Prediction of Structural Non-Coding RNAs by Comparative Sequence Analysis

DISSERTATION
Zur Erlangung des akademischen Grades
DOCTOR RERUM NATURALIUM

Vorgelegt der
FAKULTÄT FÜR LEBENSWISSENSCHAFTEN
der UNIVERSITÄT WIEN

von
Mag. Stefan Washietl

Institut für Theoretische Chemie

Wien, Oktober 2005

Meiner Mutter

Abstract

Non-coding RNAs (ncRNAs) are transcripts that function directly as RNA molecule without ever being translated to protein. Facing the ever-growing list of newly discovered ncRNAs, it can be expected that further types of ncRNAs are still hidden in recently completed genomes. Unlike protein coding genes, ncRNAs lack any statistically significant characteristics in primary sequence that could be exploited for reliable prediction. Therefore, *de novo* prediction of ncRNAs is still one of the most challenging (but largely unsolved) problem in bioinformatics. Since many functional ncRNAs depend on a defined secondary structure, algorithms based on secondary structure prediction seem to be the most promising.

In the first part of the thesis, we show that thermodynamic stability is a characteristic feature of functional ncRNAs but, if computed for a single sequence, generally not significant enough to reliably distinguish native ncRNAs from the genomic background. However, functional structures are often evolutionary conserved. Using a comparative approach, we could demonstrate that the prediction of a consensus secondary structure of homologous sequences, which considers thermodynamical stability and covariance information, can be a significant measure. We introduced a novel method to assess multiple sequence alignments for thermodynamically stable and evolutionary conserved RNA secondary structures. The method is highly accurate but since it depends on a time-consuming random shuffling algorithm it is not suitable for screens of large genomes.

We therefore developed an alternative algorithm in the second part of the thesis. It consists of two basic components: (i) a novel measure for structure conservation based on consensus structure prediction and (ii) a measure for thermodynamic stability, which, in the spirit of a z-score, is normalized with respect to both sequence length and base composition but can be calculated without sampling from shuffled sequences. With the help of a support vector machine learning algorithm, both scores are combined into a composite score that efficiently detects functional secondary structures in sequence alignments. Our approach was implemented in the program `RNAz`. Benchmarking tests showed that `RNAz` clearly outperforms any other available programs both in terms of accuracy and speed.

In the last part of the thesis, we used `RNAz` to conduct the first comprehensive screen for conserved RNA structures in the human genome. We screened alignments of conserved non-coding DNA of several mammals/vertebrates and predict more than 30,000 putative structural RNA elements throughout the human genome. Our screen recovers hundreds of known structural ncRNAs, it identifies additional members of known ncRNA families, and detects previously undescribed conserved structural elements in some known ncRNAs. Most of the detected RNA structures, however, are of completely unknown function. Our computational results point to thousands of previously undetected functional ncRNAs in the human genome. It provides a strong basis for further theoretical and experimental studies.

Zusammenfassung

Als nicht-kodierende RNAs (ncRNAs) bezeichnet man Transkripte, die im Gegensatz zu mRNAs nicht in Protein übersetzt werden, sondern direkt als RNA-Molekül ihre Funktion ausüben. Die immer länger werdende Liste von neu entdeckten ncRNAs lässt vermuten, dass in vorhandenen Genomdaten noch viele weitere, bisher unbekannte ncRNAs zu finden sind. Im Gegensatz zu Protein-kodierenden Genen haben ncRNAs jedoch keine charakteristischen Eigenschaften in ihrer Primärsequenz, die man für einen effizienten Suchalgorithmus ausnützen könnte. Die Vorhersage von ncRNAs ist daher immer noch eine der größten Herausforderungen der Bioinformatik.

Im ersten Teil der Dissertation wird gezeigt, dass thermodynamische Stabilität der Sekundärstruktur eine charakteristische Eigenschaft für viele funktionelle ncRNAs ist. Die Berechnung der Stabilität einer einzelnen Sequenz reicht jedoch im Normalfall nicht aus, um ncRNAs vom genomischen Hintergrund zu unterscheiden. Funktionelle Sekundärstrukturen sind in vielen Fällen evolutionär konserviert. Basierend auf der Berechnung einer Konsensus-Sekundärstruktur mehrerer homologer Sequenzen, die sowohl die Thermodynamik als auch Covarianz Information berücksichtigt, wurde eine neue Methode entwickelt, um multiple Sequenzalignments auf funktionelle RNAs zu testen. Dieser vergleichende Ansatz führt im Gegensatz zu Einzelsequenzanalysen zu statistisch signifikanten Resultaten. Die Methode beruht jedoch auf einem rechnerisch aufwändigen Zufallsalgorithmus und erscheint daher nicht geeignet für die Analyse großer Genome.

Im zweiten Teil der Dissertation wurde daher ein alternativer Algorithmus entwickelt. Er basiert auf zwei Komponenten: (i) Eine neues Maß für strukturelle Konservierung und (ii) ein Maß für thermodynamische Stabilität, das — gleich einem z-score — normalisiert bezüglich Länge und Basenzusammensetzung ist, aber ohne Zufallsalgorithmus effizient berechnet werden kann. Mit Hilfe eines Support Vector Machine Algorithmus werden beide Komponenten zu einer Größe kombiniert, welche schließlich verwendet wird, um funktionelle RNAs in Sequenzalignments zu detektieren. Die Methode wurde im Programm `RNAz` implementiert. Tests zeigen, dass `RNAz` sowohl in Bezug auf Geschwindigkeit als auch Genauigkeit anderen Programmen klar überlegen ist.

Der letzten Teil der Dissertation beschreibt die erste umfassende Suche nach konservierten RNA Strukturen im menschlichen Genom. Mit Hilfe von `RNAz` wurden Sequenzalignments konservierter nicht-kodierender DNA von Säugetier-/Vertebratengenomen analysiert und 30,000 potentielle strukturelle RNA Elemente vorhergesagt. Die Vorhersage beinhaltet hunderte bekannter ncRNAs, neue Vertreter bekannter ncRNA Klassen und bisher unbeschriebene konservierte RNA Strukturen in einigen bekannten ncRNAs. Die meisten der vorhergesagten RNAs sind jedoch nicht zuordenbar. Die Resultate dieser theoretischen Analyse geben Hinweise auf tausende bisher unbekannte funktionelle ncRNAs im menschlichen Genom und bieten eine gute Grundlage für weitere theoretische und experimentelle Analysen.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 This thesis . . . 3

2 Background **4**

2.1 RNA secondary structure and its prediction . . . 4

2.1.1 General aspects of RNA structure . . . 4

2.1.2 Formal definition and representation of RNA secondary structures . . . 4

2.1.3 The loop based energy model . . . 6

2.1.4 Nussinov algorithm for maximizing base pairs . . . 8

2.1.5 Energy minimization algorithm in the loop based energy model . . . 9

2.1.6 Consensus secondary structure prediction . . . 9

2.2 Secondary structure prediction using stochastic context free grammars . . . 13

2.3 Well known classes of ncRNAs . . . 15

2.3.1 Transfer RNAs and ribosomal RNAs . . . 15

2.3.2 Spliceosomal RNAs . . . 16

2.3.3 Other snRNA-like molecules . . . 16

2.3.4 Small nucleolar RNAs . . . 18

2.3.5 MicroRNAs . . . 18

2.3.6 Other classes of ncRNAs . . . 19

2.3.7 mRNA-like ncRNAs . . . 20

2.3.8 Antisense RNAs . . . 21

2.3.9 Natural ribozymes . . . 21

2.3.10 Functional RNA motifs in UTRs of mRNAs . . . 22

2.4 Genomics of ncRNAs 23

2.4.1 Experimental detection of new ncRNAs 23

2.4.2 Computational detection of ncRNAs 24

2.5 Machine learning with Support Vector Machine algorithms 27

2.5.1 Introduction 27

2.5.2 Hyperplane classifiers 27

2.5.3 The kernel trick 31

2.5.4 Support vector machine regression 32

2.5.5 Support vector machines in practice 33

3 Development, benchmarking and implementation of new algorithms 35

3.1 Free energy of single sequences for detection of ncRNAs 35

3.2 Well-definedness of secondary structure for detection of ncRNAs 37

3.3 Consensus folding as a new measure for detecting ncRNAs 39

3.3.1 Randomizing multiple sequence alignments 40

3.3.2 Creation of test sets 41

3.3.3 Results on Rfam test sets 41

3.3.4 Distribution and significance of *z*-scores 43

3.3.5 Dependence on sequence divergence and alignment method 45

3.3.6 Tests on ncRNAs from *Caenorhabditis elegans* 47

3.3.7 Tests on ncRNAs from *Saccharomyces cerevisiae* 48

3.3.8 Perl implementation: `Alifoldz` and `Shuffle-aln` 50

3.4 An efficient algorithm without shuffling 51

3.4.1 The structure conservation index 51

3.4.2 Empirical estimation of z-scores using linear regression 54

3.4.3 Combining SCI and estimated z-scores 56

3.4.4 Accurate estimation of z-scores using SVM regression 57

3.4.5 SVM classification based on SCI and z-scores 60

3.4.6 Benchmarks on Rfam alignments 60

3.4.7 Comparison to other methods 65

3.4.8 `C` Implementation: `RNAz` 65

4 Screening the Human Genome 68

4.1 Overview 68

4.2 Methods and screen design 68

4.2.1 Selection of most conserved regions 68

4.2.2 RNAz screen 70

4.2.3 Estimating specificity 71

4.2.4 Estimating sensitivity on microRNAs and snoRNAs 71

4.2.5 Non-coding RNA annotation 71

4.2.6 Annotation relative to protein coding genes 72

4.2.7 Comparison with tiling array transcriptional maps 72

4.3 An integrated database system 73

4.4 Results 74

4.4.1 General statistics and specificity 74

4.4.2 Detection performance on known ncRNAs 79

4.4.3 New members of known ncRNA families 81

4.4.4 Structures conserved across all vertebrates 85

4.4.5 Comparison with protein-gene annotations and transcriptional maps . 85

5 Review of other applications 89

5.1 `Alifoldz` prediction and verification of cyanobacterial ncRNAs 89

5.2 Benchmarking of sequence alignment programs upon structural RNAs 89

5.3 `RNAz` predicted miRNA precursors in the miRNAMap database 90
5.4 `RNAz` screens of urochordate and nematode genomes 91

6 Discussion **93**

6.1 Protein prediction vs. ncRNA prediction 93
  6.1.1 Protein prediction, a standard procedure in genome annotation 93
  6.1.2 The “RNA revolution” and the need for ncRNA prediction 93
  6.1.3 ncRNA prediction, an ill-defined problem 94
6.2 Predicting structural ncRNAs 94
  6.2.1 Limitations of available structural ncRNA finders 94
  6.2.2 A next generation of structural ncRNA finders 95
  6.2.3 Limitations of `RNAz` 96
6.3 A prototype screen for ncRNAs of the human genome 97
  6.3.1 Pre-selection of candidates on the basis of sequence conservation . . . 97
  6.3.2 Promising results on known ncRNAs 98
  6.3.3 The real false-positive rate remains uncertain 98
  6.3.4 Can experimental verification help? 99
  6.3.5 Potential functions of newly discovered RNA structures 100
  6.3.6 How many ncRNAs in human? 101
6.4 Conclusion 102

1 Introduction

1.1 Motivation

About twenty years ago, when the plan was announced to determine the complete DNA sequence of the human genome, this ambitious goal was considered to be one the major scientific endeavors in the history of mankind. With the first draft sequence of human published in 2001, the goal was reached much faster than anticipated and it soon became clear that having a complete mammalian genome in hand raises far more questions than it answers. Today, after some years of the "post-genomic" era being reality, we face new challenges. The identification and characterization of all functional elements encoded in a genome has moved into the focus and "functional genomics" has become an important new discipline in current biology. Using both high-throughput experimental and computational techniques, the goal is to find protein-coding genes, non-coding RNA genes (ncRNAs), gene regulatory elements, sequences that mediate chromosome structure/dynamics, and, possibly, functional elements which have not been described so far.

The subject of this thesis is to detect ncRNAs in genomic data. ncRNAs are transcripts that function as RNA molecules without being translated to protein. In the past years, we have seen a series of studies with partly striking and unexpected results. In particular, we want to mention the following observations which have motivated this work:

i. The number of protein-coding genes in human is much smaller than estimated

For more than twenty years, the number of protein coding genes in the human genome has been estimated to be in the order of 100,000. Even in the late 1990s, estimates of up to 150,000 genes were discussed. The first analysis of the draft sequence in 2001 predicted only 35,000 genes and, since then, the number has constantly declined. At the time this thesis was written, 22,000–25,000 proteins were estimated [213], a number which is now well-founded and generally accepted. It shows that the protein repertoire of highly complex organisms like human is comparable to that of much simpler organized organisms. The nematode *Caenorhabditis elegans*, for example, has approximately 20,000 protein genes.

ii. The number of described functional ncRNAs is constantly growing

"Classical" ncRNAs such as tRNAs, rRNAs, or the signal recognition particle RNA, have been known for a long time. Recently, new classes of ncRNAs have been discovered in various organisms suggesting that ncRNA function is much more widespread than believed. For example, the discovering of microRNAs [121, 126] has led to a new paradigm of RNA-directed gene expression regulation. Databases of ncRNAs [70, 139, 174, 69] are constantly growing and ncRNAs are implicated in many cellular pathways and diseases.

iii. The mammalian transcriptome is much more complex than expected

Several independent lines of evidence (cDNA cloning [24], tiling micro-arrays [31], mapping of transcription factor binding sites [27]) indicate that a much higher fraction of the mammalian genome is transcribed than one could explain by known protein-coding genes. The latest large scale cDNA sequencing project reports a striking number of 62.5% of the mouse genome to be transcribed [24]. The vast majority (98% as estimated in reference [156]) does not code for proteins and consists of intronic sequences and other non-coding transcripts. Moreover, detailed analysis of newly detected transcripts show that transcriptional patterns are generally much more complex than previously thought. Extensive use of alternative splicing, different transcriptional starts and ends, and antisense transcription make up an interlaced network of transcription [31, 107, 24, 60] blurring our traditional understanding of a “gene”.

iv. Highly conserved non-coding regions await functional annotation

The fast progress in sequencing technology has it made possible to systematically sequence closely related species of all major model organisms such as bacteria, yeasts, nematodes, insects and even mammals. The power of “comparative genomics” [78] opens new perspectives which can help to better understand the relevant parts in a genome. One of the first results of comparative genomic studies showed that there is a large number of non-coding regions which are highly conserved in evolution. The evidence for purifying selection implies that these regions have some function. This observation has excited lots of interest in the community, probably best reflected in the number of papers and different names describing such regions: Ultra conserved elements (UCE, [10]), highly conserved elements (HCE, [204]), multiple species conserved sequences (MCS, [214]), conserved non-coding sequences (CNS, [77], conserved non-genic sequences (CNG, [48]). The criteria for defining such regions are rather arbitrary but they have at least one thing in common: their function remains enigmatic. One must expect that these conserved non-coding regions correspond to different classes of functional elements, including ncRNAs.

All these points make clear that a complete understanding of the biological processes that constitute a complex organism is impossible if we only consider proteins as important functional entities. A picture is emerging that the complexity we see for example in mammals is the result of the non-coding RNA output which forms a hidden layer of regulation [156, 157].

There is need for experimental and computational techniques to detect and analyze ncRNAs. In this thesis, we develop and apply computational methods to predict ncRNAs in large-scale genomic screens. More precisely, we use comparative sequence analysis techniques to detect evolutionary conserved RNA secondary structures, which are characteristic signals for functional ncRNAs and also regulatory elements in mRNA.

In contrast to protein-gene prediction, *de novo* prediction algorithms for ncRNAs are still in their infancy. We are convinced that a new generation of such algorithms will be helpful to address currently widely discussed questions. Are the thousands of newly discovered non-coding transcripts functional or are they just “transcriptional noise”? Which of the conserved non protein-coding sequences are candidates for functional RNAs? Are there new classes of ncRNAs?

1.2 This thesis

This thesis is based on the following four journal articles as well as unpublished observations. Tables, figures and text passages taken from these articles are used throughout the thesis without further notice.

- Washietl S., Hofacker I.L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342**:19-30 (2004)
- Washietl S., Hofacker I.L., Stadler P.F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **102**:2454-2454 (2005)
- Washietl S., Hofacker I.L., Lukasser M., Hüttenhofer A., Stadler P.F. Genome wide mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in human. *Nat. Biotech.*, in press
- Bompfünewerer A.F., Flamm C., Fried C., Fritzsch G., Hofacker I.L., Lehmann J., Missal K., Mosig A., Müller B., Prohaska S.J., Stadler B.M.R., Stadler P.F., Tanzer A., Washietl S., Witwer C. Evolutionary patterns of non-coding RNAs. *Theor. Biosci.* **123**:301-369 (2005).

The thesis is organized as follows. In section 2, we give a short introduction into the topics relevant for this work. We describe formal aspects of RNA secondary structure and prediction algorithms, we review the current knowledge on ncRNAs, and explain the theory behind support vector machine algorithms. In section 3, we describe the development, benchmarking and implementation of novel algorithms for the detection of conserved RNA secondary structures. This section basically consists of two parts describing two different approaches, which resulted in the two programs `Alifoldz` and `RNAz`. In a first large-scale application, we used `RNAz` to screen the human genome for conserved RNA secondary structures. This screen and the results are described in section 4. Some other applications of our programs were published while this thesis was being written. Five of these papers are briefly reviewed in section 5. We close with a discussion of our methods and the results of the human screen (section 6).

2 Background

2.1 RNA secondary structure and its prediction

The computational biology of RNA structure has a long tradition. First attempts to predict RNA structure were made more than 30 years ago [217]. In this section, we want to give a short overview of the formal aspects of RNA secondary structure and the algorithms for its prediction. We focus on the topics most relevant for this work, namely energy based folding algorithms and their variants. We follow, in part, the presentation of Hofacker & Stadler [84].

2.1.1 General aspects of RNA structure

RNA is a polymer made of covalently linked ribonucleotides. The succession of the four different types of ribonucleotides (adenine, guanine, cytosine and uracil) defines the *primary structure* of the molecule. RNA is generally single stranded but complementary regions in the molecule can fold back onto itself and form double helices similar to DNA. In RNA, we usually find Watson-Crick pairs CG and AU as well as GU "wobble pairs". Although all other combinations of non-standard base pairs can occur, they are generally neglected for the purpose of secondary structure prediction. The intra-molecular base pairing results in a pattern of double helical stretches interspersed with loops which is called the *secondary structure*. The arrangement of secondary structure elements in space finally forms the three-dimensional *tertiary structure*.

The folding of an RNA molecule into its spatial structure can be seen as hierarchical process (Fig. 1). Most of the stabilizing energy of the structure is contributed by secondary structure interactions. Secondary structure forms before and independently of tertiary structure. Tertiary structure usually does not induce changes in secondary structure.

Although the function of an RNA molecule is ultimately dependent on its tertiary structure, secondary structure can be seen as a coarse-grained approximation and is thus a useful level on which to understand RNA function.

2.1.2 Formal definition and representation of RNA secondary structures

A secondary structure is a list of base pairs (i, j) fulfilling the following constraints:

1. A base may participate in at most one base pair.
2. Paired bases must be separated by at least 3 bases

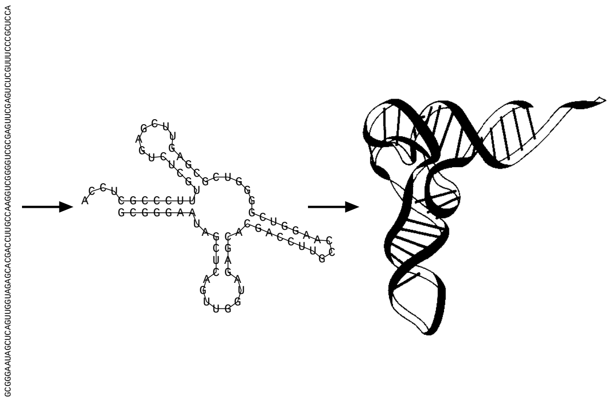

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA

Fig. 1. Hierarchical folding of a tRNA molecule. Complementary regions in the primary sequence (left) can form intra-molecular base pairs which define a pattern of loops and helices, the secondary structure (middle). Secondary structure elements interact with each other in space and form the tertiary structure, the three-dimensional spatial structure of the molecule (right).

3. No two base pairs (i, j) and (k, l) “cross” in the sense that $i < k < j < l$

The first condition excludes tertiary structures motifs such as base triplets and G-quartets. The second constraint defines a minimum loop size of three, which takes into account that the RNA backbone cannot bend too sharply. The third condition excludes pseudoknots. Although pseudoknots are important structural elements in many natural RNAs, they are here (rather arbitrarily) classified as tertiary structure mainly because dynamic programming algorithms cannot deal with them.

With no pseudoknots allowed, a secondary structure can be represented as an outer-planar graph. This means one can draw the secondary structure by placing the backbone on a circle and drawing a chord for every base pair such that no two chords intersect (Fig. 2 a). The most frequently used secondary structure representation is only a more realistic layout of this outer-planar graph (Fig. 2 b).

The so-called mountain representation is another form of drawing a secondary structure, which is in particular well suited for large structures and comparison of structures (Fig. 2 c). In the mountain representation, a single secondary structure is represented in a two dimensional graph, in which the x-coordinate is the position k of a nucleotide in the sequence and the y-coordinate the number $m(k)$ of base pairs that encloses nucleotide k.

In a dot plot representation, each base pair (i, j) is represented by a dot or box in row i and column j of a rectangular grid. This kind of plot can be used to visualize thermodynamic ensembles of a structure. Equilibrium base pair probabilities p_{ij} as computed by McCaskill’s algorithm (section 2.1.5) are usually shown as boxes in the matrix with area proportional to p_{ij} (Fig. 2 d).

Finally, secondary structures can be represented in a very simple and compact string format (Fig. 2 e). For any pair between positions i and j ($i < j$) we place an open bracket "(" at position i and a closed bracket ")" at position j, while unpaired positions in the molecule are represented by a dot ".". Since base pairs may not cross, the representation is unambiguous.

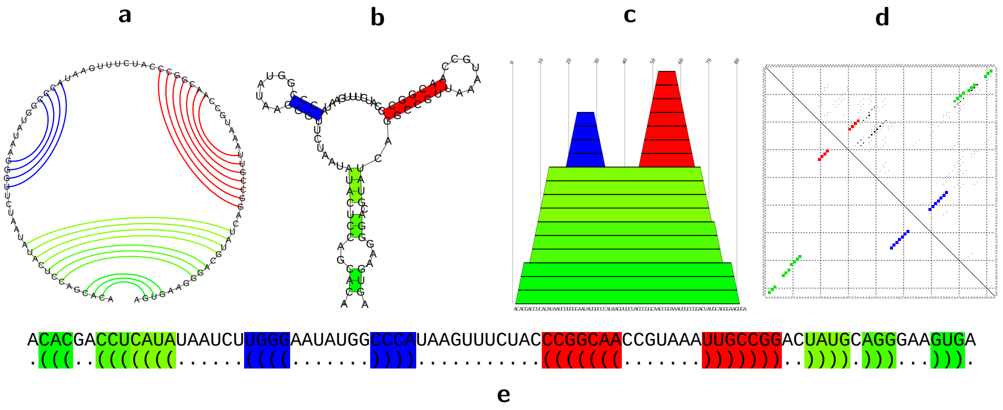


Fig. 2. Representation of RNA secondary structures. (a) Circle plot (b) Conventional secondary structure graph representation (c) Mountain plot (d) Dotplot (e) "Dot/bracket" string notation. The structure shown is a purine riboswitch. Adopted from [84].

2.1.3 The loop based energy model

RNA secondary structures can be uniquely decomposed into *loops*, i.e. the faces of the planar drawing of the structure. More formally, we call a position k immediately interior of the pair (i, j) if $i < k < j$ and there exists no other base pair (p, q) such that $i < p < k < q < j$. A loop then consists of all positions immediately interior of (i, j).

Fig. 3 shows the major types of loops that occur in RNA secondary structures. A loop is characterized by its length, i.e. the number of unpaired nucleotides in the loop, and its degree, given by the number of base pairs delimiting the loop (including the closing pair). Loops of degree 1 are called hairpin loops, interior loops have degree 2, and loops with degree > 2 are called multi-loops.

The loop decomposition forms the basis of the standard energy model for RNA secondary structures which assumes that the energy E of a structure $\mathcal{S}$ can be obtained as the sum over the energies of its constituent loops.

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

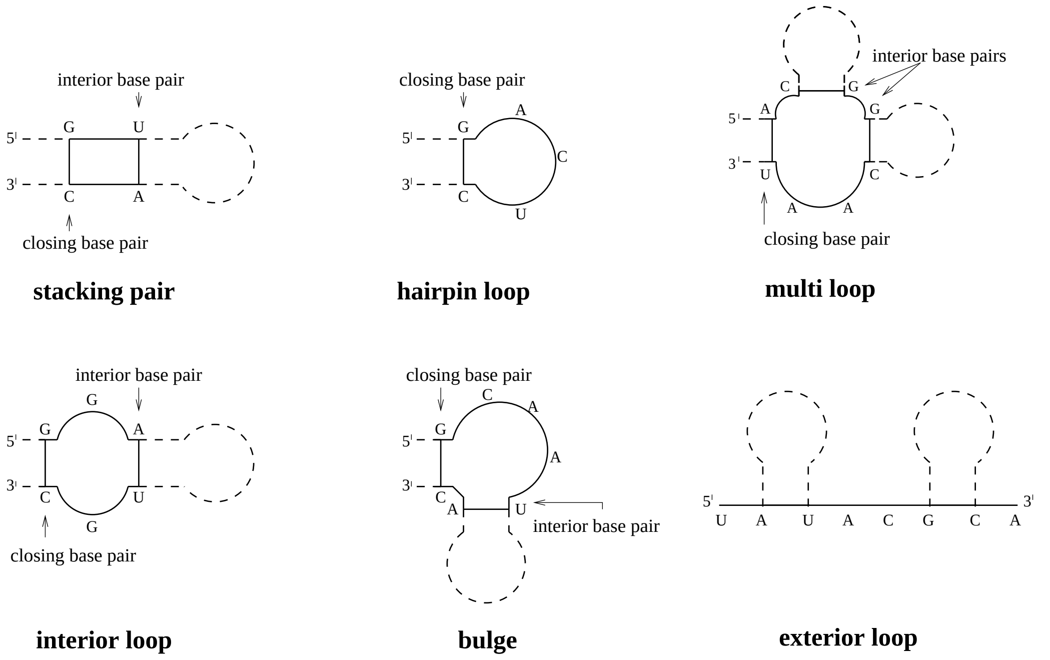


Fig. 3. Classification of loop types in RNA secondary structures.

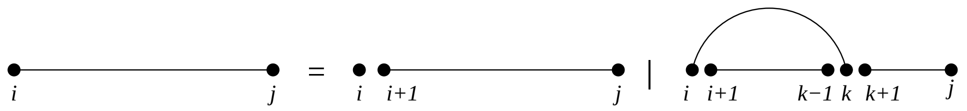
Qualitatively, the major energy contributions are base stacking, hydrogen bonds and loop entropies. While hydrogen bonds and stacking energies can in principle be computed using quantum chemistry, the secondary structure model is solely based on empirically established energy parameters. Essentially, it considers energy differences between folded and unfolded states which are measured by melting experiments. To date, an extensive collection of standard energy parameters measured in a buffer of 1M NaCl at 37^oC has been made available [153, 235, 152].

Stacked base pairs confer most of the stabilizing energy to the secondary structure. Stacking energies are the most carefully measured parameters and tabulated for all possible base pair combinations. With the exception of some other small loop types which are tabulated exhaustively, a simplified model is used for most other loop types. To keep the number of parameters manageable, loop energies are generally split in two terms, describing the size and sequence dependency, respectively. The sequence dependent part only considers the base pairs delimiting the loop and unpaired positions adjacent to these pairs. The size dependent part is extrapolated logarithmically for hairpin loops. In the case of interior loops, loop asymmetry is also taken into account. Multi-loops are modeled linearly in loop size and loop degree to allow for efficient dynamic programming algorithms (see section 2.1.5).

2.1.4 Nussinov algorithm for maximizing base pairs

First attempts to predict secondary structure tried to find the structure having the maximum number of base pairs. Although this approach rarely yields reasonable predictions, it captures the important concepts that current algorithms are built on.

Based on earlier work by Waterman [228, 229], Nussinov proposed the first dynamic programming algorithm solving the maximum base pair problem [168]. Denote $E_{i,j}$ the maximum number of base pairs in a secondary structure of a subsequence $x[i..j]$. One can calculate $E_{i,j}$ in a simple recursive manner. There are only two distinct ways on how the optimal structure of $x[i..j]$ can be formed from a shorter subsequence $x[i+1..j]$:



Either the newly added nucleotide does not pair, in which case the maximum number of base pairs in $x[i..j]$ is the same as in $x[i+1..j]$, or the newly added nucleotide pairs with some partner base k. In the latter case, the maximum number of base pairs in $[x..j]$ is the sum of the base pairs in the two subsequences $x[i+1..k-1]$ and $x[k+1..j]$ plus the newly added pair. Since base pairs may not cross, the two sub-sequences can be treated independently. One can write this recursion as follows:

$$E_{ij} = \max \left\{ E_{i+1,j}, \max_{\substack{i+1 \leq k \leq j \\ \Pi_{ik}=1}} \{E_{i+1,k-1} + E_{k+1,j} + \beta_{ik}\} \right\}$$

where Π is a base pairing matrix with the entries $\Pi_{ij} = 1$ if sequence positions i and j can form a base pair, i.e., if (i,j) is in the set of allowed base-pairs $B = \{\mathsf{GC}, \mathsf{CG}, \mathsf{AU}, \mathsf{UA}, \mathsf{GU}, \mathsf{UG}\}$, and $\Pi_{ij} = 0$ if positions i and j cannot pair. In addition, one can assign a weight β_{ij} to a base pair depending on the type of the base pair (i,j). In the simplest case of finding the maximum number of base pairs, one simply sets $\beta_{ij} = 1$ for all types of pairs.

Finding the optimal structure is a two step process. First, the recursion rules are used to incrementally fill a matrix with the optimal values for all subsequences which eventually leads to the maximum number of base pairs for the whole sequence. In the second step, the optimal structure is found by a *backtracking* procedure. In essence, the path through the matrix which leads to the optimum is reconstructed building up the corresponding structure. The algorithmic complexity of this procedure is $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ in memory and CPU, respectively.

2.1.5 Energy minimization algorithm in the loop based energy model

In order to achieve reasonable prediction accuracies it is essential to use the more sophisticated loop based energy model as described in section 2.1.3. Since loop energies are additive, the optimal structure, now in the sense of the structure of minimum energy, can also be found using a recursive dynamic programming algorithm. Zuker & Stiegler [241] first formulated the recursions for the loop based energy model. Although the recursions are substantially more complicated as for the simple base pair rules, the memory and CPU requirements of the algorithm are still $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$. Fig. 4 illustrates the rationale behind the recursive loop decomposition. Unlike in the simple case of base pair dependent energies where a single matrix is filled, four matrices are required for loop-based energies to find the minimum free energy.

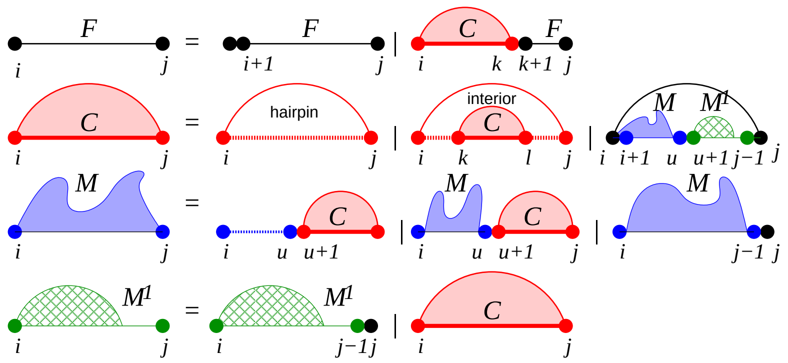
At room temperature, the folding of an RNA molecule is not restricted to a single structure. The molecule will fluctuate between many alternative conformations forming an *ensemble* of structures. Although the minimum free energy structure will be the most likely structure in the ensemble, one generally should not neglect suboptimal structures. The minimum free energy algorithm outlined in Fig. 4 can be extended to calculate also suboptimal structures within a defined range of energy as shown by Zuker [239] and Wuchty *et al.* [234]. Moreover, McCaskill [158] demonstrated that the partition function over all secondary structures $Z = \sum_S \exp(-\Delta G(S)/kT)$ can be calculated by dynamic programming as well. From the partition function, one can calculate the frequency of a base pair occurring in the Boltzmann weighted ensemble of all possible structures, which can conveniently be visualized in a dot-plot (see Fig. 2 d).

Various implementations of the here described prediction algorithms exist. We want to mention the currently most popular ones, the `mfold` package [240] and the `Vienna RNA` package [88]. In this work we used the programs and libraries of the latter one.

2.1.6 Consensus secondary structure prediction

Functional RNA secondary structures are often conserved in evolution. If we have an alignment of related sequences that share a common fold but have diverged in sequence, covariation can be used to improve secondary structure prediction. Hofacker *et al.* elegantly extended the standard energy based algorithm by covariance information yielding fast and accurate consensus structure predictions [87]. The methods developed in this thesis make use of the these algorithms, and it is therefore explained in detail here.

Assume that we are given a multiple sequence alignment $\mathbb{A}$ of N sequences. By $\mathbb{A}_i$ we denote the i-th column of the alignment, while a_i^α is the entry in the α-th row of the i-th column. The length of $\mathbb{A}$, i.e. the number of columns, is n. Furthermore, let $f_i(\mathsf{X})$ be the



$$F_{ij} = \min\left\{F_{i+1,j},\quad \min_{i<k\leq j} C_{ik} + F_{k+1,j}\right\}$$

$$C_{ij} = \min\left\{\mathcal{H}(i,j),\quad \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l),\quad \min_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a\right\}$$

$$M_{ij} = \min\left\{\min_{i<u<j} (u-i+1)c + C_{u+1,j} + b,\quad \min_{i<u<j} M_{i,u} + C_{u+1,j} + b,\quad M_{i,j-1} + c\right\}$$

$$M^1_{ij} = \min\left\{M^1_{i,j-1} + c,\quad C_{ij} + b\right\}$$

F_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$
C_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$ subject to the constraint that i and j form a basepair
M_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$ subject to the constraint that that $x[i..j]$ is part of a multiloop and has at least one component.
M^1_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$ subject to the constraint that that $x[i..j]$ is part of a multiloop and has exactly one component, which has the closing pair i, h for some h satisfying $i \leq h < j$.

Fig. 4. Recursive structure decomposition in the loop-based energy model. The main difference to the simple base pair dependent rules, is that we now have to distinguish between different types of loops (top). Thus we have to further decompose the set of substructures enclosed by the base pair (i, k) according to the loop types: hairpin loop, interior loop, and multi(branched) loops. The hairpin and interior loop cases are simple since they reduce again to the same decomposition step. The multiloop case is more complicated, however, since the multiloop energy depends explicitly on the number of substructures ("components") that emanate from the loop. We therefore need to decompose the structures within the multiloop in such a way that we can at least implicitly keep track of the number of components. To this end we represent a substructure within a multiloop as a concatenation of two components: An arbitrary 5' part that contains *at least* one component and a 3' part that starts with a base pair and contains only a single component. These two types of multiloop substructures are now decomposed further into parts that we already know: unpaired intervals, structures enclosed by a base pair, and (shorter) multiloops substructures. It is not too hard to check that this decomposition really accounts for all possible structures and that each secondary structure has a unique decomposition. Given the recursive decomposition of the structures, the associated recursion formulas (middle) for the energy minimization algorithm can be derived which uses four different matrices (bottom). $\mathcal{H}(i, j)$ denotes the energy of a hairpin loop closed by the pair (i, j), $\mathcal{I}(i, j; k, l)$ denotes the energy of an interior loop determined by the two base pairs (i, j) and (k, l). Multiloop energies are assumed of the form $E_{\mathrm{ML}} = a + b \cdot \mathrm{degree} + c \cdot \mathrm{size}$. Adopted from [84].

the frequency of base X at aligned position i and let $f_{ij}(\mathsf{XY})$ be the frequency of finding X in i *and* Y in j.

The most common way of quantifying sequence covariation for the purpose of RNA secondary structure determination is the mutual information score [33, 74]

$$M_{ij} = \sum_{\mathsf{X},\mathsf{Y}} f_{ij}(\mathsf{XY}) \log \frac{f_{ij}(\mathsf{XY})}{f_i(\mathsf{X}) f_j(\mathsf{Y})} \tag{1}$$

This score is useful if a large set of sequences is available. It does not rely on base pairing rules and, therefore, also non-canonical base pairs or tertiary structural interactions are considered. However, if available data sets are sparse, the signal from the mutual information score is usually too weak. In particular, mutual information does not account for consistent non-compensatory mutations (e.g. $\mathsf{GC} \rightarrow \mathsf{GU}$) which also indicate stabilizing selection on the structure. Hofacker *et al.* introduced a new covariance score C_{ij} considering both compensatory and consistent mutations.

$$C_{ij} = \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^{\alpha} \Pi_{ij}^{\beta} \tag{2}$$

with

$$d_{ij}^{\alpha,\beta} = 2 - \delta(a_i^{\alpha}, a_i^{\beta}) - \delta(a_j^{\alpha}, a_j^{\beta}) \tag{3}$$

where $\delta(a', a'') = 0$ if $a' = a''$ and 0 otherwise. Thus $d_{ij}^{\alpha\beta} = 0$ if the sequences α and β coincide in both aligned positions i and j, $d_{ij}^{\alpha\beta} = 1$ if they differ in one position, and $d_{ij}^{\alpha\beta} = 2$ differ in both positions. Π is again a base pairing matrix with the entries $\Pi_{ij} = 1$ if sequence positions i and j can form a base pair and $\Pi_{ij} = 0$ if positions i and j cannot pair.

