549, 1956.

E Rivas and S R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285:2053–68, 1999.

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.

E Rivas and S R Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8–8, 2001.

D Rose, J Hackermüller, S Washietl, K Reiche, J Hertel, S Findeiss, P F Stadler, and SJ Prohaska. Computational RNomics of drosophilids. *BMC Genomics*, 8:406, 2007.

D Rose, J Jöris, J Hackermüller, K Reiche, Q Li, and P F Stadler. Duplicated RNA genes in teleost fish genomes. *J Bioinform Comput Biol*, 6(6):1157–1175, 2008.

R Salari, C Aksay, E Karakoc, P J Unrau, I Hajirasouliha, and S C Sahinalp. smyRNA: a novel *ab initio* ncRNA gene finder. *PLoS One*, 4(5), 2009.

D Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

P Schattner. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res*, 30(9): 2076–2082, 2002.

P Schattner, W A Decatur, C A Davis, M Ares, M J Fournier, and T M Lowe. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*, 32(14):4281–4296, 2004.

B Schölkopf and A J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

G Schweikert, A Zien, G Zeller, J Behr, C Dieterich, C S Ong, P Philips, F De Bona, L Hartmann, A Bohlen, N Krüger, S Sonnenburg, and G Rätsch. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res*, 19(11):2133–2143, 2009.

W G Scott, J T Finch, and A Klug. The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, 81:991–1002, 1995.

S E Seemann, M J Gilchrist, I L Hofacker, P F Stadler, and J Gorodkin. Detection of RNA structures in porcine EST data and related mammals. *BMC Genomics*, 8:316–316, 2007.

W Seffens and D Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*, 27(7):1578–1584, 1999.

M T Shamim, M Anwaruddin, and H A Nagarajaram. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, 23(24):3320–3327, 2007.

B A Shapiro. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci*, 4(3):387–393, 1988.

P A Sharp. The centrality of RNA. *Cell*, 136(4):577–580, 2009.

J Y Shi, S W Zhang, Q Pan, Y M Cheng, and J Xie. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, 33(1):69–74, 2007.

Y Shi, G W Tyson, and E F DeLong. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature*, 459(7244):266–269, 2009.

T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147 (1):195–197, 1981.

T F Smith, M S Waterman, and W M Fitch. Comparative biosequence metrics. *J Mol Evol*, 18(1): 38–46, 1981.

D Song and Z Deng. A novel ncRNA gene prediction approach based on fuzzy neural networks with structure learning. *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1 –5, 2010.

D Song, Y Yang, B Yu, B Zheng, Z Deng, B L Lu, X Chen, and T Jiang. Computational prediction of novel non-coding RNAs in *Arabidopsis thaliana*. *BMC Bioinformatics*, 10 Suppl 1, 2009.

S Sonnenburg, G Rätsch, A K Jagota, and K-R Müller. New methods for splice site recognition. In *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks*, pages 329–336, London, UK, 2002. Springer-Verlag.

S Sonnenburg, A Zien, and G Rätsch. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):472–480, 2006.

E Sonnleitner, T Sorger-Domenigg, M J Madej, S Findeiss, J Hackermüller, A Hüttenhofer, P F Stadler, U Bläsi, and I Moll. Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools. *Microbiology*, 154(Pt 10):3175–3187, 2008.

J Sridhar, S R Narmada, R Sabarinathan, H Y Ou, Z Deng, K Sekar, Z A Rafi, and K Rajakumar. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*, 5(8), 2010.

P F Stadler, J J Chen, J Hackermüller, S Hoffmann, F Horn, P Khaitovich, A K Kretzschmar, A Mosig, S J Prohaska, X Qi, K Schutt, and K Ullmann. Evolution of vault RNAs. *Mol Biol Evol*, 26(9): 1975–1991, 2009.

