RNAs Everywhere: Genome-Wide Annotation of Structured RNAs

Athanasius F. Bompfünewerer [a,b], Rolf Backofen [c], Stephan H. Bernhart [a] Christoph Flamm [a], Claudia Fried [d], Guido Fritzsch [h], Jörg Hackermüller [i], Jana Hertel [h], Ivo L. Hofacker [a], Kristin Missal [d], Axel Mosig [e,f], Sonja J. Prohaska [g], Dominic Rose [d], Peter F. Stadler [d,h,a,j,*], Andrea Tanzer [h,a], Stefan Washietl [a], Sebastian Will [c]

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

[b] *Zentralfriedhof Wien, 3. Tor Simmeringer Haupstraße, A-1110 Wien, Austria*

[c] *Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee, Geb. 106, D-79110 Freiburg, Germany*

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

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

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

[g] *Biomedical Informatics, Arizona State University, Tempe, PO-Box 878809, AZ 85287, USA*

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

[i] *Fraunhofer Institut für Zelltherapie und Immunologie — IZI Deutscher Platz 5e, D-04103 Leipzig, Germany*

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

Abstract

Starting with the discovery of microRNAs and the advent of genome-wide transcriptomics, non-protein-coding transcripts have moved from a fringe topic to a central field research in molecular biology. In this contribution we review the state of the art of "computational RNomics", i.e., the bioinformatics approaches to genome-wide RNA annotation. Instead of rehashing results from recently published surveys in detail, we focus here on *the* open problem in the field, namely (functional) annotation of the plethora of putative RNAs. A series of exploratory studies are used to provide non-trivial examples for the discussion of some of the difficulties.

Key words: Noncoding RNA, RNA Secondary Structure, Genome Annotation

1 Introduction

A series of recent studies of the mammalian transcriptome have dramatically changed our perception of genome organization. Experimental studies using a variety of different techniques, from tiling arrays (Bertone *et al.*, 2004; Kampa *et al.*, 2004; Johnson *et al.*, 2005; Cheng *et al.*, 2005), to cDNA sequencing (Okazaki *et al.*, 2002; Imanishi and *et al.*, 2004; Carninci *et al.*, 2005; Ravasi *et al.*, 2006), and unbiased mapping of transcription factor binding sites (Cawley *et al.*, 2004) agree that a substantial fraction of the genome is transcribed and that non-protein-coding RNAs (ncRNAs), Tab. 1, are the dominating component of the transcriptome. It remains unclear, however, to what extent these ncRNAs are functional; alternatively they might be "transcriptional noise"(Hüttenhofer *et al.*, 2005) or they could be the by-product of transcriptional activity that takes place in order to regulate the gene expression at adjacent loci. As shown in (Ravasi *et al.*, 2006), however, many of the non-coding cDNA clones are "derived from genuine transcripts of unknown function whose expression appears to be regulated".

Non-coding RNAs form a very heterogeneous group of transcripts: Besides the well characterized "ancient" classes (such as the spliceosomal RNAs and tRNAs), the function of several pol-III transcripts remains unknown. Vault RNAs (Mossink *et al.*, 2003; van Zon *et al.*, 2003) seem to play a critical role in multi drug resistance (Gopinath *et al.*, 2005) and Y RNA (Maraia *et al.*, 1994; Farris *et al.*, 1999; Stein *et al.*, 2005) control activity of RNA chaperones as Ro60 and La (Belisova *et al.*, 2005; Stein *et al.*, 2005)

Several ncRNAs exhibit more or less strong similarity to retro-elements. In mammals, SINEs are derived from tRNAs and 7SL RNAs and LINEs from tRNAs (Deininger and Batzer, 2002; Kramerov and Vassetzky, 2005). Both are able to serve as source for new ncRNAs, as shown for a set of microRNAs (Smalheiser and Torvik, 2005; Tanzer *et al.*, 2005) as well as 4.5SI RNA (Gogolevskaya *et al.*, 2005) and 4.5SH RNA (Gogolevskaya and Kramerov, 2002) in rodents. Interestingly, the ncRNAs are derived from the long terminal repeats (LTRs) of LINEs, not from their protein coding regions. The small RNA generating loci in *Arabidopsis* follow a similar principle: inverted duplication of target genes leads to new miRNAs (Allen *et al.*, 2004a).

Table 1
Functional RNAs in eukaryotes.

ancient RNAs	rRNAs, tRNAs, SRP RNA, RNase P
repeat associated	miRNAs, rasiRNAs, 4.5SH RNA, 4.5SI RNA, LINEs, SINEs
mRNA-like	H19, AncR-1, Ntab, U87HG, BIC, Evf-1
mRNA-like associated	miRNAs, snoRNAs
Pol-III transcripts	snRNA, vRNA, Y RNA, tRNAs, MRP, U6, H1, 7SK, 7SL
small RNAs	miRNAs, siRNAs, rasiRNAs

Genes annotated as e.g. "Putative ORF" are good candidates for so-called mRNA-like-ncRNAs (mlncRNAs). These transcripts are processed just as normal mRNAs, but carry only very small ORFs or no ORFs at all. Transcriptional control, (Berteaux *et al.*, 2004, 2005; Carninci *et al.*, 2005), tissue specific differential expression (French *et al.*, 2001), alternative splicing and polyadenylation (Sawata *et al.*, 2004) of mlncRNAs does not seem to differ from those of protein coding polymerase II products. Some of them remain in the nucleus (Sawata *et al.*, 2004). If amino-acid sequences are predicted for such transcripts they are usually not conserved within a genus (Inagaki *et al.*, 2005). Several miRNAs (Rodriguez *et al.*, 2004; Baskerville and Bartel, 2005) and snoRNAs (Pelczar and Filipowicz, 1998; Makarova and Kramerov, 2005) reside in introns and even exons of mlncRNAs. A few examples of functional ncRNAs that changed their host genes have been reported, see e.g. (Rodriguez *et al.*, 2004; Bompfünewerer *et al.*, 2005). Only a hand full of mlncRNAs are annotated in databases as in Y2K (Erdmann *et al.*, 1999, 2000) or RNAdb (Pang *et al.*, 2005), while most cDNAs that lack a CDS remain functionally unassigned (Carninci *et al.*, 2005).

Small ncRNAs of only about 20nt in length seem to serve as the exchangeable RNA module in protein complexes allowing them to bind DNA and RNA in a sequence specific way. MicroRNAs, one of the most prominent classes of ncRNA, are found in plants (Jones-Rhoades *et al.*, 2006), animals (Berezikov and Plasterk, 2005) and play a fundamental role in virus infections (Sullivan and Ganem, 2005). They differ slightly from siRNAs (Du and Zamore, 2005). High expression of repeat associated small RNAs (rasiRNAs) was detected during embryogenesis of *D. melanogaster* (Aravin *et al.*, 2003) and later also in *Danio rerio* (Chen *et al.*, 2005b).

Recently, genome wide surveys for non-coding RNAs have provided evidence for tens of thousands of previously undescribed evolutionary conserved RNAs with distinctive secondary structures (Washietl *et al.*, 2005a; Pedersen *et al.*, 2006). The conservation of structure indicates that the molecule functions (also) as an RNA. Taken together, both the experimental and computational data provide strong evidence that ncRNAs are an important, functional component of the mammalian transcriptome. The elucidation of these functions, however, remains elusive in almost all cases.

In this contribution we discuss the currently available techniques for finding structured RNAs and we focus in particular on current approaches towards their annotation. To this end we discuss here a series of pilot studies and partial results.

2 Computational ncRNA Detection

Large, highly conserved ncRNAs, in particular ribosomal RNAs, can easily be found using `blast` (Altschul *et al.*, 1990). Similarly, `blast` can be used to find orthologous ncRNAs in closely related species, e.g. (Tanzer and Stadler, 2004; Weber, 2005). 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.

Specialized programs have been developed to detect members of particular ncRNA families. Examples of this approach include `miRseeker` for microRNAs (Lai *et al.*, 2003), `BRUCE` for tmRNAs (Laslett *et al.*, 2002), `tRNAscan` for tRNAs (Lowe and Eddy, 1997), `snoScan` (Lowe and Eddy, 1999) and `SNO.pl` (Fedorov *et al.*, 2005) for box C/D snoRNAs, `fisher` (Edvardsson *et al.*, 2003) and `snoGPS` (Schattner *et al.*, 2004) for box H/ACA snoRNAs, as well as a heuristic for SRP RNAs (Regalia *et al.*, 2002; Rosenblad *et al.*, 2004).