One can write C_{ij} as

$$C_{ij} = \frac{1}{\binom{N}{2}} \sum_{\mathsf{XY},\mathsf{X'Y'}} f_{ij}(\mathsf{XY}) \mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} f_{ij}(\mathsf{X'Y'}) \tag{4}$$

where the 16×16 matrix $\mathbf{D}$ has entries $\mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} = d_H(\mathsf{XY}, \mathsf{X'Y'})$ if both $\mathsf{XY} \in \mathcal{B}$ and $\mathsf{X'Y'} \in \mathcal{B}$ and $\mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} = 0$ otherwise. This equation can be reformulated as a scalar product, $C_{ij} = \langle f_{ij} \mathbf{D} f_{ij} \rangle$, and hence efficiently evaluated.

This score rewards compensatory and consistent mutations but it does not penalize inconsistent mutations, i.e. sequences in the alignment that do not form a base pair at the positions

i, j. `RNAalifold` uses a score dealing with inconsistent mutation as follows

$$q_{ij} = 1 - \frac{1}{N} \sum_{\alpha} \Pi_{ij}^{\alpha} \tag{5}$$

This simply counts the number of inconsistent mutations, where a nucleotide paired with a gap is counted as inconsistent and gap/gap combinations are ignored.

This score q_{ij} is combined with C_{ij} to

$$B_{ij} = C_{ij} - \phi_1 q_{ij} \tag{6}$$

where ϕ_1 is a scaling factor controlling the contribution of inconsistent mutations relative to the covariance contribution.

B_{ij} can now be used to extend the minimum free energy folding algorithm with covariance information. For the sake of simplicity, we show this for the simple base pair dependent energy rules (see section 2.1.4). The procedure of finding the optimal consensus structure of the alignment is essentially the same as finding the optimal structure for a single sequence. In the case of a single sequence, each possible base pair is assigned a weight dependent on the type of the base pair. In the case of folding an alignment, a weight is assigned to each pair of columns by averaging the contributions of the single sequences and considering the covariance score described above:

$$\beta_{ij}^{\mathbb{A}} = \frac{1}{N} \sum_{\alpha} \epsilon(a_i^{\alpha}, a_j^{\alpha}) - \phi_2 B_{ij} \tag{7}$$

where $\epsilon(a_i^{\alpha}, a_j^{\alpha})$ is the contribution for a $(a_i^{\alpha}, a_j^{\alpha})$ pair in sequence α. ϕ_2 is another scaling factor controlling the energy contribution relative to the covariance contribution.

The folding algorithm depends on a pairing matrix to decide which columns in the alignment can pair and which not. In a multiple sequence alignment we have to expect alignment or even sequencing errors. Also non-standard base pairs cannot be excluded. It would be too restrictive to simply mark a pair of columns as non-pairing if a single sequence in the alignment cannot pair. Therefore, a threshold value B^* for B_{ij} is used resulting in pairing matrix of the form

$$\Pi_{ij}^{\mathbb{A}} = \begin{cases} 0 & \text{if} \quad B_{ij} < B^* \\ 1 & \text{if} \quad B_{ij} \geq B^* \end{cases} \tag{8}$$

This approach of consensus folding is implemented by `RNAalifold` for the full loop-based energy model. `RNAalifold` calculates a consensus MFE which is not a free energy in a strict physical sense, but rather a “pseudo-MFE” which consists of an energy term and a covariance term. Only three additional parameters (ϕ_1, ϕ_2, B^*) must be set. Using the standard configuration, one compensatory mutation has approximately the same effect as extending a helix by one base pair.

`RNAalifold` performs well in benchmarking tests [61] and, together with `Pfold` [116], it is probably one of best algorithms for predicting consensus secondary structures at present.

2.2 Secondary structure prediction using stochastic context free grammars

All prediction algorithms we have discussed so far are based on energy minimization. Although we exclusively use these methods in our work, it must be mentioned that there also exist different approaches. Stochastic context free grammars represent an alternative and also widely used framework to model RNA secondary structures. We introduce the concept of SCFGs briefly (and rather informally) in this section.

A grammar consists of a number of *symbols* and *production rules*. There are two kind of symbols: abstract *nonterminal* symbols and *terminal* symbols that actually appear in an observed string. The production rules build up a string which conforms to the grammar. For example, the production rules of a simple “palindrome grammar” that generates palindromic words of as and bs can be written as: $S \rightarrow aSa$, $S \rightarrow bSb$, $S \rightarrow \epsilon$. There is one nonterminal S and two terminals a and b. ϵ is a “null string” used as and ending production. The palindrome $aabbaa$ is generated by the production series: $S \Rightarrow aSa \Rightarrow aaSaa \Rightarrow aabSbaa \Rightarrow aabbaa$. This special kind of grammar which allows nested pairwise correlations between terminal symbols is called *context free grammar (CFG)*.

In a similar way, we can write the production rules for a simple RNA secondary structure CFG[1]:

$S \rightarrow aSu \mid uSa \mid cSg \mid gSc$ (base pair)

$S \rightarrow aS \mid cS \mid gS \mid uS \mid Sa \mid Sc \mid Sg \mid Su$ (single nucleotide)

$S \rightarrow SS$ (bifurcation)

$S \rightarrow \epsilon$ (end)

A sequence can be generated in different ways by these rules, corresponding to different secondary structures. The alignment of a sequence to the production rules of CFG can be

---

[1] To save space, alternative productions are written in one line, where “|” means “or”

represented as *parse tree* (Fig. 5).

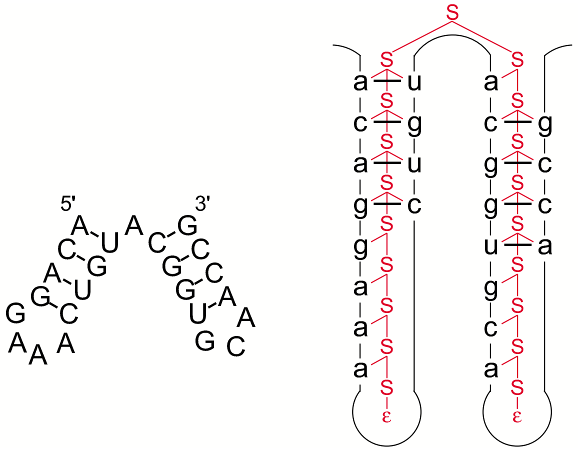


Fig. 5. CFG parse tree of a simple RNA structure element. The grammar is described in the text. Each node in the tree corresponds to a production rule of the grammar, which in turn corresponds to a structure element of the RNA (base pair, single nucleotide, bifurcation). Since the grammar is structurally ambiguous, the parse tree shown here is only one possible way to produce this RNA structure within the rules of this grammar. Adopted from [51].

Simple CFGs can only decide if a sequence can be produced by the production rules or not. This is useful for example in pattern matching applications but much more powerful models can be generated by enhancing CFGs with probability information. In the case of stochastic context free grammars (SCFGs) each production rule is assigned a probability. Each parse tree has thus a different overall probability. Using various standard algorithms, which we will not discuss here, the RNA folding problem can be solved by probabilistic "machine learning":

1. Formulate a SCFG describing the RNA folding model.
2. From a trusted training set of sequences/structures (e.g. tRNAs or rRNAs) estimate the optimal probability parameters for the SCFG.
3. Calculate the maximum probability parse tree of a sequence to the parameterized SCFG.

Since a parse tree corresponds to a secondary structure, we get the optimal secondary structure in step 3. A variety of grammars of different complexity have been developed, including a SCFG counterpart of the complete loop-based energy model [186]. Some of the grammars have prediction accuracies near the performance of current energy minimization programs [116, 51].

Within the framework of SCFGs one can address other RNA structure related problems. Covariance models [52] have become popular to describe RNA families with sequence/structure profiles. `Pfold` [116], the probabilistic counterpart to `RNAalifold`, predicts consensus secondary structures for aligned sequences. In addition, SCFG based methods were used for structure based alignments [93] and, most notably, ncRNA gene finding [187] which will be discussed in more detail later.

2.3 Well known classes of ncRNAs

The number of observed ncRNAs described in the literature is constantly growing. Most of the newly discovered ncRNAs could not be assigned a function. In the rare cases a function is known, the underlying molecular mechanisms are often poorly understood. In this section, we want to give an overview over the current knowledge on ncRNAs. We briefly review all major classes of ncRNAs with the focus on the "classical" ncRNAs which have been known for some time and therefore are best characterized.

2.3.1 Transfer RNAs and ribosomal RNAs

Transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) are the basic components of the protein synthesis machinery and can be found in all domains of life.

Specific types of tRNAs transfer amino acids to the growing polypeptide chain during translation. Comparative sequence analysis of tRNA by means of statistical geometry provides strong evidence that tRNA sequences diverged long before the divergence of archaea and eubacteria [54]. Multiple copies of functional tRNA genes, the existence of numerous pseudogenes and tRNA-derived repeats are general characteristics of tRNA evolution [59].

The ribosome is a high molecular complex made of proteins and rRNAs. In prokaryotes, there are three different rRNAs, the 23S and the 5S rRNA in the large, and the 16S rRNA in the small ribosomal subunit. In eukaryotes, we find four types of rRNAs. The 28S and the two short 5S and 5.8S rRNAs are components of the large subunit, while the 18S rRNA is the RNA component of the small subunit. Evidence from both *in vitro* studies [167] and the analysis of the atomic structure [183] reveals that the ribosome is in fact a ribozyme in which only rRNA is involved in the positioning of the A- site and P-site substrates, and only RNA is in a position to chemically facilitate peptide-bond formation [207]. Due to its ubiquity, size, and generally slow rate of evolution, the small-subunit ribosomal RNA has become the most sequenced of all genes and an invaluable tool for molecular phylogenetics [171].

Most organisms have multiple copies of their rRNA genes. In *Escherichia coli*, for instance, there are seven operons encoding rRNAs 16S, 23S, and 5S. Typical Eukaryotes contain tandemly repeated arrays of rRNAs genes each of which contains three of the four ribosomal

RNA components separated by two “internally transcribed spacers” (18S/ITS1/5.8S/ITS2/28S). In most species the fourth rRNA gene, 5S rRNA, is also contained in this array.

One can observe that paralogous genes in the same sequence are more similar than orthologous sequences of different species [66]. This is the result of “concerted evolution”, i.e. the tendency of different genes in a gene family cluster to evolve “in concert”. This phenomenon can be explained by different molecular mechanisms, for example gene conversion events or frequent duplications and losses with the gene family.

2.3.2 Spliceosomal RNAs

Most genes in higher eukaryotes contain introns that must be excised from the primary transcript to yield a mature mRNA. Intron removal and ligation of the exons occurs in a massive ribonucleoparticle (RNP), the *spliceosome*. Recently, there has been mounting evidence that main catalytic function in the spliceosome are indeed performed by its RNA components, i.e., that the spliceosome, like the ribosome, is essentially a ribozyme [221]. The spliceosomal RNA U1 has an additional function in the regulation of transcriptional initiation [120].

There are three distinct splicing mechanisms that are all dependent on a small set of RNA components of the spliceosome: The major-spliceosome (U1, U2, U4, U5, and U6) is the predominant mechanism e.g. in vertebrates, plants, and yeasts, which splices introns with the “canonical” GT-AG boundaries. The minor-spliceosome (U11, U12, U4atac, U5, U6atac) processes introns with non-canonical boundaries [176], predominantly AT-AC. In some species “trans-splicing” is observed, a mechanism that joins a small non-coding exon derived from the SL RNA to each coding exon of the pre-mRNA and to produce multiple mature mRNAs from a single poly-cistronic pre-mRNA [180].

Both the pol-II transcribed spliceosomal RNAs U1, U2, U4, and U5 and the pol-III transcribed U6 snRNA appear in multiple copies in many vertebrates and are known to be subject to concerted evolution in some species [134, 166]. Divergent paralogs are also known in some species: For example, Xenopus has distinct embryonic and somatic classes of U1 snRNAs [43].

2.3.3 Other snRNA-like molecules

U7 RNA Replication dependent histone pre-mRNAs, in contrast to all other mRNAs, are not polyadenylated. Instead, their 3’-end is cleaved by a ribonucleoprotein (RNP) complex, which consists of the U7 snRNA and three protein components [198]. The U7 snRNA is relatively short (60–70 nt) and forms a well-conserved stem-loop. This ncRNA is specific for the metazoan lineage.

SRP RNA The signal recognition particle (SRP) is responsible for targeting nascent proteins to the ER membrane [109]. The SRP RNP can be found in all three domains of life [189] and contains a ncRNA component which in higher metazoan is also known as 7SL RNA. The secondary structure of SRP RNAs is highly conserved between eukarya and archaea and consists of two domains, the Alu- and S-domain. Protozoan and fungal SRP RNAs deviate considerably, and only the S-domain is present in most bacterial sequences [190].

RNAse P and RNAse MRP The RNase P and RNase MRP RNAs are the catalytically active components of their respective RNPs, which both act as endonucleases. RNase P is essential for the maturation of tRNAs in Bacteria, Eukarya, and Archaea [181]. MRP RNA, in contrast, has been found only in Eukarya where it cleaves the primers necessary for the initiation of mitochondrial DNA replication [165], but also has nuclear functions. RNase P and MRP appear to be ancient paralogs, albeit it remains unclear whether MRP RNA is a eukaryote innovation or an older invention [38]. The absence of structural homology between bacterial and archaeal/eukaryotic RNase P proteins suggests that RNase P once was a pure ribozyme that pursued completely different strategies in the recruitment of protein subunits in the two different lineages [80].

7SK RNA Despite its abundance in mammalian cells, the function of the 7SK RNP has remained unknown until recent studies implicated 7SK RNA as component of the splicing apparatus [120] as well as in the regulation of transcriptional elongation [15]. The 7SK RNA is well conserved across vertebrates but divergent homologues have been also reported in some invertebrate species [118].

Y RNAs Y RNAs are small eukaryotic RNAs that are part of the Ro ribonucleoprotein (Ro RNP) complex, whose function is not known at present. Four families of Y RNAs, Y1, Y3, Y4, and Y5, have been described in human and frog. Their secondary structure is very well conserved among vertebrates [169, 212]. It consists of at least three stems, two of which form a stem-loop structure separated by a relatively short interior loop. The sequences in the stems, as well as parts of the loop regions, are highly conserved and probably serve as binding sites to the Ro60 protein in the Ro RNP complex and/or other cellular nucleic acids.

Vault RNAs Vault RNAs belong to a class of pol-III transcribed RNA genes with poorly understood function. Vaults are cytoplasmic ribonucleoprotein particles believed to be involved in multidrug resistance. The complex contains several small untranslated RNA molecules [222]. So far, vault RNAs have been described only for a few vertebrate species. Vault particles, however, are known also in the slime mold *Dictyostelium discoideum* [223], suggesting that vault RNAs are at least as old as Eukaryotes. The human genome contains at least 4 distinct vaultRNA genes, three of which are located in a small cluster and share external promoter elements [222].

2.3.4 Small nucleolar RNAs

Nascent rRNA transcript are matured in both eukarya and archaea [47, 172] with the help of a large number ribonucleoparticles that modify bases and direct cleavage. The human rRNAs, for instance, together contain more than 200 modified nucleotides [147]. The position of the snoRNA function is determined by the formation of a local snoRNA-rRNA duplex. Two major classes of snoRNA can be distinguished: The C/D box snoRNAs direct 2'-O-methylation of the ribose, while the H/ACA box snoRNAs guide the conversion of uridine nucleotides to pseudouridine. Both classes are characterized by typical primary sequence motifs. H/ACA box snoRNAs fold into a characteristic bipartite stem-loop structure while the secondary structure of C/D box snoRNAs is less pronounced and usually limited to a short base paired region connecting the 3'- and 5'-ends [7].

Besides their canonical roles in rRNA maturation, snoRNAs also target spliceosomal RNA. These snoRNAs perform their function in the Cajal bodies; for this reason they are sometime referred to as scaRNAs ("small Cajal-body associated RNAs") [114]. Most recently, three novel C/D box snoRNAs targeting U2, U4, and U12 snRNAs were identified, that, in contrast to all other known metazoan snoRNAs are independently transcribed [218]. In archaea, tRNAs are also targeted for modification [210], in trypanosomatids the spliced leader SL RNA is modified as well [133, 219]. An intriguing representative of this group is U85, a hybrid snoRNA that has both a functional C/D box and a functional H/ACA box domain that simultaneously modify the U5 snRNA [103]. Some snoRNAs lack complementarity to rRNAs or snRNAs. A small group of "orphan snoRNAs" (U3, U8, U22 and yeast snR10) directs rRNA cleavage instead of modification. The C/D box snoRNA U14, as well as the H/ACA box snoRNAs U17 (also called E1, and homologous to yeast sn30), E2 and E3, are both functional modification guides and play an additional role in pre-rRNA cleavage [56]. An increasing number of recently identified snoRNAs exhibits tissue-specific expression patterns in contrast to all snoRNAs that are known to modify rRNA or snRNA [26]. The genes of these, mostly brain-specific, RNAs are subject to genomic imprinting.

2.3.5 MicroRNAs

MicroRNAs (miRNAs) are a class of small, typically 22–25 nt long single stranded RNAs regulating gene expression [126, 136, 122]. miRNAs function as part of a general RNA mediated silencing pathway. Long, primary miRNA transcripts (pri-miRNAs) are recognized by the nuclear RNase Drosha and processed to a hairpin precursor form (pre-miRNAs), which are then exported from the nucleus to the cytoplasm by exportin-5. The RNase Dicer produce the mature single-stranded miRNA, which is then loaded into the RNA-induced silencing complex (RISC). The miRNA mediates sequence specific interaction of the RISC complex with target mRNAs resulting in translational inhibition or mRNA degradation.

The family of miRNAs is currently rapidly expanding not only in terms of the number of newly discovered members. A wealth of new functions are described for this family implicating miRNAs in development, proliferation, apoptosis, and stress response. To date, miRNAs have been found in animals and plants but also in some viruses (see the `miRNA Registry`, a dedicated database collecting all described miRNAs [69]). All aspects of miRNAs are currently under heavy research. This includes evolution, biogenesis, molecular functions and genomics (detection of miRNAs and their targets). At this place, we only refer to some recent reviews on these topics [175, 209, 42, 19, 81].

2.3.6 Other classes of ncRNAs

Telomerase RNA Telomeres are specialized protein-DNA complexes that cap chromosome ends that are essential for genome stability and cellular proliferation [58]. Sequence loss during replication is counteracted by specialized mechanism(s) in organisms with linear chromosomes [138]. In most organisms, the telomerase RNP extends chromosome ends by iterative reverse transcription of its RNA template, the telomerase RNA [111]. The secondary structures of the telomerase RNAs from vertebrates, ciliates, and yeast vary dramatically in sequence composition and in their size but share a common core structure [30, 44] that hints at an ancient origin. The vertebrate telomerase RNA apparently has co-opted a H/ACA box snoRNA domain [163] during its evolution, shares evolutionary conserved proteins with H/ACA snoRNPs, and contains a Cajal body specific localization signal that is shares with a Cajal body specific subclass of H/ACA snoRNPs [102].

Guide RNAs in trypanosomes RNA editing in trypanosome mitochondria is a unique post-transcriptional maturation process in which uridine residues are inserted and/or deleted at precise sites of mitochondrial mRNAs [67]. Guide RNAs which are usually transcribed from the kinetoplast DNA minicircles [94], provide the information for the editing. A phylogenetic analysis of U-insertion editing [124] suggests that extensive editing is a primitive genetic phenomenon that has disappeared in more modern organism [205].

tmRNA Probably the best-understood bacteria-specific non-coding RNA is the tmRNA, which is part of a ribonucleoprotein complex and combines the functions of tRNAs and mRNAs in order to rescue stalled ribosomes [75]. Usually tmRNA is a single molecule. At least three isolated clades in alpha-proteobacteria [110], cyanobacteria [63, 231], and beta-proteobacteria [203] have two-component tmRNAs, while jakobids have lost the mRNA-like region in their mitochondrial tmRNAs [101].

Prokaryotic sRNAs Prokaryotes contain a diverse set of small non-coding sRNAs. For example, a number of small (40–400 nt) RNAs that neither encode proteins nor function as tRNAs or rRNAs, have been characterized in *E. coli* [83, 224]. The functions of many of these RNAs remain to be determined, while some of them are known to play crucial regulatory roles. There appear to be three general mechanisms: some are integral parts of RNP complexes, such as the 4.5S component of the signal recognition particle and RNase P RNA. A few, such as the 6S RNA, which regulated RNA polymerase activity [164], and the CsrB and CsrC RNAs mimic the structures of other nucleic acids, while a third class, reviewed in [208], acts by specific base pairing with other RNAs. The co-evolution of the small RNA *micF* and its target mRNA *ompF* in enterobacteria was studied in some detail [45]. A curious case are the MCS4 RNAs in mycoplasmas, which have a sequence similarity with eukaryotic U6 snRNAs. Homologs in other bacteria do not seem to exist [220], so that horizontal gene transfer from the host organism is a plausible explanation. Otherwise, very little is know about the origin and evolutionary relationships of the small ncRNAs in prokaryotes.

Viral ncRNAs An increasing number of viral noncoding RNAs have been reported as well. Examples include the recently discovered viral microRNAs [12, 173, 179], the well-known VA1 RNA of adenoviruses [154], which is capable of inhibiting RNAi in human cells [142], the pRNA component of the packaging motor in some bacteriophages [8, 73]. One might suspect that at least some of the conserved RNA structure elements that were discovered in computational surveys of RNA virus genomes [91, 216, 232] are also non-coding RNAs rather than cis-acting elements.

2.3.7 mRNA-like ncRNAs

mRNA-like ncRNAs are RNA transcripts that are polyadenylated and spliced. In contrast to translated genes, they lack long ORFs. This class of ncRNAs is rapidly expanding and many new examples of mRNA-like ncRNAs are collected in ncRNA databases [139, 174]. The best-known mammalian representatives are *H19* and *Xist*. Some of these large ncRNAs, including mammalian *Xist* and *Air*, and *roX* in Drosophila, have distinct roles in epigenetic gene regulation they are performed by means of chromatin modifications. A number of plant specific mRNA-like ncRNAs are known experimentally; additional candidates were detected in a computational survey of *Arabidopsis thaliana* ESTs [145].

2.3.8 Antisense RNAs

Antisense RNAs predominantly act as post-transcriptional downregulators of gene expression [127]. Indeed, some of the RNA families discussed above can be viewed as antisense RNAs since they exert their function by binding complimentary to their target RNAs; examples are the microRNAs, snoRNAs, as well as many of the bacterial small RNAs [225]. The analysis of genomic sequence data, however, has revealed that a substantial fraction of transcribed DNA does not code for proteins and often derives from the anti-sense strand [237]. Antisense transcripts thus emerge as a common mechanism of regulating gene expression in eukaryotic cells [127].

Mechanistically, there are three major pathways: The formation of *double-stranded RNA* may trigger the RNAi pathway and lead to degradation of the sense transcript [76]. Binding of sense and anti-sense transcript may prevent the binding of other trans-acting factors (*RNA masking*). *Transcriptional interference* is the inhibition of transcriptional elongation due to a collision of the RNA Pol-II complexes on overlapping transcriptional units located at opposite strands [182]. Antisense RNAs are transcribed either *in cis* from the opposite strand, or *in trans* from a different genomic locus.

2.3.9 Natural ribozymes

Until about 20 years ago, it was believed that proteins were the only catalytic macromolecules in biology. The discovery of the first catalytic RNA molecules, or ribozymes, in the early 1980s, however, has changed this picture considerably. We have already encountered several examples: RNase P, the spliceosome, and the ribosome are essentially ribozymes. In most cases, ribozymes serve an RNA-processing function using RNA as substrates.

A number of natural ribozymes are not independently stable ncRNAs but rather are part of larger RNA molecules. For example, there are four distinct groups of nucleolytic ribozymes: hammerhead and hairpin ribozymes are mostly found in plant viruses, the Varkud satellite (VS) ribozyme was found in fungal mitochondria, and hepatitis delta virus contains another ribozyme. A recent study suggests a common origin of hammerhead, hairpin, and hepatitis delta ribozymes [79], although convergent evolution cannot be ruled out.

The second large class of naturally occurring ribozymes is involved in the self-splicing of introns in a wide range of species; these molecules belong to one of two structural classes known as *group I* and *group II* ribozymes. All these ribozymes perform different kinds of phosphoryl transfer reactions, in which a transesterification reaction results in breakage of the backbone in the first step [135].

2.3.10 Functional RNA motifs in UTRs of mRNAs

Cis-acting elements in the untranslated regions of mature mRNA bind trans-acting factors and control in this way translational efficiency, mRNA stability and subcellular localization [160]. These elements represent an important group of "functional ncRNAs" albeit they are not independent ncRNAs in the sense of an "RNA-gene". Since secondary structures are involved in various known UTR-elements, they are of interest for our work. A selection of examples of such regulatory motifs in UTRs will be given here.

Iron response elements (IRE) IRE elements are short hairpin structures with an internal loop and a conserved sequence in the hairpin loop, which are observed in 5'-UTRs of ferritin mRNAs in 3'-UTRs of of transferrin receptor mRNAs [82]. They can be classified in two slightly different instances, the first containing an internal loop of length three, which is replaced by a bulge loop in the second. Both have the primary consensus motif CNNNNNCAGWGH [178]. The IRE motif can be readily described with regular grammars; because of the highly redundant sequence pattern and frequent, simple secondary structure one has to expect a large number of false positives, however.

Internal ribosome entry site (IRES) IRES elements were first described in the 5'-untranslated region of picornavirus RNA [104]. The IRES element enables cap-independent initiation of translation starting at an internal initiation codon. In addition to several types of viruses, which contain an IRES element, a small group of eukaryotic mRNA can be translated by internal ribosome entry. IRES-containing mRNAs mostly encode regulatory proteins such as, e.g., growth factors and transcription factors. Several studies have reported that under stress conditions, where cap-dependent translation is blocked, translation of specific mRNAs is enabled through IRES elements [151]. Another function of IRESs involves the control of alternative initiation of translation. For example, the human fibroblast growth factor 2 contains 5 translation initiation codons. Translation initiation of the codon proximal to the 5'-end is initiated by a cap-dependent process, whereas initiation of the remaining codons depends on the IRES [16]. IRES elements are defined by functional criteria and cannot yet be predicted by the presence of characteristic RNA sequence or structural motifs. In general, there are no significant similarities between individual IRESs unless they are from related sources.

Selenocysteine insertion sequences (SECIS) SECIS elements can be found in the coding region of some eubacterial mRNAs and in 3' untranslated regions of some mRNAs in archaea and eukaryotes [119]. In eubacteria, it forms a hairpin structure of conserved length with the selenocysteine codon in the outer helix. In archaea, the primary rather than the secondary structure is conserved. The consensus is a hairpin structure that differs in stem length, occurrence of internal loops and size of the hairpin loop, but it has a very conserved sequence motif in the helix beneath the apical loop. In eukaryotes, the secondary structure contains most of the information while only small sequence motifs are conserved. The core secondary structure is composed of a long hairpin structure consisting of two (type 1) or three (type 2) consecutive helices [57, 119].

2.4 Genomics of ncRNAs

2.4.1 Experimental detection of new ncRNAs

Large scale approaches to characterize the transcriptional output of complete genomes have uncovered a large number of ncRNA candidates. In mouse and human, extensive cDNA libraries derived from the poly-A RNA fraction have been sequenced mainly with the goal to describe the complement of coding mRNAs in these organisms [170, 99]. The analysis of these cDNAs revealed, however, that a substantial fraction does not contain a long open reading frame. The "Fantom" project in mouse described 15,000 of such cDNAs with reduced coding capacity. The "Human Invitational-Project" reports more than 2,000 transcripts with ORFs <80 that passed several additional filters designed to exclude likely protein-coding genes.

In addition, tiling arrays [106] have been used to directly map the transcriptional output of genomes. Also these studies find a significant fraction of transcription not associated with known protein coding genes. To mention only one recent example, Cheng *et al.* constructed a map of 10 human chromosomes at 5 nucleotide resolution from 8 different cell lines [31]. More than 10% of the non-repeat regions in the human genome was represented in the polyA fraction of one or more cell lines. More than 50% of these regions do not overlap with well annotated coding exons, mRNAs or ESTs.

Using a similar tiling array technique in combination with chromatin immunoprecipitation it is possible to map transcription factor binding sites. A detailed map of human chromosomes 21 and 22 suggest that the human genome contains tens of thousands non-coding genes that are bound by common transcription factors and regulated by common environmental signals [27].

All these large scale transcriptional studies are not specifically designed to detect novel ncRNAs. Most of them are limited to polyA transcripts and small ncRNAs are likely to get lost in the experimental procedure. More suitable cDNA cloning techniques for small non-mRNA like ncRNAs have been developed [96]. This approach was applied to various model organisms and in all cases dozens of new species of small (< 500) ncRNAs could be sequenced and verified by northern blot analysis [97, 149, 224, 238, 46]. However, one must suspect random cDNA sequencing to be biased towards strongly expressed ncRNAs.

It appears impossible to quantitatively fish underrepresented RNA species from a pool of highly abundant rRNAs, tRNAs and snoRNAs, even if enrichment strategies are used [46].

2.4.2 Computational detection of ncRNAs

Detection of ncRNAs of known classes Large, highly conserved ncRNAs, in particular ribosomal RNAs, can easily be found using `Blast`. Similarly, `Blast` can be used to find orthologous ncRNAs in closely related species, e.g. [211, 230]. In most cases, however, this approach is limited by the relatively fast evolution of most ncRNAs. Since RNA sequence often evolves much faster than structure, the sensitivity of search tools can be greatly improved by using both sequence and secondary structure information.

The simplest class of search tools uses regular or context free grammars to describe RNA motifs that are explicitly known to the user. There is no possibility to adapt the model to variations of the instance, and it is also very difficult for a user to define production rules for complicated motifs with a large number of exceptions.

With probabilistic models, such as SCFGs (section 2.2), the user is able to assign probability distributions to production rules; noise in the dataset is handled easily because the model can adapt itself to variations. The main drawback of stochastic context free grammars is that most of the available implementations demand large computational resources.

Hybrid languages, like `HyPaL` [68] or the language used in `RNAMotif` [146], connect pattern languages with user defined approximative rules, which rank the results according to their distance to the motif. Their advantage lies in a faster processing compared to SCFG. Nevertheless, the definition of approximative rules also requires explicit knowledge, at least to some extent.

`ERPIN` [64] is an example of tools that do not need an explicit definition of a descriptor to search for homologs of a motif. From a sequence alignment annotated with helix regions it extracts frequencies of nucleotides in single strands and base pair frequencies in helices. Those frequencies are compared to expected base frequencies in the target database by calculating log-odds ratios. The sum of log-odds ratios over all positions of a target sequence gives the final score.

A number of large-scale surveys have been performed using one of the general purpose tools mentioned above. An non-exhaustive list includes a microRNA survey using `ERPIN` [131], a search for U5 snRNA and RNase P using `RNAmotif` [37], and a survey of RNase P RNAs in bacterial genomes [132].

Specialized programs have been developed to detect members of particular ncRNA families. Examples of this approach include `miRseeker` for microRNAs [123], `BRUCE` for tmRNAs [125], `tRNAscan` for tRNAs [140], `snoScan` for box C/D snoRNAs [141], `fisher` for box H/ACA snoRNAs [53], as well as a heuristic for SRP RNAs [184, 190]. An improved method for box C/D snoRNAs was recently presented by Accardo *et al.* [2]: starting from yeast rRNA methylation sites, they first identified homologous positions in *D. melanogaster* rRNAs and then use `snoScan` [141] to search for putative snoRNAs with binding motifs complementary to the putative methylation sites.

De novo prediction Detecting novel ncRNAs without any prior knowledge of sequence or structure is still a largely unsolved issue. In contrast to protein-coding genes, which show strong statistical signals like open reading frames or codon bias, ncRNAs lack any comparable signals in primary sequence that could be used for reliable detection.

Only in very special cases can ncRNAs be identified based on a significant bias in base composition. AT-rich hyper-thermophiles were successfully screened for ncRNAs simply by searching for GC rich regions [115, 194]. MicroRNAs can be detected based on their increased thermodynamic stability [17]. Carter *et al.* used machine learning techniques to extract common sequence features of known ncRNAs including GC content in *E. coli* [25].

Most ncRNAs do, however, depend on a well-defined structure for their function. This has led to various attempts to predict functional RNAs using predicted secondary structures. It was first suggested by Maizel and co-workers that functional RNA elements should have a more stable secondary structure than expected by chance [128, 29]. However, Rivas and Eddy had to conclude in an in-depth study on the subject that thermodynamic stability alone is generally not statistically significant enough for reliable ncRNA detection [186]. Some other characteristic measures derived from secondary structure predictions have been proposed [197, 130, 129] which, however, are also of limited value in the context of genome wide ncRNA prediction. A combination of gene expression data and high level sequence conservation was successful in discovering novel ncRNAs in the intergenic regions of the *E. coli* genome [227].

The reason for the limited success of these approaches is that the presence of secondary structure in itself does not indicate any functional significance, because almost all RNA molecules form secondary structures. In fact, most compelling evidence for functional significance comes from comparative studies that demonstrate evolutionary conservation of structure.

Extensive computer simulations [199, 71, 72, 98], showed that a small number of point mutations is very likely to cause large changes in the secondary structures. It follows that structural features will be preserved in RNA molecules with less than some 80% of sequence identity only if these features are under stabilizing selection, i.e., when they are functional.

This fact is exploited by the `Alidot` [86] algorithm for searching conserved secondary structure patterns in large RNAs. Secondary structures are predicted independently for each sequence, typically using McCaskill's algorithm [158], which yields a list of thermodynamically plausible base pairs with their equilibrium probabilities. Next, a conventional multiple sequence alignment is computed, e.g. using `ClustalW`. By copying the gaps from the multiple sequence alignment into the predicted structures, a list of homologous base pairs is obtained. This list is then sorted by means of hierarchical credibility criteria that explicitly take into account both thermodynamic information and sequence covariation. A detailed description of the method can be found in [86, 90]. A similar approach is taken by the `ConStruct` tool [144, 143], which also features a graphical tool for manipulating the sequence alignment in order to achieve a better consensus structure. `Alidot` does not pre-suppose the existence of a global conserved structure. It is therefore particularly well suited when the sequences are expected to contain only small structurally conserved regions, as is the case for example in RNA viruses. `Alidot` does not provide a measure of significance for its predictions, making it difficult to use it for scanning large non-viral genomes.

There are a few other programs available for the detection of conserved RNA secondary structures. `QRNA` [187] is the most widely known program of this kind. It classifies pairwise sequence alignments as ncRNA, protein coding, or anything else. This program compares the score of three distinct models of sequence evolution to decide which one describes best the given alignment: a pair SCFG is used to model the evolution of secondary structure, a pair hidden Markov model (HMM) describes the evolution of protein coding sequence, and a different pair HMM implements the null model of a non-coding sequence. `QRNA` was successfully used to predict ncRNAs candidates in *E. coli* and *S. cerevisiae* [188, 159], some of which could be verified experimentally.

`MSARi` [40] uses McCaskill's algorithm to detect probable base paired regions in the single sequences of an alignment and then employs a statistical procedure to asses the significance of reverse complementarity.

`ddbRNA` [49], in contrast, does not rely on secondary structures predicted by energetic rules or SCGFs. It simply counts the compensatory mutations in all possible conserved stem loops in the alignment and compares it to a background signal obtained from shuffled alignments.

In this thesis, we adapt and extend the `RNAalifold` approach (section 2.1.6) for the detection of conserved secondary structures resulting in the new program `RNAz`.

2.5 Machine learning with Support Vector Machine algorithms

2.5.1 Introduction

Machine learning algorithms automatically improve by the analysis of data sets, i.e. they "learn" by experience. Speech or handwriting recognition are typical applications of machine learning approaches. Also in computational biology various machine learning techniques have been successfully used, for example neural networks for detection of signal peptides in proteins [11], Hidden Markov models for protein homology detection [108] and, as briefly discussed in section 2.2, stochastic context free grammars for modeling and prediction of RNA secondary structures [52].

In the past years, Support Vector Machines (SVMs), a new class of learning algorithms, have become increasingly popular in computational biology [22]. SVM algorithms come with a number of attractive features:

- Since the training involves optimization of a convex cost function, there is no problem of local minima and the optimal solution can always be found.
- The algorithms are computationally efficient and perform well on noisy and high dimensional data.
- SVM algorithms are modular in their design which simplifies their implementation and analysis.
- SVM algorithms are properly motivated by learning theory.

There exists a large body of literature on SVMs, e.g. [41, 196, 13, 21]. In this work we used SVM techniques for classification and high-dimensional regression tasks. Therefore, we give here a brief overview on the theoretical principles of SVMs.

2.5.2 Hyperplane classifiers

Assume a binary classification task. We have a set of l training points $\{\mathbf{x}_i, y_i\}$ with $i = 1, \ldots, l,\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in \{-1, +1\}$. Each training point consists of a vector of n features and belongs to the positive or negative class. We want to find a classification function $f(\mathbf{x})$ such that $\mathbf{x}$ is assigned to the positive class if $f(\mathbf{x}) \geq 0$, and otherwise to the negative class.

In the simplest case, if the the training points are linearly separable, a hyperplane

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \text{ with } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R} \tag{9}$$

can be constructed which separates the positive from the negative training examples (Fig. 6). However, an infinite number of such hyperplanes exist. Which one will best classify unseen examples, i.e. generalizes well? Intuitively, one will choose the hyperplane with the maximum margin between any training point and the hyperplane (Fig. 7). There are arguments from learning theory supporting this intuition.