A Stark, M F Lin, P Kheradpour, J S Pedersen, L Parts, J W Carlson, M A Crosby, M D Rasmussen, S Roy, A N Deoras, J G Ruby, J Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, E Hodges, A S Hinrichs, A Caspi, B Paten, S W Park, M V Han, M L Maeder, B J Polansky, B E Robson, S Aerts, J van Helden, B Hassan, D G Gilbert, D A Eastman, M Rice, M Weir, M W Hahn, Y Park, C N Dewey, L Pachter, W J Kent, D Haussler, E C Lai, D P Bartel, G J Hannon, T C Kaufman, M B Eisen, A G Clark, D Smith, S E Celniker, W M Gelbart, and M Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450:219–32, 2007.

S Steigele, W Huber, C Stocsits, P F Stadler, and K Nieselt. Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biol*, 5:25–25, 2007.

A J Stein, G Fuchs, C Fu, S L Wolin, and K M Reinisch. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell*, 121(4):529–539, 2005.

X D Sun and R B Huang. Prediction of protein structural classes using support vector machines. *Amino Acids*, 30(4):469–475, 2006.

The FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature*, 409:685–90, 2001.

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

I Tinoco, O C Uhlenbeck, and M D Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

I Tinoco, Jr and C Bustamante. How RNA folds. *J Mol Biol*, 293:271–81, 1999.

M Towsey, P Timms, J Hogan, and S A Mathews. The cross-species prediction of bacterial promoters using a support vector machine. *Comput Biol Chem*, 32(5):359–366, 2008.

M C Tsai, O Manor, Y Wan, N Mosammaparast, J K Wang, F Lan, Y Shi, E Segal, and H Y Chang. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992): 689–693, 2010.

J L Tupy, A M Bailey, G Dailey, M Evans-Holm, C W Siebel, S Misra, S E Celniker, and G M Rubin. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*, 102:5495–500, 2005.

R Upadhyay, P Bawankar, D Malhotra, and S Patankar. A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 144(2):149–158, 2005.

A V Uzilov, J M Keegan, and D H Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:173–173, 2006.

S Valadkhan. Role of the snRNAs in spliceosomal active site. *RNA Biol*, 7(3):345–353, 2010.

C Wang, C Ding, R F Meraz, and S R Holbrook. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22(21):2590–2596, 2006.

S Washietl and I L Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, 342(1):19–30, 2004.

S Washietl, I L Hofacker, M Lukasser, A Hüttenhofer, and P F Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol*, 23(11):1383–1390, 2005a.

S Washietl, I L Hofacker, and P F Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, 2005b.

S Washietl, J S Pedersen, J O Korbel, C Stocsits, A R Gruber, J Hackermüller, J Hertel, M Lindemeyer, K Reiche, A Tanzer, C Ucla, C Wyss, S E Antonarakis, F Denoeud, J Lagarde, J Drenkow, P Kapranov, T R Gingeras, R Guigó, M Snyder, M B Gerstein, A Reymond, I L Hofacker, and P F Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*, 17(6):852–864, 2007.

D A Wassarman and J A Steitz. Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. *Mol Cell Biol*, 11(7):3432–3445, 1991.

L S Waters and G Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–628, 2009.

J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

K C Wiese, E Glen, and A Vasudevan. JViz.Rna-a Java tool for RNA secondary structure visualization. *IEEE Trans Nanobioscience*, 4(3):212–218, 2005.

S Will, K Reiche, I L Hofacker, P F Stadler, and R Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4), 2007.

C R Woese, L J Magrum, R Gupta, R B Siegel, D A Stahl, J Kop, N Crawford, J Brosius, R Gutell, J J Hogan, and H F Noller. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res*, 8(10):2275–2293, 1980.

C Workman and A Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822, 1999.

T Xia, J SantaLucia, M E Burkard, R Kierzek, S J Schroeder, X Jiao, C Cox, and D H Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–14735, 1998.

J H Xu, F Li, and Q F Sun. Identification of microRNA precursors with support vector machine and string kernel. *Genomics Proteomics Bioinformatics*, 6(2):121–128, 2008.