MicroRNAs in plants can be found by extracting those hairpin structures that contain sequence motifs complementary to a mRNA, which is then a putative target (Jones-Roades and Bartel, 2004; Bonnet *et al.*, 2004; Adai *et al.*, 2005). In animals, on the other hand, the situation is more complicated since miRNAs do not bind with perfect complementarity to their target. A large array of different approaches, summarized in Table 2, has recently been developed to detect microRNAs, among them our own tool `RNAmicro` that has been designed specifically to analyze large-scale comparative genomics data (Hertel and Stadler, 2006).

A wide variety of different approaches to perform homology searches based on both sequence and structure have been proposed in the last few years in order to utilize the strong conservation of secondary structure in many ncRNA families, see (Bompfünewerer *et al.*, 2005) for a recent more extensive summary of this topic.

Table 2. Overview of published miRNA detection methods.

Approach	miRNA only	sequence	structure	homology	mach.learning	web tool	download	Ref.	Remark
`miRseeker`	Y	Y	Y	Y	n	n	n	(Lai *et al.*, 2003)	drosophilids only
`MiRscan`	Y	Y	Y	Y	n	Y	n	(Lim *et al.*, 2003a,b)	scores hairpins
`ERPIN`	n	Y	Y	Y	—	Y	Y	(Gautheret and Lambert, 2001; Legendre *et al.*, 2005)	scores structure profiles
`HARVESTER`	Y	Y	Y	Y	n	Y	Y	(Dezulian *et al.*, 2006)	plants only
`ProMiR`	Y	Y	Y	Y	HMM	n	n	(Nam *et al.*, 2005)	
`PalGrade`	Y	n	Y	n	n	n	n	(Bentwich *et al.*, 2005)	scores hairpins
`mir-abela`	Y	Y	Y	n*	SVM	Y	n	(Sewer *et al.*, 2005)	scores hairpins, clusters
`Vmir`	Y	n	Y	n	—	n	n	(Grundhoff *et al.*, 2006)	scores hairpins, clusters
`---`	Y	Y	Y	Y	n	n	n	(Berezikov *et al.*, 2005)	Phylogenetic Shadowing
`BayesMIRfinder`	Y	Y	Y	Y	NBS	Y	n	(Yousef *et al.*, 2006)	NBS first, then comparative
`RNAz+RNAmicro`	*	Y	Y	Y	SVM	n	Y	(Washietl *et al.*, 2005b; Hertel and Stadler, 2006)	investigates alignments

HMM: Hidden Markov Model, SVM: Support Vector Machine, NBS: Naive Bayses Score

Stochastic context free grammars (SCFGs) can be used to construct covariance models from a multiple alignment with structural annotation as in `infernal` (Eddy, 2002). The consensus model can then be used to search for homologs. Similar in spirit, `ERPIN` also uses multiple structure-annotated alignments as input to construct a descriptor for homology search. `Rsearch` (Klein and Eddy, 2003) is a local alignment algorithm which considers structural and sequence constraints. It uses both single nucleotide and base pair substitution matrices to define alignment scores. `FastR` (Bafna and Zhang, 2004) combines a pairwise alignment algorithm with a filtering step to improve performance. It uses both single nucleotide and base pair substitution matrices to define alignment scores. Backofen and Will (2004) introduced an efficient local sequence-structure alignment method based on predicted structures. Beside sequence-local motifs (i.e., motifs that consist of a subsequence in each molecule), it is able to find also structure-local motifs, i.e. motifs that are connected substructures such as a helix without the connecting hairpin loop.

Simple description languages have been proposed to allow users to define combined sequence/structure for genome-wide searches. Such approaches are implemented e.g. in `RNAmot` (Gautheret *et al.*, 1990) and Sean Eddy's `rnabob`. Hybrid languages, like `HyPaL` (Gräf *et al.*, 2001) or the language used in `RNAMotif` (Macke *et al.*, 2001), connect pattern languages with user defined approximate rules, which rank the results according to their distance to the motif.

A number of large-scale surveys have been performed using one or more of the general purpose tools mentioned above, including a microRNA survey using `ERPIN` (Legendre *et al.*, 2005), a search for U5 snRNA and RNase P using `RNAmotif` (Collins *et al.*, 2004), and a survey of RNase P RNAs in bacterial genomes (Li and Altman, 2004).

Attempts to predict novel functional RNAs are in general based on predicted secondary structures. However, since most RNA sequences will form extensive structures, the problem of distinguishing incidental from functional structures is non-trivial. It was first suggested by Maizel and co-workers that functional RNA elements should have a secondary structure that is energetically more stable than expected by chance (Le *et al.*, 1988). 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 (Rivas and Eddy, 2000).

Therefore, all current approaches to *de novo* prediction of structured RNAs work comparatively, requiring two or more related sequences as input, typically in the form of a multiple sequence alignment. The first reasonably successful attempt to predict structured RNAs from sequence alignments was `qrna` (Rivas and Eddy, 2001). This program compares the score of three distinct models

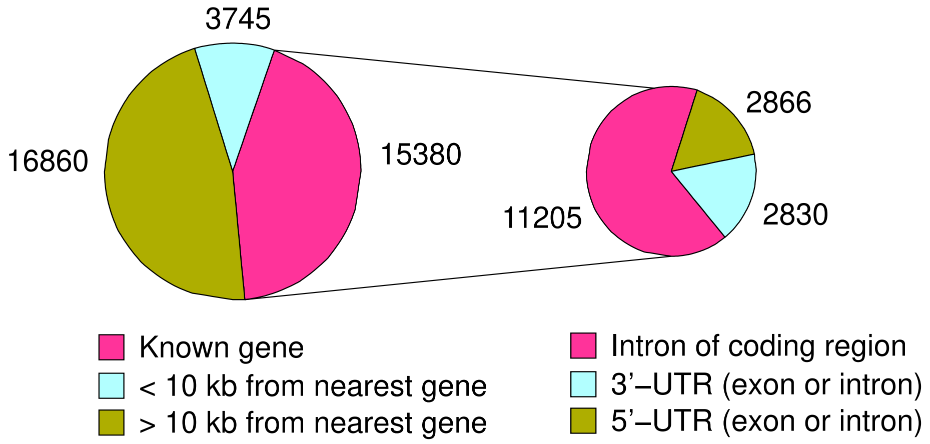

Fig. 1. Summary of a comparative screen of vertebrate genomes, which evaluated conserved genomic DNA sequences for signatures of structural conservation of base pairing patterns and exceptional thermodynamic stability using the `RNAz` program. (Adapted from (Washietl *et al.*, 2005a)). About half of the structured RNA motifs are found far away from known coding regions, the other half is located within known protein-coding genes. Two thirds of the latter motifs are intronic, one sixth each is located in the UTRs of the mRNAs.

of sequence evolution to decide which one describes best the given alignment: a pair stochastic context free grammar (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* (Rivas *et al.*, 2001; McCutcheon and Eddy, 2003), some of which could be verified experimentally. `EvoFold` (Pedersen *et al.*, 2006) is essentially an extension of the `qrna` approach to multiple sequence alignments. The program combines SCFGs for RNA structure modeling with phylogenetic models that describe the substitution process along the branches of a tree.

The `RNAz` algorithm, in contrast, is based on thermodynamic RNA folding (Washietl *et al.*, 2005b). It uses two independent criteria for classification: a z-score measuring thermodynamic stability of individual sequences, and a *structure conservation index* obtained by comparing folding energies of the individual sequences with the predicted consensus folding. The two criteria are combined by a support vector machine that detects conserved and stable RNA secondary structures with high sensitivity and specificity. Other recent programs for detecting conserved RNA secondary structures include `ddbRNA` (di Bernardo *et al.*, 2003) and `MSARi` (Coventry *et al.*, 2004).