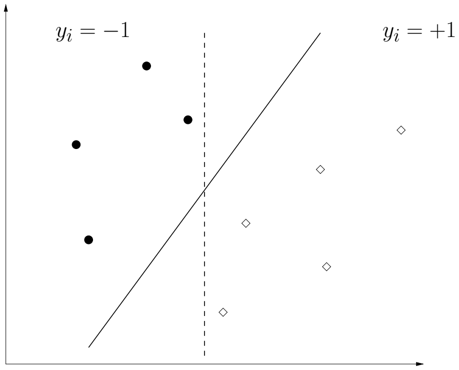


Fig. 6. A binary classification problem. The goal is to find a decision function which optimally separates black circles from open diamonds. In the linearly separable case (as shown here), a linear decision function in the form of a hyperplane can be found that classifies all examples without error. There exists an infinite number of such hyperplanes. Two possible hyperplanes are shown (drawn as solid and broken lines). It is plausible that the solid line represents a better classifier because it has a wider "margin" (this is the distance of the closest point to the hyperplane).

To find the hyperplane with the largest margin, one implicitly rescales $\mathbf{w}$ and b such that the points closest to the hyperplane satisfy $\langle \mathbf{w} \cdot x_i \rangle + b = 1$ $(i = 1, \ldots, l)$ and thus $y_i(\langle \mathbf{w} \cdot x_i \rangle + b) \geq 1$ for all data points $i = 1, \ldots, l$. In this case, the margin equals $2/\|\mathbf{w}\|$. Maximizing the margin is equivalent to the following constrained optimization problem:

$$\underset{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \; \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{10}$$

$$\text{subject to } y_i(\langle \mathbf{w} \cdot x_i \rangle + b) - 1 \geq 0 \text{ for all } i = 1, \ldots, l \tag{11}$$

Optimization problems of this kind are dealt with using *Lagrangian* theory. The constraint equation is multiplied by positive Lagrange multipliers $\alpha_i \geq 0, i = 1, \ldots, l$ and subtracted from the objective function yielding the Lagrangian:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1) \tag{12}$$

The Lagrangian L has to be minimized with respect to $\mathbf{w}$ and b and maximized with re-

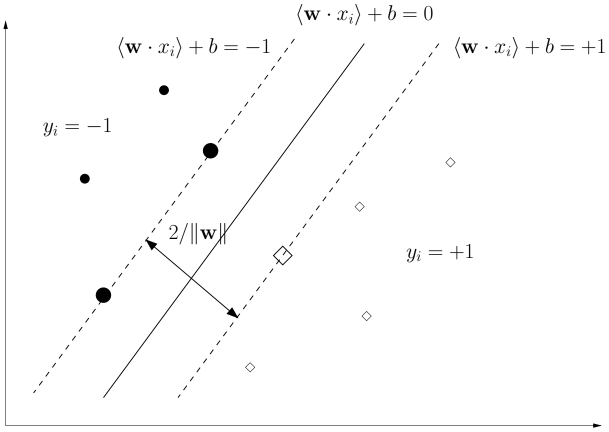

$\langle \mathbf{w} \cdot x_i \rangle + b = 0$
$\langle \mathbf{w} \cdot x_i \rangle + b = -1$
$\langle \mathbf{w} \cdot x_i \rangle + b = +1$
$y_i = -1$
$2/\|\mathbf{w}\|$
$y_i = +1$

Fig. 7. Constructing the optimal separating hyperplane. Three "canonical" hyperplanes are shown which define the maximum margin of the classification problem. They are scaled in a way that the maximum margin is $2/\|\mathbf{w}\|$. The points nearest to the optimal separating hyperplane are called the "support vectors" (drawn larger than the other points).

spect to α_i. This type of quadratic optimization problem has been studied extensively and numerous algorithms are available offering efficient numerical solution. Although one could directly solve this form of the optimization problem, for the effectiveness of SVM algorithms it is crucial to transform the problem into its so-called *dual* form. To this end, the primal Lagrangian function is differentiated with respect to the *primal* variables $\mathbf{w}$ and b which must vanish: $\frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha) = 0$ and $\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha) = 0$

This leads to

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{13}$$

and

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i. \tag{14}$$

Substitution of (13) and (14) into the Lagrangian (12) eliminates the primal variables $\mathbf{w}$ and b yielding the *dual* optimization problem that is usually solved:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \tag{15}$$

$$\text{subject to } \alpha_i \geq 0 \text{ for all } i = 1, \ldots, l \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{16}$$

Maximizing W is equivalent to minimizing the primal Lagrangian L. This reformulation, however, unveils important properties of the SVM algorithm. Each training point has a corresponding α_i. The solution vector $\mathbf{w}$ is an expansion in terms of a subset of the training points (14). Only those training points with non-zero α_i contribute to the solution, all other training points are not relevant. The hyperplane is thus completely determined by the data points closest to it, the so-called *Support Vectors*. More precisely, the SVs lie on the margin which can be seen through the Karush-Kuhn-Tucker complementary condition which is satisfied in this kind of constrained optimization problem:

$$\alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1] = 0 \text{ for all } \; i = 1, \ldots, l \tag{17}$$

Furthermore, the training data only appears in the form of dot products between the vectors in the optimization problem. This property allows for the elegant extension to non-linear problems as described in the next section.

Finding the solution for our initial problem of binary classification in the linearly separable case can be summarized as follows: (i) Solve the optimization problem in (15) (i.e. find the optimal α using the training data and an appropriate numerical algorithm. (ii) Recover b by making use of the condition in equation (17). (iii) Rewrite the hyperplane decision function (9) as:

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \langle \mathbf{x} \cdot \mathbf{x}_i \rangle + b \tag{18}$$

In practice, the training data will not always be linearly separable and the algorithm will fail to find a separating hyperplane. A simple modification that allows classification errors can help. One introduces slack variables $\xi_i \geq 0$ for all $i = 1, \ldots, l$ in order to relax the strict constraints in (11). Any training point falling on the wrong side of its supporting hyperplane is considered to be an error. Now we want to find the hyperplane with the largest margin and the smallest error arriving at a new optimization problem:

$$\underset{\mathbf{w}, \xi \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \; \tau(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{l} \xi_i \tag{19}$$

$$\text{subject to } y_i(\langle \mathbf{w} \cdot x_i \rangle + b) - 1 \geq \xi_i \text{ for all } i = 1, \ldots, l \tag{20}$$

The constant $C > 0$ determines the trade-off between margin maximization and training error minimization. The *dual* form of this optimization problem turns out to be exactly the same as before (15). The constraints of the Lagrange multipliers, however, have an upper bound bound $0 < \alpha_i < C$ for all $i = 1, \ldots, l$. This variant of the hyperplane classifier is generally used in SVM algorithms and known as *soft margin classifiers*.

2.5.3 The kernel trick

Learning algorithms based on hyperplane classifiers as outlined in the previous section have been known since the 1950s [191]. However, this type of classifier cannot be used if there is no linear relationship between the data points (Fig. 8). In combination with another rather old mathematical method, the so-called kernel trick [3], hyperplane classifiers can be extended to non-linear problems in a surprisingly straightforward way. This combination was introduced as *Support Vector Machine* in 1992 by Vapnik and co-workers [18].

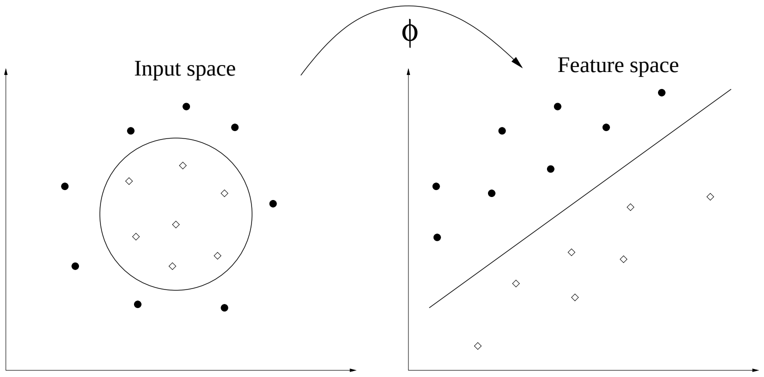


Fig. 8. Mapping input data in to a higher dimensional features space. Problems that are not linearly separable in input space can be linearly separable in feature space.

A linear classification algorithm can be converted to a non-linear algorithm by adding additional attributes to the data that are non-linear functions of the original data. In the example in Fig. 8, introducing a quadratic term could solve the classification problem. More generally, one needs to map the data from the original *input space* X into a higher dimensional *feature space* F: $\Phi : X \mapsto F : \mathbf{x} \mapsto \phi(\mathbf{x})$.

The training algorithm (17) and the resulting decision function (18) is unaffected by this mapping. The decision function (18) with mapped input data:

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i) \rangle + b \tag{21}$$

The only major drawback is that the algorithm becomes computationally infeasible if the data has to be explicitly mapped to a high dimensional feature space. An important feature of the dual representation is that the training vectors only appear as dot product in the algorithm. So we only need to evaluate $\langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i) \rangle$.Through the use of a *kernel* function it is possible to compute this dot product in feature space directly as a function of the original input points.

A kernel function is a function K, such that for all $\mathbf{x}, \mathbf{z} \in X$

$$K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle,$$

where ϕ is a mapping from the input space X to a (dot product) feature space F.

By substituting the dot products $\langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i) \rangle$ in the learning algorithm (17) by a kernel function $K(\mathbf{x}, \mathbf{x}_i)$ we can now solve a non-linear classification problem by using a linear algorithm which works in a high (probably infinite) dimensional feature space. Since we do not need to explicitly map the data into feature space (we do not even need to know the map ϕ) the high dimensionality does not impose a performance problem.

Without going into detail, one can show that essentially each function that gives rise to a positive matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{ij}$ is a kernel of some feature space. For specific applications one can design kernel functions which best represent the properties of the particular data. In practice, standard kernel functions yield good results for most types of applications. Some commonly used kernel functions are listed here:

- linear: $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x} \cdot \mathbf{z} \rangle$
- polynomial: $K(\mathbf{x}, \mathbf{z}) = (\gamma \langle \mathbf{x} \cdot \mathbf{z} \rangle + c_0)^d$
- radial basis: $K(\mathbf{x}, \mathbf{z}) = exp(-\gamma \|\mathbf{x} - \mathbf{z}\|^2)$

2.5.4 Support vector machine regression

The idea of finding a hyperplane as decision function for classification tasks can be easily generalized to regression estimation, which we only discuss briefly here.

In the case of regression, the training points are of the form $(\mathbf{x}, y)$ with $y \in \mathbb{R}$. The algorithm tries to construct a linear function $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ such that the training points lie within a specified distance $\varepsilon > 0$ (Fig. 9). Points that lie outside this "ε-tube" are penalized similar to the soft margin classifier (cf. equation 19).

Again, this can be formulated as constrained optimization problem that can be solved using kernel functions. The ε bound has to be specified *a priori*. In a more recent formulation of the algorithm, which we will use in section 3.4.4, a parameter $0 \leq \nu \leq 1$ can be set as the fraction of points allowed to lie outside the "tube" and the corresponding ε is computed automatically.

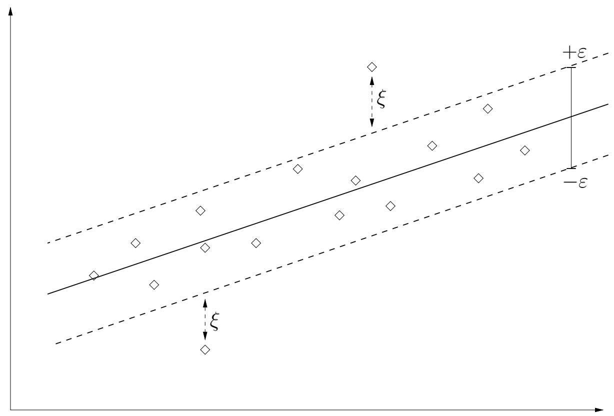


Fig. 9. Principle of SVM regression. A hyperplane is constructed that fits the training data. Points within a specified range ϵ are not penalized while any points outside this range are considered as training errors and penalized with a cost proportional to ξ.

2.5.5 Support vector machines in practice

The application of SVM algorithms to every-day problems has been facilitated considerably by various easy-to-use software packages. We used `Libsvm` [28] throughout this work. The `Libsvm` package provides a `C` and `Java` library which implements all major algorithms for classification and regression. There are language bindings for all major scripting languages such as `Perl`, as well as plugins for mathematical and statistical software packages (e.g. `Matlab` and `R`). For standard applications, one can get reasonable results by following these guidelines (this assumes that one of the standard kernels described above is used):

1. Compile a representative test set of positive and negative examples.
2. Define the input vector, i.e. choose characteristics of the examples which are relevant for classification and, if necessary, convert them to real values.
3. Uniformly scale all elements of the input vector (e.g. linearly to an interval [0,1]).
4. Choose a kernel and optimize C and the kernel parameter(s) by self-consistency tests.
5. Apply the model to unknown examples.

The first two points are problem-specific. Finding a reasonable large test set is not always easy if available data is sparse. Depending on the application, defining and extracting the characteristic features can either be straightforward, or one of the most challenging problems in the whole process (e.g. image recognition). The other steps are problem-independent and, in some SVM packages, partly automatized. Choosing the kernel and the parameters is an essentially empirical process which can be guided by self-consistency tests. Self-consistency tests give an impression of the performance of a model in terms of sensitivity and specificity. Given that the test set is representative, one can expect similar classification performance also for new examples.

3 Development, benchmarking and implementation of new algorithms

3.1 Free energy of single sequences for detection of ncRNAs

In previous chapters we highlighted that many known functional RNAs are tied to a defined secondary structure and we presented well established algorithms for their prediction. Intuitively, MFE predictions appear to be a straightforward measure also for the detection of functional RNAs. However, prediction programs readily calculate MFE structures for arbitrary random sequences. The question arises, if natural RNAs are more stable (have lower MFE) than random sequences.

A simple statistical procedure can be used to address this question. To assess the significance of a MFE for a given sequence, we generate a number of randomized sequences by shuffling the positions [117]. This gives us sequences of the same length and base composition. We calculate the mean μ and standard deviation σ of the MFEs of these random sequences and compare it to the MFE m of the native sequence by calculating a normalized z-score $z = (m - \mu)/\sigma$. Negative z-scores indicate that the native sequence is more stable than one could expect by chance.

We tested six structural RNA families (tRNA, 5S rRNA, Hammerhead ribozyme type III, Group II catalytic intron, Signal recognition particle RNA, U5 spliceosomal RNA). We used `RNAfold` with standard parameters for the MFE prediction and calculated z-scores from a sample of 100 random sequences. The results are shown in Tab. 1. On average, the structural RNAs have all z-scores clearly below zero, meaning they have lower folding energy than the random samples. Is this significant enough to reliably distinguish single sequences from the random background? Fig. 10 illustrates this for the tRNA test set. The histogram shows the distribution of z-scores for 579 tRNAs together with the z-scores of 579 random sequences (one shuffled version for each tRNA). If we use a conservative limit of -4 to define a significant z-score, we can only detect 2% of the tRNAs. To detect half of all tRNAs we would have to lower the cutoff to -1.8. Then, however, we would encounter 4% of false positives. For genome-wide screens where a huge number of candidates has to be scored, the specificity is too low (especially for a corresponding sensitivity of only 50%). Some of the tested families form more stable structures (e.g. Group II catalytic intron: average z=-3.88, Hammerhead ribozyme III: z=-3.08) but generally the native sequences are not efficiently separated from the bulk of random sequences.

An additional point seems noteworthy regarding these experiments. Workman & Krogh [233] pointed out that dinucleotide content influences secondary structure predictions, because of the energy contributions of stacked base pairs. A correct randomization procedure should,

Tab. 1. z-scores of MFEs of various functional RNAs

ncRNA Type	n	z_{mono}	z_{di}
tRNA	579	-1.84	-1.71
5S rRNA	606	-1.62	-1.71
Hammerh. III	251	-3.08	-3.17
Gr. II Intron	116	-3.88	-3.77
SRP RNA	73	-3.37	-3.09
U5	199	-2.73	-2.38

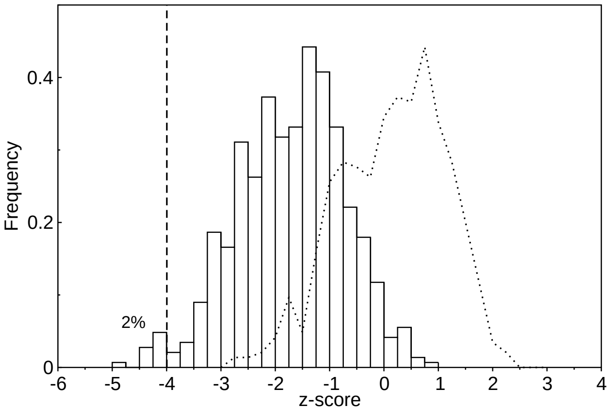


Fig. 10. Distribution of MFE z-scores of 579 tRNAs. Solid bars: native RNAs, dashed line: random controls. Only 2% of the tRNAs have significant z-scores below -4.

therefore, generate random sequences of the same dinucleotide content. It is impossible to consider this in the randomization of multiple sequence alignments (see section 3.3). For single sequences, however, we performed the z-score calculations with both mono- and dinucleotide shuffled random sequences. For dinucleotide shuffling we used a recent implementation [36] of an algorithm developed by Altschul & Erickson [4]. The results (Tab. 1) show that a systematic bias is not recognizable for our test sets. The values differ only minimally and the mononucleotide-shuffled z-scores are not necessarily below the dinucleotide-shuffled score. Thus, while dinucleotide composition was important in the study of Workman & Krogh where mRNAs are tested for an (obviously non-existent) subtle bias towards lower folding energies, it can be neglected in our case.

We can conclude from these results, that folding energy is indeed a characteristic signal of (structural) ncRNAs, but is in itself not sufficient for a reliable detection. This finding is consistent with similar studies on the subject [186, 36] and the reason why any efforts to build a general RNA gene finder based only on MFE predictions have failed so far.

3.2 Well-definedness of secondary structure for detection of ncRNAs

Since thermodynamic stability alone is not significant enough, we have to look for additional characteristics of ncRNAs. Various measures have been proposed to capture such characteristics of naturally evolved, functional RNA structures [130, 129, 197]. However, none of them appeared to be a useful measure efficiently complementing the simple MFE calculations described above.

Here we briefly address the question whether a measure for "well-definedness" as defined by Hofacker (2003, unpublished) can be used for detection of functional RNAs.

At room temperature an RNA molecule will fluctuate between different secondary structures. In equilibrium (i.e. given there are no high energy barriers in the landscape) the ensemble of visited structures is described by the Boltzmann distribution. For some sequences this ensemble will be dominated by the ground state, MFE structure, in which case we call the structure well-defined, for other sequences the ensemble contains several dis-similar structures with near equal frequencies.

The simplest measure of well-definedness is given by the probability of the MFE structure $\mathcal{S}_{\min}$ in the ensemble

$$P(\mathcal{S}_{\min}) = \exp(-E_{\min}/RT)/Z. \tag{22}$$

This measure, however, does not take into account whether the structures in the ensemble are diverse or just trivial variations of one another. Ideally, we would like to measure the structural variety of the ensemble as the mean distance between structures in the ensemble. There are many possible distance measures. One of the simplest is the so-called base pair distance which counts the number of base pairs that are either in $\mathcal{S}_1$ or in $\mathcal{S}_2$ but not in both. This measure is especially well-suited for comparing structures on the same sequence (as in our case), since it corresponds to the minimal number of base opening/closing moves necessary for re-folding from one of the structures to the other. As demonstrated by Hofacker, the mean base pair distance in the ensemble $\langle D \rangle$ can be computed directly from the pair probabilities as:

$$\langle D \rangle = \sum_{i<j} p_{ij} - p_{ij}^2 \tag{23}$$

The concept of "well-definedness" is illustrated in Fig. 11 which shows the base pair probability dotplots of a two secondary structures with different mean base pair distance.

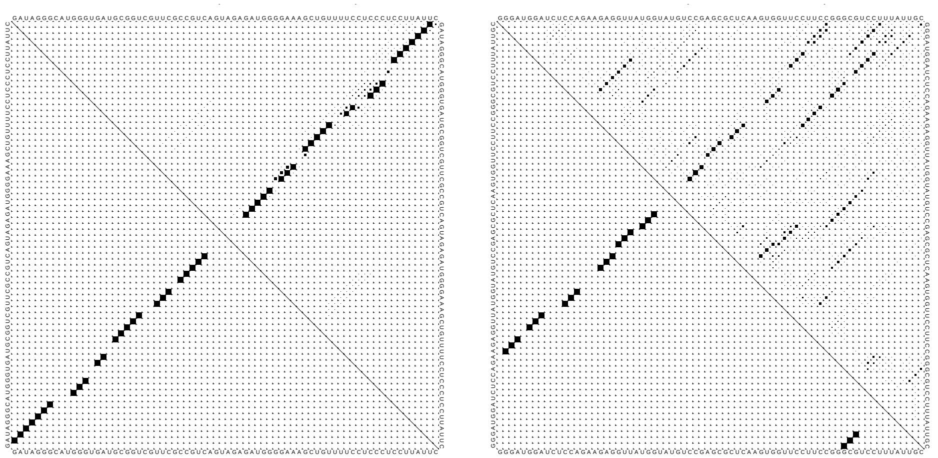


Fig. 11. "Well-definedness" of a secondary structure. Base pairing probabilities for two artificial sequences of the same length were calculated by McCaskill's algorithm. The pairing probabilities are shown in the upper right triangle of the dot plot while the lower right triangle shows the pairing matrix of the MFE structure. Both structures have similar MFE structures. The left structure is rather well-defined in a sense that there a relatively few alternative structures in the ensemble. In contrast, the right structure is not well-defined because many partly completely different structures exist in the ensemble with reasonable probabilities. The left and right structures have mean base pair distances of 3.27 and 15.7, respectively.

In analogy to the statistical tests of MFEs described in the previous section, we calculated z-scores of mean pair distances (Tab. 2). Also here, the z-scores are below zero in all cases indicating that natural ncRNAs have on average a lower mean base pair distance in their thermodynamical ensemble than the random controls. However, the effect is not as pronounced as in the case of MFE z-scores. Moreover, we observed that to some degree both measures, MFE and well-definedness, are correlated as exemplified in Fig. 12 again on the tRNA test set. The scatter plot shows that the native sequences are clearly better separated from the controls on the MFE axis than on the mean base pair distance axis.

Without elaborating on statistical details, we concluded that also the measure of well-definedness is not of immediate interest for the purpose of developing an efficient ncRNA gene finding algorithm. Therefore, we did not pursue this topic any further.

Tab. 2. z-scores of mean base pair distances for various functional RNAs

ncRNA Type	n	z
tRNA	579	−0.5
5S rRNA	606	−0.7
Hammerh. III	251	−1.5
Gr. II Intron	116	−1.2
U5	199	−2.73

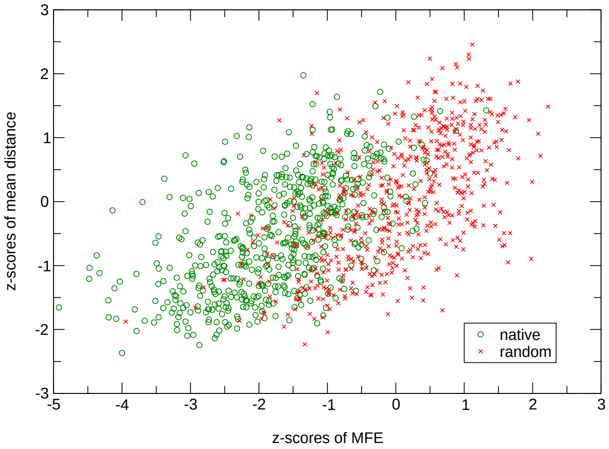


Fig. 12. Scatter plot of z-scores of MFE and mean base pair distances on the tRNA test set. The mean base pair distance measuring the "well-definedness" is partly correlated with the MFE and does not significantly improve the classification.

3.3 Consensus folding as a new measure for detecting ncRNAs

The results so far show that single sequence predictions are of limited statistical significance. Given the wide availability of comparative data mentioned in the introduction, we wondered how to efficiently make use of this information. We use the program `RNAalifold`, which was originally developed to predict consensus secondary structures of aligned sequences (section 2.1.6). `RNAalifold` calculates an averaged MFE for the alignment, incorporating covariance information into the energy model. We consider `RNAalifold`-MFEs to be a good measure for the existence of a conserved fold and a good alternative for the probabilistic approach implemented in `QRNA`. `RNAalifold` makes use the standard energy model for RNA secondary structures, and thus reduces to simple MFE structure prediction in the case of single sequences. For an alignment of several sequences the energy model is augmented through covariance information. `RNAalifold` is not limited in the number of input sequences.

It seems to be straightforward to apply the ideas described in the previous sections and to calculate z-scores of `RNAalifold`-MFEs in order to test whether ncRNAs can be more efficiently detected by including comparative information from homologous sequences. To this end, we developed an algorithm for randomizing multiple sequence alignments and created test sets from sequences in `Rfam`, as described in the next two sections.

3.3.1 Randomizing multiple sequence alignments

The randomization procedure is of crucial importance for the calculation of meaningful z-scores. A straightforward algorithm would simply shuffle the columns of the alignment. This would result in an alignment of the same length, the same base composition and the same overall conservation. However, the gap structure and the local conservation pattern would be different. Possible consequences for consensus folding and z-score calculations are illustrated in Fig. 13. If there is for example a gap of length 10 in the alignment, the shuffling probably would produce 10 gaps of length 1. This can result in artefactual low z-scores since many gaps spread over the complete alignment can remarkably impair the consensus folding, while one long gap probably does not. The same is true for local conservation patterns, meaning that a well conserved column AAAAAGG should not be shuffled with a less conserved column AGUACUA, but rather with a column CCCCCAA of the same pattern. We considered this in our shuffling algorithm: First we collect all columns which have the same gap structure and local conservation pattern into individual groups of columns. We memorize which column of the initial alignment has which pattern. Subsequently, we shuffle the groups individually using a standard procedure [117]. Finally, we reassemble the alignment. Since the shuffling procedure of the individual sets is provably random and independent from each other, all possible alignments are sampled with the same probability.

It must be pointed out that we here only shuffle columns with exactly the same pattern of nucleotide succession (i.e. we shuffle AAAAAGG with CCCCCAA but no with CCAAAAA). Alternatively, one might shuffle columns of the same *degree* of conservation but different pattern (this option is implemented in the program `Shuffle-aln`, see section 3.3.8). While we cannot think of a possible scenario where this could introduce randomization artifacts, we decided to use the more restrictive version here.

As the conservative shuffling procedure restricts the possible number of permutations, the question arises if it is effective enough to destroy a secondary structure. It is known that if only a small fraction (around 10%) of a sequence is randomly mutated this leads almost certainly to unrelated structures [199]. These theoretical considerations, as well as our computational results, suggest that the shuffling procedure is effective enough to destroy any native secondary structures.

3.3.2 Creation of test sets

Most of the RNA sequences used in this work were taken from the `Rfam` database release 5.0 [70]. We took the sequences from the *full* alignments of Hammerhead ribozyme III (RF00008), Group II catalytic intron (RF00029) and U5 spliceosomal RNA (RF00020). For tRNA (RF00005) and 5S rRNA (RF00001) we used the sequences from the *seed* alignment. In the case of tRNA, the number of the sequences in the seed alignment was reduced to 579 (we removed every second of the 1161 sequences). The signal recognition particle RNA test set was taken from the SRP database [189]. We used the 73 eukaryotic sequences that could be found in the database as of January 2004.

To get a reasonable number of non-redundant alignments of different size N (2 to 4 sequences) within a defined range of mean pairwise identity (65% to 85%) and ideally with all sequences of the test set equally represented, we used the following procedure: First, we roughly clustered the sequences using `BlastClust` and created clusters with approximate pairwise identities between 60% and 95%. Within those clusters we computed all possible combinations for a given N. From each cluster we randomly chose a varying number of combinations taking into account the size of the cluster. This should avoid that the resulting alignments are made up just by a fraction of the sequences of the initial test set (which can easily happen because the number of possible combinations can get very large). In the next step, the collected sequence combinations were realigned using `ClustalW` [215] and the mean pairwise identities were calculated. For the experiments shown in Tab. 3 and Figs. 15 and 14, we eventually used alignments with mean pairwise identities between 65% and 85%. To estimate the false positive rate, we generated a shuffled version of each of the alignments. Here we used all alignments generated by the procedure above. This set consisted of 5930 alignments with mean pairwise identities between 30% and 100%, GC-content between 30% and 70% and length between 50 and 350 columns. The test set included 3280 pairwise alignments, 1701 alignments with N=3 and 949 alignments with N=4.

3.3.3 Results on Rfam test sets

The results for the z-score calculations using our randomization algorithm are summarized in Tab. 3 and Fig. 15. If we compare the average z-score from the single sequences to the average z-scores of the pairwise alignments (N=2), we observe in all cases that the average z-score drops by almost 2. It further drops for the alignments consisting of three and four sequences. We want to recall that the units of z-scores are standard deviations, so that even

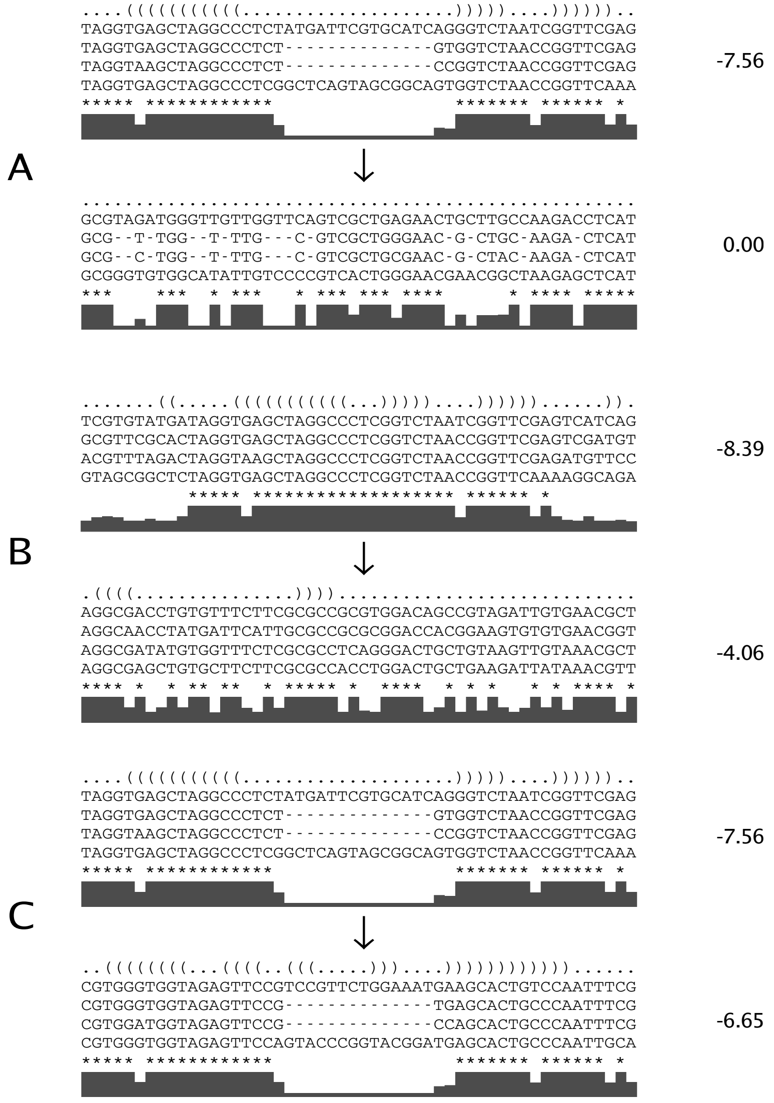

A
....(((((((((((....................)))))....))))))..
TAGGTGAGCTAGGCCCTCTATGATTCGTGCATCAGGGTCTAATCGGTTCGAG
TAGGTGAGCTAGGCCCTCT--------------GTGGTCTAACCGGTTCGAG
TAGGTAAGCTAGGCCCTCT--------------CCGGTCTAACCGGTTCGAG
TAGGTGAGCTAGGCCCTCGGCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
***** ************   ******* ****** *
-7.56
....
GCGTAGATGGGTTGTTGGTTCAGTCGCTGAGAACTGCTTGCCAAGACCTCAT
GCG--T-TGG--T-TTG---C-GTCGCTGGGAAC-G-CTGC-AAGA-CTCAT
GCG--C-TGG--T-TTG---C-GTCGCTGCGAAC-G-CTAC-AAGA-CTCAT
GCGGGTGTGGCATATTGTCCCCGTCACTGGGAACGAACGGCTAAGAGCTCAT
***   *** * ***   * *** *** ****   * **** *****
0.00
B
.......((.....((((((((((((...)))))....)))))))......)).
TCGTGTATGATAGGTGAGCTAGGCCCTCGGTCTAATCGGTTCGAGTCATCAG
GCGTTCGCACTAGGTGAGCTAGGCCCTCGGTCTAACCGGTTCGAGTCGATGT
ACGTTTAGACTAGGTAAGCTAGGCCCTCGGTCTAACCGGTTCGAGATGTTCC
GTAGCGGCTCTAGGTGAGCTAGGCCCTCGGTCTAACCGGTTCAAAAGGCAGA
          ***** ******************* ****** *
-8.39
.((((..................)))).........................
AGGCGACCTGTGTTTCTTCGCGCCGCGTGGACAGCCGTAGATTGTGAACGCT
AGGCAACCTATGATTCATTGCGCCGCGCGGACCACGGAAGTGTGTGAACGGT
AGGCGATATGTGGTTTCTCGCGCCTCAGGGACTGCTGTAAGTTGTAAACGCT
AGGCGAGCTGTGCTTCTTCGCGCCACCTGGACTGCTGAAGATTATAAACGTT
**** *   * ** **   * ***** *   ****   * * *   * * **** *
-4.06
C
....(((((((((((....................)))))....))))))..
TAGGTGAGCTAGGCCCTCTATGATTCGTGCATCAGGGTCTAATCGGTTCGAG
TAGGTGAGCTAGGCCCTCT--------------GTGGTCTAACCGGTTCGAG
TAGGTAAGCTAGGCCCTCT--------------CCGGTCTAACCGGTTCGAG
TAGGTGAGCTAGGCCCTCGGCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
***** ************   ******* ****** *
-7.56
..((((((((...((((..(((.....)))....)))))))))))))......
CGTGGGTGGTAGAGTTCCGTCCGTTCTGGAAATGAAGCACTGTCCAATTTCG
CGTGGGTGGTAGAGTTCCG--------------TGAGCACTGCCCAATTTCG
CGTGGATGGTAGAGTTCCG--------------CCAGCACTGCCCAATTTCG
CGTGGGTGGTAGAGTTCCAGTACCCGGTACGGATGAGCACTGCCCAATTGCA
***** ************   ******* ****** *
-6.65

Fig. 13. Randomization of multiple sequence alignments. Three examples of shuffled alignments are shown. In A and B, the alignments are randomized by simply shuffling the columns. In C, only columns of the same gap pattern and local conservation pattern are shuffled. The degree of conservation is illustrated by black bars of varying size and asterisk for perfectly conserved columns. Each alignment was folded using `RNAalifold`. The consensus secondary structure prediction is shown in dot/bracket-notation in the first line. The `RNAalifold`-MFE is shown next to the alignment. (A) The alignment has one long gap in the middle which is spread over the whole length of the alignment after shuffling. In the resulting random alignment, `RNAalifold` cannot predict a consensus secondary structure (MFE=0.0). This results in significant low z-scores (-4.1 in this special case) although there is no unusually stable structure in the initial alignment (see C). (B) A highly conserved block is embedded in a less conserved region. Shuffling destroys this block and the consensus structure of the resulting random alignment is thus more unstable. Artifacts of this kind can lead to low z-scores and thus false positives. (C) The same alignment as in A is shuffled using our conservative algorithm. The randomized alignment retains the gap pattern and local conservation pattern of the initial alignment. It has a comparable MFE although the consensus structure is completely different (they do not have a single base pair in common). Using this shuffling procedure, we obtain a meaningful z-score of -0.8.

small changes shift the sensitivity significantly (for fixed z-score threshold). In Tab. 3 we calculated detection sensitivities for a threshold of -4. In Fig. 14 the z-score distribution is shown for the tRNA alignments with varying N. Folding of pairwise alignments instead of single sequences improves sensitivity from 2.1% to 71.1%. For $N = 4$, the native alignments are completely separated from the random alignments and almost all score below -4 (98.4%).

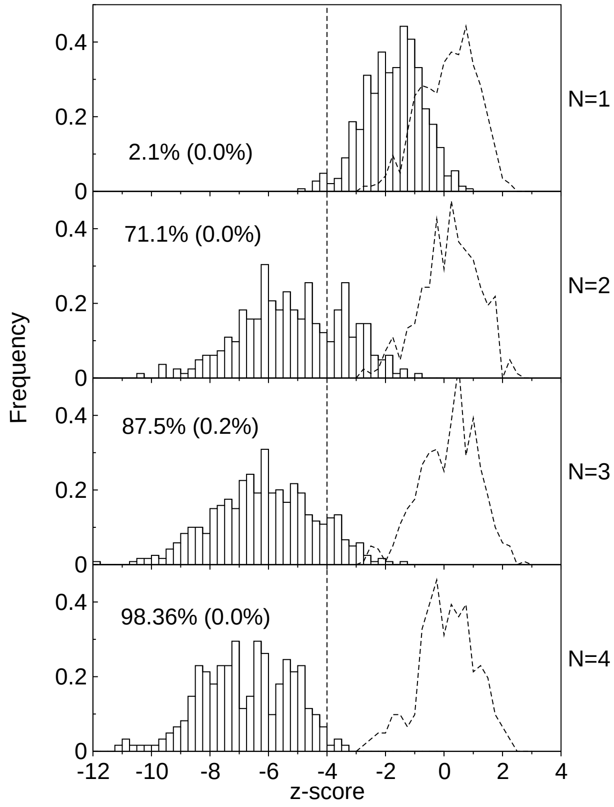


Fig. 14. Distribution of z-scores for the tRNA test sets. The distribution of native z-scores are shown as bars. The distribution of z-scores of the corresponding random sequences are shown as dashed line. N is the number of sequences in the alignment. $N = 1$ means `RNAfold` predictions for single sequences. The sensitivity (percent of native alignments with a z-score below a threshold of -4) and the false positive rate (percent of random alignments with z-scores below -4) are shown for each set.