X Xu, Y Ji, and G D Stormo. Discovering cis-regulatory RNAs in shewanella genomes by support vector machines. *PLoS Comput Biol*, 5(4), 2009.

C Xue, F Li, T He, G P Liu, Y Li, and X Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310–310, 2005.

C Xue, F Li, and F Li. Finding noncoding RNA transcripts from low abundance expressed sequence tags. *Cell Res*, 18(6):695–700, 2008.

N Yachie, K Numata, R Saito, A Kanai, and M Tomita. Prediction of non-coding and antisense RNA genes in *Escherichia coli* with gapped Markov model. *Gene*, 372:171–181, 2006.

J H Yang, X C Zhang, Z P Huang, H Zhou, M B Huang, S Zhang, Y Q Chen, and L H Qu. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res*, 34(18):5112–5123, 2006.

Z Yang, Q Zhu, K Luo, and Q Zhou. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature*, 414(6861):317–322, 2001.

G Yuan, C Klämbt, J P Bachellerie, J Brosius, and A Hüttenhofer. RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res*, 31:2495–507, 2003.

D Yusuf, M Marz, P F Stadler, and I L Hofacker. Bcheck: a wrapper tool for detecting RNase P RNA genes. *BMC Genomics*, 11:432–432, 2010.

C Zhong, H Tang, and S Zhang. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res*, 38:e176, 2010.

A Zien, G Rätsch, S Mika, B Schölkopf, T Lengauer, and K R Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.

G Zieve and S Penman. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell*, 8(1):19–31, 1976.

M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.

Curriculum Vitae

Andreas Gruber
Institute for Theoretical Chemistry
Währingerstrasse 17
1090 Wien, Austria

E-mail: agruber@tbi.univie.ac.at
WWW: www.tbi.univie.ac.at/~agruber
Office: +43 1 4277 52731

Date of birth: 05/01/1981
Nationality: Austria

Education

Dr. rer. nat. (equivalent to Ph.D.) in Molecular Biology
University of Vienna, 10/2007 – present
Thesis: Computational noncoding RNA detection
Supervisor: Ivo Hofacker
Expected date of completion: January 2011

DI (equivalent to M.Sc.) in Scientific Computing
University of Vienna, 11/2006 – present
75/120 ECTS completed

Mag. rer. nat. (equivalent to M.Sc.) in Molecular Biology
University of Vienna, 10/2000 – 09/2007
Thesis: Strategies for measuring structural conservation of RNA secondary structures.
Supervisor: Ivo Hofacker

BA in Bioinformatics
Technical University of Vienna, 03/2005 – 11/2006

Employment

Research Assistant, University of Vienna, 10/2009 – present
Project: Structural and functional analysis of mRNA molecules targeted by the RNA-binding protein Tristetraprolin

Research Assistant, University of Leipzig, 06/2009 – 09/2009
Project: Sensory and regulatory RNAs in prokaryotes (DFG SPP 1258)

Research Assistant, University of Vienna, 10/2007 – 05/2009
Project: Bioinformatics Integration Network II – C5 RNA related bioinformatics tools

Publications

Gruber AR, Fallmann J, Kratochvill F, Kovarik P, Hofacker IL. **ARESite: a database for the comprehensive investigation of AU rich elements.**
Nucleic Acids Res. (2011) (Database issue) *in press*

Gruber AR, Bernhart SH, You Zhou, Hofacker IL. **RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures.**
In Proceedings of the *German Conference on Bioinformatics GCB'10. Braunschweig, Germany, pp. 12–21.*

Boria I, Gruber AR, Tanzer A, Bernhart SH, Lorenz R, Mueller MM, Hofacker IL, Stadler PF. **Nematode sbRNAs: Homologs of Vertebrate Y RNAs.**
J. Mol. Evol. (2010) 70(4):346-58

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. **RNAz 2.0: Improved noncoding RNA detection.**
Pac. Symp. Biocomput. (2010) 15:69-79