Both `RNAz` and `Evofold` have been applied to surveying the human genome providing evidence for tens of thousands of genomic loci with signatures of evolutionarily conserved secondary structure (Washietl *et al.*, 2005b; Pedersen *et al.*, 2006). Further `RNAz` surveys have been conducted for urochordates (Missal *et al.*, 2005), nematodes (Missal *et al.*, 2006), and yeasts (Steigele *et al.*, 2006). These investigations have produced extensive lists of candidates for functional RNAs without using (or providing) information on membership in a particular class of RNAs, see figure 1.

Approaches based on pairwise or multiple sequence alignments are of course limited by the quality of the input alignment. In regions with sequence similarity below some 50-60%, sequence alignments will in general not be structurally correct, making accurate prediction of consensus structures impossible (Washietl and Hofacker, 2004). This problem can in principle be overcome by computing structural alignments, albeit at significantly higher computational cost. Most recently, Uzilov *et al.* (2006) presented a classification method based on an updated version of `dynalign` (Mathews and Turner, 2002), a restricted variant of the Sankoff-algorithm for the simultaneous computation of alignment and consensus structure (Sankoff, 1985). Using a similar approach based on their `foldalign` variant of the Sankoff algorithm, Torarinsson *et al.* (2006) screened a significant fraction of the non-alignable DNA that could be identified as homologous between man and mouse by virtue of alignable flanking sequences. These authors reported several thousands regions without significant sequence conservation that show evidence for a conserved RNA secondary structure.

A major limitation of these Sankoff based algorithms is their enormous computational cost. The computations required to evaluate the ≈100,000 genomic regions as described in (Torarinsson *et al.*, 2006) took 5 months on 70 CPUs with 2 gigabytes of RAM. Also Uzilov *et al.* (2006) estimate that it would take several months on a similar sized computing cluster to screen all human/mouse regions with pairwise identity below 50% using `dynalign`. Although such approaches are feasible given sufficient computational resources, the CPU requirements render them impracticable for extended analysis tasks. Moreover, since both `dynalign` and `foldalign` perform pairwise alignments only, one would have to screen many different pairwise combinations (e.g. human/mouse and human/rat).

All other methods like `qrna`, `RNAz` and `EvoFold` that rely on a given input alignment are of the same algorithmic time complexity (essentially cubic with alignment length). Although effective run time of the different programs may vary considerably, all three programs seem to be fast enough to allow routine analysis of even large mammalian genomes.

3 Limitations of Sequence Alignments: An RNAz Screen of Stramenopiles

To-date, genomic screens for non-coding RNAs have been applied mostly to fairly closely related organisms, e.g. vertebrates (with a focus on mammals), rhabditid nematodes, or ascidians. In principle, however, the applicability of `RNAz` is limited only by the quality of the input alignments, so that highly conserved structures from distant organisms might still be detectable.

As an example of an `RNAz` screen of phylogenetically very distant organisms we summarize a survey of the three currently available stramenopile sequences. Heterokonts, or stramenopiles, form a major clade within the eukaryote kingdom chromista, see e.g. (Yoon *et al.*, 2002). Most are algae, ranging from the giant multicellular kelp to the unicellular diatoms. However some are colorless and superficially resemble fungi. Three complete genomes have been sequenced: data are available for two closely related oomycetes *Phytophthora sojae* [1], *Phytophthora ramorum* [2] (Gajendran *et al.*, 2006), and the diatom *Thalassiosira pseudonana* [3] (Armbrust *et al.*, 2004). Our protocol closely follows the approach taken in (Missal *et al.*, 2005, 2006).

In the first step an annotation track for *P. sojae* was constructed by mapping the available mRNA and protein sequences back to the genome using `blat`. Using `blast` with a $E < 10^{-10}$, all non-protein-coding loci were compared to the entire *P. ramorum* genome. We combine `blast` alignments that are separated not more than 30nt provided they pass several consistency checks detailed in (Missal *et al.*, 2005, 2006). This leaves 149375 conserved loci with an average length of 195nt. Since the two phytophthora sequences are too similar, we estimate an unacceptably high estimated false discovery rate for the pairwise `RNAz` screen. We therefore compare these loci to the much more distant diatom *T. pseudonana* and obtain 903 homologous non-coding loci with an average length of about 80nt. We re-aligned the blast-hits using `clustalw` (Thompson *et al.*, 1994) and screened them using `RNAz` with window-length 120nt in steps of 50nt (for alignments longer than 120nt). For some loci more than one alignment is found. These are combined if possible; otherwise only the alignment with the largest `RNAz` p-score is retained so that each genomic locus is covered by at most one RNA prediction. Details of this procedure are given in (Missal *et al.*, 2006). In order to estimate the false positive rate and the false discovery rate, the `RNAz` screen was repeated with shuffled alignments as described in (Washietl *et al.*, 2005a; Missal *et al.*, 2006). The results are summarized in Tab. 3.

[1] `http://genome.jgi-psf.org/sojae1/sojae1.home.html`
[2] `http://genome.jgi-psf.org/ramorum1/ramorum1.home.html`
[3] `http://genome.jgi-psf.org/thaps1/thaps1.home.html`

Table 3
Summary statistics of the RNAz screen of stramenopiles.

Threshold	$p > 0.50$	$p > 0.90$	$p > 0.98$	$p > 0.99$
Specificity per test	0.9861	0.9949	0.9985	0.9996
candidate alignments	115	60	42	35
randomized	37	14	4	1
False discovery rate	32 %	23 %	10 %	3 %
Distinct loci (*P. sojae*)	44	17	12	11

The 115 RNAz slices that are classified with $p_{\texttt{RNAz}} > 0.5$ map to only 44 distinct loci in the *P. sojae* genome. 20 of these can be identified as tRNA genes. A comparison with the updated annotation at the JGI *P. sojae* site shows that the remaining loci map to protein-coding regions. Given the data in Tab. 3 we expect a substantial false discovery rate. Furthermore, there is growing evidence for evolutionarily conserved secondary structure also within the coding parts of mRNAs (Steigele *et al.*, 2006; Meyer and Miklós, 2005), so that some of these signals could well be real.

Due to the high degree of sequence divergence most of the known ncRNAs do not lead to significant alignments of sufficient length between *P. sojae* and *T. pseudonana*. This set includes about 60 loci in the *P. sojae* genome that can be identified by comparison with the noncode database. Among them are 13 U2, 30 U4, 1 U5, 1 U6 snRNA, and 1 SRP RNA. In addition tRNAscanSE predicts 235 tRNA loci.

The low sensitivity of the screen on this data set highlights the limitations of approaches that are based on sequence alignments. With genome sizes of 33-87Mb, using a structure-based approach (e.g. dynalign or foldalign) require excessive computational resources. As more sequenced genomes become available, however, the scope of sequence-alignment based methods expands for two reasons: (1) The specificity of methods such as RNAz increases dramatically with the number of aligned sequences. (2) Additional genomes in a suitable evolutionary distance from the currently available ones can give very good results already from pairwise comparisons as demonstrated in the case of Ascidians (Missal *et al.*, 2005) and Nematodes (Missal *et al.*, 2006).

4 The Importance of Being Local

A comprehensive understanding of structured RNAs requires the analysis not only of ncRNAs with an often globally conserved structure, but also of local

RNA motifs in larger molecules. Examples of the latter class are IRES (internal ribosome entry sites), SECIS (selenocystein insertion elements), or the *Rho*-independent termination signal in *E.coli*.

From a computational genomics point of view, there is actually little difference between these two classes of RNA structures. In a large-scale screen of genomic sequence, the transcript structure is typically unknown. As a consequence, both ncRNAs and structural mRNA motifs appear as local features in the genomic input sequence. The ability to compute *locally stable* secondary structures is thus a necessary prerequisite for any genome wide analysis of structured RNA for both computational and biological reasons: (i) Long-range base pairs in large transcripts are disfavored kinetically (Flamm *et al.*, 2000) and predicted poorly (Doshi *et al.*, 2004) relative to short-range pairs. (ii) Global approaches to RNA folding are limited to sequence length ≤ 20000 on most hardware because of memory consumption. (iii) In general, the exact boundaries of the transcript are unknown, so that global folds cannot add to the accuracy of the structure prediction relative to folding individual sequence windows.