3.3.4 Distribution and significance of z-scores

Sensitivity and specificity depend on a predefined z-score threshold. To estimate the false positive rate for our test set, we also scored a shuffled random control for each alignment in the set. The distribution of 5930 random z-scores is shown in Fig. 16. Three alignments had z-scores below -4. This means that the sensitivities shown in Tab. 3 have a corresponding false positive rate of 0.05%. The form of the distribution is of particular interest. It can

Tab. 3. z-scores and detection sensitivities for single and aligned sequences of various functional RNAs

					Number of sequences in alignment											
	Single sequence				2				3				4			
ncRNA Type	n	z_{mono}	z_{di}	S	n	ID	z	S	n	ID	z	S	n	ID	z	S
tRNA	579	-1.84	-1.71	2.24	329	76.60	-5.15	71.12	479	73.29	-6.13	84.47	244	75.65	-6.76	98.36
5S rRNA	606	-1.62	-1.71	5.11	87	77.34	-3.89	40.23	81	80.03	-5.26	70.37	102	79.24	-5.12	69.61
Hammerh. III	251	-3.08	-3.17	8.80	94	76.07	-5.50	80.85	120	78.44	-6.10	93.33	130	79.74	-6.11	98.46
Gr. II Intron	116	-3.88	-3.77	44.82	109	75.98	-5.79	89.91	138	76.26	-7.00	94.20	134	76.06	-7.03	96.27
SRP RNA	73	-3.37	-3.09	34.24	135	77.29	-6.52	89.63	55	78.42	-7.09	90.91	50	78.75	-7.59	92.00
U5	199	-2.73	-2.38	17.58	110	74.32	-4.36	49.09	125	74.88	-5.14	64.80	127	74.57	-5.43	71.65

n … number of sequences/alignments scored, ID … average mean pairwise identity, z … average z-score, S … sensitivity (% below -4).

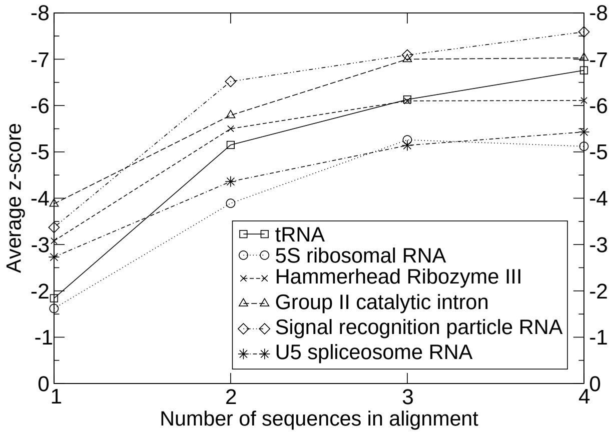


Fig. 15. Mean z-scores of various RNA types dependent on the number of sequences in the alignment. $N = 1$ means `RNAfold` predictions for single sequences. Mean pairwise identities of the alignments are between 65% and 85%. See Tab. 3 for more details.

be fairly well approximated by a standard normal distribution. However, the distribution is slightly skewed with a negative tail: There are apparently more z-scores below -3 than z-scores above $+3$. This tail is not due to our shuffling algorithm. Single sequences (whether mono- and dinucleotide shuffled) show the same skew in the distribution (not shown), as noted also in other studies [186]. A possible explanation might be that we select the *minimum* free energy from random sequences and one could therefore expect behavior similar to an extreme value distribution. We also attempted a fit to an extreme value distribution but our data can much better be explained by a normal distribution although it underestimates the negative tail.

In any case, the significance of a given cutoff has to be estimated empirically. Especially for genome-wide studies it cannot be assumed that the genomic background behaves exactly like random alignments and it might be possible that various inhomogeneities cause more false positives than experienced here. The false-positive rate will depend on preparation of the data (e.g. masking of repeats and low complexity regions) and the quality of the alignments. This is exactly what we find for automatically generated yeast alignments (see section 3.3.7) where the -4 cutoff has a much higher false-positive rate (0.25%) compared to our test set of `ClustalW` alignments of `Rfam` sequences.

3.3.5 Dependence on sequence divergence and alignment method

`RNAalifold` takes a multiple sequence alignment as input. It can predict an existing consensus structure only if the sequence alignment reflects common structural properties. Ideally, one would like to feed `RNAalifold` with structurally aligned sequences. However, existing algorithms [193, 85], are much too slow to make this a feasible alternative for a large number

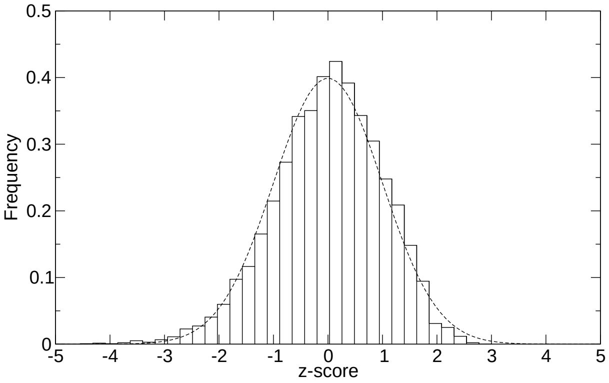


Fig. 16. Frequency distribution of random z-scores. The solid bars show the distribution of z-scores of 5930 random alignments (mean 0.003 and standard deviation 0.989). The dashed line shows a standard normal distribution.

of alignments, so that typically alignments based on sequence similarity alone will be used. To test to which extent the performance of our method depends on the alignment method, we did the following experiment: We took 73 eukaryotic SRP-RNAs and generated 2083 pairwise alignments with a wide variety of pairwise identities. For this test set, manually curated structural alignments exist [189]. We calculated z-scores for structurally aligned pairs and for `ClustalW` aligned pairs (Fig. 17). The detection performance for the structural alignments constantly increases with increasing sequence divergence over the full range of pairwise identities. This is exactly what could have been expected, since higher sequence divergence means more information-rich covariances. From appr. 60% to 100% pairwise identity, the z-scores of the sequence based alignments are essentially the same. Below 60%, the detection performance drops remarkably. Extrapolating from this example, we can conclude that there is obviously no need for structural alignments above 65% pairwise identity and that our method scores best somewhere between 60% and 70%.

Although the sensitivity will vary at different degrees of conservation, the practicability of our method is not limited to a specific interval of pairwise identities. Since the z-score combines both energy contribution and the covariance contribution, we can detect stable structures even at 100% conservation. On the other hand, structures which are not exceptionally stable can be detected on the basis of covariance information if there is enough variation in the sequences. It must be pointed out that the specificity is constant in any case and for all pairwise identities. In contrast, `QRNA` for example shows good performance around 85% pairwise identity, but above this value the false positive rate increases dramatically.

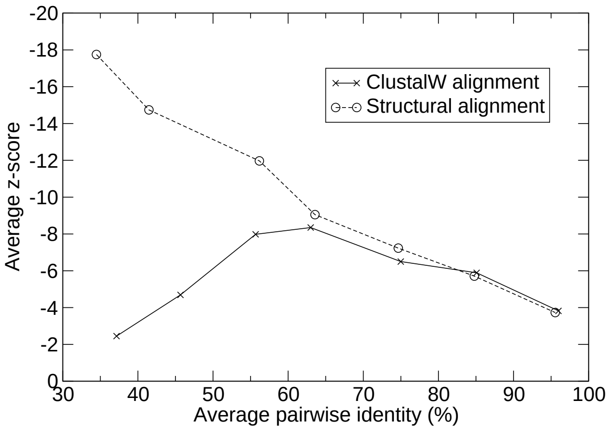


Fig. 17. Average z-scores of structural and sequence-based pairwise alignments of SRP RNAs versus pairwise identity. 2083 alignments were scored and average z-scores where calculated for seven intervals of pairwise identities between 30% an 100%. The average z-scores are plotted against the average pairwise identities calculated for each interval.

3.3.6 Tests on ncRNAs from Caenorhabditis elegans

The results so far show that detection sensitivity highly depends on the quality of the available data. A large number of homologous sequences with high divergence (but still alignable) is desirable. However, in real-life applications, such ideal data sets will not always be found. To test our method on more realistic data we created pairwise alignments of known ncRNAs from *C. elegans* [23] and *C. briggsae* [206].

We tried to take one example of each ncRNA family (excluding tRNAs and rRNAs) reported in reference [206]. If available, sequences were simply taken from the respective `Rfam` family. *C. elegans* RNA genes which could not be found in `Rfam` were taken from `Wormbase` release 117 (`www.wormbase.org`) and the corresponding *C. briggsae* homologs were searched using `Blast`. We could not find annotated sequences of RNase P and U3 snoRNA although they have been reported to exist [206].

We calculated the z-scores for this test set (Tab. 4). For scanning whole genomes it will not be feasible to predict structures longer than appr. 200 nucleotides. We therefore scored alignments longer than 150 columns using a sliding window (size 150, slide 20) and report the lowest z-score obtained. To estimate the contribution of secondary structure stability alone, we also scored single sequences from *C. elegans* alone.

We found that the ncRNA sequences are highly conserved between *C. elegans* and *C. briggsae*. Pairwise identities are above 90% in most cases. Still, most genes score well below -4. Some of them (e.g. SRP RNA or let-7 pre-miRNA) form exceptionally stable structures that can also be detected by single sequence predictions without problems. However, the alignment scores are more significant in all cases with values below the single scores in the order of appr. one standard deviation. Only the spliceosome RNAs U4 and U6 cannot be detected. This shows the inherent limitation of this method. U6 for example is known to form extensive *inter*molecular interactions with U4 rather than forming a stable *intra*molecular secondary structure. U6 only features a short 5'-stem loop. Although predicted by `RNAalifold` in the native alignment, this loop is too short to be significantly different from the random background.

Tab. 4. z-scores of ncRNAs in *C. elegans* aligned to homologs of *C. briggsae*

				z-score	
ncRNA Type	No. of Seqs.	Identity (%)	Length	Single	Alignment
SRP RNA	2	83.8	296	−5.5	−7.9
U1 spliceosome RNA	2	91.5	165	−4.6	−5.0
U2 spliceosome RNA	2	94.5	193	−5.0	−5.9
U4 spliceosome RNA	2	99.3	139	+0.7	+0.2
U5 spliceosome RNA	2	92.7	123	−2.3	−5.0
U6 spliceosome RNA	2	98.0	102	−0.8	−0.4
let-7 pre-miRNA	2	89.0	73	−7.5	−8.4
lin-4 pre-miRNA	2	90.0	70	−4.1	−4.8
SL2 RNA	2	91.3	103	−2.5	−3.6

3.3.7 Tests on ncRNAs from Saccharomyces cerevisiae

Pairwise alignments can easily be obtained by `Blast`. However, if more than two genomes are available, multiple sequence alignments have to be generated. The generation of high quality multiple sequence alignments on a genome wide scale is a difficult task and still subject of heavy research. We evaluated the performance of our method on automatically generated alignments on the genome of *S. cerevisiae* to draft sequences of six related yeast species [35, 112]: *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *S. kluyveri*.

We chose `MultiPipMaker` [200] to generate the alignments. At the time these experiments where conducted this was the only program available capable of aligning a reference sequence to unassembled contigs on a genome wide level off-the-shelf.

To estimate the sensitivity for screening the yeast genome we used the following procedure: We generated multiple sequence alignments of all 16 chromosomes. We then extracted the regions of annotated ncRNAs according to the annotation table from the Saccharomyces Genome Database (`www.yeastgenome.org`, July 2003). Since `MultiPipMaker` could not find homologous sequences in all species for all ncRNAs and sometimes only dubious fragments could be found, the alignments were automatically refined before scoring. The pairwise identity of the reference sequence from *S. cerevisiae* to all other sequences in the alignment was calculated. If it was below 60% the sequence was dropped. We included the gap character in the calculation of the similarity and thus excluded sequences that did not have a match in the region (i.e. only gaps) and also sequences that had only some short fragments aligned by `MultiPipmaker`. After this selection, we removed the gaps in the remaining sequences and re-aligned them using `ClustalW`. Finally, we calculated the z-scores for the re-aligned window. Since there is no sense to calculate a z-score if there is no stable secondary structure even in the native alignment, we only considered alignments which had a `RNAalifold` MFE below -15.

As before, we scanned the rough alignments in windows of size 150 and slide 20. If we encountered a window (after the refinement steps) having a z-score below -4 we regarded the ncRNA as detected.

Tab. 5 summarizes the results for the different ncRNA classes. Tab. 6 shows detailed predictions for selected ncRNAs. The alignment characteristics and z-scores are shown for the best scoring 150 column window in each of the genes.

Only a small fraction of the tRNAs can be found. This is due to the high conservation ($> 95\%$) of this class of RNAs. As expected, also the ribosomal RNAs are highly conserved between the closely related yeast species. Still, the large 18S and 25S subunits, which have obviously extremely stable local secondary structures, can be detected even at 100% conservation. As seen for *C. elegans*, RNA genes lacking a stable secondary structure are missed. This is true for some small nuclear RNAs and all C/D-type small nucleolar RNAs. The H/ACA type snoRNAs on the other hand have a typical two stem loop secondary structure and therefore 14 of 20 can be detected. The six ones that are missed score around -3. Also here, the high conservation (around 90%) probably hinders a more efficient detection.

All other known RNAs (SRP-, RNaseP-, RNase MRP-, and Telomerase-RNA) can be detected. Also the RNAs of unknown function (RUF) that have been identified by a `QRNA` screen [159] can be found with our method. Initially, 8 RUFs have been detected. However, in additional experiments the expression of RUF4, RUF6, RUF7 could not be verified and RUF8 has been found to be a coding mRNA (a correction has been issued for the original paper[2]). This is consistent with our predictions: Only RUF1, RUF2, RUF3 and the two copies of RUF5 have z-scores below -4. We do not find significantly conserved secondary structures in RUF4, RUF6, RUF7 and RUF8.

To conclude, our method has good sensitivity in this test screen. Most of the structured RNAs which show some variation in sequence (i.e. which are not too conserved) could be detected.

To assess the false-positive rate for the cutoff of -4 in this experiment, we repeated the

---

[2] `ftp://ftp.genetics.wustl.edu/pub/eddy/papers/2003-mccutcheon-yeast/correction_long.pdf`

screen in exactly the same way, but shuffled the windows before calculating the z-score. In the 313 genes, we scored 807 randomized windows and encountered two windows scoring below -4. This is a false positive rate of 0.25% per alignment.

To estimate the false positive rate not only on known ncRNAs but also on coding genes and other conserved regions, we randomized the complete chromosome 5 by shuffling the alignment in non-overlapping windows of length 150. We then scanned the random chromosome also in non-overlapping windows of length 150 in the same way as before. We scanned both the forward direction and the reverse complement. We finally found 2217 conserved blocks which have a `RNAalifold` MFE below -15 after the re-alignment step. Out of these, five had a z-score below -4, which is a false positive rate of 0.23% per alignment. This is approximately the same as we found for the randomized ncRNAs. The chromosome 5 is 574,860 base pairs long. So we can expect around 8 to 10 false positives per megabase of the yeast genome in such a screen. However, this number of *statistical* false positives will also depend on how many overlapping windows we score.

This number also does not include *biological* false positives as for example inverted repeats which could be interpreted as stable hairpins. Also pseudogenes could be a problem here. However, we expect our method to be quite robust to distinguish real ncRNAs from pseudogenes. Unlike other methods which search for sequence patterns, our method only relies on the conservation of a secondary structure. It is known that only a small number of random mutations destroy secondary structures [199] and it is thus unlikely that pseudogenes retain a conserved structure without evolutionary pressure.

Tab. 5. Sensitivity on known ncRNAs in *S. cerevisiae*

ncRNA Type	Annotated genes	Detected genes ($z < -4$)	Sensitivity
tRNA	275	28	10.2%
rRNA	11	6	55.5%
snRNA	6	4	66.7%
C/D snoRNA	46	5	10.9%
H/ACA snoRNA	20	14	70.0%
other ncRNAs of known function	4	4	100.0%
ncRNAs of unknown function (RUF)	5	5	100.0%

3.3.8 Perl implementation: `Alifoldz` and `Shuffle-aln`

The algorithms described in the previous sections were implemented in the `Perl 5` script `Alifoldz`. The script takes an alignment in `ClustalW` or `Fasta` format and calculates z-scores for the whole alignment or in sliding windows. Fig. 18 shows a sample output of `Alifoldz`.

Also the shuffling algorithm was implemented in a `Perl` script. The script `Shuffle-aln` is

Tab. 6. z-scores of selected ncRNAs in *S. cerevisiae*

				z-score	
ncRNA Type	Gene Name	No. of Seqs.	Identity (%)	Single	Alignment
Signal recognition particle RNA	SCR1	4	85.6	−2.2	−4.2
RNAase P RNA	RPR1	4	85.0	−3.7	−6.5
RNAse MRP RNA	NME1	6	69.1	−4.8	−11.1
Telomerase RNA	TLC1	3	71.1	−4.5	−7.4
U1 spliceosomal RNA	snR19	6	74.3	−3.4	−8.5
U2 spliceosomal RNA	LSR1	3	74.9	−6.3	−6.5
U4 spliceosomal RNA	snR14	6	87.0	−1.8	−3.0
U5 spliceosomal RNA	snR7-L	5	80.6	−3.6	−5.5
	snR7-S	5	79.4	−3.4	−4.4
U6 spliceosomal RNA	snR6	6	90.9	−1.9	−2.3
RNAs of unknown function	RUF1	4	75.8	−4.4	−7.3
	RUF2	4	80.9	−4.0	−8.9
	RUF3	4	77.3	−3.9	−6.7
	RUF5-1	4	66.8	−3.0	−4.5
	RUF5-1	4	66.7	−2.4	−4.4

used throughout this thesis to generate randomized alignments.

3.4 An efficient algorithm without shuffling

Despite the promising results, the limitations of the shuffling approach are obvious. To calculate one z-score, we have to fold the sequences 100 times to estimate the random background. This of course imposes a serious performance problem. Moreover, since the algorithm depends on a random variable, the results vary from run to run.

Ideally, a program that is used in every-day sequence analysis, should calculate a deterministic score within reasonable time. In the next sections we describe the development of a more efficient approach that meets these criteria. We first demonstrate the basic concepts using simplified models and mathematical methods, which are then optimized using support vector machine approaches that, eventually, lead to the final implementation of our structural RNA finding program `RNAz`.

3.4.1 The structure conservation index

It is clear by now that the consensus MFE calculated by `RNAalifold` is not only a useless by-product in the process of predicting a consensus secondary structure, but provides valuable information on whether there is a conserved fold in a sequence alignment or not. However, it is difficult to interpret such a consensus MFE in absolute terms, since it depends on the alignment length, the base composition, the degree of conservation and the gap-pattern. To devise a reasonable *ad hoc* measure which considers all these factors seems impossible. The only remedy so far was the time-consuming shuffling approach.

```
###################################################################
# alifoldz.pl
#
#           Input: 4 sequences of 342 columns
#   Sample Number: 100
#          Window: 120
#           Slide: 40
#          Strand: forward and reverse
#   MFE threshold: -3
#    Re-alignment: OFF
#  Random control: OFF
#    Program call: RNAalifold
#
###################################################################

  From      To    Strand    Native MFE    Mean MFE     STDV      Z
 ------------------------------------------------------------------
     1     120       +        -24.68       -16.14       3.05    -2.8
     1     120       -        -16.25       -10.50       3.02    -1.9
    41     160       +        -23.40        -9.64       2.65    -5.2
    41     160       -        -15.02        -6.17       2.50    -3.5
    81     200       +        -18.42        -4.77       2.26    -6.0
    81     200       -         -9.60        -2.11       1.90    -3.9
   121     240       +        -10.81        -3.22       2.29    -3.3
   121     240       -         -1.11
   161     280       +         -7.94        -2.23       1.67    -3.4
   161     280       -         -0.57
   201     320       +         -6.18        -4.47       2.25    -0.8
   201     320       -         -3.75        -4.42       2.03     0.3
   241     342       +         -3.64        -3.41       1.81    -0.1
   241     342       -         -4.07        -5.93       2.21     0.8

-6.0
```

Fig. 18. Output of `Alifoldz` on an alignment of four RNAseP sequences. Both strands are scanned in overlapping windows and for each window a z-score is calculated.

A much more efficient normalization can be achieved, however, by comparing the consensus MFE with the MFEs of each individual sequence in the alignment. To this end, we fold the alignment and calculate the consensus MFE E_A of the alignment using `RNAalifold`. If the sequences in the alignment fold into a conserved common structure, the average $\bar{E}$ of the individual MFEs will be close to the MFE of the alignment, $E_A \approx \bar{E}$. Otherwise, the MFE of the alignment will be much higher (indicating a less stable structure) than the average of the individual sequences, $E_A \gg \bar{E}$. We therefore define the *structure conservation index* (SCI) as

$$\mathrm{SCI} = E_A/\bar{E}$$

A SCI close to zero indicates that `RNAalifold` does not find a consensus structure while a set of perfectly conserved structures has SCI $\approx$ 1. A SCI larger than 1 indicates a perfectly conserved secondary structure which is in addition supported by compensatory and/or consistent mutations, which contribute a covariance score to E_A.

We tested the SCI on a simple test set of alignments with three sequences each and mean pairwise identities between 60%–90%. The test set was created as described in section 3.3.2 and consisted of 1344 alignments from six ncRNA families of different length and base composition: tRNA (741), 5S rRNA (100), Group II catalytic intron (148), U5 spliceosomal RNA (134), SRP-RNA (68) and Hammerhead III ribozyme (143). For each native RNA alignment one random alignment was generated as control. The SCI was calculated for the native alignments and the random alignments. Fig. 19 shows the frequency distribution of the SCIs. The native alignments are clearly separated from the random controls. Fig. 19 resembles the z-score histograms shown before. It must be emphasized that here we do not calculate z-scores depending on a sampled random background, but simple calculate one measure which normalizes for length and base composition.

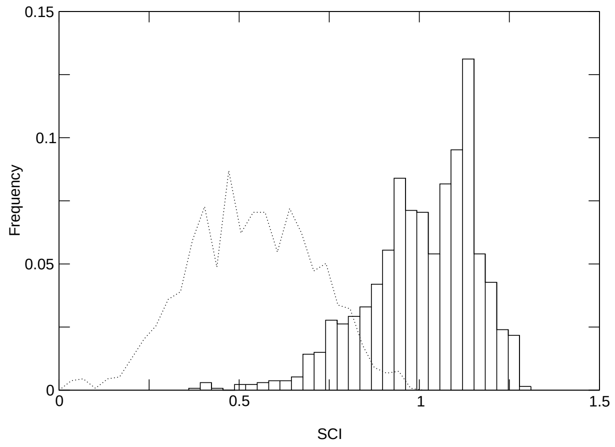


Fig. 19. Distribution of the SCIs for a test set of 1344 alignments. Solid bar: native RNAs, dashed line: random control

3.4.2 Empirical estimation of z-scores using linear regression

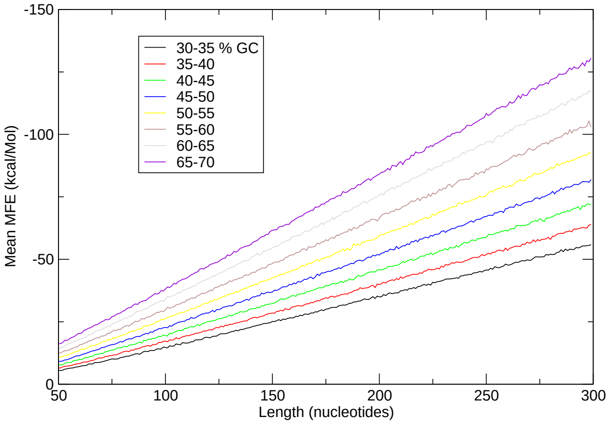
The SCI is obviously a suitable measure to assess a multiple sequence alignment for a conserved fold. However, it does not give information on whether this fold is exceptionally stable or not. In section 3.1 we demonstrated that natural ncRNAs are indeed more stable than one could expect by chance. Although we had to conclude that this is not statistically significant enough for reliable detection of ncRNAs, it is valuable additional information that should be included in a search algorithm. The impressive results we obtained by calculating z-scores of `RNAalifold` MFEs can only be explained by the fact the both structure conservation and thermodynamical stability are combined in one score.

Since the aspect of structural conservation is now covered by the SCI, there is need for an additional score measuring the thermodynamic stability. However, the time-consuming calculation of z-scores by random sampling is currently the only available method for this purpose. This approach was introduced 16 years ago [128] and it is still widely used today [17, 10, 9]. Here we ask whether it is possible to empirically estimate z-scores.

A z-score is calculated from the mean μ and the standard deviation σ of MFEs of a large sample of random sequences of the same base composition and length. To obtain a z-score without sampling we need to know the mean and the standard deviations which are, by construction, functions of length and base composition.

To model these functions we generated 800,000 synthetic random sequences of different length (50–300 nucleotides) and GC-content (30%-70%). In a single-stranded RNA sequence not only GC-content matters but also how the number of As/Us and Gs/Cs are distributed. For the sake of simplicity, we here focused on the GC content, and sampled the other distributions uniformly between 40% and 60%. MFEs were calculated for all sequences using `RNAfold`. Fig. 20 shows the mean and standard deviation of the MFEs in dependence on the length for different GC-intervals (appr. 300 samples for each length where evaluated). For a given GC-interval, mean and standard deviation are linearly dependent on the length. Therefore, for a given GC-interval the behavior can easily and accurately be approximated by linear regression. The behavior of mean and standard deviation with respect to GC-content is not linear (Fig. 21) but also shows a predictable behavior. For the time being, we do not perform multidimensional regression but rather calculate different linear regressions for eight GC-content intervals. Using this simplified approach, a z-score for a given sequence can be estimated by first determining the GC-content and the corresponding regression parameters, and then calculating mean and standard deviation from the length.

To compare this empirical estimation procedure with the traditional sampling method, we estimated and sampled (N=1000) z-scores of 200 ncRNAs of different length and families (Fig. 22). We find that this simple procedure based on linear regression can approximate z-scores with fair accuracy.



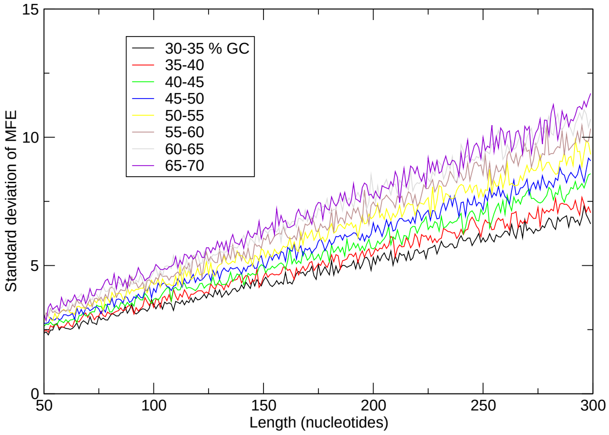


Fig. 20. Mean and standard deviations of MFEs for random sequences of different length and intervals of GC content.

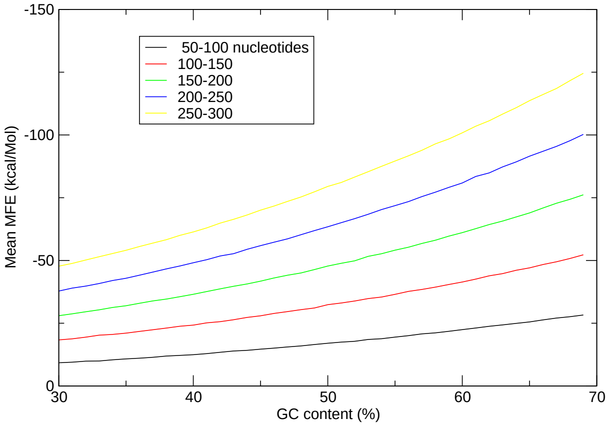


Fig. 21. Mean MFE for random sequences in dependence on the GC-content for various length intervals.

In practice it is of interest if the estimated z-scores, which can now be seen as a deterministic measure for the stability of a sequence, can distinguish real RNAs from random controls. Using the same test set as before, we calculated the average z-score of the sequences in each of the alignments. Fig. 23 shows the distribution of the averaged z-scores in the 1334 alignments. Also here, the real RNAs are clearly separated from the random control.

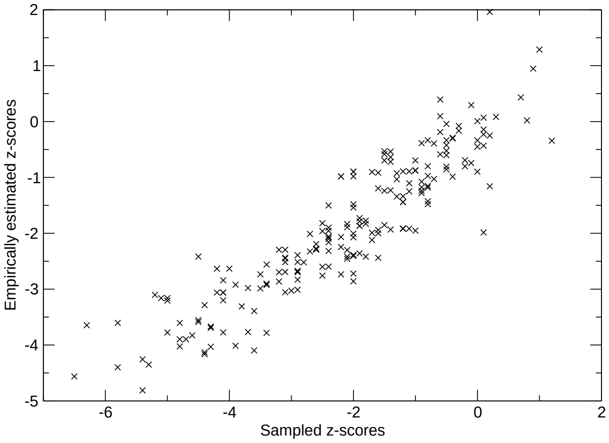


Fig. 22. Empirically estimated z-scores vs. traditionally sampled z-scores ($N = 1000$) of 200 ncRNAs.

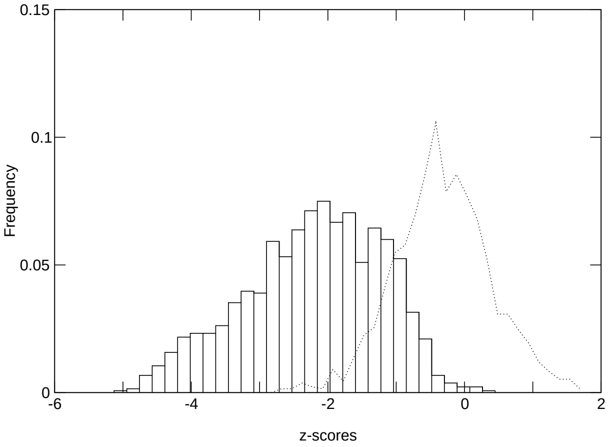


Fig. 23. Distribution of empirically estimated z-scores for a test set of 1344 alignments. Solid bar: native RNAs, dashed line: random control

3.4.3 Combining SCI and estimated z-scores

We now have two measures characteristic for functional RNA secondary structures: (i) evolutionary conservation and (ii) thermodynamic stability. A scatter plot of the two scores on our test sets indicates that the scores are independent of each other (Fig. 24). The point clouds are separated in two dimensions, which means that together both scores can much more efficiently distinguish the functional RNAs from the random controls.

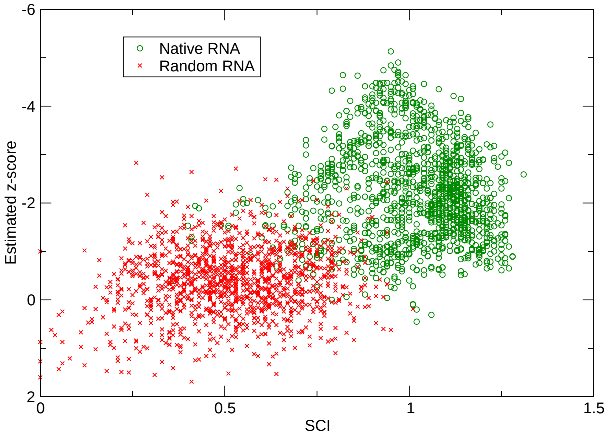


Fig. 24. Two dimensional separation of native RNAs from random controls based on SCI and estimated z scores.

In practice, it will be necessary to combine both scores into a composite score which can be used as a decision criterion in course of classification. Finding an optimal combination of both scores corresponds to the problem of finding the separatrix between functional RNAs and the random controls in the SCI/z-score plane. In an naive approach, one could consider a linear classification function. The slope of the linear function has an intuitive interpretation; it corresponds to the weighting of conservation vs. stability in the classification.

To demonstrate that even simple linear rules can result in effective classification, we used a linear score which roughly weights both components equally: combined score $= (-z) + 4 \times \mathrm{SCI}$. The distribution of this score for our test-set is shown in Fig. 25. For a cutoff of 5 one can observe a sensitivity of 84.0% at at specifity of 99.2%. Cutoff 6 has still 62.1% sensitivity at a specificity of 99.9%. Although less accurate than the sampling method (93.5% sensitivity at 99.6% specificity for this test set), the results show that it is worthwhile to elaborate this approach in a more systematic manner.

3.4.4 Accurate estimation of z-scores using SVM regression

The mean μ and standard deviation σ which are needed to calculate the z-score, are functions of the length and base composition. To accurately estimate these functions, we have to solve a five dimensional regression problem:

$$\mu, \sigma(N, \frac{N}{n_\mathsf{G} + n_\mathsf{C}}, \frac{n_\mathsf{A}}{n_\mathsf{A} + n_\mathsf{T}}, \frac{n_\mathsf{C}}{n_\mathsf{G} + n_\mathsf{C}})$$

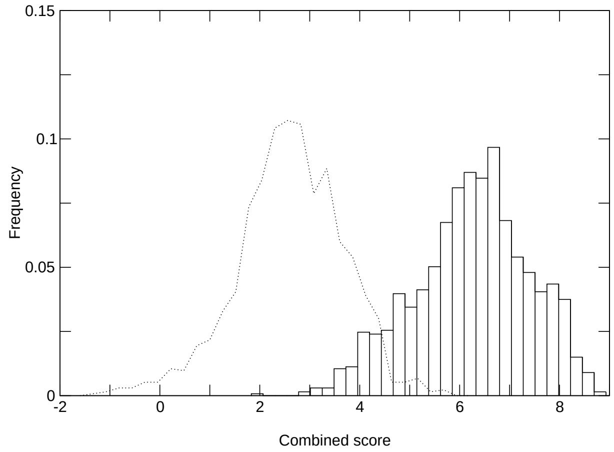


Fig. 25. Distribution of the combined scores (see text) for the test set of 1344 alignments. Solid bar: native RNAs, dashed line random control.

where N is the length of the sequence and n_{X} the number of Xs in the sequence. Here we do not only consider GC content but the frequencies of all nucleotides.

We generated a large test set of synthetic sequences of different length and base composition. The length ranged from 50 to 400 nucleotides in steps of 50. The base composition ratios ranged from 0.25 to 0.75 in steps of 0.05. This resulted in 10,648 points in the four-dimensional space of the independent variables. For each of these points we calculated the mean and standard deviation of the MFE of 1,000 random sequences, representing the dependent variables in our regression.

We trained a SVM regression model on this test set. We used the `Libsvm` library and chose the ν variant of regression (section 2.5.4) and a radial basis function kernel (section 2.5.3). Empirical testing of different parameter combinations in self-consistency tests found $\nu = 0.5$, $C = 5$ and $\gamma = 1$ to yield the best results for both the mean and standard deviation model.

The accuracy of our SVM model was verified by comparing z-scores from the SVM approach with z-scores obtained by standard sampling (Fig. 26). As test sequences we chose 100 sequences from random locations in the human genome and 100 known ncRNAs from the `Rfam` database. We found that the correlation between sampled values and SVM values was as good as two independently sampled z-scores at a sample size of 1,000. This is a marked improvement compared to the simple approach shown before (Fig. 22). Moreover, this result clearly shows that we can replace the time-consuming sampling procedure by the SVM estimate without any loss of accuracy, while saving about a factor of 1,000 in CPU time.

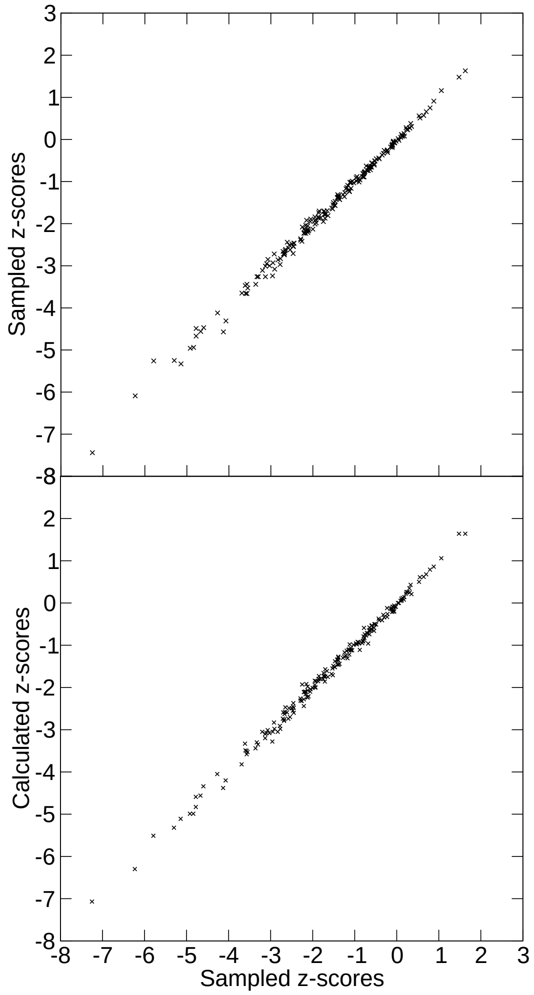


Fig. 26. z-scores calculated by SVM regression in comparison to z-scores determined from 1000 random samples for each data point. Upper panel: Correlation of z-scores from two independent samplings (mean squared error: 0.00990). Lower panel: Correlation of calculated z-scores and sampled z-scores (mean squared error: 0.00998)

3.4.5 SVM classification based on SCI and z-scores

Also the classification problem can be much more efficiently solved using a SVM algorithm. SVMs have been developed exactly for this kind of problems.

We therefore trained a binary classification SVM on test sets from 12 `Rfam` families encompassing all major classes of ncRNAs. Test alignments with mean pairwise identity between appr. 50% and 100% and 2–6 sequences per alignment were generated as described in section 3.3.2. For each of the native alignments we created one shuffled random control. Our final test set consisted of 4101 positive and 4101 negative examples.