Gansterer WN, Gruber AR, and Pacher C. **Non-Splitting Tridiagonalization of Complex Symmetric Matrices.**
*ICCS (*2009) 1:481-490

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. **RNAalifold: improved consensus structure prediction for RNA alignments.**
BMC Bioinformatics (2008) 9:474

Gruber AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF. **Arthropod 7SK RNA.**
Mol. Biol. Evol. (2008) 25(9):1923-30

Gruber AR, Lorenz R, Bernhart S, Neuböck R, Hofacker IL. **The Vienna RNA Websuite.**
Nucleic Acids Res. (2008) 36(Web Server issue):W70-4

Gruber AR, Bernhart S, Hofacker IL, Washietl S. **Strategies for measuring evolutionary conservation of RNA secondary structures.**
BMC Bioinformatics (2008) 9:122

Gruber AR, Koper-Emde D, Marz M, Tafer H, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ. **Invertebrate 7SK snRNAs.** *J. Mol. Evol. (*2008) 66(2):107-15

Gruber AR, Neuböck R, Hofacker IL, Washietl S. **The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures.** *Nucleic Acids Res.* (2007) 35(Web Server issue):W335-8

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF. **Structured RNAs in the ENCODE selected regions of the human genome.** *Genome Res.* (2007) 17(6):852-64

Papers in Preparation

Nickel A, Lechner M, Beckmann B, Sharma C, Gruber AR, Vogel J, Hartmann RK, Marz M. **Genome-wide comparison and novel ncRNAs in Aquificiales.**

Tafer H, Rose D, Marz M, Hertel J, Bartschat S, Kehr S, Otto W, Donath W, Tanzer A, Bermudez-Santana C, Gruber AR, Juhling F, Engelhardt J, Busch A, Hiller M, Stadler PF, Dieterich C. **Comparative Analysis of Non-Coding RNAs in Nematodes.**

Kratochvill F, Machacek C, Gruber AR, Vogl C, Hartweger H, Lang R, Hofacker IL, Kovarik P. **Negative feedback determines elimination of unstable inflammation-induced mRNA molecules.**

Professional Activities

Reviewer for Bioinformatics, Pacific Symposium on Biocomputing, Journal of Bioinformatics and Computational Biology, ECCB, Evolutionary Bioinformatics

Co-founder and **developer** of the open source reference manager **Paperpile**

Teaching

Instructor, **UE Computer course for bioinformatic problems in chemistry and biology**, University of Vienna, Fall 2010
Teaching assistant, **UE Exercises for Structural Biology and Theoretical Chemistry**, University of Vienna, Spring 2008
Teaching assistant, **VU Algorithms and Programming in Scientific Computing**, University of Vienna, Spring 2008
Teaching assistant, **VU Introduction into Scientific Computing - Algorithms and Applications**, University of Vienna, Fall 2008
Teaching assistant, **UE Exercises for Theoretical Chemistry and Structural Biology I**, University of Vienna, Spring 2007
Teaching assistant, **UE Laboratory Course: Structural Biology II**, University of Vienna, Fall 2006

Honors and Scholarships

Travel Award, 2010, University of Vienna
PSB 2010 Travel Award, 2010, NIH
Short Term Grant, 2009, University of Vienna
Award for best talk, 2007, BIN II PhD Workshop
Performance Scholarship, 2006, NOEL
Performance Scholarship (bm:bwk), 2004, University of Vienna

Conference Participation

German Conference on Bioinformatics 2010, Braunschweig, Germany (Presenter)
Pacific Symposium on Biocomputing 2010, Hawaii, USA (Presenter)
German Conference on Bioinformatics 2009, Halle, Germany (Poster)
Summer School on Cache-oblivious Algorithms 2008, Aarhus, Denmark (Attendee)
Summer School on Chemoinformatics: CheminfoS3 2008, Obernai, France (Attendee)

Programming Skills

Perl, C, Fortran, Java, bash, Matlab/Octave, R, HTML, Javascript, SVG, SQL, LaTeX