Local folds can trivially be obtained by folding sub-sequences of length L in a window sliding along the genomic sequence nucleotide by nucleotide. In practice, however, the sequence windows have to be shifted by a substantial fraction of L in order to keep the CPU requirements manageable. It is well known, however, that the predicted structures depend strongly on the flanking context, i.e., on the exact window position. In fact, a recent algorithm for microRNA detection is based upon the idea to consider the stability of secondary structure against changes in the immediate environment (Sewer *et al.*, 2005). A large step size for the window implies poor sampling of the plausible local structures, hence small step sizes are important for accuracy.

Combining a global folding algorithm with a sliding window is also problematic in the context of ncRNA detection using tools such as `RNAz`: Large window sizes are preferable in order to detect larger ncRNAs, but may actually be detrimental for detecting small RNA structures, since the flanking regions interfere with the signal from the small structured RNA.

Two modifications of the global RNA folding algorithm have been developed to address this problem. `RNALfold` computes local minimum free energy structures with base pairs spanning no more than L bases in $\mathcal{O}(N \times L^2)$ time (Hofacker *et al.*, 2004b). This is equivalent to folding all windows of size L, while saving a factor L in CPU time compared to the naive approach. A partition function variant with the same time complexity, `RNAplfold`, computes the probability of a base pair (i, j) occurring in the structural ensemble, averaged over all sequence windows with a given size W (Bernhart *et al.*, 2006). To get robust statistics, the size of the averaging window W should be chosen somewhat larger than the maximum span L, resulting in an algorithm with complexity $\mathcal{O}(N \times W^2)$, see Fig. 2.

Both `RNALfold` and `RNAplfold` are true "scanning algorithms"; requiring only $\mathcal{O}(N + L^2)$ memory, and are therefore suitable for genome-wide surveys. Together with a z-score for the energy of a sequence window, which could be cheaply computed in the course of the algorithm (Washietl *et al.*, 2005b), `RNAplfold` may be used as a first crude indicator whether stable RNA secondary structures can be expected in a given part of the genome.

In principle, it is straight-forward to generalize the prediction of consensus structures from aligned sequences from global structures to local structures. `RNAalifold` (Hofacker *et al.*, 2002) computes the most stable structure that is common to a collection of aligned sequences. Algorithmically, only the energy model changes: one simply has to add up the contributions of aligned sequence intervals instead of evaluating a single sequence. This modification can easily be implemented also in the scanning programs described in the previous paragraph. The resulting `RNALalifold` program will be available with the next release of the `Vienna RNA` package. An example for the partition function case is given in Fig. 2, where the conservation of the miRNA precursor structures, as opposed to any other structural features present is shown.

`Foldalign` (Hull Havgaard *et al.*, 2005; Torarinsson *et al.*, 2006) implements a scanning local alignment algorithm. More precisely, this version of the Sankoff algorithm mutually scans two sequences of arbitrary length for common local structures with a maximum motif length. While the restriction to local motifs speeds up the algorithm, it is still computationally demanding.

In principle, also the SCFG based algorithms `qrna` and `EvoFold` predict local secondary structures. However, both are not implemented as true "scanning algorithms", and thus still require a sliding window approach.

5 Who is Who? — Approaches towards RNA Annotation

5.1 The Problem

With the exception of a small number of evolutionarily very well conserved RNAs (in particular rRNAs, tRNAs (Lowe and Eddy, 1997), the U5 snRNA (Collins *et al.*, 2004), RNAse P and MRP (Piccinelli *et al.*, 2005)), most ncRNAs are not only hard to discover *de novo* in large genomes, but they are also surprisingly hard to recognize if presented without annotation. While **homologs** of known sequences can often be reliably recognized already by

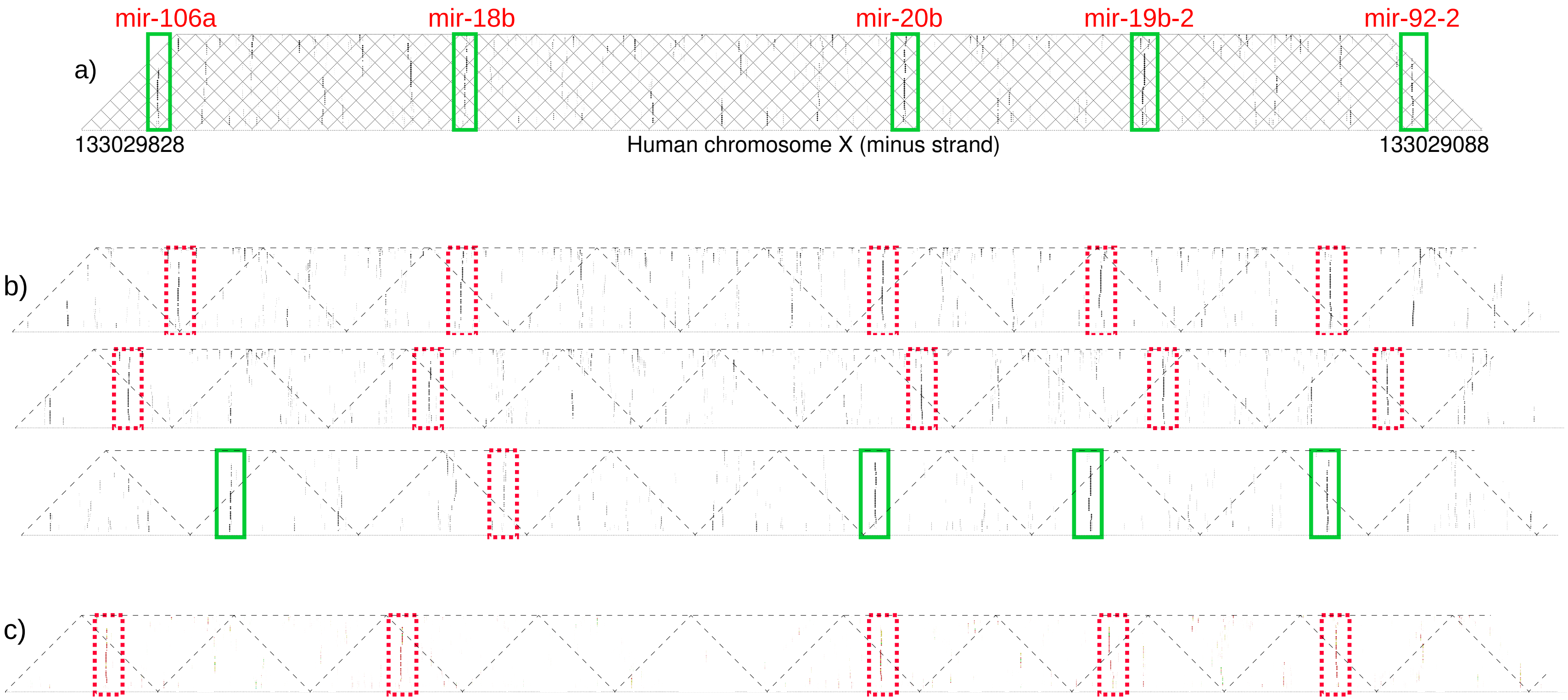

Fig. 2. a) Local pair probabilities of a human miRNA cluster
b) homologs in dog (upper), opossum (middle) and mouse (lower) to the miRNA cluster in a). MiRNAs annotated in miRBase are outlined in full lines, putative miRNAs not annotated in dashed lines. Note that every putative miRNA almost perfectly aligns to the human counterpart (with one mutation at most).
Dog, opossum: $W = L = 100$, Mouse: $W = 150$, max. base pair span $L = 100$. The base pairs on the upper edge of the mouse plot (long range base pairs) are less probable than in the other three because of the more robust statistics.
c) `RNALalifold` partition function of an alignment of the homologous miRNA clusters in a) and b). The noise is considerably reduced as opposed to the single folds shown above.

blast or at least with the help of Rfam alignments and infernal or ERPIN, determining class membership of novel examples is a much harder problem.

Given an alignment not more than a few hundred nucleotides in length that is known to contain a conserved secondary structure, it should be very easy to decide whether these sequences belong to a known class of ncRNAs or not. Conceptually, this is a very simple classification task that should be solvable efficiently by most machine learning techniques.