Fig. 27 shows the z-score/SCI scatter plots for the 12 ncRNA classes. For the SVM classification, we did not only use z-score and SCI as input parameters but also included mean pairwise identity and the number of sequences in the alignment. This refinement is necessary because the information content of a multiple alignment strongly depends on these parameters: in the extreme case, an alignment of identical sequences has SCI $= 1$ but does not contain any information about structural conservation at all. In this case, the classification algorithm should not trust the high SCI but rather give more weight to the stability score. On the other hand, if there is high sequence divergence in an alignment, say six sequences with 60% mean pairwise identity, even a low SCI of about 0.5 can indicate structural conservation. Since we use a randomized control which has the same number of sequences and the same pairwise sequence conservation together with each positive example, the calibration process is not biased by these additional variables.

We scaled all parameters linearly from -1 and 1. Again using `Libsvm`, we trained a binary SVM with a radial basis kernel and the parameters $\gamma = 2$ and $C = 32$. For the final calibration of the SVM in the current implementation of `RNAz` (see section 3.4.8) we used all classes of ncRNA with the exception of tmRNAs and U70 snoRNAs.

The significance of SVM classification is generally quantified by an abstract "decision-value", the result of the decision function which corresponds to the distance of a test point to the hyperplane (cf. section 2.5.2). `Libsvm` implements a method to estimate class probabilities from decision values. We make use of this option of the package and, from now on, use the class probability p as significance measure. Fig. 28 illustrates the SVM approach we follow to accomplish non-linear classification and significance estimation.

3.4.6 Benchmarks on Rfam alignments

We tested the accuracy of the classification on all `Rfam` families in our test set. We emphasize that, although we use here a machine learning approach for classification, we do *not* train the SVM on specific sequences, sequence patterns, structure motifs, conservation patterns, or base-compositions. We use the SVM solely as a guide to interpret the SCI and z-score which represent two diagnostic features that do not contain any information that is specific for a particular class of ncRNAs. In fact, it would be interesting to replace the SVM by a direct statistical model.

In order to demonstrate that our classification procedure is generally applicable and not biased towards ncRNA classes of the training set, we used here a "one-leave-out" strategy for all benchmarking tests. We trained the SVM excluding particular classes of ncRNAs and used those models to classify the excluded ncRNAs and their randomized controls.

The results are summarized in Tab. 7 shows the sensitivity and specificity for detecting different ncRNA classes at different probability cutoffs. We used alignments with mean pairwise sequence identities between 60% and 100% and 2–4 sequences per alignment. At a cutoff of $p = 0.9$, we can detect on average 75.27% at a specificity of 98.93%.

The accuracy of the classification depends quite strongly on the type of the ncRNA. We can find most RNA classes with high sensitivities in the range of 80%–100%. Only two of the twelve classes in our test set (U70 snoRNA and tmRNA) are difficult to detect. The scatter-plots in Fig. 27 show that the U70 is quite stable but not very well conserved, whereas the tmRNA has a conserved secondary structure that is obviously not very stable and moreover contains pseudo-knots. Alignments with more sequences are needed to detect also these two RNA classes quantitatively.

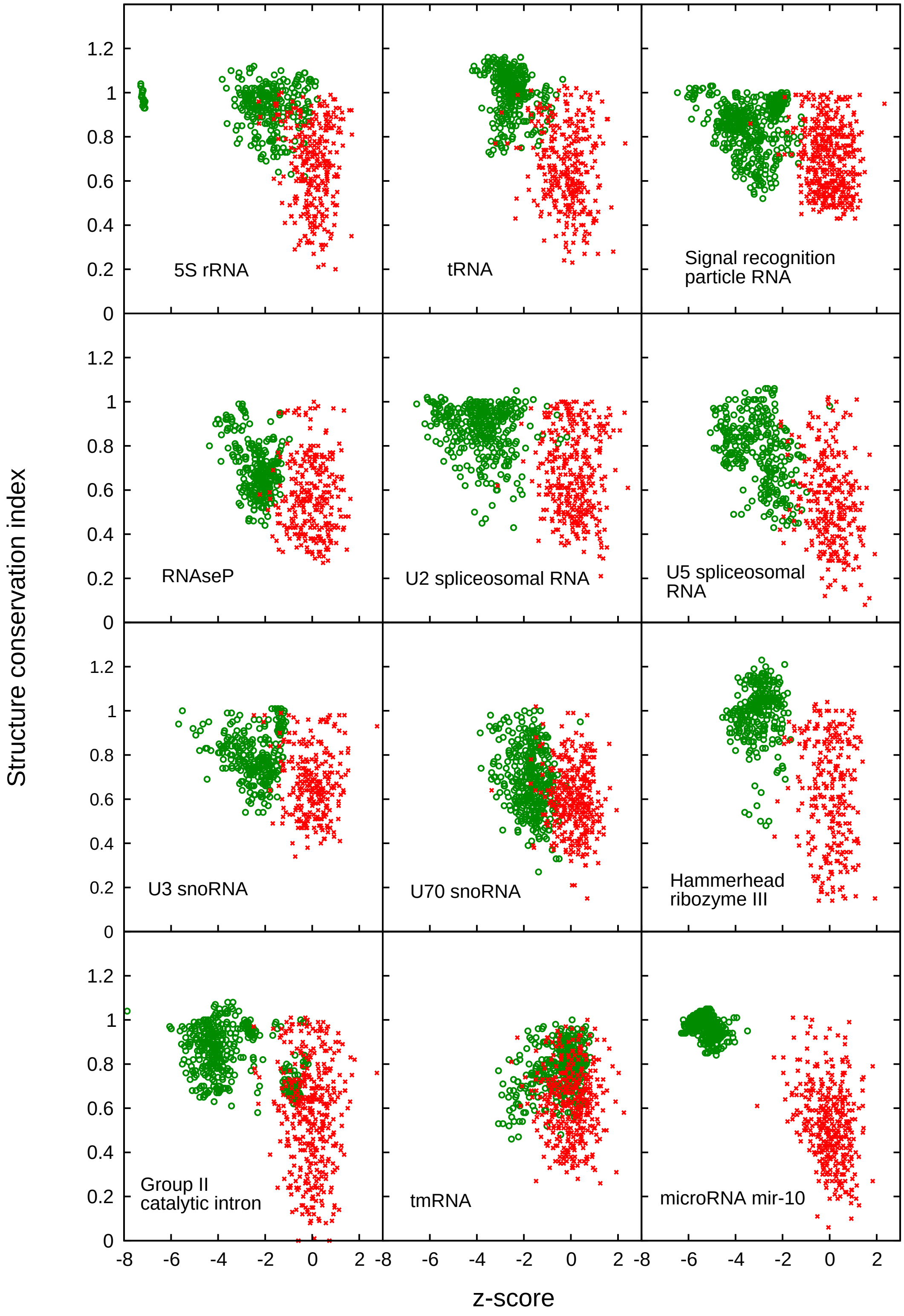


Fig. 27. Separation of native alignments (green) from random controls (red) for various classes of ncRNAs. The test sets are the same as used in Tab. 7 with mean pairwise identities between 60% and 100% and 2–4 sequences per alignment.

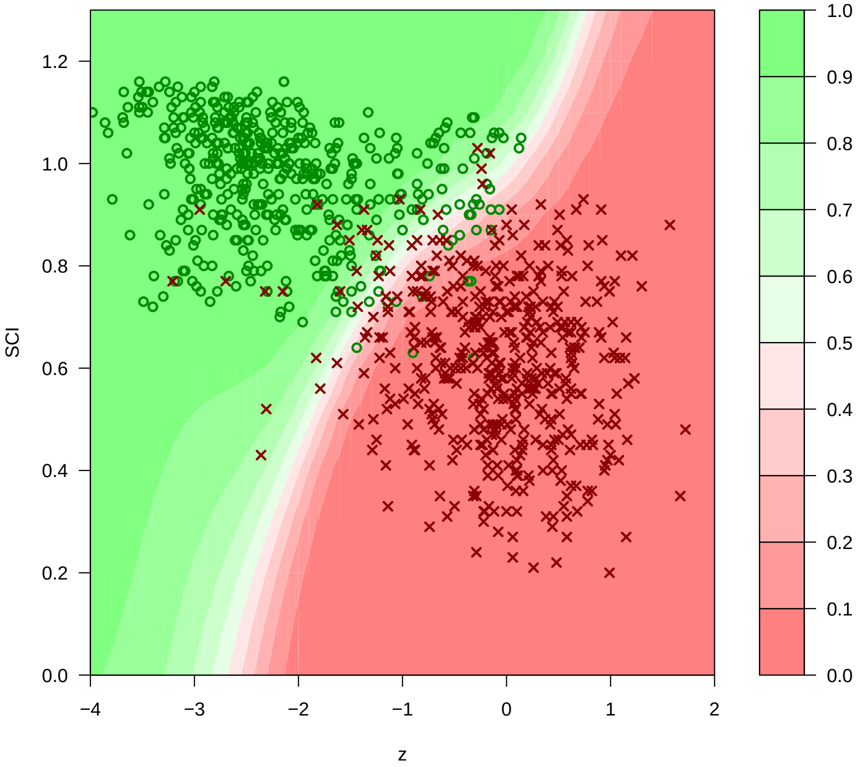


Fig. 28. Classification based on z-scores and SCI using a support vector machine. Alignments of tRNAs and 5S-rRNAs with 2–4 sequences per alignment and mean pairwise identities between 60% and 90% are shown. Green circles represent native alignments, red crosses represent shuffled random controls. The background color ranging from red to green indicates the RNA-class probability for different regions of the z–SCI-plane.

Tab. 7. Detection performance for different classes of ncRNAs

		Cutoff					
		0.5		0.9		0.99	
ncRNA Type	N	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
5S ribosomal RNA	297	81.48 (242)	96.63 (10)	68.69 (204)	99.33 (2)	33.00 (98)	100.00 (0)
tRNA	329	94.83 (312)	93.62 (21)	90.27 (297)	97.87 (7)	75.68 (249)	99.70 (1)
SRP RNA	464	100.00 (464)	96.55 (16)	96.55 (448)	98.92 (5)	66.16 (307)	100.00 (0)
RNAse P	291	98.97 (288)	96.22 (11)	84.19 (245)	99.31 (2)	56.70 (165)	100.00 (0)
U2 spliceosomal RNA	351	98.58 (346)	97.72 (8)	95.44 (335)	99.15 (3)	66.67 (234)	99.72 (1)
U5 spliceosomal RNA	285	91.58 (261)	98.25 (5)	81.75 (233)	100.00 (0)	70.53 (201)	100.00 (0)
U3 snoRNA	277	83.75 (232)	98.56 (4)	62.82 (174)	99.28 (2)	44.40 (123)	99.64 (1)
U70 snoRNA	363	61.16 (222)	96.69 (12)	35.54 (129)	98.90 (4)	17.91 (65)	99.72 (1)
Hammerhead III ribozyme	271	100.00 (271)	95.20 (13)	98.15 (266)	98.89 (3)	89.67 (243)	99.26 (2)
Group II catalytic intron	407	78.62 (320)	96.31 (15)	76.90 (313)	98.53 (6)	25.31 (103)	100.00 (0)
tmRNA	386	24.87 (96)	96.37 (14)	18.65 (72)	98.19 (7)	8.55 (33)	99.48 (2)
micro RNA mir-10	380	100.00 (380)	95.26 (18)	97.63 (371)	99.21 (3)	62.37 (237)	100.00 (0)
Total	4101	84.17 (3452)	96.42 (147)	75.27 (3087)	98.93 (44)	50.18 (2058)	99.80 (8)

Results for alignments with 2–4 sequences and mean pairwise identities between 60% and 100% are shown. N is the number of alignments in the test set. For each native alignment, one randomized alignment was produced, and randomized alignments classified as ncRNA were counted as false positives. Sensitivity and specificity are shown in percent for three cutoffs of the RNA class probability predicted by the SVM. Absolute numbers of true positives and false negatives are shown in brackets.

3.4.7 Comparison to other methods

RNAseP and SRP RNAs have repeatedly been used for benchmarking ncRNA detection algorithms [187, 40]. We therefore use these datasets here as well. For the comparison to `QRNA` and `ddbRNA` we used pairwise and three-way alignments with mean pairwise identities between 60% and 90%, respectively. In contrast to the previous section we exclude alignments with identities higher than 90% since both `QRNA` and `ddbRNA` are known to perform poorly on such input data. We used a cut-off of $p = 0.9$ for our method and chose the cut-offs for the other programs in a way that the specificity is at least 90%. Results are summarized in Tab. 8. We find that our approach is substantially more sensitive on both pairwise and three-way alignments than `QRNA` and `ddbRNA` and at the same time has a larger specificity.

We also tested our method on larger alignments with 10 sequences as used for benchmarking `MSARi`. We generated 150 alignments which had mean pairwise identities between 50% and 70%. Our SVM classification model is currently trained only for up to six sequences so we did not use it for the classification of this test set. It turns out, however, that the simple rule SCI ≥ 0.3 and $z \leq -1.5$ perfectly separates the native alignments from the controls with 100% sensitivity and 100% specificity using *either* of the two scores without help of a SVM. Although the alignments produced by `ClustalW` are, at this level of sequence similarity, structurally not perfectly correct, our consensus folding algorithm still finds the correct common structure and the SCI is still significant, albeit at lower levels.

At the time this thesis was written, no executable version of `MSARi` was available so we can only cite the published results: according to [40] `MSARi` achieves at best a sensitivity of 56% at 100% specificity for `ClustalW` alignments of $N = 10$ RNAseP or SRP sequences.

3.4.8 `C` Implementation: `RNAz`

The procedure described here was implemented in the `ANSI C` programming language using the `Vienna RNA` and the `Libsvm` libraries. The current version of the program `RNAz` takes a `ClustalW` formatted sequence alignment of up to 6 sequences and 400 columns in length. It calculates the SCI from the MFEs of the single sequences and the consensus MFE. From the length and base composition, the z-scores of the sequences (without gaps) are calculated using the SVM regression model. Finally, The SCI, the mean z-score of the sequences, the mean pairwise identity and the number of sequences are used to classify an alignment as functional RNA or not. Fig. 29 shows the output of `RNAz` on an alignment 4 tRNA sequences.

The time complexity of the program is $\mathcal{O}(N \times n^3)$, where N is the number of sequences and n is the length of the alignment. Tab. 9 compares the runtime for pairwise alignments of different lengths between `RNAz` and the alternative methods: `RNAz` is not only more accurate

Tab. 8. Detection performance (Sensitivity/Specificity) for SRP- and RNAseP-alignments with mean pairwise identities between 60% and 90%

	Number of sequences in alignment		
Program	2	3	10
`QRNA`	42.9/92.9	—	—
`ddbRNA`	45.4/98.5	58.0/94.5	—
`MSARi`	—	—	appr. 56/100
`RNAz`	87.8/99.5	94.1/99.6	100/100

```
###########################  RNAz 0.1.1  #############################

 Sequences: 4
 Slice: 1 to 74
 Columns: 74
 Strand: forward
 Mean pairwise identity:  74.38
 Mean single sequence MFE: -33.10
 Consensus MFE: -30.98
 Energy contribution: -27.72
 Covariance contribution:  -3.25
 Mean z-score:  -2.68
 Structure conservation index:   0.94
 SVM decision value:   3.87
 SVM RNA-class probability: 0.999672
 Prediction: RNA

######################################################################

>AF041468.1-40566_40494
GGGGGUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGGUUCGAGUCCCCUUACCUCCA
(((((((..((((........)))).(((((.......))))).....(((((.......))))))))))))). ( -31.60)
>X54300.1-105_177
GGGGGUAUAGCUUAGUUGGUAGAGCGCUGCUUUUGCAAGGCAGAUGUCAGCGGUUCGAAUCCGCUUACCUCCA
(((((((..((((.((.(((((....)))))...)).)))).......(((((.......))))))))))))). ( -27.90)
>L00194.1-685_756
GGGGCCAUAGCUCAGUUGGUAGAGCGCCUGCUUUGCAAGCAGGUGUCGUCGGUUCGAAUCCGUCUGGCUCCA
(((((((..((((........))))(((((((.....)))))))...(.(((.......))).)))))))). ( -32.50)
>AY017179.1-1528_1601
GGGCCGGUAGCUCAGCCUGGGAGAGCGUCGGCUUUGCAAGCCGAAGGCCCCGGGUUCGAAUCCCGGCCGGUCCA
((((((((...((((((((((...((.((((((.....))))))..)))))))))).))......)))))))). ( -40.40)
>consensus
GGGGCUAUAGCUCAGU_UGGUAGAGCGCCGCCUUUGCAAGGCAGAUGUCAGCGGUUCGAAUCCCCUUACCUCCA
(((((((..((((.........)))).((((((.....)))))).....(((((.......))))))))))))). (-30.98 = -27.72 +  -3.25)
```

Fig. 29. Output of `RNAz` on an alignment of four tRNAs.

but also significantly faster than the other methods. (A comparison with `MSARi` was not possible since no implementation was available. It should have similar run times as `RNAz`, however, since it also uses the RNA folding routines of the `Vienna RNA` package as the rate limiting step.)

Tab. 9. CPU-time in seconds for 1000 alignments on an Intel 2.4 GHz Pentium 4

	Alignment length		
Program	100	200	300
`QRNA`	485	4044	14777
`ddbRNA`	741	921	1522
`RNAz`	163	375	754

4 Screening the Human Genome

4.1 Overview

As a first application of `RNAz`, we screened the human genome for conserved RNA secondary structures. With the availability of several mammalian/vertebrate genomes and a method showing reasonable accuracy, a screen of this kind has come into reach. Recently, high levels of sequence conservation of non-coding DNA regions have been reported [201, 202, 148, 48, 204]. Here we screen the complete collection of conserved non-coding DNA sequences from mammalian genomes and provide a first annotation of the complement of structurally conserved RNAs in the human genome. We limit our screen to the "most conserved" regions as annotated by the `PhastCons` program [204]. This program tries to estimate by a two-state phylogenetic hidden Markov model whether a region is under purifying selection or not. It has been estimated that about 5% of the human genome is under selective pressure [100, 39] (but may be even higher [202]). The `PhastCons` program was tuned in such a way that about 5% of the human genome was annotated as "most conserved". Since we are interested in non-coding RNAs, we removed all annotated coding exons from this set and retained only regions which are conserved at least in the four eutherian mammals (human, mouse, rat, dog). These regions were screened using `RNAz`. Fig. 30 illustrates our strategy with some screenshots from the UCSC genome browser [113]. In the next section we describe the technical details of our screen.

4.2 Methods and screen design

4.2.1 Selection of most conserved regions

We started from the "Most Conserved" track generated by the PhastCons program. This track was edited as follows:

1. Adjacent conserved regions that are separated by <50 nucleotides were joined because many known ncRNAs are not conserved over the full length but only contain shorter fragments of highly conserved regions (in microRNA precursors, for example, the two sides of the stems are detected as conserved while the loop region in between is not).
2. Conserved regions (after the joining step) with a length <50 nucleotides were removed because shorter RNA secondary structures are below the detection limit of `RNAz`.
3. All regions with any overlap with annotated coding exons according to the "Known Genes" and "RefSeq Genes" annotation tracks were removed.

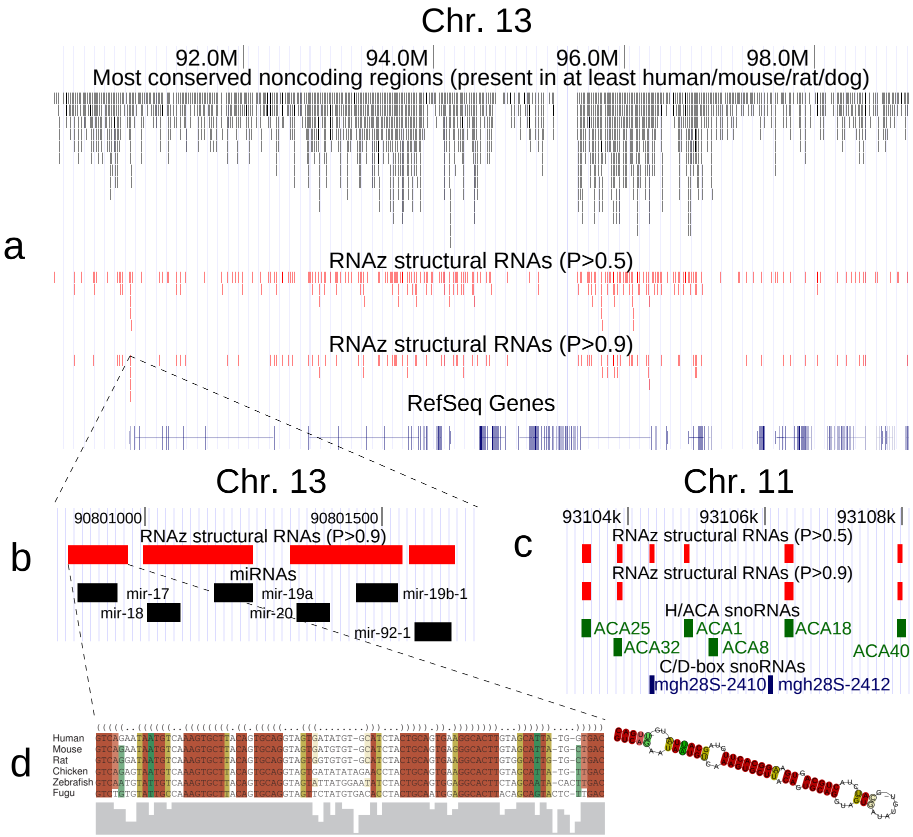

Chr. 13
92.0M
94.0M
96.0M
98.0M
Most conserved noncoding regions (present in at least human/mouse/rat/dog)
a
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
RefSeq Genes
b
Chr. 13
90801000
90801500
RNAz structural RNAs (P>0.9)
miRNAs
mir-17
mir-18
mir-19a
mir-20
mir-19b-1
mir-92-1
c
Chr. 11
93104k
93106k
93108k
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
H/ACA snoRNAs
ACA25
ACA1
ACA18
ACA32
ACA8
ACA40
C/D-box snoRNAs
mgh28S-2410
mgh28S-2412
d
((((..((((((..((((((((.((.(((((...(((........)))...))))).)).))))))))...))))))....)))))
Human GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATGT-GCATCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-GTGAC
Mouse GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATGTGT-GCATCTACTGCAGTGAGGGCACTTGTAGCATTA-TG-CTGAC
Rat GTCAGGATAATGTCAAAGTGCTTACAGTGCAGGTAGTGGTGTGT-GCATCTACTGCAGTGAAGGCACTTGTGGCATTG-TG-CTGAC
Chicken GTCAGAGTAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATATAGAACCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-TTGAC
Zebrafish GTCAATGTATTGTCAAAGTGCTTACAGTGCAGGTAGTATTATGGAATATCTACTGCAGTGGAGGCACTTCTAGCAATA-CACTTGAC
Fugu GTCTGTGTATTGCCAAAGTGCTTACAGTGCAGGTAGTTCTATGTGACACCTACTGCAATGGAGGCACTTACAGCAGTACTC-TTGAC

Fig. 30. Overview of our annotation strategy. We filter the 5% of the most conserved non-coding DNA for potentially functional RNAs using `RNAz`. (a) A 10 MB region on chromosome 13 is shown with all conserved elements in black and the predicted RNAs (for two levels of confidence) in red. The RefSeq protein-coding genes of these region are shown at the bottom. The detected RNAs contain many known ncRNAs as shown here for a cluster of miRNAs (b) and a snoRNA cluster on chromosome 11 (c). In the process of ncRNA detection also a consensus secondary structure model is generated which allows further analysis. Here we show the typical hairpin secondary structure of a microRNA.

The initial set of alignments consisted of all `Multiz` alignments corresponding to regions in the modified “Most Conserved” track. After the processing steps described below, we only considered alignments which were conserved at least in the four mammals (“input alignments”).

4.2.2 RNAz screen

The input alignments where screened for structural RNAs using `RNAz`. Alignments with $<$200 columns were used as a single block. Alignments with length $>$200 were screened in sliding windows of length 120 and slide 40. This window size, on the one hand, appears long enough to detect local secondary within long ncRNAs and, on the other hand, is small enough to detect short ncRNAs (appr. 50–70 nucleotides) without loosing the signal in a much too big window.

The individual alignment block presented to `RNAz` were further processed in the following way:

1. We discarded alignments in which the human sequence contained masked positions by `RepeatMasker`. The vast majority of repeats was already filtered out in the input alignments: either they were not aligned by `Multiz` or not detected by `PhastCons`.
2. Some alignments in the input set contained a large fraction of gaps resulting from a documented problem of `PhastCons` when treating missing data. We therefore further edited the alignments and removed sequences with more than 25% gaps. The region was regarded as not conserved in this species. If the human reference sequence contained more than 25% gaps, the complete alignment was discarded.
3. The classification model of `RNAz` is currently only trained for up to six sequences. Therefore, we removed one sequence from alignments which were conserved in all seven species. One of the two sequences in the most similar pair of sequences in the alignment was removed because this pair provides the least comparative information. For the same reason only one representative was retained if two or more sequences in the alignment were 100% identical.
4. Columns of gaps were removed from the reduced alignments.

The resulting alignments were scored with `RNAz` using standard parameters. All alignments with classification score $p > 0.5$ were stored. Finally, overlapping hits (resulting from hits in overlapping windows and/or hits in both the forward and reverse strand) were combined into clusters. The corresponding region in the human sequence was annotated as “structured RNA” with the maximum p-value of the single hits in the cluster.

4.2.3 Estimating specificity

The specificity of `RNAz` tested on shuffled alignments was found to be $\approx 99\%$ and $\approx 96\%$, for $p = 0.9$ and $p = 0.5$, respectively (Tab. 7). For benchmarking `RNAz` we used a defined set of high quality `ClustalW` alignments of 2–4 sequences and 60%–100% mean pairwise identity. In this screen, however, we used automatically generated genome-wide alignments essentially based on `Blast` hits. It was therefore not clear if the specificity is the same on these alignments and how other parameters (e.g. the sliding window) affects the false positive rate. We therefore estimated the false-positive rate for this particular special screen. To this end, we repeated the complete screen in exactly the same manner on randomized alignments. Alignments <200 columns were randomized as a whole, alignments >200 were randomized in non-overlapping windows of 200 before they were sliced in windows for scoring as described above for the true data.

Our shuffling procedure is very conservative and we found that it cannot remove the signal in all cases. The number of possible permutations is reduced if all relevant alignment characteristics are strictly preserved. Furthermore, the typical mutation pattern of non-coding RNAs is not removed by shuffling of the columns. The number of "compatible" columns which can form a base pair in the consensus structure remains the same. This might be one reason why we observe a number of random hits overlapping with native hits (Tab. 12). Another reason for this effect might be that some alignments display special properties which cause an increased false positive rate. We observe this for highly conserved alignments with little covariance information.

4.2.4 Estimating sensitivity on microRNAs and snoRNAs

We used the "sno/miRNA" track created from the microRNA Registry [69] and the snoRNA-LBME-DB maintained at the *Laboratoire de Biologie Moléculaire Eucaryote*. The track contained 207 unique microRNA loci, 86 H/ACA snoRNA, and 256 C/D snoRNAs. We compared our predictions with the annotation tracks using the "Table browser" feature of the UCSC Genome Browser. Loci overlapping with our predictions were counted as detected. Loci that did not show any overlap with our input alignments were counted as "Not in input set".

4.2.5 Non-coding RNA annotation

We compared all hits to available databases of non-coding RNAs:

- `Rfam` (release 6.1, August 2004) [70]

- `RNAdb` (August 2004) [174]
- `NONCODE` (release 1.0, March 2004) [139]
- `microRNA registry` (release 5.0, September 2004) [69]
- `UTRdb` (April 2004) [178]

We generated `Blast` libraries for each of the databases and matched the human sequence of all the detected `RNAz` clusters against them. In case of the `UTRdb` we used the EMBL formatted files from `ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data/` and extracted all annotated UTR elements >20 with flanking regions of 30 to build the `Blast` library. We considered `Blast` hits with E-values $E < 10^{-6}$ (see Tab. 15).

4.2.6 Annotation relative to protein coding genes

For annotating the `RNAz` hits relative to known protein coding genes (Fig. 35 d) , we used the "Known Genes" and "RefSeq Genes" annotation tables from UCSC genome browser. The UTR annotation is partly ambiguous. As a result, some hits in the second pie chart in Fig. 35 d are classified both as intron of a coding region and UTR. Counting only unambiguous annotations, 9825, 2095 and 1987 hits are annotated as intron of coding region, 3'-UTR and 5'-UTR, respectively.

4.2.7 Comparison with tiling array transcriptional maps

We compared our results with recently published tiling array data from [31]. We downloaded the 11 "transfrag" annotation tracks for all cell-lines and RNA fractions from `http://transcriptome.affymetrix.com`. The annotation tracks were combined into one and the coordinates were converted from the "hg15" assembly to "hg17" using the `LiftOver` tool and chain-files provided by UCSC (`http://hgdownload.cse.ucsc.edu/downloads.html`). We then compared our annotation (Set 1, $p > 0.9$) on chromosomes 6, 7, 13, 14, 19, 20, 21, 22, X and Y with the transcription map using the "Table browser" and determined the fraction of overlapping annotations. To estimate the significance, we generated randomized annotation tracks: For each predicted structural RNA we randomly chose a non-repeat region of the same length, on the same chromosome with the same annotation. We distinguished the following three annotation types: intergenic <10 kb from the nearest gene, intergenic >10 kb from the nearest gene and intronic. We did not consider regions in UTRs for this comparison. We compared five of such random tracks with the transcriptional map and found on average 29.6% overlapping annotations (the maximum overlap of all five tracks was 30.0%). To assess the detection performance on known miRNAs and snoRNAs we used the annotation tracks described above.

4.3 An integrated database system

The large amount of data that is generated and that needs to be analyzed in such a large-scale screen requires appropriate computational means for storing and processing. For this task we developed an integrated system based on a relational database (Fig. 31).

A `MySQL` database (`www.mysql.com`) stores the downloaded raw data (sequences and alignments) as well as the pre-processed data (e.g. slices of most conserved regions). `RNAz` is used to score the pre-processed alignments and the results are stored in the database. To speed up the computation we used a client/server approach that distributes the `RNAz` processes over many cluster computers which are connected to the database by TCP/IP. The complete screen took appr. 7 hours on 20 processors of different speed (Intel Pentium 3/4, 700 MHz–2.8 Ghz). The results were annotated according to the UCSC annotation tables. The UCSC tables could be directly imported since the UCSC genome browser is also based on `MySQL` and dumps of the tables are freely available for download. In addition, the results of the `Blast` queries against the RNA databases were stored in custom-made annotation tables. The results can be retrieved from the database in different ways. Either the database is directly queried by SQL commands or the functions from a `Perl` module are used. On top of the `Perl` module, a web-interface was developed making it possible to interactively browse and query the database in an intuitive manner (Figs. 32–34). This web-interface is tightly linked to the UCSC genome-browser and one can analyze the detected hits immediately in the context of the wealth of genomic annotations available through the UCSC browser.

The system can be easily adapted/extended to annotate different organisms whenever alignments and annotation data are available. However, the system was mainly conceived as a tool to facilitate the screening process and the analysis of the results. It is, in its current form, not publicly available because we do not consider it to meet the requirements of a stable and reliable database resource for the community. Issues like security, performance and future maintenance could not be adequately addressed as part of this thesis.

However, we have made publicly available a static “snapshot” of our database which can be accessed from within the UCSC browser through a “custom track” (`http://genome.ucsc.edu`) or directly from our website: `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/`. Some screenshots of the database are shown in Figs. 32–34.

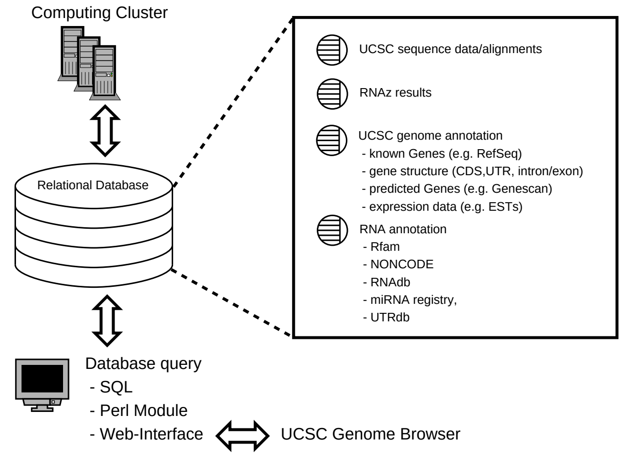


Fig. 31. Schematic overview of the analysis and annotation pipeline. For details see text.

4.4 Results

4.4.1 General statistics and specificity

We started from the `PhastCons` most conserved regions which cover 4.81% of the human genome. The filtering process reduced this initial set to 438,788 alignments covering 82,64 MB or 2.88% of the genome (Tab. 10). These alignments were screened as described. The results are summarized in Tab. 11 and Fig. 35 a,b. In the complete genome, we detected 91,676 (15.1% of the conserved sequence) independent RNA structures on the $p = 0.5$ level and 35,985 (6.6%) structures on the $p = 0.9$ level.

The specificity of `RNAz` is generally high, $\approx$ 99% for the $p = 0.9$ cutoff. Due to the large number of input alignments, however, we have to expect a non-negligible number of false positives. We therefore repeated the complete screen with shuffled alignments (Tab. 12). We obtain a false positive rate of 28.9% ($p = 0.5$) and 19.2% ($p = 0.9$), respectively. As expected, the hits in the randomized dataset are on average smaller than the native ones, reducing the false positive rates to 25.7% ($p = 0.5$) and 16.3% ($p = 0.9$) in terms of sequence length. The estimate for the false positive rate implies lower bounds of $65,000$ ($p = 0.5$) and $29,000$ ($p = 0.9$) for the number of structural RNA elements in the human genome. On average, we predict 21 ($p = 0.5$) and 10 ($p = 0.9$) structural elements per megabase.

Furthermore, we observed that many of the hits in randomized alignments overlap with native predictions. This might indicate that our shuffling process does not effectively remove the signal in all cases. We also observed that the random hits are clearly enriched in highly conserved alignments. The false positive rate of `RNAz` is higher in this case, because these alignments contain little covariance information so that the classification is dominated by

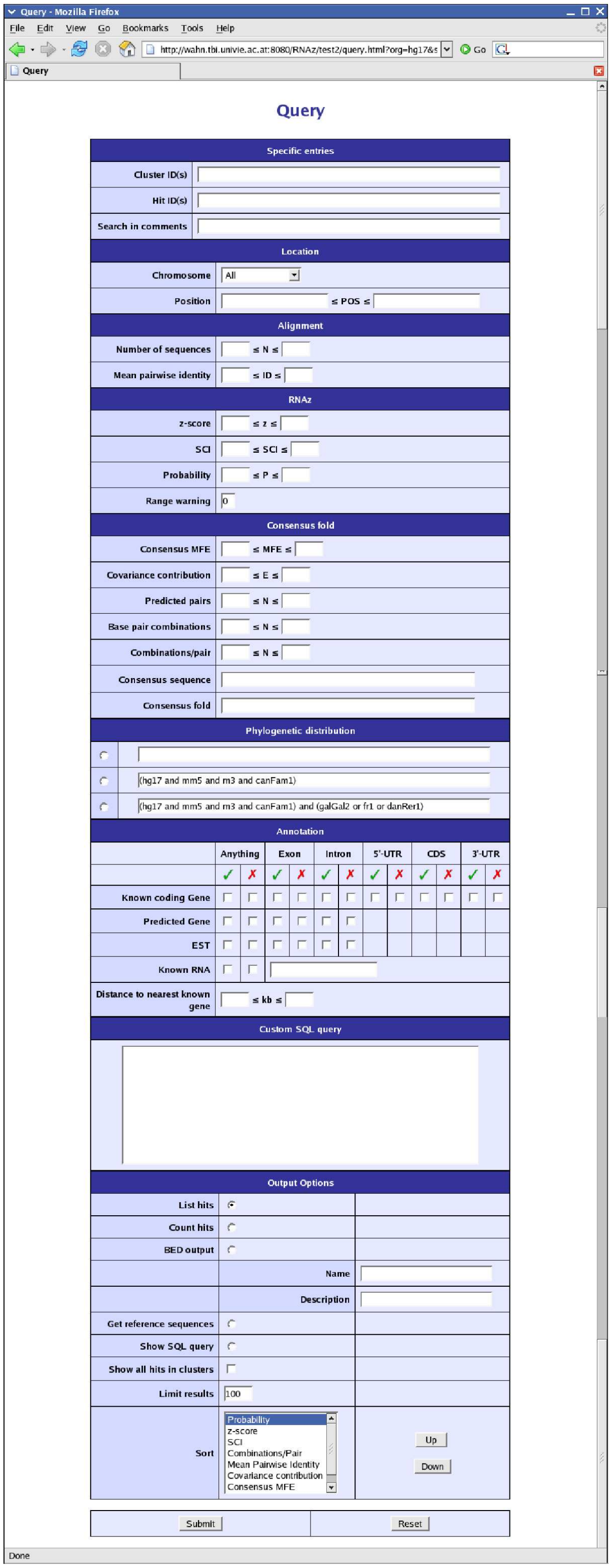
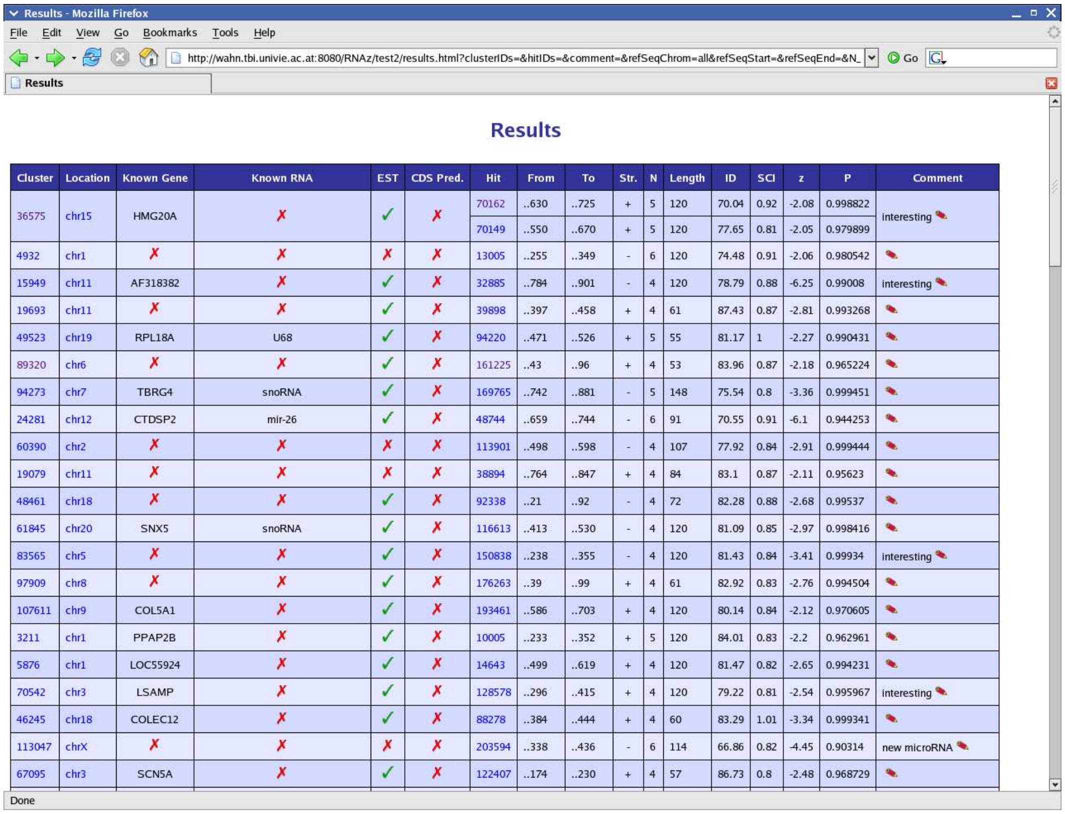


Fig. 32. Screenshot of the database web-interface. The query form is shown which allows to formulate sophisticated database queries. The hits can be filtered and sorted by significance, phylogenetic distribution and genomic annotations.