In the case of non-coding RNAs, however, machine learning approaches severely suffer from the very limited amount of available positive training data and the fact that negative training data are almost never known at all. Even for the most benign case, microRNA precursors, there is only a few hundred independent known examples, namely the miRNA families listed in the mir-base (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2005; Hertel *et al.*, 2006). Overtraining is thus a serious problem. As a consequence, it is necessary to restrict oneself to a small set of descriptors. These constraints, however, make the choice of the descriptors a crucial task.

5.2 Which Direction?

A relatively simple example for such a classification task is the problem of strand prediction: Large parts of the RNA energy model, in particular the stacking energies for Watson-Crick pairs, are symmetric when forming the reverse complement of a sequence and it's structure. Asymmetry is introduced in particular by GU pairs that map to a non-canonical AC in the reverse complement. Nevertheless, plus and minus strand of a sequence often exhibit similar folding energies. In computational screens for ncRNAs one will usually look at both the forward and reverse version of any given alignment, and often a significant signal for a structural RNA is detected on both strands. RNAz so far simply estimated the reading direction as the one that achieved a higher classification probability for "structured RNA". This method is quite inaccurate, however, in particular when the differences in classification probability are small.

An efficient strand detector to be used in conjunction with RNAz can be constructed from only six descriptors which moreover are already computed by RNAz: The difference of the SCI value for plus and minus strand, the difference of the RNAalifold consensus energy, the difference in mean folding energies, and the difference in mean z-score. In addition, the average sequence conservation and the length of the alignment is used. A support vector machine can then predict the correct strand with over 96% accuracy. This method is implemented also as a stand-alone tool RNAstrand (Missal and Stadler, 2006)

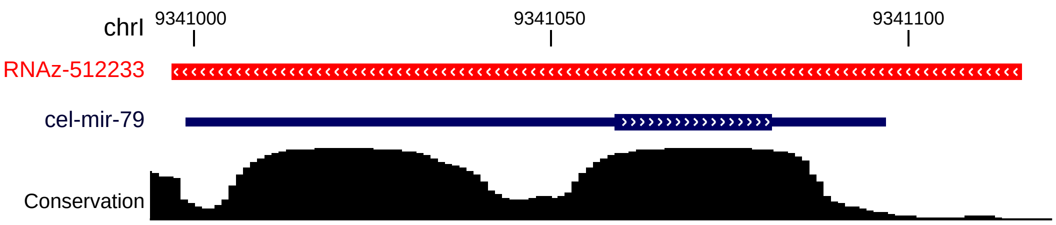

Fig. 3. The `RNAz` prediction *Ce-512233* (Missal *et al.*, 2006) coincides with *pre-mir-79*. The `RNAz` prediction favours the minus strand (top). The correct reading direction is on the plus-strand, however. `RNAstrand` computes a score of $D = -0.82$ for the `RNAz` hit, indicating the that the direction predicted by `RNAz` is incorrect. The `RNAstrand` classification coincides here with the correct reading direction of the microRNA precursor.

that can be used to re-evaluate earlier ncRNA screens. Fig. 3 shows one such example.

5.3 Family Membership: H/ACA-Box snoRNAs

In order to assign predicted ncRNAs to a particular ncRNA family, it seems natural to include structural descriptors in the classification procedure. RNA structure prediction, however, is less than perfect even when co-variation information from an alignment can be used (Hofacker *et al.*, 2002). This is true in particular when the exact ends of structured sequences within the multiple sequence alignment are not known. Furthermore, most ncRNAs can tolerate deviations from the "typical" structure without loss of function. The microRNA precursor structure may for example contain small branching helices, instead of forming a single stem-loop. These limitations restrict the usefulness of structure description languages in particular when one is interested in ncRNAs that are not members of one of the few well-known families.

Thus, structural descriptors have to be sufficiently fuzzy to allow for imperfect structure prediction and structural variation. The `RNAmicro` program for example uses 12 descriptors, only two of which are derived from the structure, namely the length of the stem and hairpin loop region of the miRNA stem-loop structure. Four descriptors measure sequence conservation in the loop and stem regions (the loop tends to be very variable, while the mature miRNA is highly conserved), another five descriptors measure the thermodynamic stability, and one measures sequence composition. This approach was quite successful, see e.g. figure 4 for an example.

It seems thus natural to extend this approach to other classes of ncRNAs. During the work on `RNAmicro` we observed that H/ACA-box snoRNAs, which

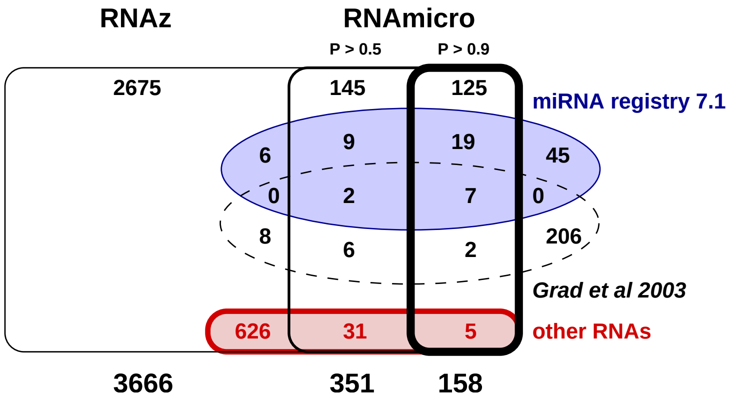

Fig. 4. `RNAmicro` annotation (Hertel and Stadler, 2006) of a `RNAz` survey of nematode genomes (Missal *et al.*, 2006). About half of the known *C. elegans* microRNAs are not conserved in *C. briggsae* and are hence not detected by comparative genomics. (Adapted from (Hertel and Stadler, 2006)).

also form hairpin-like structures, formed a particularly resilient group of false positives. This suggests to use the same set of descriptors and simply train the system with multiple sequence alignments of H/ACA-box snoRNAs as positive training set, while a sample of randomized hairpins, microRNAs, as well as known stem-loop structures from other ncRNA classes are used a negative training set. This yields a sensitivity of only about 63% at a specificity of about 75%. This suggests to include a small number of additional descriptors that are geared towards specific structural properties of box H/ACA snoRNAs discussed e.g. in (Henras *et al.*, 2004).

The `snowReport` classificator uses 9 descriptors, among them the same quantities for assessing folding thermodynamics as in `RNAmicro`: energy z-score, structure conservation index, average folding energy of the individual aligned sequences, ratio of folding energy and GC content. The stem-loop structures of snoRNAs are significantly shorter than those of miRNAs. Thus we include the number of stacked pairs and the length of the hairpin loop. Furthermore, H/ACA box snoRNAs have a single large interior loop which is (nearly) symmetric. We hence add the average symmetry (absolute value of the length difference between the 3' and 5' unpaired stretches of all interior loops) as well as the length of the longest interior loop as additional descriptors. We use `libsvm 2.8.2` with the same settings as `RNAmicro`, `RNAz` and `RNAstrand`.

In contrast to the microRNAs, a sufficiently large set of snoRNA alignments is not available, albeit there are several examples in the `Rfam` (Griffiths-Jones *et al.*, 2005) and `snoRNA-LBME-db` (Lestrade and Weber, 2006) databases. We thus searched all available vertebrate genomes for homologs on the known human H/ACA-box snoRNAs following the protocol for microRNA homology search used in (Hertel *et al.*, 2006). This yields 395 alignments containing 2 to 18 sequences per alignment.

Table 4
Application of `snowReport` on sea squirts, nematodes, and vertebrates. Numbers to distinct loci in the genomes of *C. intestinalis*, *C. elegans*, and *H. sapiens*, resp. The false positives in urochordates and nematodes are tRNAs and rRNAs, which were excluded from the training data.

candidates	urochordata	nematoda	vertebrata
`snowReport`	1553	1833	5519
true negatives	203	310	3
distance constraint	14	255	17
H-box	14	204	12
known HACA snoRNA	13	10	65
`snowReport`	9	9	24

Using half of this set as positive training set and half of the microRNA alignments reported by Hertel *et al.* (2006) as negative training set resulted in a sensitivity of 81% and a specificity of 87%. Afterwards, the test data was used for retraining the SVM and the training data for testing it. This resulted in a sensitivity of 91% and a specificity of 81%. For application to previously reported `RNAz` screens we used the full sets of positive and negative examples for retraining. Unfortunately, `snowReport` misclassifies a large number of tRNAs and 5S rRNAs as snoRNAs. Closer inspections shows that in these cases only structures are recognized. Since the false positives appear to restricted almost exclusively to these well known ncRNAs they do not present a serious problem since they are reliably identified by sequence homology or `tRNAscanSE`.

Metazoan box H/ACA usually are composed of two stem-loop structures (Henras *et al.*, 2004); we hence classify only those `RNAz` hits as putative box H/ACA snoRNAs in which (1) two hairpins separated by not more than 20nt are classified positively by `snowReport`, and which (2) contain the H-box motif. Results are summarized in Tab. 4. The most interesting observation of this preliminary screen is the large number of plausible candidates in nematodes in contrast to both the urochordate and mammalian data. It is interesting to note in this context that the a recent experimental screen by Deng *et al.* (2006) identified dozens of putative snoRNAs in *C. elegans*.

In contrast to `snoGPS` (Schattner *et al.*, 2004), we do not rely on the existence of a known or suspected target site in an rRNA or snRNA. Our approach thus predicts a few plausible candidates of “orphan” snoRNAs, i.e., snoRNAs within unknown modification target site.

In contrast to `RNAmicro`, the SVM-based classification of box H/ACA snoRNAs was only moderately successful. The most significant problem appears to be the generally low quality of the predicted consensus structures, which seems to be at least in part a consequence of problems in the underlying sequence alignments. Reliable methods for structure-based or structure-assisted multiple sequence alignments are thus a necessary pre-requisite for the successful application of structural descriptors in automatic ncRNA annotation. Although several approaches exist, reviewed e.g. in a comparison of techniques for consensus structure prediction by (Gardner and Giegerich, 2004), their suitability for the purpose of ncRNA annotation has not yet been studied systematically.

5.4 snRNA-like Candidates

A recent experimental survey of *C. elegans* genome (Deng *et al.*, 2006) identified a class of snRNA-like ncRNAs that are characterized by a recognizable SMN-binding site. We have therefore re-analyzed the results of the `RNAz` screen of urochordates (Missal *et al.*, 2005) to identify potential SMN binding sites in these structured RNA candidates.

We use `RNAbob` to search for the sequence motif `AUUUYUS` followed by a hairpin of rather variable stem and loop length. This pattern is a common generalization of the SMN binding sites in the known *Ciona intestinalis* snRNAs. In our analysis we require that the pattern occurs in aligned positions of the *Ciona intestinalis* and *Ciona savignyi* ncRNA candidates. This procedure recovers many of the known snRNAs that we found by the `RNAz` screen and in additions identifies 28 plausible candidates (as well as five copies of tRNA-Ile and one probable protein coding transcript). One example is described in some detail in Fig. 5.

6 New Kids on the Block

6.1 Sequence-based Clusters

The simplest approach to identifying multi-gene families is `blastclust`. A re-investigation of the urochordate `RNAz` screen (Missal *et al.*, 2005) shows that about a third of the candidates have at least one related sequence in the candidate set, Tab. 5. As one would expect, individual tRNA and snRNA families are identified by this approach. In addition, however, we find three very large families of candidates. They do not show significant homology out-

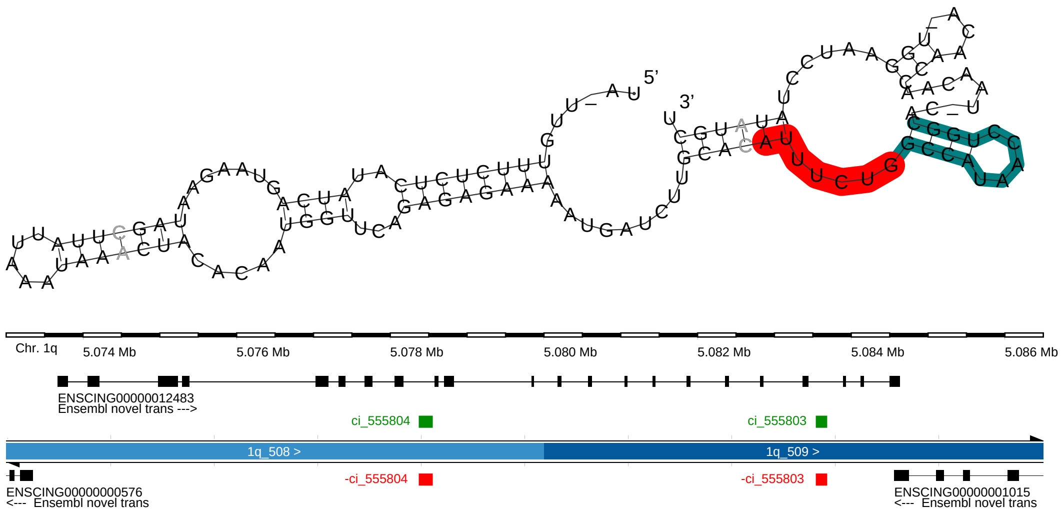

Fig. 5. Secondary structure (top) and genomic location (below) of a putative snRNA-like RNA in *Ciona intestinalis*. The `RNAz` predictions 555803 and 555804 are located within two introns of a ENSEMBL gene, which match a single locus in the *Ciona savignyi* genome It is the reverse complement of these two sequences, however, which contain the putative SMN binding site, which is highlighted in the secondary structure.
Trimming the alignment to the three distinct sequences, two from *C. intestinalis* and a single one from *C. savignyi* so that only the well-conserved region is retained and re-scoring with `RNAz` yields $p_+ = 0.709774$ and $p_- = 0.961678$. `RNAstrand` returns a decision value of $p = -0.999999$, i.e., an unambiguous vote for the negative strand.

side the Ascidians and they are not associated with a known or predicted family of protein coding genes. The cluster members are not uniformly distributed across the genome but appear concentrated at a few genomic loci. This pattern is reminiscent of many groups of vertebrate ncRNAs, including in particular tRNAs and snRNAs, which appear in multiple, often genomically clustered, copies. Lineage specific examples of functional repeat-derived ncRNAs include e.g. the mouse B2-element (Allen *et al.*, 2004b). Since the `RNAz` classification values are very high for most members of these classes, we speculate that these groups contain functional ncRNAs that are associated with an ascidian-specific repeat family.

A similar pattern was observed in *C. elegans* (Missal *et al.*, 2006). With slightly different `blastclust` setting, 148 clusters containing a total of 916 `RNAz` signals as well as 2756 non-clustered sequences were found. In contrast to the urochordate data, however, all large clusters could be annotated. It is not clear at this point whether this difference is biologically meaningful, or whether sequences with high copy numbers have been excluded more effectively from the nematode screen as a consequence of a more complete exclusions of repetitive

Table 5
Sequence-based clustering of *Ciona intestinalis* ncRNA candidates.
Here we used `blastclust` requiring a sequence overlap of $\geq$ 50%, 80% identity in the overlap region, and word size of 20, i.e., much less stringent settings than the defaults. Numbers in parentheses refer to sequence families for which consensus sequences are provided in the electronic supplement.

size	Annotation	size	Annotation	size	Annotation
197	(1)	13	tRNA Arg:CCT	8	tRNA Thr:AGT
160	(2)	13	(16)	8	tRNA Val:AAC
144	(3)	12	tRNA Gly:GCC	8	
32	5S RNA	11	tRNA Gly:TCC	8	
26	tRNA Ile:YAT	11	(19)		
22	(6)	11	(20)	**size**	**frequency**
20	(7)	9		7	7
18	tRNA Pro:HGG	9	U5	6	1
17	(9)	9		5	10
17	(10)	9		4	13
14	tRNA Leu:WAG	9		3	22
13	U3	8	tRNA Ala:WGC/Ser:GCT	2	83
13	tRNA Arg:ACG	8		1	2065
13	tRNA Leu:TAA	8	tRNA Asn:GTT		

DNA. Not surprisingly, no large sequence-based clusters were found in the mammalian screen (Washietl *et al.*, 2005a) since in this case the input alignment were already devoid of multi-copy genes including tRNAs and snRNAs.