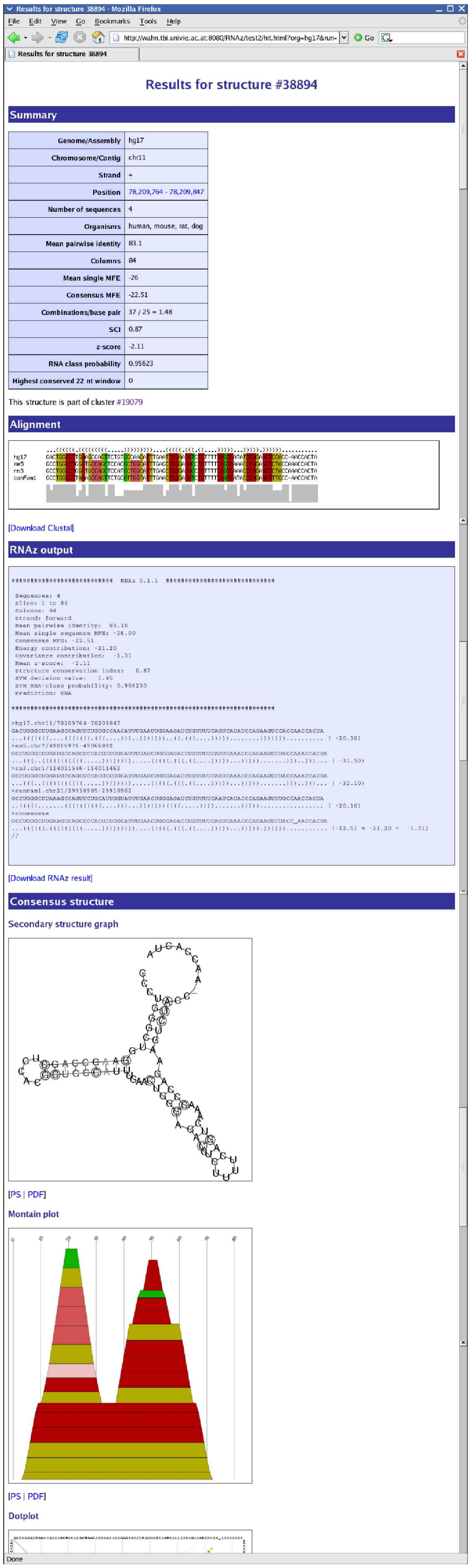


http://wahn.tbi.univie.ac.at:8080/RNAz/test2/results.html?clusterIDs=&hitIDs=&comment=&refSeqChrom=all&refSeqStart=&refSeqEnd=&N_

Results

Cluster	Location	Known Gene	Known RNA	EST	CDS Pred.	Hit	From	To	Str.	N	Length	ID	SCI	z	P	Comment
36575	chr15	HMG20A	✗	✓	✗	70162	..630	..725	+	5	120	70.04	0.92	-2.08	0.998822	interesting ✎
						70149	..550	..670	+	5	120	77.65	0.81	-2.05	0.979899	
4932	chr1	✗	✗	✗	✗	13005	..255	..349	-	6	120	74.48	0.91	-2.06	0.980542	✎
15949	chr11	AF318382	✗	✓	✗	32885	..784	..901	-	4	120	78.79	0.88	-6.25	0.99008	interesting ✎
19693	chr11	✗	✗	✓	✗	39898	..397	..458	+	4	61	87.43	0.87	-2.81	0.993268	✎
49523	chr19	RPL18A	U68	✓	✗	94220	..471	..526	+	5	55	81.17	1	-2.27	0.990431	✎
89320	chr6	✗	✗	✓	✗	161225	..43	..96	+	4	53	83.96	0.87	-2.18	0.965224	✎
94273	chr7	TBRG4	snoRNA	✓	✗	169765	..742	..881	-	5	148	75.54	0.8	-3.36	0.999451	✎
24281	chr12	CTDSP2	mir-26	✓	✗	48744	..659	..744	-	6	91	70.55	0.91	-6.1	0.944253	✎
60390	chr2	✗	✗	✗	✗	113901	..498	..598	-	4	107	77.92	0.84	-2.91	0.999444	✎
19079	chr11	✗	✗	✗	✗	38894	..764	..847	+	4	84	83.1	0.87	-2.11	0.95623	✎
48461	chr18	✗	✗	✓	✗	92338	..21	..92	-	4	72	82.28	0.88	-2.68	0.99537	✎
61845	chr20	SNX5	snoRNA	✓	✗	116613	..413	..530	-	4	120	81.09	0.85	-2.97	0.998416	✎
83565	chr5	✗	✗	✓	✗	150838	..238	..355	-	4	120	81.43	0.84	-3.41	0.99934	interesting ✎
97909	chr8	✗	✗	✓	✗	176263	..39	..99	+	4	61	82.92	0.83	-2.76	0.994504	✎
107611	chr9	COL5A1	✗	✓	✗	193461	..586	..703	+	4	120	80.14	0.84	-2.12	0.970605	✎
3211	chr1	PPAP2B	✗	✓	✗	10005	..233	..352	+	5	120	84.01	0.83	-2.2	0.962961	✎
5876	chr1	LOC55924	✗	✓	✗	14643	..499	..619	+	4	120	81.47	0.82	-2.65	0.994231	✎
70542	chr3	LSAMP	✗	✓	✗	128578	..296	..415	+	4	120	79.22	0.81	-2.54	0.995967	interesting ✎
46245	chr18	COLEC12	✗	✓	✗	88278	..384	..444	+	4	60	83.29	1.01	-3.34	0.999341	✎
113047	chrX	✗	✗	✗	✗	203594	..338	..436	-	6	114	66.86	0.82	-4.45	0.90314	new microRNA ✎
67095	chr3	SCN5A	✗	✓	✗	122407	..174	..230	+	4	57	86.73	0.8	-2.48	0.968729	✎

Fig. 33. Screenshot of the database web-interface. The results of a database query is shown. All hits that matched a query are listed together with the most important annotation information.


Results for structure 38894 - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://wahn.tbi.univie.ac.at:8080/RNAz/test2/hit.html?org=hg17&run=
Go
Results for structure 38894

Results for structure #38894

Summary

Genome/Assembly	hg17
Chromosome/Contig	chr11
Strand	+
Position	78,209,764 - 78,209,847
Number of sequences	4
Organisms	human, mouse, rat, dog
Mean pairwise identity	83.1
Columns	84
Mean single MFE	-26
Consensus MFE	-22.51
Combinations/base pair	37 / 25 = 1.48
SCI	0.87
z-score	-2.11
RNA class probability	0.95623
Highest conserved 22 nt window	0

This structure is part of cluster #19079

Alignment

```
        ...((((((.(((((((((.....)))))))))....(((((.(((.((....)))))...))))).))))))...........
hg17    GACTGGGCTTGAAGCCAGTTCTGTGCCAACATTTGAATTGGAAGATCTGTTTTCAGTCATATCCAGAAGTCCACC-AACCACTA
mm5     GCCTGGGCTGGATGCCAGCTCCACGCTGGCATTTGAGCTGGGAGACCTGTTTTCAGTCAAACCCAGAAGTCTACCAAACCACTA
rn3     GCCTGGGCTGGATGCCAGCTCCATGCTGGCATTTGAGCTGGGAGACCTGTTTTCAGTCAAACCCAGAAGTCTACCAAACCACTA
canFam1 GCCTGGGCTTAAAGCCAGTTCTGCATTGGTATTTGAACTGGGAGATCTGTTTTCAATCATACCCAGAAGTTTGCC-AACCACTA
```

[Download Clustal]

RNAz output

```
###########################  RNAz 0.1.1  ##############################

 Sequences: 4
 Slice: 1 to 84
 Columns: 84
 Strand: forward
 Mean pairwise identity:  83.10
 Mean single sequence MFE: -26.00
 Consensus MFE: -22.51
 Energy contribution: -21.20
 Covariance contribution:  -1.31
 Mean z-score:  -2.11
 Structure conservation index:   0.87
 SVM decision value:   1.45
 SVM RNA-class probability: 0.956230
 Prediction: RNA

######################################################################

>hg17.chr11/78209764-78209847
GACUGGGCUUGAAGCCAGUUCUGUGCCAACAUUUGAAUUGGAAGAUCUGUUUUCAGUCAUAUCCAGAAGUCCACCAACCACUA
...((((((....((((((((.(((....)))..))))))))..((.(((....))))).........))))))........... ( -20.30)
>mm5.chr7/49066976-49066892
GCCUGGGCUGGAUGCCAGCUCCACGCUGGCAUUUGAGCUGGGAGACCUGUUUUCAGUCAAACCCAGAAGUCUACCAAACCACUA
...(((..((((((((((((.....))))))))))).....(((((.(((.((....)))))...))))).......)))..)))... ( -31.50)
>rn3.chr1/114011546-114011462
GCCUGGGCUGGAUGCCAGCUCCAUGCUGGCAUUUGAGCUGGGAGACCUGUUUUCAGUCAAACCCAGAAGUCUACCAAACCACUA
...(((..((((((((((((.....))))))))))).....(((((.(((.((....)))))...))))).......)))..)))... ( -32.10)
>canFam1.chr21/29918985-29918902
GCCUGGGCUUAAAGCCAGUUCUGCAUUGGUAUUUGAACUGGGAGAUCUGUUUUCAAUCAUACCCAGAAGUUUGCCAACCACUA
..(((((.......(((((((((((....)))...))))))))(((.....))))........))))).................. ( -20.10)
>consensus
GCCUGGGCUGGAAGCCAGCUCCACGCUGGCAUUUGAACUGGGAGACCUGUUUUCAGUCAAACCCAGAAGUCUACC_AACCACUA
...((((((.(((((((((.....)))))))))....(((((.(((.((....)))))...))))).))))))........... (-22.51 = -21.20 +  -1.31)
//
```

[Download RNAz result]

Consensus structure

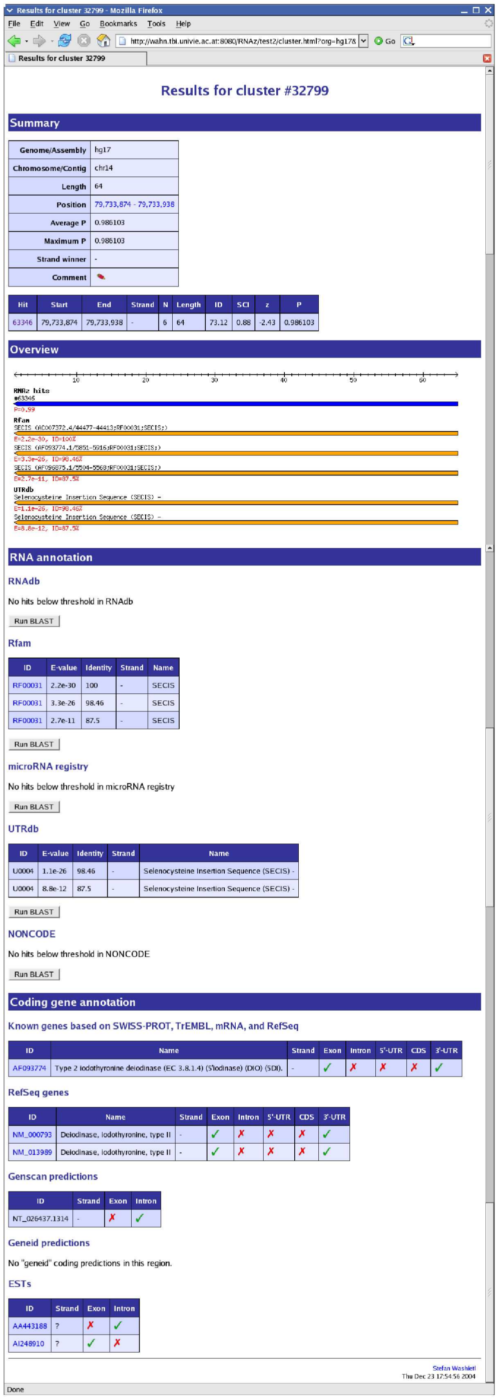
Secondary structure graph

[PS | PDF]

Montain plot

[PS | PDF]

Dotplot


Results for cluster 32799 - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://wahn.tbi.univie.ac.at:8080/RNAz/test2/cluster.html?org=hg17&
Go
Results for cluster 32799

Results for cluster #32799

Summary

Genome/Assembly	hg17
Chromosome/Contig	chr14
Length	64
Position	79,733,874 - 79,733,938
Average P	0.986103
Maximum P	0.986103
Strand winner	-
Comment	

Hit	Start	End	Strand	N	Length	ID	SCI	z	P
63346	79,733,874	79,733,938	-	6	64	73.12	0.88	-2.43	0.986103

Overview

10
20
30
40
50
60
RNAz hits
#63346
P=0.99
Rfam
SECIS (AC007372.4/44477-44413;RF00031;SECIS;)
E=2.2e-30, ID=100%
SECIS (AF093774.1/5851-5916;RF00031;SECIS;)
E=3.3e-26, ID=98.46%
SECIS (AF096875.1/5504-5568;RF00031;SECIS;)
E=2.7e-11, ID=87.5%
UTRdb
Selenocysteine Insertion Sequence (SECIS) -
E=1.1e-26, ID=98.46%
Selenocysteine Insertion Sequence (SECIS) -
E=8.8e-12, ID=87.5%

RNA annotation

RNAdb

No hits below threshold in RNAdb

Rfam

ID	E-value	Identity	Strand	Name
RF00031	2.2e-30	100	-	SECIS
RF00031	3.3e-26	98.46	-	SECIS
RF00031	2.7e-11	87.5	-	SECIS

microRNA registry

No hits below threshold in microRNA registry

UTRdb

ID	E-value	Identity	Strand	Name
U0004	1.1e-26	98.46	-	Selenocysteine Insertion Sequence (SECIS) -
U0004	8.8e-12	87.5	-	Selenocysteine Insertion Sequence (SECIS) -

NONCODE

No hits below threshold in NONCODE

Coding gene annotation

Known genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq

ID	Name	Strand	Exon	Intron	5'-UTR	CDS	3'-UTR
AF093774	Type 2 iodothyronine deiodinase (EC 3.8.1.4) (5'iodinase) (DIO) (5DI).	-	✓	✗	✗	✗	✓

RefSeq genes

ID	Name	Strand	Exon	Intron	5'-UTR	CDS	3'-UTR
NM_000793	Deiodinase, iodothyronine, type II	-	✓	✗	✗	✗	✓
NM_013989	Deiodinase, iodothyronine, type II	-	✓	✗	✗	✗	✓

Genscan predictions

ID	Strand	Exon	Intron
NT_026437.1314	-	✗	✓

Geneid predictions

No "geneid" coding predictions in this region.

ESTs

ID	Strand	Exon	Intron
AA443188	?	✗	✓
AI248910	?	✓	✗

Stefan Washietl
Thu Dec 23 17:54:56 2004

Fig. 34. Screenshots of the database web-interface. Two detailed results page are shown that summarize the characteristics of the detected structures (left) and the annotation of the corresponding genomic location (right).

Tab. 10. Genomic coverage and filtering steps of input alignments

	Genome Size (MB)	Coverage Fraction (%)	Alignments Number
Human genome	3,095.02	100.00	–
`PhastCons` most conserved	137.85	4.81	1,601,903
without coding regions	110.04	3.84	1,291,385
without alignments $< 50nt$	103.83	3.33	564,455
Set 1: 4 Mammals	82.64	2.88	438,788
Set 2: + Chicken	24.00	0.85	104,266
Set 3: + Fugu or zebrafish	6.86	0.24	30,896

Tab. 11. Results on the native alignments

Set		clusters	size (MB)	% of input	% of genome	cluster length average	maximum
A: Set 1	$p > 0.5$	91,676	12.47	15.09	0.44	136	1320
B: Set 1	$p > 0.9$	35,985	5.48	6.62	0.19	152	1320
C: Set 2	$p > 0.5$	20,391	2.80	11.52	0.10	137	665
D: Set 2	$p > 0.9$	8,802	1.34	5.50	0.05	152	665
E: Set 3	$p > 0.5$	2,916	0.38	5.57	0.01	131	488
F: Set 3	$p > 0.9$	996	0.14	2.03	0.00	139	488

Set 1: human/mouse/rat/dog, Set 2 = Set 1 + chicken, Set 3 = Set 2 + fugu or zebrafish
"cluster" refers to clustered regions of overlapping `RNAz`.

Tab. 12. Results on the randomized alignments

Set		clusters	Overlap native A	size (MB)	% of input	% of genome	Cluster length average	maximum
Set 1	$p > 0.5$	26,508	9039	3.20	3.87	0.11	121	496
Set 1	$p > 0.9$	6,898	2555	0.89	1.08	0.03	130	496
Set 2	$p > 0.5$	6,551	2158	0.81	3.35	0.03	124	394
Set 2	$p > 0.9$	2,281	881	0.31	1.26	0.01	134	394
Set 3	$p > 0.5$	795	179	0.096	1.40	0.00	121	338
Set 3	$p > 0.9$	208	63	0.026	0.38	0.00	127	279

the thermodynamic stability alone. Since many known ncRNAs are contained in this set we decided against removing highly conserved alignments from our survey despite the increased false positive rate.

4.4.2 Detection performance on known ncRNAs

A comprehensive annotation of ncRNAs in the human genome is not available, thus it is impossible to determine the overall sensitivity of our screen. For miRNAs and snoRNAs, however, a comprehensive annotation is provided in the UCSC browser.

There are 207 annotated miRNA loci of which 45 loci are not in our set of input alignments. We detect 157 (96.9%) of the remaining 162 miRNAs. The effective sensitivity is 75.8% for miRNA precursors, which are among the most easy-to-find ncRNAs (Fig. 35 c).

22 of the 86 annotated H/ACA-box snoRNAs are not included in the input set mostly because they are not detected by `PhastCons`. We recover 55 of the remaining 64 sequences (85.9%). We can thus relatively accurately detect this class of ncRNAs which have resisted computational prediction so far. (Effective sensitivity: 64.0%)

Our screen performs poorly on C/D-Box snoRNAs, however. Out of the 256 known C/D snoRNAs about a half (129) are missing in the input alignments. Even though we detect 39.4% of C/D snoRNAs in our set, the effective sensitivity is only 19.5%. C/D-Box snoRNAs are hard to detect computationally even with specialized approaches [2].

From these examples we estimate that the overall sensitivity of the combination of the `Multiz`/`PhastCons` alignments and `RNAz` is on the order of 30%.

We found that most of the miRNAs and snoRNAs are missed in our screen because they are not in our input set. To optimize future screens, and in particular sub-screens for miRNAs and H/ACA snoRNAs, we investigated in detail why miRNAs and H/ACA snoRNAs were missed in our selection of input alignments (Tabs. 13 and 14). miRNAs are mainly missed because they overlap with repeats or because they are not strictly conserved in all four mammals (It is more likely that the corresponding sequences are simply missing in one of the unfinished draft assemblies, in particular of the rat genome.) H/ACA snoRNAs are not well conserved on sequence level and `PhastCons` cannot detect conserved regions >50 nucleotides in many of them. In the case of C/D snoRNAs the problem is even more pronounced. Out of the 129 C/D snoRNAs not in our set, 63 are completely missed by `PhastCons`, in most of the other cases only short regions <50 are detected. Moreover, many snoRNAs which are contained in our set are not conserved over the full length. Given the fact the C/D snoRNAs in general do not exhibit very stable structures, the detection for `RNAz` is even more difficult if significant portions of the structure are missing in the input alignments.

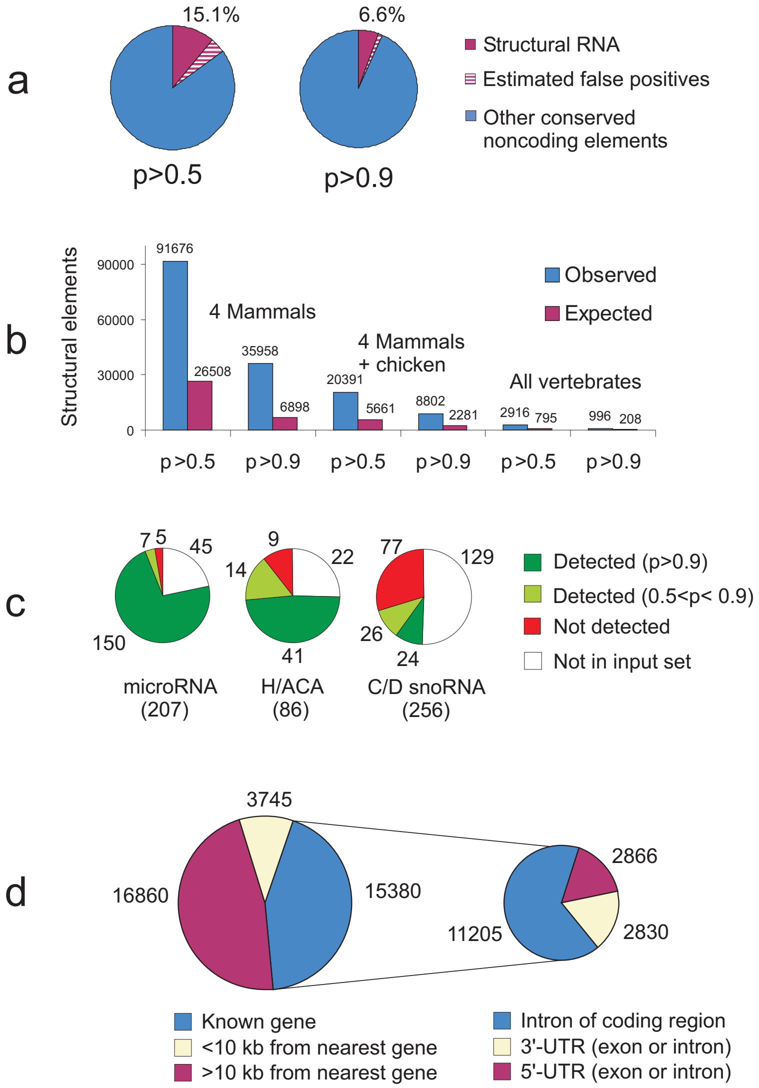


Fig. 35. Statistical analysis of predicted structural RNAs. (a) Observed and expected fraction of structural RNAs in the complete set of conserved non-coding elements. (b) Observed and expected number of structural elements for different phylogenetic distributions. (c) Detection performance on known miRNA and snoRNA genes. (d) Comparison with current protein-gene annotations.

We compared all hits with available databases of known ncRNAs (Tab. 15). Most of the "classical" structured ncRNAs, such as tRNAs and most snRNAs, are not contained in the input alignments because they are marked as repetitive DNA by `RepeatMasker` and were therefore excluded from the `Multiz` alignments. We do, however, detect all snRNAs of the minor spliceosome (U4atac, U6atac, U11, and U12), as well as very well conserved (although not very stable) structures within the RNAse P. We miss RNAse MRP and telomerase RNA, presumably because of the pseudoknotted structures [32, 137], which are not taken into account by `RNAz`.

We find local secondary structure motifs in various other documented ncRNAs which do not appear to have conserved global structures (Tab. 16). The *Xist* gene, a 17 kb ncRNA which plays a key role in dosage compensation and X chromosome inactivation [5] contains three independent conserved RNA secondary structures. Intriguingly, we find 8 `RNAz` hits in the human genome with significant sequence similarity to the *Air* RNA. This antisense transcript regulates imprinted gene expression in mouse [192] but is not conserved over its full length (≈1000 kb) in human. 7 of the 8 hits correspond to the same local secondary structure motif in *Air*. One of them can be found in an intron of *HERC2*, a locus located near the Prader-Willi imprinting center which in turn is regulated by antisense transcripts.

The `RNAdb` [174] compiles collections of expressed sequences with reduced protein coding capacity. A comparison of our `RNAz` hits with the `RNAdb` identifies conserved structured elements in many of these transcripts, thereby supporting that they function as ncRNAs (Tab. 15).

4.4.3 New members of known ncRNA families

A number of signals are novel ncRNAs that can be associated with known ncRNAs or ncRNA families through sequence similarity. Some of these are additional paralogs or orthologs of known RNA genes. For example, we found more than 100 hits with sequence similarity to snoRNAs. Some of these are most likely functional snoRNAs since they are human homologs of mouse snoRNAs described in reference [97].

Another class of signals are novel members of one of the large, well-described classes of ncRNAs. A simple subscreen was performed to identify putative H/ACA box snoRNAs. We selected all `RNAz` hits with two stems at least 15 pairs in length and separated by an unpaired hinge, which in addition have the motif ACA in the consensus sequence in the last 20 nt. We found 137 structures, of which 28 were known snoRNAs. Visual inspection shows that 30–40 additional clusters have typical H/ACA snoRNA-like secondary structure of which 15 also have the canonical H-box sequence ANANNA (Fig. 36 c,d). In many known snoRNAs, only short parts of the stem are conserved in the predicted consensus structure and/or only parts of the complete structure are detected as conserved structural element.

Tab. 13. microRNAs missing from the input set of alignments

Name	Conservation	Repeat	Other
hsa-let-7g	rat missing		
hsa-let-7i	gap in dog		
hsa-mir-9-1		simple Repeat	
hsa-mir-15a	rat missing		
hsa-mir-16-1	rat missing		
hsa-mir-22			overlap with coding region
hsa-mir-23a			`PhastCons` artifact[1]
hsa-mir-28		LINE	
hsa-mir-95		LINE	
hsa-mir-130b		SINE	
hsa-mir-133a-2			overlap with coding region
hsa-mir-135a-1	part of rat sequence missing		
hsa-mir-138-1	mouse missing		
hsa-mir-147	`PhastCons` region <50		
hsa-mir-148a	rat missing		
hsa-mir-149	rat missing		
hsa-mir-150			overlap with coding region
hsa-mir-151		LINE	
hsa-mir-155	rat missing		
hsa-mir-182	part of rat sequence missing		
hsa-mir-197	long gap in mouse		
hsa-mir-198	rat missing		
hsa-mir-199b	rat missing		
hsa-mir-203			`PhastCons` artifact[1]
hsa-mir-205			overlap with coding region
hsa-mir-212		low complexity	
hsa-mir-302a	rat missing		
hsa-mir-302b	rat missing		
hsa-mir-302c	rat missing		
hsa-mir-302d	rat missing		
hsa-mir-321		tRNA	
hsa-mir-325		LINE	
hsa-mir-326		Arthur 1	
hsa-mir-328	`PhastCons` region <50		
hsa-mir-330		SINE	
hsa-mir-335	rat missing	SINE	
hsa-mir-337	dog missing		
hsa-mir-340	rat missing	MARNA	
hsa-mir-345		SINE	
hsa-mir-367	rat missing		
hsa-mir-370		SINE	
hsa-mir-371	`PhastCons` region <50		
hsa-mir-372	`PhastCons` region <50		
hsa-mir-373	rat and mouse missing		
hsa-mir-374	gaps in mouse and rat	LINE	

[1] `PhastCons` region extends into the very gap-rich surrounding of the miRNA. Alignment discarded because it contains too many gaps.

Tab. 14. H/ACA snoRNAs missing from the input set of alignments

Name	Conservation	Repeat	Other
ACA2A	gap in mouse and rat		
ACA5	`PhastCons` region <50		
ACA5b	`PhastCons` region <50		
ACA10	`PhastCons` region <50		
ACA11	gap in mouse		
ACA29			alignment artifact[1]
ACA33	`PhastCons` region <50		
ACA39	`PhastCons` region <50		
ACA42	not detected by `PhastCons`		
ACA48	not detected by `PhastCons`		
ACA56	rat missing		
ACA59 (Chr. 1)		SINE	
ACA59 (Chr. 17)		SINE	
ACA67	`PhastCons` region <50		
U17a		other	
U17b		other	
U64			alignment artifact[1]
U66	`PhastCons` region <50		
U71a	`PhastCons` region <50		
U71b	rat missing		
U98b	`PhastCons` region <50		

[1] The sequence in chicken is much longer and opens up long gaps in the other sequences, which are thus discarded.

Tab. 15. Comparison of predicted RNAs with ncRNAs from the literature

Database	Ref.	$p > 0.5$	$p > 0.9$
`Rfam`	[70]	267	189
`NONCODE`	[139]	273	177
`RNAdb`	[174]	446	327
`miRNA Registry`	[69]	176	168
`UTRdb`	[178]	388	159
Curated		984	563
`hinv`	[99]	478	205
`Fantom`	[170]	1908	781
`chr7`	[195]	180	90
antisense pipeline	[174]	149	59
cDNA collections		2539	1056
Total		3441	1585

Tab. 16. Selected ncRNAs from literature with conserved RNA structures

Name	Type	max p	hits	Comment
U11	snRNA	0.98	1	
U12	snRNA	0.94	2	
U4atac	snRNA	0.71	3	
U6atac	snRNA	0.98	12	
RNAseP	Ribozyme	0.57	1	
UM 9(5)	Transcript of unknown function	1.0	8	Transcript was found to be differentially expressed in the brain, 7 of the 8 hits match the same region of this long (1241nt) transcript
HUC-1	Other functional transcript	0.95	1	Tissue specific transcript that enhances H19 transcription (an antisense transcript for imprinting)
MALAT-1	transcript of unknown function	1.0	3	three independent hits along this 8 kb transcript, which was identified in lung cancer cells as ncRNA
NCRMS	Other functional transcript	0.90	3	three independent hits in this 1.8 kb transcript; identified in rhabdomyosarcoma (RMS); host gene of mir-135a-2
BCMS	Other functional transcript	0.71	1	B-cell neoplasia associated transcript
aHIF	antisense transcript	0.98	1	aHIF is complementary to the 3' untranslated region of HIF1alpha mRNA, which encodes a protein known to stabilize p53 protein during hypoxia and to act as a transcription factor for hypoxia inducible genes
Air	Antisense transcript	0.96	8	Classical mouse model for imprinted antisense transcription.
CNS1	Other functional transcript	0.83	1	Expression of CNS1 accompanies the induction of the hyperacetylation of histone H3 on nucleosomes associated with the interleukin (IL)-4, IL-13 and IL-5 genes in developing Th2 cells
HOXA11 AS	Antisense transcript	0.53	1	
GA3824	Transcript of unknown function	0.74	1	Homo sapiens noncoding RNA GA3824 implicated in autism
Xist	Other functional transcript	1.0	3	Three independent hits in the long transcript responsible for X-inactivation in mammals
TTTY11	Transcript of unknown function	0.98	12	Identified in testis
TTTY3	Transcript of unknown function	0.86	1	Identified in testis
TTTY23	Transcript of unknown function	0.54	1	Identified in testis
His-1	Transcript of unknown function	1.0	2	Two independent hits on the same transcript; activation of this transcript leads to carcinogenesis

As a consequence, this subscreen is not exhaustive and a more detailed analysis can be expected to bring up even more candidates.

Berezikov and co-workers [14] identified 975 miRNA candidates in mouse/human and mouse/rat comparisons by means of a combination of phylogenetic shadowing and selection of stable stem-loop structures. Our set of input alignments contains 642 of these candidates, 472 overlap with our predictions ($p > 0.9$). Not all these stem-loops, which are stable as single sequences, are structurally conserved in all four mammals: some of them lack a stable consensus structure. A simple filter requiring a stem with at least 20 base pairs in the consensus structure, a mean z-score < -3.5 and a 22nt window with more than 0.95% pairwise sequence identity (the prospective mature miRNA sequence) retains 312 candidates, among them 109 known human miRNAs. Some of the unknown candidates show the typical mutation pattern of miRNA, see Figure 36 a,b. Others exhibit clear structural conservation but show a very different mutation pattern. We speculate that these sequences are not miRNAs but belong to different, so far undescribed, classes of ncRNAs.

4.4.4 Structures conserved across all vertebrates

The most highly conserved structures are of particular interest. We find 996 `RNAz` signals that are conserved in all 4 mammals, chicken and at least one of the two fish genomes (fugu, *Takifugu rubripes*, and zebrafish, *Danio rerio*). Of these, 152 can be at least partially annotated: 52 are miRNAs, 16 are snoRNAs, 28 are known elements in untranslated regions (UTRs), and 56 are similar to other described RNAs. 42 detected regions are contained within one of the different cDNA collections. 38 overlap with one of the 481 “ultraconserved elements” (segments longer than 200 base pairs that are identical between human, mouse and rat genomes) reported by Bejerano *et al.* [10]. A few of these can be identified as potential RNAs because of the substitution pattern in the fish and chicken sequences. For most of them, however, we cannot give a definitive classification because there is too little sequence variation in this special set of extremely conserved sequences.

4.4.5 Comparison with protein-gene annotations and transcriptional maps

The majority of the 35,989 structured RNA features detected with $p > 0.9$ is of completely unknown function (see Fig. 37 for a few selected examples). We compared the location of the hits with the protein coding gene annotations provided by the UCSC genome browser. About half of the predicted structures are located far away from any known protein coding gene, the other half is associated with known genes. Two thirds of the latter are located in introns. One sixth can be mapped to annotated UTRs.

In a recent study, sites of transcription of polyadenylated and non-polyadenylated RNAs

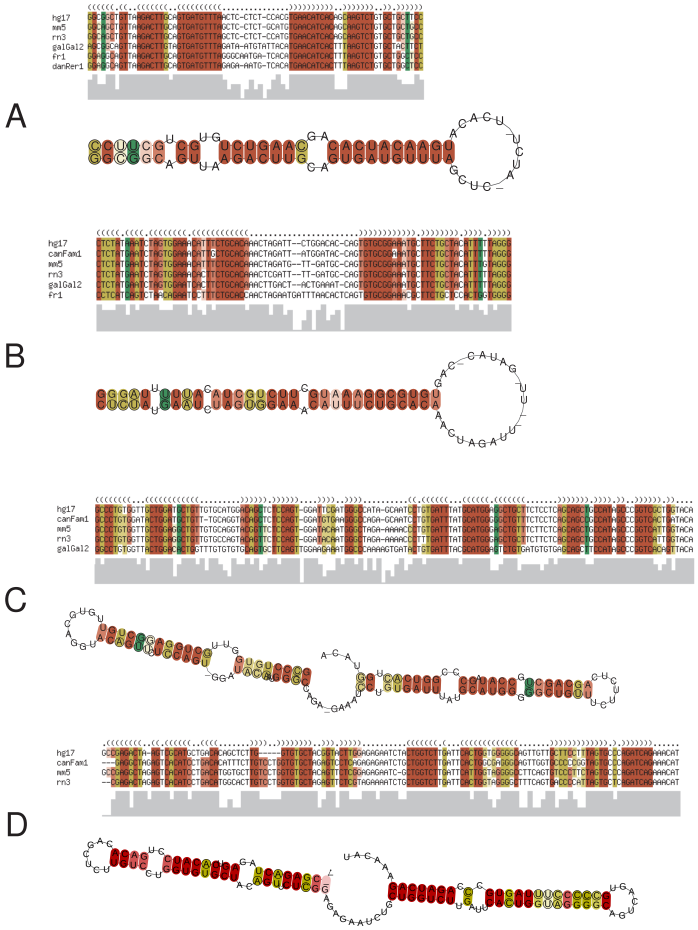


Fig. 36. Examples of microRNA and H/ACA snoRNA candidates detected with $p > 0.9$. The miRNA candidates (a,b) exhibit several characteristic features: (i) a stable hairpin consensus structure; (ii) the sequence of one arm of the stem is highly conserved over 22 nt (the putative mature miRNA); (iii) the opposite stem is also conserved but not that strictly; (iv) the loop sequence is diverged due to the absence of functional constraints in this region; (v) compensatory, or at least consistent, mutations are found in the outer parts of the stem where only structure but not sequence is important for function. The sequence in a is located on human chr.20 (pos. 33,041,857) in an intron of a mysine protein gene (AB040945). The position of candidate b is chr.15:43,512,536, in the UTR region of FOAP-11 (AF228422). The H/ACA snoRNA candidates (c,d) fold into the typical bipartite hairpin secondary structure. We observe H-box motifs ANANNA in the hinge regions and ACA motifs in the tail regions. Candidate c is located at chr.9:92,134,300 in an intron of Isoleucine-tRNA synthetase (D28473). Candidate d is located at chr16:2,786,411 and not associated with any known protein coding gene. Primary sequence motifs and secondary structure strongly suggest a role as classical pseudouridylation guides for these `RNAz` hits. Both candidates could be experimentally verified by Northern blot analysis [226]. Species abbreviations: `hg17` human, `mm5` mouse, `rn3` rat, `canFam1` dog, `galGal2` chicken, `fr1` fugu, `danRer1` zebrafish.

for 10 human chromosomes were mapped at 5-bp resolution in eight cell lines using tiling array technology [31]. We compared our predictions located on the 10 chromosomes with the cumulative “1 in 8” map, in which a positive probe need appear in at least one of eight cell lines. We found 40.7% of the predicted RNAs to overlap with detected sites of transcription (45.0% including signals in exons or introns of known UTRs). This is significantly (appr. 10%) higher than the background (see Methods) and comparable to the detection rate of well known ncRNAs: 45.2% of known microRNAs and 56.7% of known snoRNAs are detected on this transcriptional map.