6.2 Structure-Based Clustering

A more general approach to assign ncRNAs to families is based purely on structural similarity. Given a set of predicted ncRNAs one may use a structural alignment method to compute all pairwise alignments, and subsequently cluster all ncRNAs by similarity. In principle this should allow not only to assign predicted ncRNAs to known families, but even to define complete new ncRNA families. For the pairwise alignment step one would ideally use a variant of the Sankoff algorithm which simultaneously computes sequence alignment and consensus structure, but is computationally expensive (Sankoff, 1985). Per-

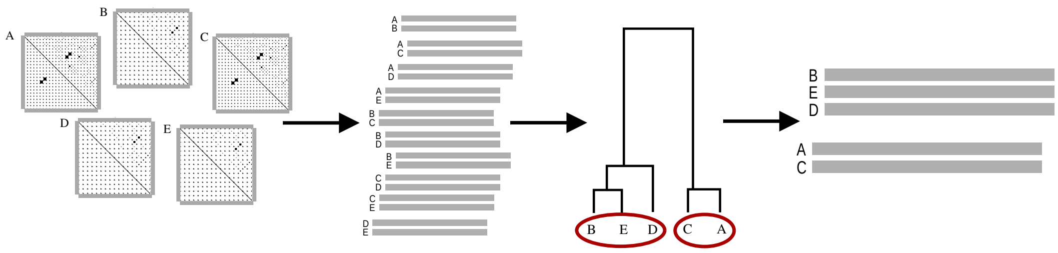

Fig. 6. Pipeline for clustering a set of ncRNAs A,B,C,D, and E. Starting from the RNAs with pair probability matrices, compute all pairwise alignments, determine clusters, and multiply align the RNAs of each cluster.

forming structural alignments for all pairs of ncRNA candidates in a set of several ten thousand is therefore still problematic. Moreover, most existing implementations can use only two sequences (no profile alignments) and compute global instead of local similarity. A local variant is described in (Hull Havgaard *et al.*, 2005).

The goal of annotation tools that classify family membership in results of other surveys is different from the direct search for RNA family members in genomic data. In the latter case one is interested in a "short list" of candidates that contains as few false positives as possible (e.g., for use in experimental verification). In post-processing data such as those from `RNAz` we are interested in a more balanced trade-off between sensitivity and specificity similar to that of annotating protein motifs in known predicted protein coding genes.

For the purpose of a structural clustering of ncRNA candidates we suggest a pipeline consisting of the following three major steps:

- generate all pairwise local sequence/structure alignments
- based on this information, hierarchically cluster the ncRNAs using WPGMA (or any other suited hierarchical clustering method) into a tree,
- finally, extract relevant clusters and construct multiple alignments of the ncRNA candidates in each cluster

Recent developments in pairwise sequence-structure alignment allow us to get very close to the ideal of using Sankoff's algorithm and in the same time increase the efficiency dramatically. Hofacker *et al.* (2004a) proposed a (global) scoring scheme that is based on all base pair probabilities (in the structure ensemble) of the two RNAs. Such probabilities can be reasonably predicted using McCaskill's pair probability algorithm (McCaskill, 1990). Since probabilities reflect thermodynamical properties of the RNAs, the new scoring scheme factors in thermodynamics without the need of computing a full energy model during alignment. It turns out that this idea can be used to design an even

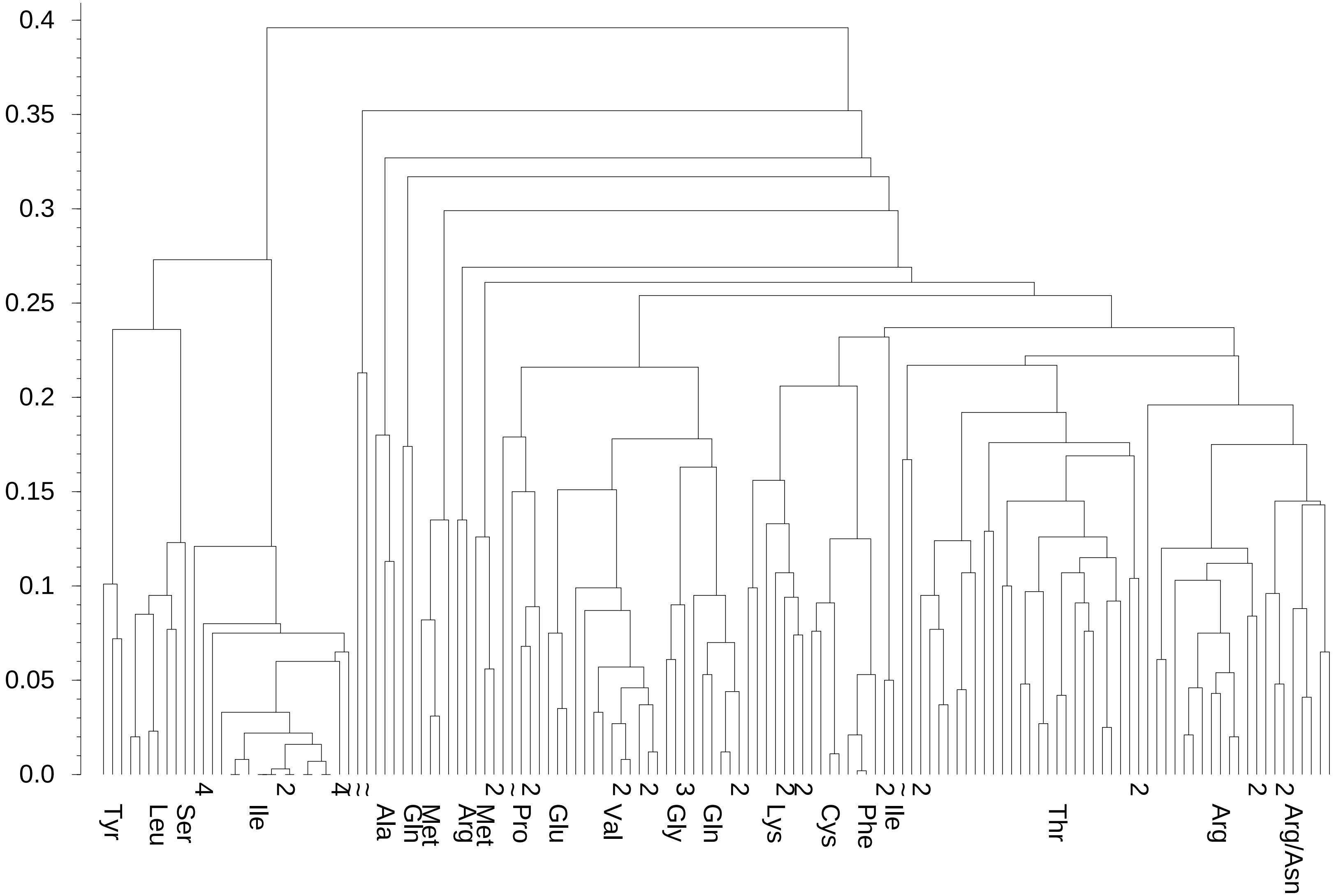

Fig. 7. Structure-based clustering of the 3332 *Ciona intestinalis* ncRNAs candidates predicted by `RNAz` (Missal *et al.*, 2005) yields (among others) this cluster comprising 157 of the 301 detected tRNAs that code for 17 different amino acids. Some nodes represent groups of almost identical (> 90% identity) sequences of which only one representative has been used for clustering.

more time and space efficient algorithm that can also be extended to local alignment (Will *et al.*, 2006). The resulting new algorithm `LocARNA` is ready to manage the envisioned ten thousands of RNAs. As a test case we consider the 3332 *Ciona intestinalis* ncRNA candidates from (Missal *et al.*, 2005). In contrast to the previous section, we used here a very stringent sequence-homology based pre-processing set that identifies sequences with more than 90% identity. Pairwise structural alignments of the resulting 2804 distinct sequences can be computed in about two days on 10 dual core CPUs.