A list compiling 50 examples of `RNAz` hits that are transcribed according to the tiling array experiment and that are strong candidates for independent ncRNAs can be found here: `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/lists/affy.html`. All these conserved RNA structures are at least 10 kb away from the nearest known gene and also do not appear to be part of other plausibly predicted protein coding genes.

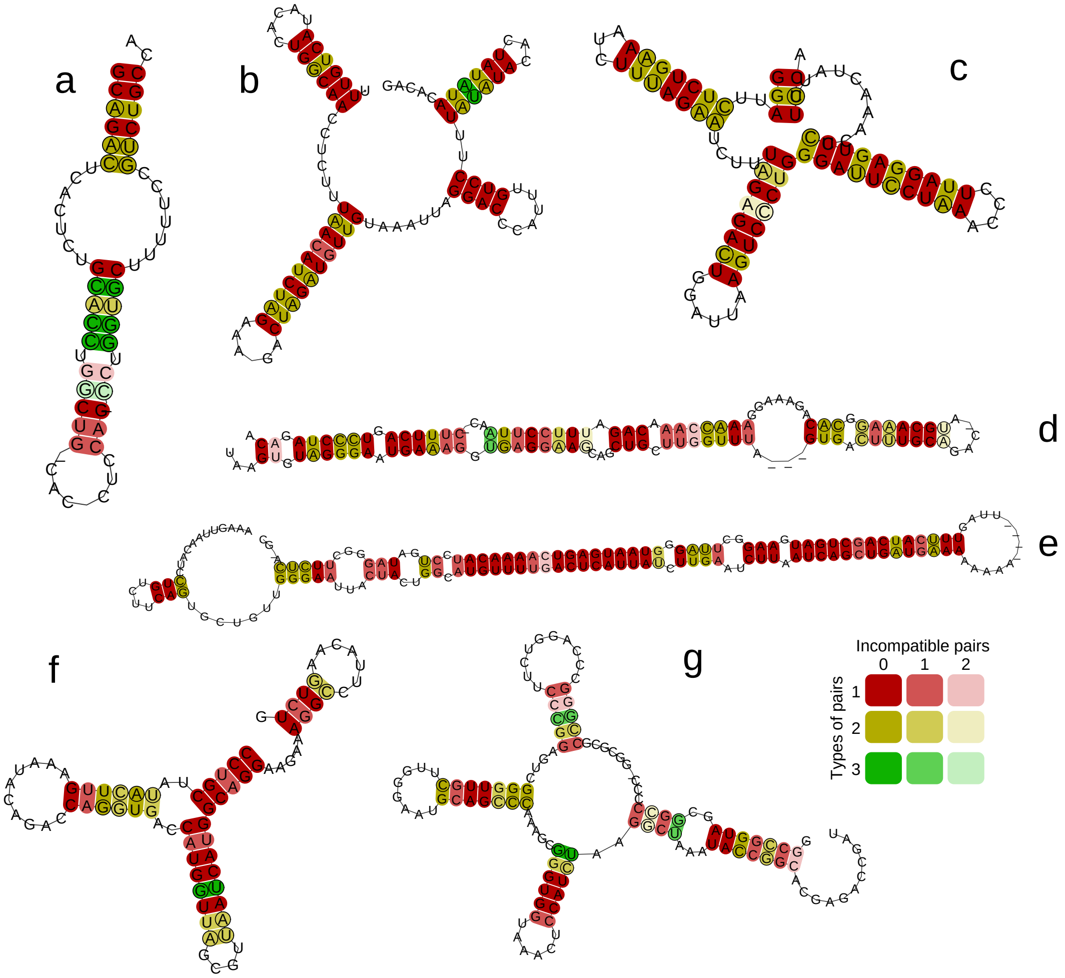


Fig. 37. Selected examples of candidates for novel structural RNAs detected with $p > 0.9$. For more examples see `http://www.tbi.univie.ac.at/Papers/SUPPLEMENTS/ncRNA`. Structure a is conserved across all vertebrates. It is located near an intron/exon boundary and EST data suggests alternative splicing events in this region. The sequence of structure b has similarity to a transcript in the Chr. 7 set of RNAdb and is conserved in mammals. We found more than 50 conserved secondary structures throughout the genome with sequence similarity to this transcript. Within these hits, we could identify this structural motif 7 times by visual inspection. Structures d and e form highly conserved stem loops. Structure d is located in an intron of a coding gene and conserved across all vertebrates. The stem loop is longer than a typical microRNA precursor and also shows a different mutational signature. Additional `RNAz` hits in the close vicinity suggest that this is a local substructure of a longer RNA. Also the stem loop in structure c is almost twice as long as a typical microRNA precursor. Structures c,f,g are locally highly structured. Element c is conserved in mammals, f and g are conserved across all vertebrates. Genomic locations of all examples: (a) chr.22:18,488,478 (in intron of RAN binding protein 1, D38076) (b) chr.12:74,595,654 (intergenic), (c) chr.7: 133,840,516 (intergenic), (d) chr.2: 104,445,752 (intergenic), (f) cqhr.8:57,457,661 (intergenic), (h) chr.5:32,415,412, (intron of zinc finger RNA binding protein, AJ314790) (g) chr.10:32,739,292 (intergenic)

5 Review of other applications

5.1 `Alifoldz` prediction and verification of cyanobacterial ncRNAs

Ilka Axmann, Philip Kensche and colleagues conducted a computational screen coupled with experimental methods to detect and characterize ncRNAs in marine cyanobacteria [6]. Using `Blast` they generated alignments from the intergenic regions of four sequenced strains and scored them with `Alifoldz`. The analysis was focused on the highest scoring regions and 7 novel ncRNAs were detected and described in detail. The new ncRNAs were designated Yfr1–7 (Yfr for cyanobacterial functional RNA-coding gene). The authors note a very high rate of verification in this set of highly scoring elements and expect additional ncRNAs among the hits with lower scores that have not been tested.

Some ncRNAs could only be detected in two or three of the four strains, whereas Yfr2–Yfr5 are structurally highly related and are encoded by a rapidly evolving gene family as their genes exist in different copy number and at different sites in the four strains. Yfr7 has been shown to be present in at least seven other cyanobacteria and the authors suspect it to be a homologue of the γ-proteobacterial 6S RNA on the basis of structural similarities. Although no direct function for the novel ncRNAs was shown in this paper, the authors assume regulatory roles for most of the ncRNAs similar to the regulatory small ncRNAs that could be detected in other eubacteria. This could explain how the few regulatory proteins in the compact genomes of cyanobacteria can efficiently sustain the lifestyle of an ecologically successful marine microorganism.

In addition to the characterization of the ncRNAs, the analysis revealed evidence for control elements for several ribosomal operons as well as riboswitches for thiamine pyrophosphate and cobalamin.

5.2 Benchmarking of sequence alignment programs upon structural RNAs

The use of multiple sequence alignments is an essential step for many RNA sequence analysis methods, including RNA structure analysis, RNA homology search, RNA based phylogenetic inference and, of course, ncRNA detection as described in this thesis. There exist a large number of different alignment programs. The performance of alignment programs are carefully benchmarked on proteins. However, the results of these benchmarks are not necessarily informative for somebody who is interested in optimally aligning structural RNAs. Together with Paul Gardner and Andreas Wilm, a comprehensive benchmark of multiple sequence analysis programs was conducted [62]. This study was inspired by the effect of alignment accuracy on consensus structure prediction as shown in Fig.17 (page 47).

Using the same methodology as described in section 3.3.2, we generated a large representative test set of alignments of several structural ncRNAs families. We then used 14 different alignment programs to align the test cases. The accuracy of the alignment was assessed by two scores: The SCI and the sum-of-pair score (SPS), which measures the accuracy of a given alignment against a reference alignment (in our case manually curated seed alignment from `Rfam`). The results of this benchmark study can be summarized as follows:

- The two independent measures of global alignment accuracy SPS and SCI are generally in agreement. The SCI is independent of a reference alignment. Since the SCI does not consider the alignment quality of the sequences in the loops (they are not relevant for the consensus structure), the SPS is preferable if trusted structural alignments are available as reference.
- The relative performance of multiple sequence alignment programs on RNA alignments can differ remarkably from the performance observed on protein alignments.
- The multiple sequence alignment algorithms, such as `ClustalW`, `MUSCLE`, `PCMA`, `POA`, `ProAlign` and `Prrn` perform well on high- to medium-homology datasets.
- `ClustalW`, `ProAlign` and `POA` consistently ranked in the top 10 across all homology ranges.
- The "twilight zone" of ncRNA alignment is in the 50%–60% sequence-identity range.
- Below this limit, algorithms incorporating structural information (`Dynalign`, `Foldalign`, `PMcomp` and `Stemloc`) outperform pure sequence-based methods. However, these algorithms are computationally demanding which severely limits their use in practice.

The test sets have been made available online (`http://www.binf.ku.dk/users/pgardner/bralibase/`). The standardized test sets and accuracy measures provided there, can help to test new alignment programs and optimize existing ones.

5.3 `RNAz` predicted miRNA precursors in the miRNAMap database

Wei-Che Hsu *et al.* developed an integrated database that collects miRNAs genes, miRNA targets and their regulatory relationships [95]. The database contains miRNAs from the `miRNA registry` as well as predicted miRNAs from our human screen. The candidate miRNA precursors that where identified in our simple subscreen (see section 4.4.3) where used as starting point for further analysis. Using a machine learning approach, mature miRNAs were predicted in the candidate stem loops. 464 human mature microRNAs were predicted from the the initial set of 2681 putative miRNA precursors. Using `miRanda` [105], potential targets of the known and predicted miRNAs in conserved UTRs were predicted.

This information is augmented with expression profiles of known miRNAs, cross-species comparisons, gene annotations and various cross-links to other biological databases.

5.4 `RNAz` screens of urochordate and nematode genomes

Kristin Missal, Dominic Rose and Peter F. Stadler used `RNAz` to screen urochordate and nematode genomes for functional RNAs [161, 162].

Urochordates can be regarded as the sister group of vertebrates. This lineage is of particular interest because it does not share the genome duplications that shaped the vertebrate genomes [92]. One cannot simply include urochordate sequences in the ncRNA screen for vertebrates because the large evolutionary distance makes reliable sequence alignments impossible. However, the the complete genomes of two ascidians *Ciona intestinalis* and *Ciona savignyi* as well as incomplete shotgun traces of the larvacean *Oikopleura dioica* made an independent comparative study possible. Using a `Blast` based alignment protocol, pairwise and three-way alignments were generated and screened with `RNAz`. In the pairwise comparison of the two ascidians, about 15 MB of non-coding sequence could be aligned with an E-value cut-off of 10^{-3}. 3332 hits (2.6% of the input sequences) and 2109 hits (1.7% of the input sequences) were predicted as ncRNA candidate on the $p > 0.5$ and $p > 0.9$ significance levels, respectively. The authors estimate a false-positive rate of 17.1% ($p > 0.5$) and 11.4% ($p > 0.9$) based on similar shuffling controls as in our human screen. Data on ncRNAs is sparse in these organisms, which makes it difficult to annotate the predicted RNAs. Using `tRNAscan-SE` [140], about 300 hits could be annotated as tRNAs. In addition, some 100 snRNAs, a few microRNAs and snoRNAs could be identified in the prediction set by sequence similarity searches.

A similar screen was conducted for the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* [162]. Here, about 13 MB of noncoding DNA could be aligned by `Blast` and 3672 hits (3.2% of the input sequences) and 2366 hits (2.1% of the input sequences) were predicted by `RNAz` on the $p > 0.5$ and $p > 0.9$ significance levels, respectively. The false positive rate was estimated to be 49% ($p > 0.5$) and 33% ($p > 0.9$). 679 hits can be identified as known ncRNAs or as clear homologs of known ncRNAs. Most of the known ncRNAs were tRNAs (483). Further analysis of the results found several predicted ncRNAs to be associated with characteristic upstream motifs. In addition, clustering of the hits on basis of sequence similarity could identify several interesting sequence families some of which also cluster with respect to the genomic location. Based on the sensitivity on known ncRNAs and the false positive rate, the authors estimate that there are 3000–4000 RNA with evolutionary conserved secondary structures in nematode genomes. This is in the same order of magnitude as other estimates. In a recent study of small ncRNAs in *C. elegans*, Deng *et al.* [46] estimated between 1600 and 4100 ncRNAs based on different considerations (intron conservation, upstream motif conservation, cloning frequencies/northern signals).

6 Discussion

6.1 Protein prediction vs. ncRNA prediction

6.1.1 Protein prediction, a standard procedure in genome annotation

Today, a number of protein-gene finders are available built on well established algorithms. Although these programs come with their own sets of problems, protein-gene prediction has become an invaluable tool for genome annotation. The problem of predicting protein genes is by no means trivial, but at least rather well defined: Given a genomic sequence, find those regions which correspond to open reading frames coding for a protein. Computational solutions for this problem have grown historically together with the available data. Long before complete genomes were sequenced, GenBank had already collected large numbers of sequenced mRNAs and genomic regions. Protein-gene prediction did always benefit from the detailed biological knowledge on proteins and their genes. Also in the post-genomic era with complete genomes available, large scale projects like extensive cDNA or EST libraries provide reasonable amounts of experimental data to constantly improve the methods. In organized collaborations, computational biologists team up with experimentalists to critically assess these methods and test predictions of novel genes [1, 55].

6.1.2 The "RNA revolution" and the need for ncRNA prediction

ncRNA finders are still in their infancy. Only a few years ago, the existence of ncRNAs beyond the well known text-book examples was not a subject of mainstream biology. Researchers and their experimental methods were focused on protein biology in a way that evidence for ncRNAs encountered in an experiment was simply ignored or excluded by the experimental setup from the beginning. For example, standard genetic experiments based on mutagenesis which work fine for a protein-gene where a single point mutation can completely destroy function, may fail for ncRNAs where a single substitution or deletion may have no effect. Most cloning protocols discard any RNAs which are not polyadenylated and thus exclude ncRNAs from many functional screens. With the discovering of new ncRNAs in various organisms, in particular microRNAs, the situation has changed. As outlined in the introduction, different lines of evidence ranging from simple theoretical considerations [157] to multi-million dollar experiments [31, 24] suggest that ncRNAs are much more prevalent than previously assumed. This "RNA revolution" hit the community unprepared. Our understanding of ncRNA biology is limited and we cannot rely on a large body of research which has grown over the past decades. Instead we face fundamental questions which cannot be definitely answered by any currently available approaches. How many ncRNAs are there and what are their functions? The recent finishing of the human genome sequence emphasizes the “need for reliable experimental and computational methods for comprehensive identification of non-coding RNAs” [213].

6.1.3 ncRNA prediction, an ill-defined problem

The main reason that hinders systematic computational screens for ncRNAs as seen for protein-coding genes is that there are no common statistically significant features in primary sequence. There are no start/stop codons, no open reading frame, no codon bias, no typical splicing signals. It is even not clear what we define as “ncRNA”. There is no doubt that independent “RNA genes” with a defined molecular function such as tRNAs or microRNAs should be called ncRNAs. But the situation is not always that clear. The transcriptional activity of at least mammalian genomes is much more complex than anticipated and in the light of recent studies the concept of a gene becomes blurred [60]. We see mRNA-like ncRNAs, non-polyadenylated RNAs from both intronic and intergenic regions, overlapping transcripts, extensive antisense transcription, and transcribed pseudogenes. In addition there is a recent example of a ncRNA that only is expressed to interfere with and downregulate the transcription of a neighboring gene but the produced RNA molecule itself does not have any obvious function [150]. There is even an example of a functional RNA encoding a protein [34]. The spectrum of ncRNAs and their mode of action is very heterogeneous. One can safely assume that the full spectrum of functions is not yet discovered.

6.2 Predicting structural ncRNAs

We have to accept that a general ncRNA finder is an unrealistic goal even in the long term. In this work, we focused on a more precisely defined problem. We set out to predict ncRNAs for which the secondary structure is of functional importance. Although this group covers only a subset of all ncRNAs, we are convinced that RNA secondary structure is currently the only known feature of ncRNAs which allows reasonable computational prediction.

6.2.1 Limitations of available structural ncRNA finders

We have shown in section 3.1 that the stability of RNA secondary structure is of limited statistical significance if only single sequences are considered. Also measures other than pure thermodynamic stability (e.g. “well-definedness”, section 3.2) cannot help and, therefore, we soon moved our focus to comparative approaches, i.e. finding evolutionary conserved RNA secondary structures. A few programs for this task already existed or were introduced while this thesis was written (see section 2.4.2). None of them, however, appeared to meet the requirements for automatic annotation of large eukaryotic genomes. The `ddbRNA` algorithm only relies on the existence of compensatory mutations and does not consider any RNA folding model beyond complementary regions. The sensitivity of `ddbRNA` is low on realistic datasets, where the number of compensatory mutations is not high enough to generate a meaningful signal. Similarly, `MSARi` depends on a large number of sequences (10–15) of high divergence to perform well. To our knowledge, none of both programs have been used for any real-life applications so far. The most established program is without doubt `QRNA`, which was successfully used to detect ncRNAs in prokaryotic and yeast genomes. However, based on the published results and our own tests we found that the `QRNA` approach will not easily scale to large genomes. Specificity is low if the input sequences lie outside of the optimal range of evolutionary distance, it is somewhat slow and hence requires high computational resources, and it is limited to pairwise alignments. A new version, `eQRNA`, has been released recently which uses a more sophisticated evolutionary model [185] and one can expect significant better classification accuracy. However, it is still limited to pairwise alignments and therefore it cannot make use of the additional information contained in a multiple alignment of more than two sequences.

6.2.2 A next generation of structural ncRNA finders

There is need for a next generation of ncRNA prediction algorithms which can handle multiple sequence alignments and make use of an accurate folding model. One could consider to extend available SCFG approaches. Based on previous work [116], Pedersen and colleagues work on a program called `EvoFold` which combines probabilistic models of RNA secondary structure and primary sequence evolution. `EvoFold` was not available and unpublished at the time this thesis was written.

We chose the alternative, energy-based approach. MFE algorithms for the prediction of secondary structure are well established and still most accurate. With `RNAalifold`, there is an effective extension for consensus folding of alignments. In the first part of the thesis, we showed that the `RNAalifold` score, an averaged energy score augmented with covariance information, cannot only be used to *predict* a consensus secondary structure but also to *detect* conserved RNA secondary structures in multiple sequence alignments. We observed an impressive improvement in the detection performance compared to single sequence predictions. To assess the statistical significance, we used a shuffling approach and calculated z-scores normalized for all relevant parameters such as length, base composition, and mean pairwise-identity. Although a properly normalized score combining both stability and conservation is exactly what is needed, the sampling procedure that is necessary to calculate the score, makes it impractical for general use.

In the second part of the thesis, we proposed an alternative approach which overcomes the limitations of the shuffling approach but, as it turned out, shows comparable accuracy.

Although conceptually simple, the SCI proved to be a convenient and effective measure of structural conservation. Moreover, as a consensus of several independent sequences in an alignment, also stability can be a significant measure. In this context, we have demonstrated that a properly normalized stability measure can be directly calculated without the need for time consuming sampling of shuffled sequences. We used a SVM regression algorithm and could solve the problem for mononucleotide shuffled sequences almost perfectly, i.e. without loss of accuracy compared to the traditional sampling method. Interestingly, Clote and colleagues independently developed the concept of approximating z-scores [36]. In particular, they proved a theorem of the existence of an asymptotic limit for mean and standard deviation of minimum free energy per nucleotide for random RNA. By proving this basic assumption, which we took for granted, our regression approach is now put on solid theoretical grounds.

6.2.3 Limitations of `RNAz`

Our program `RNAz` shows unprecedented accuracy in the benchmarking tests and clearly outperforms any other available programs. However, there are still a number of limitations which should be addressed in future versions.

The SVM classification of `RNAz` is limited to alignments of up to six sequences. In principle, it is of course possible to train a SVM on alignments with more than six sequences. We failed, however, to generate reasonable test sets of non-redundant alignments from the currently available sequences in `Rfam`. For many current applications this limit will not cause any problems, but with more and more genomes sequenced much larger data sets can be expected. We want to mention the ENCODE project [55], with one of its goals to sequence targeted regions from the human genome in dozens of related species. For such applications `RNAz` is currently not prepared.

Another major problem is that `RNAz` scores a given alignment *globally*. Large alignments are scanned in overlapping windows. Ideally, one likes to detect conserved RNA secondary structures *locally* by scanning smoothly over large alignments and reporting the most significant local structures. For single sequences, a local MFE folding program `RNAlfold` [89] is available. One could consider to apply this simple variant of the standard folding algorithm to `RNAalifold`/`RNAz`. Due to the quality of current genome-wide alignments this is, unfortunately, not straightforward. Missing data, spurious and gap-rich matchings, and low complexity regions are characteristics of typical genome-wide alignments. In our human screen, we found it essential to filter and pre-process the alignments before any efforts to predict a consensus secondary structure make sense. Although a local version of `RNAz`, which scans over megabases of automatically generated alignments, is highly desirable, much additional work needs to be done to achieve this goal.

Finally, we want to mention a more fundamental limitation of our approach, especially compared to SCFG based methods. Our strategy to detect conserved RNA secondary structures does not consider any model of sequence evolution and no direct statistical interpretation of the results is possible. Phylogenetic-SCFGs aim to describe the problem by combining full probabilistic models of sequence evolution and RNA folding. Of course, also these methods depend on training data and *ad hoc* assumptions in their models, but they provide a rigorous mathematical framework for the complete problem. In contrast, our work is a combination of several independent components. It depends (i) on the MFE folding algorithm which in turn depends on hundreds of empirically found energy parameters, (ii) a covariance score which incorporates structural conservation, and (iii) a SVM learning algorithm for classification. It is obvious that there is no mathematical framework which could describe our strategy as a whole.

A phylogenetic component is still missing, although the performance of `RNAz` could probably be enhanced by considering an underlying phylogenetic tree. Especially for unbalanced data, for example in cases where we have three closely related mammalian and one distant vertebrate sequence in an alignment, a weighting of the sequence contributions according to a phylogenetic tree is desirable. Preliminary work has been done by Hofacker *et al.* [87] but these ideas have not been elaborated in this thesis.

6.3 A prototype screen for ncRNAs of the human genome

We decided not to go into too much theoretical details without testing `RNAz` in large scale screens. Only in a realistic scenario the performance and probably even more shortcomings can become evident. Moreover, we felt that, even in the very first version, `RNAz` has the potential to give important biological insights. As a prototype study we conducted a screen of the human genome. Indeed, screening a large mammalian genome is not an easy undertaking but, for obvious reasons, it is one of the most attractive and challenging goals. With four mammalian and two additional vertebrate genomes available, such a screen has become possible.

6.3.1 Pre-selection of candidates on the basis of sequence conservation

The first crucial step of our screen was the pre-selection of the regions which should be scored. Only a fraction of the 3,000 megabases in the human genome can be aligned to the other genomes. We chose the most conserved non-coding regions. A large body of literature is available on the theory of evolutionary sequence conservation. There is an ongoing debate over the fraction of conserved regions in the human genome and whether these regions are functional or not. Following the most common consensus in the community, one can note that at least 5% of the human genome is more conserved than one could expect by some reasonable model of neutral sequence evolution. These regions might appear only "conserved" because of reduced mutation rates, but as the main reason for the effect one assumes purifying selection, immediately implying some function.

We did not question any of these assumptions and used the results of the new `PhastCons` program which was the best program available to detect conserved regions in complete genomes.

6.3.2 Promising results on known ncRNAs

The most conserved non-coding DNA was screened with `RNAz` for functional RNAs. At the highest significance level, we predict structural RNA elements in 6.6% of these regions (appr. 36,000 structural elements throughout the genome). The initial analysis underlines the value of this prediction. Our screen recovers hundreds of known structural RNAs (both ncRNAs and structural elements in UTRs of mRNAs), it identifies additional members of known ncRNA families, and detects previously undescribed conserved structural elements in some known ncRNAs. To our knowledge this is the first attempt of a genome wide annotation of ncRNAs in human. At the time this thesis was written, we could not think of any alternative approaches which could yield comparable or even better results.

The most intriguing but, at the same time, also the most arguable result of our study is the number of predicted structures which could not be assigned to known RNAs. In this context two questions arise (i) What is the true false positive rate not covered by our simple null-hypothesis of shuffled alignments (ii) What are the functions of the detected structures?

6.3.3 The real false-positive rate remains uncertain

We tried to estimate the false-positive rate using randomized controls. We estimate a false positive rate of 1.1% and thus observe an overall signal-to-noise ration of 6:1, implying that the majority of the predictions are biologically relevant. Clearly, any such approach can only approximate the true genomic background and hence cannot rule out the possibility that non-random sequence patterns could cause spurious hits resembling stable and conserved RNA structures. It is conceivable that local inhomogeneities of base composition or low complexity regions (e.g. repeats of single nucleotides) could bias the z-score calculation. Such "dubious" alignments usually did not pass the filtering steps through `RepeatMasker` and `PhastCons`. Nevertheless, there are alignments where the significance of the `PhastCons` prediction and/or our `RNAz` prediction is highly questionable. We also observed that inhomogeneities in the degree of conservation can cause an increased rate of false positives. For example, a highly conserved block with nearly 100% mean pairwise identity is flanked by two gap-rich regions of low identity. On average, the identity will be, say 70%. The highly conserved block results in a high SCI, which can then be mis-interpreted as significant due to the low overall mean pairwise identity. This is an obvious reason of false positives but should be covered in our false positive estimate which considers such types of artifacts in the randomization procedure (cf. Fig. 13, page 42). The same is true for alignments which are highly conserved over the full length. Here the missing covariance information can cause an increased false positive rate. In most of the cases, however, we cannot find obvious sources of false positives. In absence of any other plausible explanation we have to trust our procedure and suspect functional RNA secondary structures. Again, we cannot exclude that there exist effects which we are currently not aware of. We regard our predictions as a working hypothesis and note that the estimated number of false positives must be seen as a lower bound.

6.3.4 Can experimental verification help?

Any experimental working molecular biologist would immediately suggest to do expression studies for example by northern blot analysis to decide which fraction of the predictions is “true”. In the case of compact genomes of single cellular organisms, an experimental follow-up of the predictions by northern blot analysis can be without doubt insightful. This was demonstrated for example by Rivas *et al.* for enterobacteria [188] and, recently, by Axmann and Kensche *et al.* who experimentally verified `Alifoldz` predicted ncRNAs in cyanobacteria.

However, not primarily because this thesis reports purely computational biology, we do not consider northern blots or similar techniques an adequate means to validate a genome-wide annotation of a large mammalian genome. Without a careful evaluation of the sensitivity/specificity of the experimental procedure itself, such experiments will have little to add. In a multicellular organism, tissue specific expression and low concentrations make experimental detection of ncRNAs challenging. Nevertheless, some of our predictions were tested experimentally by Melanie Lukasser and Alexander Hüttenhofer (unpublished). We selected six candidates with strong `RNAz` signals for northern blot analysis in HeLa cells. One could observe two positive signals (one of them was very strong). The experiment was repeated and the second time only one signal was observed. What we could learn from these experiments is that (i) our predictions indeed contain northern detectable ncRNAs. (ii) The specificity of the experiment is not necessarily better than the specificity of the prediction (at least 1/6 false positives). (iii) A negative result is not definitive; it simply means that a northern on HeLa cells could not detect the RNA. Tests on different tissues with more sensitive methods like RT-PCR could give other results. However, tiling arrays and random cloning already found evidence for a relatively large fraction of the mammalian genome to produce transcripts. For example, more than 40% of our hits overlap with “transfrags” signals. Many of the transfrags can be readily “verified” by PCR based methods [31]. In the light of these results, it is questionable what information we gain from such expression experiments at all. Defining a "true" ncRNA solely on the basis of a positive northern blot (i.e. on the basis of high levels of transcription) is problematic. One can argue that this definition is useful since it applies to most known ncRNAs. On the other hand, such a definition might not include many important ncRNAs we do not know yet (probably simply because they do not meet the requirements of strong transcription). Ultimately, the most interesting question is whether a ncRNA is functional or not. Here both expression studies and computational methods can currently only give indirect evidence. A strong transcribed distinct RNA species detected on a northern blot is a good candidate for a functional RNA. Likewise, a highly conserved RNA secondary structure should not arise by chance and is thus likely to be part of some sort of functional RNA. Both experiment and computational analysis follow completely different paths. Ideally, they supplement each other, but one must not forget that also computational biology can live on its own. Provided that the results are sensibly interpreted, we are convinced that the analysis of RNA structures on a genome-wide scale can be fruitful without any direct experimental support.

6.3.5 Potential functions of newly discovered RNA structures

The detected structured elements can have different functions. If there is no obvious similarity to known ncRNA, however, a functional assignment is almost impossible. The mapping to available genomic annotations can give some insights.

Approximately one sixth of our hits can be mapped to UTR regions of known protein-coding genes (this includes the hits in introns of UTR regions). Those hits represent potential regulatory elements of the mRNA.

A third of our hits are located in introns of protein coding genes. This finding strongly supports the notion that a plethora of functional RNAs are expressed from intronic DNA [155, 157]. It is an interesting scenario, that intronic ncRNAs might have regulatory function and interact with the genomic region, the mRNA, or the protein product of the gene from which they were produced. This way of RNA based regulation allows more flexible and complex regulation networks which could not be accomplished with protein regulators alone. MicroRNAs are probably the best example of such regulatory ncRNAs. Most microRNAs are encoded in introns. Chen *et al.* [31] report many non-polyadenylated transcripts derived from introns further supporting this hypothesis.

It is also conceivable that the intronic structures play a regulatory role in the pre-spliced mRNA. There is evidence that many pre-mRNA sequences contain selected regions folding *in vivo* into well-defined secondary structures [20]. Recently, for example, a well conserved RNA secondary structure was shown to regulate alternative splicing in the homothorax gene in drosophila [65].

In our study we excluded annotated coding regions. However, we cannot rule out the possibility that undiscovered coding exons showing signs of stable secondary structures are among our hits. We compared our predictions with protein-gene prediction programs (`Geneid` and `Genscan`). Only a minor fraction (6%) of all hits overlap with coding sequence predictions. Among the 996 hits conserved across all vertebrates, the fraction of predicted coding exons was higher (17%). Some of them are likely to be coding exons but, interestingly, also several known ncRNAs (including expressed pseudogenes and miRNAs) are predicted to be coding exons in this set. The observation that protein gene finders call ncRNAs and *vice-versa*, can be interpreted as indirect evidence that they share some common features. We speculate that many ncRNAs evolved from coding mRNAs. Expressed pseudogenes are probably the best example because they represent an intermediate stage where we see ncRNA features (expressed but not translated, in some cases shown to be functional [236]), but also still clearly see the remnants of the mRNA. The connection of coding sequences and functional RNAs is definitely a topic which needs further exploration.

One half of the detected structures are located in intergenic regions at least 10 kb away from any known protein-coding gene. Given that the current protein gene annotation of the human genome is fairly complete, one can assume that most of these hits are unrelated to mRNAs of protein coding genes and thus are candidates for independent functional ncRNAs. In this context, the question arises on the number of ncRNAs which we estimate from our data.

6.3.6 How many ncRNAs in human?

Gene numbers have always been a fascinating topic for the biological community. The history of protein coding-gene estimates shows, however, that one should be very careful before attempting (and in particular publishing) any quantitative estimate [177]. The situation is even worse in the case of ncRNAs. Our data is notoriously difficult to be interpreted in terms of gene numbers for several reasons: As mentioned before, the real false-positive rate remains uncertain. It is also unclear how many ncRNAs genes in the sense of a genetic unit correspond to one detected structure. It is possible that one detected structure contains several ncRNA genes (e.g. a miRNA cluster). On the other hand, a spliced mRNA-like ncRNA can have several independently conserved structures (e.g. *Xist*). Moreover, the way the input data is processed depends on many rather arbitrary parameters some of which are embedded in third party programs (`Multiz`, `PhastCons`). We predict that the overall results will vary considerably if these parameters are changed. This is also the reason why it is not easily possible to directly compare the number of detected hits in our human screen with screens in other species (e.g. those described in section 5.4) which follow different protocols. Taken together, it would not be scientific sound to give any prediction on the number of functional ncRNAs which goes beyond an order-of-magnitude estimation. We estimate the number of functional ncRNAs in the order of ten-thousands.

6.4 Conclusion

In this work we addressed the problem of computational *de novo* prediction of non-coding RNAs, one of the most challenging problems in current bioinformatics. We used the power of comparative genomics and RNA secondary structure prediction and devised new algorithms that can detect evolutionary conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. Our program `RNAz` outperforms any other available programs and was used for a first comprehensive annotation of conserved RNA secondary structures in the human genome. We found evidence for a large number of previously undescribed RNA structures which we predict to be part of functional ncRNAs (independent ncRNAs or regulatory elements of mRNAs).

Despite the promising results, our work only describes the very first step which is necessary to elucidate ncRNA function by means of sequence analysis. In the world of protein bioinformatics and proteomics, we would have now reached a point where we can predict ORFs as candidates for potential proteins and observed that ORFs occur more frequently than one could expect by chance. For proteins, there is a big arsenal of programs to further classify and analyze the candidates. A classification of all the RNA secondary structures encoded within a genome, anticipated and dubbed "structural RNomics" few years ago [50], will be the next logical step following our analysis. This is a big challenge and we think that also here a set of new generation RNA algorithms will be necessary to handle the vast amounts of data. We hope that our programs and results also stimulate the field of experimental RNomics by helping to rationally devise new experiments.

In conclusion, we believe that our algorithms represent an important step towards reliable prediction of structural ncRNAs. We also consider the application of these algorithms to the human genome as an important contribution to its functional annotation. Although the full implications of these resuls are not yet clear, there is no doubt that they have opened a new perspective for both computational and experimental RNomics. Our results indeed challenge these fields but, at the same time, promise them a bright future.

References

1. Abbott A. **Competition boosts bid to find human genes.** *Nature*, 2005. **435**:134.
2. Accardo MC, Giordano E, Riccardo S, Digilio FA, Iazzetti G, Calogero RA, and Furia M. **A computational search for box C/D snoRNA genes in the drosophila melanogaster genome.** *Bioinformatics*, 2004. **20**:3293–301.
3. Aizerman M. **Theoretical foundations of the potential function method in pattern recognition learning.** *Automation and Remote Control*, 1964. **25**:821–837.
4. Altschul SF and Erickson BW. **Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage.** *Mol Biol Evol*, 1985. **2**:526–538.
5. Avner P and Heard E. **X-chromosome inactivation: counting, choice and initiation.** *Nat Rev Genet*, 2001. **2**:59–67.
6. Axmann I, Kensche P, Vogel J, Kohl S, Herzel H, and Hess W. **Identification of cyanobacterial non-coding RNAs by comparative genome analysis**. *Genome Biol*, 2005. **6**:R73.
7. Bachellerie JP, Cavaillé J, and Hüttenhofer A. **The expanding snoRNA world**. *Biochimie*, 2002. **84**:775–790.
8. Bailey S, Wichitwechkarn J, Johnson D, Reilly B, Anderson D, and Bodley J. **Phylogenetic analysis and secondary structure of the Bacillus subtilis bacteriophage RNA required for DNA packaging**. *J Biol Chem*, 1990. **265**:22365–22370.
9. Bejerano G, Haussler D, and Blanchette M. **Into the heart of darkness: large-scale clustering of human non-coding DNA.** *Bioinformatics*, 2004. **20 Suppl 1**:I40–I48.
10. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, and Haussler D. **Ultra-conserved elements in the human genome.** *Science*, 2004. **304**:1321–5.
11. Bendtsen JD, Nielsen H, von Heijne G, and Brunak S. **Improved prediction of signal peptides: Signalp 3.0.** *J Mol Biol*, 2004. **340**:783–95.
12. Bennasser Y, Le S, Yeung M, and Jeang K. **HIV-1 encoded candidate micro-RNAs and their cellular targets**. *Retrovirology*, 2004. **1**:43–43.
13. Bennett KP and Campbell C. **Support vector machines: Hype or hallelujah?** *SIGKDD Explorations*, 2000. **2**:1–13.
14. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, and Cuppen E. **Phylogenetic shadowing and computational identification of human microRNA genes**. *Cell*, 2005. **120**:21–24.
15. Blencowe BJ. **Transcription: surprising role for an elusive small nuclear RNA**. *Curr Biol*, 2002. **12**:R147–R149.
16. Bonnal S, Schaeffer C, Creancier L, Clamens S, Moine H, Prats AC, and Vagner S. **A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons**. *J Biol Chem*, 2003. **278**:39330–6.

17. Bonnet E, Wuyts J, Rouze P, and Van De Peer Y. **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics*, 2004. **in press**.
18. Boser B, Guyon I, and Vapnik V. **A training algorithm for optimal margin classifiers**. In **Fifth Annual Workshop on Computational Learning Theory**. ACM Press, Pittsburgh, 1992 .
19. Brown JR and Sanseau P. **A computational view of microRNAs and their targets.** *Drug Discov Today*, 2005. **10**:595–601.
20. Buratti E and Baralle FE. **Influence of RNA secondary structure on the pre-mRNA splicing process.** *Mol Cell Biol*, 2004. **24**:10505–14.
21. Burges CJ. **A tutorial on support vector machines for pattern recognition**. *Data mining and knowledge discovery*, 1998. **2**:121–167.
22. Byvatov E and Schneider G. **Support vector machine applications in bioinformatics.** *Appl Bioinformatics*, 2003. **2**:67–77.
23. C elegans Sequencing Consortium. **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science*, 1998. **282**:2012–2018.
24. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. **The transcriptional landscape of the mammalian genome.** *Science*, 2005. **309**:1559–63.
25. Carter RJ, Dubchak I, and Holbrook SR. **A computational approach to identify genes for functional RNAs in genomic sequences.** *Nucleic Acids Res*, 2001. **29**:3928–3938.
26. Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brennan CI, Horsthemke B, Bachellerie JP, and Hüttenhofer A. **Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization**. *Proc Natl Acad Sci USA*, 2000. **97**:14311–14316.
27. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al. **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas.** *Cell*, 2004. **116**:499–509.
28. Chang CC and Lin CJ. **LIBSVM: a library for support vector machines**, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
29. Chen JH, Le SY, Shapiro B, Currey KM, and Maizel Jr JV. **A computational procedure for assessing the significance of RNA secondary structure.** *Comput Appl Biosci*, 1990. **6**:7–18.
30. Chen JL and Greider CW. **An emerging consensus for telomerase RNA structure**. *Proc Natl Acad Sci USA*, 2004. **101**:14683–14684.
31. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science*, 2005. **308**:1149–1154.
32. Childs JL, Poole AW, and Turner DH. **Inhibition of Escherichia coli RNase P by oligonucleotide directed misfolding of RNA**. *RNA*, 2003. **9**:1437–1445.
33. Chiu DK and Kolodziejczak T. **Inferring consensus structure from nucleic acid sequences**. *CABIOS*, 1991. **7**:347–352.

34. Chooniedass-Kothari S, Emberley E, Hamedani MK, Troup S, Wang X, Czosnek A, Hube F, Mutawe M, Watson PH, and Leygue E. **The steroid receptor RNA activator is the first functional RNA encoding a protein.** *FEBS Lett*, 2004. **566**:43–7.

35. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, and Johnston M. **Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res*, 2001. **11**:1175–1186.

36. Clote P, Ferre F, Kranakis E, and Krizanc D. **Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency.** *RNA*, 2005. **11**:578–91.

37. Collins LJ, Macke TJ, and Penny D. **Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif**. *J Integ Bioinf*, 2004. **6**:15.

38. Collins LJ, Moulton V, and Penny D. **Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP**. *J Mol Evol*, 2000. **51**:194–204.

39. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, and Sidow A. **Characterization of evolutionary rates and constraints in three mammalian genomes.** *Genome Res*, 2004. **14**:539–48.

40. Coventry A, Kleitman DJ, and Berger B. **MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure.** *Proc Natl Acad Sci U S A*, 2004. **101**:12102–12107.

41. Cristianini N and Shawe-Taylor J. **An Introduction to Support Vector Machines**. Cambridge University Press, 2000.

42. Cullen BR. **Transcription and processing of human microRNA precursors.** *Mol Cell*, 2004. **16**:861–5.

43. Dahlberg JE and Lund E. **The genes and transcription of the major small nuclear RNAs**. In ML Birnstiel, editor, **Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles**, pages 38–70. Springer-Verlag, Berlin, 1988.

44. Dandjinou AT, Lévesque N, Larose S, Lucier JF, Elela SA, and Wellinger RJ. **A phylogenetically based secondary structure for the yeast telomerase RNA**. *Curr Biol*, 2004. **14**:1148–1158.

45. Delihas N. **Annotation and evolutionary relationships of a small regulatory rna gene micf and its target ompf in yersinia species**. *BMC Microbiol*, 2003. **3**:13–13.

46. Deng W, Zhu X, Skogerbo G, Zhao Y, Fu Z, Wang Y, He H, Cai L, Sun H, Liu C, et al. **Organisation of the Caenorhabditis elegans small noncoding transcriptome: genomic features, biogenesis and expression**, 2005. Submitted.

47. Dennis PP, Omer A, and Lowe T. **A guided tour: small RNA function in archaea**. *Mol Microbiol*, 2001. **40**:509–519.

48. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, and Antonarakis SE. **Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs)**. *Science*, 2003. **302**:1033–1035.

49. di Bernardo D, Down T, and Hubbard T. **ddbRNA: detection of conserved secondary structures in multiple alignments.** *Bioinformatics*, 2003. **19**:1606–11.

50. Doudna JA. **Structural genomics of RNA.** *Nat Struct Biol*, 2000. **7 Suppl**:954–956.

51. Dowell RD and Eddy SR. **Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction.** *BMC Bioinformatics*, 2004. **5**:71.

52. Eddy SR and Durbin R. **RNA sequence analysis using covariance models.** *Nucleic Acids Res*, 1994. **22**:2079–88.

53. Edvardsson S, Gardner PP, Poole AM, Hendy MD, Penny D, and Moulton V. **A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction.** *Bioinformatics*, 2003. **19**:865–873.

54. Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress AWM, and von Haeseler A. **How old is the genetic code? statistical geometry of tRNA provides an answer**. *Science*, 1989. **244**:673–679.

55. ENCODE consortium. **The encode (ENCyclopedia Of DNA Elements) project.** *Science*, 2004. **306**:636–40.

56. Enright CA, Maxwell ES, Elicieri GL, and Sollner-Webb B. **5'ETS rRNA processing facilitated by by four small RNAs: U14, E3, U17, and U3**. *RNA*, 1996. **2**:1094–1099.

57. Fagegaltier D, Lescure A, Walczak R, Carbon P, and Krol A. **Structural analysis of new local features in SECIS RNA hairpins**. *Nucl Acids Res*, 2000. **28**:2679–2689.

58. Ferreira MG and Miller JP Kyle Mand Cooper. **Indecent exposure: When telomeres become uncapped**. *Mol Cell*, 2004. **13**:7–18.

59. Frenkel FE, Chaley MB, Korotkov EV, and Skryabin KG. **Evolution of tRNA-like sequences and genome variability**. *Gene*, 2004. **335**:57–71.

60. Frith MC, Pheasant M, and Mattick JS. **The amazing complexity of the human transcriptome.** *Eur J Hum Genet*, 2005. **13**:894–7.

61. Gardner PP and Giegerich R. **A comprehensive comparison of comparative RNA structure prediction approaches**. *BMC Bioinformatic*, 2004. **5**:140.

62. Gardner PP, Wilm A, and Washietl S. **A benchmark of multiple sequence alignment programs upon structural RNAs.** *Nucleic Acids Res*, 2005. **33**:2433–9.

63. Gaudin C, Zhou X, Williams KP, and Felden B. **Two-piece tmRNA in cyanobacteria and its structural analysis**. *Nucl Acids Res*, 2002. **30**:2018–2024.

64. Gautheret D and Lambert A. **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.** *J Mol Biol*, 2001. **313**:1003–11.

65. Glazov EA, Pheasant M, McGraw EA, Bejerano G, and Mattick JS. **Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing.** *Genome Res*, 2005. **15**:800–8.

66. Gonzalez IL and Sylvester JE. **Human rDNA: Evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes**. *Genomics*, 2001. **73**:255–263.

67. Gott JM and Emeson RB. **Functions and mechanisms of RNA editing**. *Annu Rev Genet*, 2000. **34**:499–531.

68. Gräf S, Strothmann D, Kurtz S, and Steger G. **HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns**. *Nucl Acids Res*, 2001. **29**:196–198.

69. Griffiths-Jones S. **The microRNA registry.** *Nucleic Acids Res*, 2004. **32**:D109–11.

70. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A. **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res*, 2005. **33**:D121–D124.

71. Gruener W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, and Schuster P. **Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks**. *Monath Chem*, 1996. **127**:355–374.

72. Gruener W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, and Schuster P. **Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering**. *Monath Chem*, 1996. **127**:375–389.

73. Guo P. **Structure and function of phi29 hexameric RNA that drives the viral DNA packaging motor: review**. *Prog Nucleic Acid Res Mol Biol*, 2002. **72**:415–472.

74. Gutell RR, Power A, Hertz GZ, Putz EJ, and Stormo GD. **Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods**. *Nucl Acids Res*, 1992. **20**:5785–5795.

75. Haebel PW, Gutmann S, and Ban N. **Dial tm for rescue: tmRNA engages ribosomes stalled on defective mRNAs**. *Curr Op Struct Biol*, 2004. **14**:58–65.

76. Hannon GJ. **RNA interference**. *Nature*, 2002. **418**:244–251.

77. Hardison RC. **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends Genet*, 2000. **16**:369–72.

78. Hardison RC. **Comparative genomics.** *PLoS Biol*, 2003. **1**:E58.

79. Harris RJ and Elder D. **Ribozyme relationships: The hammerhead, hepatitis delta, and hairpin ribozymes have a common origin**. *J Mol Evol*, 2000. **51**:182–184.

80. Hartmann E and Hartmann RK. **The enigma of ribonuclease P evolution**. *Trends Genet*, 2003. **19**:561–569.

81. He L and Hannon GJ. **MicroRNAs: small RNAs with a big role in gene regulation.** *Nat Rev Genet*, 2004. **5**:522–31.

82. Hentze MW and Kühn LC. **Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress**. *Proc Natl Acad Sci USA*, 1996. **93**:8175–8182.

83. Hershberg R, Altuvia S, and Margalit H. **A survey of small RNA-encoding genes in Escherichia coli**. *Nucl Acids Res*, 2003. **31**:1813–1820.

84. Hofacker I and Stadler PF. **RNA secondary structures**, 2005. Unpublished.

85. Hofacker IL, Bernhart SH, and Stadler PF. **Alignment of RNA base pairing probability matrices.** *Bioinformatics*, 2004. **20**:2222–2227.

86. Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, and Stadler PF. **Automatic detection of conserved RNA structure elements in complete RNA virus genomes**. *Nucl Acids Res*, 1998. **26**:3825–3836.

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

88. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, and Schuster P. **Fast folding and comparison of RNA secondary structures**. *Monatsh Chem*, 1994. **125**:167–188.

89. Hofacker IL, Priwitzer B, and Stadler PF. **Prediction of locally stable RNA secondary structures for genome-wide surveys.** *Bioinformatics*, 2004. **20**:186–190.

90. Hofacker IL and Stadler PF. **Automatic detection of conserved base pairing patterns in RNA virus genomes**. *Comp & Chem*, 1999. **23**:401–414.

91. Hofacker IL, Stocsits R, and Stadler PF. **Conserved RNA secondary structures in viral genomes: A survey**. *Bioinformatics*, 2004. **20**:1495–1499.

92. Holland PWH, Garcia-Fernández J, Williams NA, and Sidow A. **Gene duplication and the origins of vertebrate development**. *Development*, 1994. **(Suppl.)**:125–133.

93. Holmes I. **Accelerated probabilistic inference of RNA structure evolution.** *BMC Bioinformatics*, 2005. **6**:73.

94. Hong M and Simpson L. **Genomic organization of Trypanosoma brucei kinetoplast DNA minicircles**. *Prostist*, 2003. **154**:265–279.

95. Hsu WC, Huang HD, Hsu SD, Lin LZ, Tsou AP, Tseng CP, Stadler PF, Washietl S, and Hofacker IL. **miRNAMap: Genomic maps for microRNA genes and their target genes in mammalian genomes**. *Nucl Acids Res*, 2005. Submitted.

96. Hüttenhofer A, Cavaille J, and Bachellerie JP. **Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms.** *Methods Mol Biol*, 2004. **265**:409–28.

97. Hüttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, and J B. **RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse**. *EMBO J*, 2001. **20**:2943–2953.

98. Huynen MA, Stadler PF, and Fontana W. **Smoothness within ruggedness: the role of neutrality in adaptation.** *Proc Natl Acad Sci U S A*, 1996. **93**:397–401.

99. Imanishi T and *et al.* **Integrative annotation of 21,037 human genes validated by full-length cDNA clones**. *PLoS Biology*, 2004. **2**:0856–0875.

100. International Mouse Genome Sequencing Consortium. **Initial sequencing and comparative analysis of the mouse genome.** *Nature*, 2002. **420**:520–562.

101. Jacob Y, Seif E, Paquet PO, and Lang FB. **Loss of the mRNA-like region in mitochondrial tmRNAs of jakobids**. *RNA*, 2004. **10**:605–614.

102. Jády BE, Bertrand E, and Kiss T. **Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body specific localization signal**. *J Cell Biol*, 2004. **164**:647–652.

103. Jády BE and Kiss T. **A small nucleolar guide RNA functions both in 2'-O-methylation and pseudouridylation of U5 spliceosomal RNA**. *EMBO J*, 2001. **20**:541–551.

104. Jang SK, Krausslich HG, Nicklin MJ, Duke GM, Palmenberg AC, and Wimmer E. **A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation.** *J Virol*, 1988. **62**:2636–43.

105. John B, Enright AJ, Aravin A, Tuschl T, Sander C, and Marks DS. **Human microRNA targets.** *PLoS Biol*, 2004. **2**:e363.

106. Johnson JM, Edwards S, Shoemaker D, and Schadt EE. **Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments**. *Trends Genet*, 2005. **21**:93–102.

107. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, and Gingeras TR. **Examples of the complex architecture of the human transcriptome revealed by race and high-density tiling arrays.** *Genome Res*, 2005. **15**:987–97.

108. Karplus K, Barrett C, and Hughey R. **Hidden markov models for detecting remote protein homologies.** *Bioinformatics*, 1998. **14**:846–56.

109. Keenan RJ, Freyman DM, Stroud RM, and Walter P. **The signal recognition particle**. *Annu Rev Biochem*, 2001. **70**:755–775.

110. Keiler KC, Shapiro L, and Williams KP. **tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: a two-piece tmRNA functions in caulobacter**. *Proc Natl Acad Sci USA*, 2000. **97**:7778–7783.

111. Kelleher C, Teixeira MT, Förstemann K, and Lingner J. **Telomerase: biochemical considerations for enzyme and substrate**. *Trends Biochem Sci*, 2002. **27**:572–579.

112. Kellis M, Patterson N, Endrizzi M, Birren B, and Lander ES. **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature*, 2003. **423**:241–254.

113. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. **The human genome browser at UCSC.** *Genome Res*, 2002. **12**:996–1006.

114. Kiss T. **Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs**. *EMBO J*, 2001. **20**:3617–3622.

115. Klein RJ, Misulovin Z, and Eddy SR. **Noncoding RNA genes identified in AT-rich hyperthermophiles.** *Proc Natl Acad Sci U S A*, 2002. **99**:7542–7.

116. Knudsen B and Hein J. **Pfold: RNA secondary structure prediction using stochastic context-free grammars**. *Nucl Acids Res*, 2003. **31**:3423–3428.

117. Knuth DE. **Fundamental Algorithms**, volume 3 of *The Art of Computer Programming*, page 237. Addison-Wesley, Reading, Massachusetts, 1973.

118. Koper-Emde D. **Phylogenetische Heterogenität der 7S-RNAs von Eukaryonten.** Ph.D. thesis, University of Bochum, 2004.

119. Krol A. **Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis**. *Biochimie*, 2002. **84**:765–774.

120. Kwek KY, Murphy S, Furger A, Thomas B, O'Gorman W, Kimura H, Proudfoot NJ, and Akoulitchev A. **U1 snRNA associates with TFIIH and regulates transcriptional initiation**. *Nat Struct Biol*, 2002. **9**:800–805.

121. Lagos-Quintana M, Rauhut R, Lendeckel W, and Tuschl T. **Identification of novel genes coding for small expressed RNAs.** *Science*, 2001. **294**:853–858.

122. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, and Tuschl T. **New microRNAs from mouse and human**. *RNA*, 2003. **9**:175–179.

123. Lai EC, Tomancak P, Williams RW, and Rubin GM. **Computational identification of Drosophila microRNA genes**. *Genome Biol*, 2003. **4**:R42.

124. Landweber LF and Gilbert W. **Phylogenetic analysis of RNA editing: a primitive genetic phenomenon**. *Proc Natl Acad Sci USA*, 1994. **91**:918–921.

125. Laslett D, Canback B, and Andersson S. **BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences.** *Nucleic Acids Res*, 2002. **30**:3449–53.

126. Lau NC, Lim LP, Weinstein EG, and Bartel DP. **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science*, 2001. **294**:858–862.

127. Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, and Casari G. **In search of antisense.** *Trends Biochem Sci*, 2004. **29**:88–94.

128. Le SV, Chen JH, Currey KM, and Maizel Jr JV. **A program for predicting significant RNA secondary structures.** *Comput Appl Biosci*, 1988. **4**:153–159.

129. Le SY, Chen JH, Konings D, and Maizel Jr JV. **Discovering well-ordered folding patterns in nucleotide sequences.** *Bioinformatics*, 2003. **19**:354–361.

130. Le SY, Zhang K, and Maizel Jr JV. **RNA molecules with structure dependent functions are uniquely folded.** *Nucleic Acids Res*, 2002. **30**:3574–3582.

131. Legendre M, Lambert A, and Gautheret D. **Profile-based detection of microRNA precursors in animal genomes**. *Bioinformatics*, 2004. Epub ahead of print.

132. Li Y and Altman S. **In search of RNase P RNA from microbial genomes**. *RNA*, 2004. **10**:1533–1540.

133. Liang XH, Xu YX, and Michaeli S. **The spliced-leader associated RNA is a trypanosome-specific sn(o)RNA that has the potential to guide pseudouridine formation on SL RNA**. *RNA*, 2002. **8**:237–246.

134. Liao D, Pavelitz T, Kidd JR, Kidd KK, and Weiner AM. **Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion**. *EMBO J*, 1997. **16**:588–598.

135. Lilley DMJ. **The origins of RNA catalysis in ribozymes**. *Trends Biochem Sci*, 2003. **28**:495–501.

136. Lim LP, Glasner ME, Yekta S, Burge CB, and Bartel DP. **Vertebrate microRNA genes.** *Science*, 2003. **299**:1540.

137. Lin J, Ly H, Hussain A, Abraham M, Pearl S, Tzfati Y, Parslow TG, and Blackburn EH. **A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein**. *Proc Natl Acad Sci USA*, 2004. **101**:14713–14718.

138. Lingner J, Cooper JP, and Cech TR. **Telomerase and DNA end replication: no longer a lagging strand problem?** *Science*, 1995. **269**:1533–1534.

139. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, and Chen R. **NONCODE: an integrated knowledge database of non-coding RNAs**. *Nucl Acids Res*, 2005. **33**:D112–D115. Database issue.

140. Lowe TM and Eddy SR. **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res*, 1997. **25**:955–964.

141. Lowe TM and Eddy SR. **A computational screen for methylation guide snoRNAs in yeast.** *Science*, 1999. **283**:1168–1171.

142. Lu S and Cullen B. **Adenovirus VA1 noncoding RNA can inhibit small interfering RNA and MicroRNA biogenesis**. *J Virol*, 2004. **78**:12868–12876.

143. Lück R, Gräf S, and Steger G. **ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure.** *Nucl Acids Res*, 1999. **27**:4208–4217.

144. Lück R, Steger G, and Riesner D. **Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein**. *J Mol Biol*, 1996. **258**:813–826.

145. MacIntosh GC, Wilkerson C, and Green PJ. **Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs**. *Plant Physiol*, 2001. **127**:765–776.

146. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, and Sampath R. **RNAMotif, an RNA secondary structure definition and search algorithm.** *Nucleic Acids Res*, 2001. **29**:4724–35.

147. Maden BEH. **The numerous modified nucleotides in eukaryotic ribosomal RNA**. *Prog Nucl Acid Res Mol Biol*, 1990. **39**:241–303.

148. Margulies EH, Blanchette M, Haussler D, and Green ED. **Identification and characterization of multi-species conserved sequences.** *Genome Res*, 2003. **13**:2507–18.

149. Marker C, Zemann A, Terhorst T, Kiefmann M, Kastenmayer JP, Green P, Bachellerie JP, Brosius J, and Hüttenhofer A. **Experimental RNomics: identification of 140 candidates for small non-messenger rnas in the plant Arabidopsis thaliana.** *Curr Biol*, 2002. **12**:2002–13.

150. Martens JA, Laprade L, and Winston F. **Intergenic transcription is required to repress the Saccheromyces cerevisiae SER3 gene**. *Nature*, 2004. **429**:571–574.

151. Martineau Y, Le Bec C, Monbrun L, Allo V, Chiu IM, Danos O, Moine H, Prats H, and Prats AC. **Internal ribosome entry site structural motifs conserved among mammalian fibroblast growth factor 1 alternatively spliced mRNAs**. *Mol Cell Biol*, 2004. **24**:7622–35.

152. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, and Turner DH. **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure.** *Proc Natl Acad Sci U S A*, 2004. **101**:7287–92.

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

154. Mathews M. **Structure, function, and evolution of adenovirus virus-associated RNAs**. *Curr Top Microbiol Immunol*, 1995. **199 (Pt 2)**:173–187.

155. Mattick JS. **Non-coding RNAs: the architects of eukaryotic complexity.** *EMBO Rep*, 2001. **2**:986–91.

156. Mattick JS. **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms.** *Bioessays*, 2003. **25**:930–939.

157. Mattick JS. **RNA regulation: a new genetics?** *Nature Rev Genetics*, 2004. **5**:316–323.

158. McCaskill JS. **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**. *Biopolymers*, 1990. **29**:1105–1119.

159. McCutcheon JP and Eddy SR. **Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics.** *Nucleic Acids Res*, 2003. **31**:4119–4128.

160. Mignone F, Gissi C, Liuni S, and Pesole G. **Untranslated regions of mRNAs.** *Genome Biol*, 2002. **3**:REVIEWS0004.

161. Missal K, Rose D, and Stadler PF. **Non-coding RNAs in the urochordate Ciona intestinalis**. In **ECCB**. 2005 In press.

162. Missal K, Zhu X, Rose D, Deng W, Skogerbo G, Chen R, and Stadler PF. **Prediction of structured non-coding rnas in the genome of the nematode Caenorhabitis elegans**, 2005. Submitted.

163. Mitchell JR, Cheng J, and K C. **A box H/ACA small nucleolar RNA-like domain at the human telomerase 3'end**. *Mol Cell Biol*, 1999. **19**:567–576.

164. Montzka Wassarman K and Storz G. **6S RNA regulates E. coli RNA polymerase activity**. *Cell*, 2000. **101**:613–623.

165. Morrissey JP and Tollervey D. **Birth of the snoRNPs: the evolution of RNase MRP and and the eukaryotic pre-rRNA-processing system**. *Trends Biol Sci*, 1995. **20**:78–82.

166. Myslinski E, Krol A, and Carbon P. **Characterization of snRNA and snRNA-type genes in the pufferfish Fugu rubripes**. *Gene*, 2004. **330**:149–158.

167. Nitta I, Kamada Y, Noda H, Ueda T, and Watanabe K. **Reconstitution of peptide bond formation with Escherichia coli 23S ribosomal RNA domains**. *Science*, 1998. **281**:666–669.

168. Nussinov R, Piecznik G, Griggs JR, and Kleitman DJ. **Algorithms for loop matching**. *SIAM J Appl Math*, 1978. **35**:68–82.

169. O'Brien CA, Margelot K, and Wolin SL. **Xenopus Ro ribonucleoproteins: Members of an evolutionarily conserved class of cytoplasmic ribonucleoproteins**. *Proc Natl Acad Sci USA*, 1993. **90**:7250–7254.

170. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs**. *Nature*, 2002. **420**:563–573.

171. Olsen GJ and Woese CR. **Ribosomal RNA: A key to phylogeny**. *FASEB J*, 1993. **7**:113–123.

172. Omer A, Lowe T, Russel A, Ebhardt H, Eddy S, and Dennis P. **Homologs of small nucleolar RNAs in Archaea**. *Science*, 2000. **288**:517–522.

173. Omoto S, Ito M, Tsutsumi Y, Ichikawa Y, Okuyama H, Brisibe E, Saksena N, and Fujii Y. **HIV-1 nef suppression by virally encoded microRNA**. *Retrovirology*, 2004. **1**:44–44.

174. Pang KC, Stephen S, Engström PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, and Mattick JS. **RNAdb — comprehensive mammalian noncoding RNA database**. *Nucl Acids Res*, 2005. **33**:D125–D130.

175. Pasquinelli AE, Hunter S, and Bracht J. **MicroRNAs: a developing story.** *Curr Opin Genet Dev*, 2005. **15**:200–5.

176. Patel AA and Steitz JA. **Splicing double: insights from the second spliceosome**. *Nat Rev Mol Cell Biol*, 2003. **4**:960–970.

177. Pennisi E. **Gene counters struggle to get the right answer.** *Science*, 2003. **301**:1040–1.

178. Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, and Saccone C. **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. update 2002.** *Nucleic Acids Res*, 2002. **30**:335–40.

179. Pfeffer S, Zavolan M, Grässer F, Chien M, Russo J, Ju J, John B, Enright A, Marks D, Sander C, et al. **Identification of virus-encoded microRNAs**. *Science*, 2004. **304**:734–736.

180. Pirotta V. **Trans-splicing in drosophila**. *Bioessays*, 2002. **24**:988–991.

181. Pitulle C, Garcia-Paris M, Zamudio KR, and Pace NR. **Comparative structural analysis of vertebrate ribonuclease P RNA**. *Nucl Acids Res*, 1998. **26**:3333–3339.

182. Precott EM and Proudfoot NJ. **Transcriptional collision between convergent genes in budding yeast**. *Proc Natl Acad Sci USA*, 2002. **99**:8796–8801.

183. Ramakrishnan V and Moore PB. **Atomic structures at last: the ribosome in 2000**. *Curr Opinions Struct Biol*, 2001. **11**:144–154.

184. Regalia M, Rosenblad MA, and Samuelsson T. **Prediction of signal recognition particle RNA genes.** *Nucleic Acids Res*, 2002. **30**:3368–77.

185. Rivas E. **Evolutionary models for insertions and deletions in a probabilistic modeling framework.** *BMC Bioinformatics*, 2005. **6**:63.

186. Rivas E and Eddy SR. **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics*, 2000. **16**:583–605.

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

188. Rivas E, Klein RJ, Jones TA, and Eddy SR. **Computational identification of noncoding RNAs in E. coli by comparative genomics.** *Curr Biol*, 2001. **11**:1369–1373.

189. Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C, and Samuelsson T. **SRPDB: Signal recognition particle database.** *Nucleic Acids Res*, 2003. **31**:363–364.

190. Rosenblad MA, Zwieb C, and Samuelson T. **Identification and comparative analysis of components from the signal recognition particle in protozoa and fungi**. *BMC Genomics*, 2004. **5**:5.

191. Rosenblatt F. **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological Review*, 1959. **65**.

192. Rougeulle C and Heard E. **Antisense RNA in imprinting: spreading silence through Air.** *Trends Genet*, 2002. **18**:434–7.

193. Sankoff D. **Simultaneous solution of the RNA folding, alignment and protosequence problems**. *SIAM J Appl Math*, 1985. **45**:810–825.

194. Schattner P. **Searching for RNA genes using base-composition statistics.** *Nucleic Acids Res*, 2002. **30**:2076–82.

195. Scherer SW and *et al.* **Human chromosome 7: DNA sequence and biology**. *Science*, 2003. **300**:767–772.

196. Schölkopf B and Smola A. **Learning with kernels.** MIT Press, Cambridge, MA, 2002.

197. Schultes EA, Hraber PT, and LaBean TH. **Estimating the contributions of selection and self-organization in RNA secondary structure.** *J Mol Evol*, 1999. **49**:76–83.

198. Schümperli D and Pillai RS. **The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein**. *Cell Mol Life Sci*, 2004. **61**:2560–2570.

199. Schuster P, Fontana W, Stadler PF, and Hofacker IL. **From sequences to shapes and back: a case study in RNA secondary structures.** *Proc R Soc Lond B Biol Sci*, 1994. **255**:279–284.

200. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, and Miller W. **MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences.** *Nucleic Acids Res*, 2003. **31**:3518–3524.

201. Shabalina SA and Kondrashov AS. **Pattern of selective constraint in C. elegans and C. briggsae genomes.** *Genet Res*, 1999. **74**:23–30.

202. Shabalina SA, Ogurtsov AY, Kondrashov VA, and Kondrashov AS. **Selective constraint in intergenic regions of human and mouse genomes.** *Trends Genet*, 2001. **17**:373–6.

203. Sharkady SM and Williams KP. **A third lineage with two-piece tmRNA**. *Nucl Acids Res*, 2004. **32**:1–8.

204. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res*, 2005. **15**:1034–50.

205. Simpson L, Thiemann OH, Savill NJ, Alfonzo JD, and Maslov DA. **Evolution of RNA editing in trypanosome mitochondria**. *Proc Natl Acad Sci USA*, 2000. **97**:6986–6993.

206. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al. **The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics.** *PLoS Biol*, 2003. **1**:E45.

207. Steitz TA and Moore PB. **RNA, the first macromolecular catalyst: the ribosome is a ribozyme**. *Trends Biochem Sci*, 2003. **28**:411–418.

208. Storz G, Opdyke JA, and Zhang A. **Controlling mRNA stability and translation with small noncoding RNAs**. *Cur Op Microbiol*, 2004. **7**:140–144.

209. Tang G. **siRNA and miRNA: an insight into RISCs.** *Trends Biochem Sci*, 2005. **30**:106–14.

210. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, and Hüttenhofer A. **Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus**. *Proc Natl Acad Sci USA*, 2002. **99**:7536–7541.

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

212. Teunissen SW, Kruithof MJ, Farris AD, Harley JB, Venrooij WJ, and Pruijn GJ. **Conserved features of Y RNAs: a comparison of experimentally derived secondary structures**. *Nucl Acids Res*, 2000. **28**:610–619.

213. The Human Genome Sequencing Consortium. **Finishing the euchromatic sequence of the human genome**. *Nature*, 2004. **431**:931–945.

214. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, et al. **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature*, 2003. **424**:788–93.

215. Thompson JD, Higgins DG, and Gibson TJ. **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res*, 1994. **22**:4673–4680.

216. Thurner C, Witwer C, Hofacker I, and Stadler PF. **Conserved RNA secondary structures in Flaviviridae genomes**. *J Gen Virol*, 2004. **85**:1113–1124.

217. Tinoco I, Uhlenbeck OC, and Levine MD. **Estimation of secondary structure in ribonucleic acids.** *Nature*, 1971. **230**:362–7.

218. Tycowski KT, Aab A, and Steitz JA. **Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa**. *Curr Biol*, 2004. **14**:1985–1995.

219. Uliel S, Liang Xh, Unger R, and Michaeli S. **Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organization, and unique functions**. *Int J Parasit*, 2004. **34**:445–454.

220. Ushida C, Yoshida A, Miyakawa Y, Ara Y, and Muto A. **Distribution of the MCS4 RNA genes in mycoplasmas belonging to the Mycoplasma mycoides cluster**. *Gene*, 2003. **314**:149–155.

221. Valadkhan S and Manley JL. **Splicing-related catalysis by protein-free snRNAs**. *Nature*, 2001. **413**:701–707.

222. van Zon A, Mossink M, Schoester M, Scheffer G, Scheper R, Sonneveld P, and Wiemer E. **Multiple human vault RNAs. Expression and association with the vault complex**. *J Biol Chem*, 2001. **276**:37715–37721.

223. Vasu SK and Rome LH. **Dictyostelium vaults: Disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein**. *J Biol Chem*, 1995. **270**:16588–16594.

224. Vogel J, Bartels V, Tang TH, Churakov G, Slagter-Jäger JG, Hüttenhofer A, and E Wagner GH. **RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria**. *Nucl Acids Res*, 2003. **31**:6435–6443.

225. Wagner EGH and Flärdh K. **Antisense RNAs everywhere?** *Trends Genet*, 2002. **18**:223–226.

226. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, and Stadler PF. **Genome-wide mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in human.** *Nat Biotech*, 2005. In press.

227. Wassarman K, Repoila F, Rosenow C, Storz G, and Gottesman S. **Identification of novel small RNAs using comparative genomics and microarrays**. *Genes Dev*, 2001. **15**:1637–1651.

228. Waterman MS. **Secondary structure of single - stranded nucleic acids**. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press NY*, 1978. **1**:167 – 212.

229. Waterman MS and Smith TF. **RNA secondary structure: A complete mathematical analysis**. *Mathematical Biosciences*, 1978. **42**:257–266.

230. Weber MJ. **New human and mouse microrna genes found by homology search.** *FEBS J*, 2005. **272**:59–73.

231. Williams KP. **Descent of a split DNA**. *Nucl Acids Res*, 2002. **30**:2025–2030.

232. Witwer C, Rauscher S, Hofacker I, and Stadler P. **Conserved RNA secondary structures in picornaviridae genomes**. *Nucl Acids Res*, 2001. **29**:5079–5089.

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

234. Wuchty S, Fontana W, Hofacker IL, and Schuster P. **Complete suboptimal folding of RNA and the stability of secondary structures.** *Biopolymers*, 1999. **49**:145–65.

235. Xia T, , Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, and Turner DH. **Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson-crick base pairs.** *Biochemistry*, 1998. **37**:14719–35.

236. Yano Y, Saito R, Yoshida N, Yoshiki A, Wynshaw-Boris A, Tomita M, and Hirotsune S. **A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene.** *J Mol Med*, 2004. **82**:414–22.

237. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, et al. **Widespread occurrence of antisense transcription in the human genome**. *Nat Biotechnol*, 2003. **21**:379–386.

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

239. Zuker M. **On finding all suboptimal foldings of an RNA molecule.** *Science*, 1989. **244**:48–52.

240. Zuker M. **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res*, 2003. **31**:3406–15.

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

Danksagung

Diese Arbeit entstand im Zeitraum von Juli 2003 bis September 2005 am Institut für Theoretische Chemie unter der Anleitung von Prof. Peter F. Stadler. Ihm gilt mein Dank für die hervorragende Betreuung während dieser Zeit — insbesondere dafür, dass er mir stets die Freiheit gab, meine eigenen Ideen zu verfolgen, er gleichzeitig aber *immer* erreichbar war und mir die notwendige Unterstützung bot.

Weiters möchte ich Dr. Ivo Hofacker danken, der ebenfalls jederzeit bereit war, mir bei meinen Fragen weiterzuhelfen. Prof. Peter Schuster gilt mein Dank für die freundliche Aufnahme an seinem Institut.

Schließlich möchte ich allen weiteren Kollegen danken, die zum Gelingen dieser Arbeit beigetragen haben: Stephan Bernhart, Lukas Endler, Christoph Flamm, Ulrike Langhammer, Rainer Machné, Ulrike Mückstein, Andreas Svrcek-Seiler, Hakim Tafer, Andrea Tanzer, Michael Wolfinger, Claudia Fried, Guido Fritzsch, Jörg Lehmann, Kristin Missal, Axel Mosig, Bettina Müller, Sonja Prohaska, Bärbel Stadler, Christine Witwer, Paul Gardner, Andreas Wilm, Melanie Lukasser, Alexander Hüttenhofer.

Lebenslauf

Persönliche Daten

Stefan Washietl

Geb. am 4. November 1978 in Stockerau

Österreichischer Staatsbürger, ledig

Schule

1985–1989	Volksschule Stockerau
1989–1997	Bundesgymnasium Stockerau
06/1997	Matura, mit Auszeichnung

Studium

10/1997	Beginn des Studiums der Biologie, Studienzweig Genetik an der Universität Wien
06/2000	1. Diplomprüfung Biologie, mit Auszeichnung
05/2002–6/2003	Diplomarbeit am Forschungsinstitut für molekulare Pathologie Wien in der Arbeitsgruppe von Dr. Frank Eisenhaber. Thema: Sequenzanalyse und Evolution von Proteinen.
02/2003	2. Diplomprüfung Biologie, mit Auszeichnung. Sponsion zum Mag.rer.nat.
07/2003–	Dissertation am Institut für theoretische Chemie, Universität Wien bei Prof. Peter F. Stadler.

Sonstiges

10/1998–10/1999	Zivildienst beim Österreichischen Roten Kreuz

Publikationen

1. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A. and Stadler PF. **Genome-wide mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in human.** *Nat Biotech*, 2005. In press.
2. Gardner PP, Wilm A, and Washietl S. **A benchmark of multiple sequence alignment programs upon structural RNAs.** *Nucleic Acids Res*, 2005. **33**:2433–9.
3. Bompfünewerer A, Flamm C, Fried C, Fritzsch G, Hofacker I, Lehmann J, Missal K, Mosig A, Müller B, Prohaska S, Stadler BMR, Stadler PF, Tanzer A, Washietl S and Witwer C. **Evolutionary patterns of non-coding RNAs**. *Theor Biosci*, 2005. **123**:301–369.
4. Washietl S, Hofacker IL, and Stadler PF. **Fast and reliable prediction of noncoding RNAs**. *Proc Natl Acad Sci USA*, 2005. **102**:2454–2459.
5. Washietl S and Hofacker IL. **Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics.** *J Mol Biol*, 2004. **342**:19–30.
6. Washietl S and Eisenhaber F. **Reannotation of the CELO genome characterizes a set of previously unassigned open reading frames and points to novel modes of host interaction in avian adenoviruses.** *BMC Bioinformatics*, 2003. **4**:55.
7. Maurer-Stroh S, Washietl S, and Eisenhaber F. **Protein prenyltransferases: anchor size, pseudogenes and parasites.** *Biol Chem*, 2003. **384**:977–89.
8. Maurer-Stroh S, Washietl S, and Eisenhaber F. **Protein prenyltransferases.** *Genome Biol*, 2003. **4**:212.