An appropriate distance measure that is based on both sequence and structure information is necessary for applying the weighted pair group method (WPGMA) or any comparable method for tree construction and cluster extraction. Since the pairwise alignments are computed together with their similarity scores, one might naively attempt to use these scores also for clustering. This is not appropriate, however, since the local scores reflect the quality of local structure prediction, not the similarity of the different alignments

We therefore used the following normalized sequence and structure similarity measures of different alignments of RNAs. On the sequence level, we use the average sequence identity between the RNA sequences of the alignment. For the scoring of the structural similarity, we use the structure conservation index (SCI), which is the ratio between the mean single minimum free energy (mfe) and the consensus mfe. The similarities are then transformed into distances and WPGMA is applied onto the resulting distance matrix to produce the final tree. Fig. 7 shows a subtree that contains to about half of the known tRNA precursors. With very few exceptions, the tRNAs are clustered according to their amino acid and anticodon, demonstrating that the procedure indeed yields plausible results.

The resulting tree is then cut at a specific threshold to generate the clusters, from which we can then extract a common motif using an appropriate multiple alignment method. Sequence identity in the identified clusters can be rather low due to the structure influence in the clustering (often below 60%). The approach is thus capable of identifying families of structured RNAs in a range where global multiple sequence alignment already yield very poor results

6.3 Interactions with mRNAs

Regulatory RNAs more often than not function by means of direct RNA-RNA binding via complementary base pairing. This mechanism underlies the post-transcriptional gene silencing pathways of microRNAs and siRNAs (reviewed e.g. in (Nelson *et al.*, 2003)) as well as RNAi (Elbashir and Tuschl, 2001), it is crucial for snoRNA-directed RNA editing (Gott and Emeson, 2000), and it is used in the gRNA directed mRNA editing in kinetoplastids (Stuart *et al.*, 1997). A wide range of ncRNA regulation in bacteria is based upon RNA duplex formation (Gottesman, 2004). Synthetic "modifier RNAs" have been used as experimental techniques for changing the gene expression patterns independent of the RNAi pathway, see e.g. (Childs *et al.*, 2002; Meisner *et al.*, 2004; Nulf and Corey, 2004; Paulus *et al.*, 2004). Recent studies of the transcriptome of various organisms have uncovered ample evidence for wide-spread anti-sense transcription (Shendure and Church, 2002; Yelin *et al.*, 2003; Chen *et al.*, 2005a; Steigele and Nieselt, 2005; Katayama *et al.*, 2005; David *et al.*, 2006). These transcripts might at least in part be involved in RNA-RNA interactions.

In order to gather evidence for a possible function of the ncRNAs candidates in ncRNA-mRNA interactions, we investigated whether the `RNAz` predictions show an increased propensity of interacting with known mRNAs. In detail, we used the following procedure:

True and shuffled `RNAz` hits from both genomic strands were aligned to human mRNA sequences from `RefSeq` (NCBI FTP server, March 7) using `NCBI blast` (version 2.2.10 with standard parameters for blastn except an E-value cutoff of $E \leq 0.1$ and filtering set to false). The resulting alignment were filtered by removing all `blast` alignment with a length of less than 20 nucleotides or less than 75% sequence identity. Furthermore, query sequences were retained only of the matched the antisense strand of the mRNA. Of the 71970 `RNAz` hit this yields 11112 (15.4%) true and 1319 (1.8%) shuffled predicted antisense interactions. In the control consisting of 71968 conserved non-coding DNA sequences that were part of the input set in the vertebrate `RNAz` screen, we find 9055 (12.6%) predicted interactions. This corresponds to an enrichment between true and random fractions of 1.22.

After `blast` search, 1396 of true, none of the shuffled, and 1108 of the control hits were removed because they overlapped the mRNA sequence that they matched. This step eliminates potential false positives, but might also exclude true *cis*-antisense transcripts. Interestingly, this step does not affect the enrichment factor of 1.22.

For each `RNAz` hit, the longest alignment was kept for further analysis. For these interacting pairs of RNAz hits and mRNAs, a coarse grained estimate of the interaction free energy was computed using `RNAduplex`. This component of the `Vienna RNA Package` computes a simplified hybridization of two RNAs which allowing only inter-molecular base pairs, see also (Rehmsmeier *et al.*, 2004; Dimitrov and Zuker, 2004).

Interaction free energy distributions of the true and shuffled `RNAz` hits were tested against the null hypothesis of a common distribution using the Kolmo-

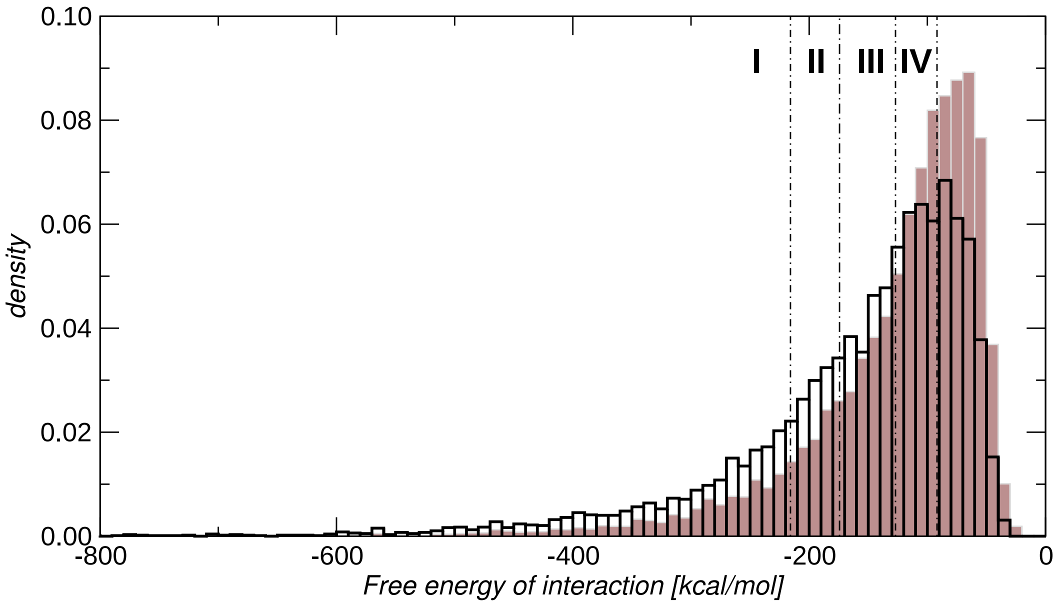

Fig. 8. Densities of interaction free energy distributions. The density of the interaction free energy distribution of true RNAz hit – mRNA interactions is shown in black, those of shuffled RNAz hits – mRNA interactions in red. Dotted blue lines indicate the energy thresholds used for classification, at −215.82, −174.28, −126.91 and −91.80 kcal/mol, corresponding to the 0.05, 0.10, 0.25, 0.50 quantiles of the randomized distribution and defining classes I to IV respectively as interactions with energy lower or equal the threshold.

gorov–Smirnov test from the statistical package `R`. The null hypothesis that both IFE distributions of shuffled and true RNAz sequence originate from a common distribution was rejected with $p < 10^{-16}$. The remaining `RNAz` sequences were co-folded with their potential target mRNAs using `RNAduplex` to determine approximate interaction free energies (IFE). Density plots of the IFE distributions are shown in Figure 8.

Interactions were classified according to their interaction free energy (IFE). Class **I** interactors have an IFE lower or equal than the empirical 0.05 quantile of the shuffled IFE distribution, class **II** interactors lower or equal than the 0.1 empirical quantile, class **III** interactors lower or equal than the 0.25 empirical quantile and class **IV** interactions lower or equal than the empirical median IFE of the shuffled RNAz hits. Results of this classification are given in table 6

Only 13 of the 11112 `RNAz` sequences retained after `Blast` search are known miRNAs, and 9 of these are classified as interacting based on IFE. Four of these sequences have an entry in `Tarbase` (Sethupathy *et al.*, 2006), listing experimentally verified miRNA target-mRNAs and in all four cases our approach would have predicted the correct target. Four of the predicted interactors are

Table 6
Interaction with mRNAs – relative to total number of RNAz hits. Absolute numbers of RNAz hit sequences matching a particular interaction class and the respective fraction of total RNAz hit sequences (71970 true, 71968 random) are given. 2451 true and 3204 randomized RNAz hits have an interaction energy smaller than the median of shuffled RNAz hit – mRNA interaction energies and are not mentioned in the table.

Interaction	**True**		**Random**		
	`RNAz` hits		`RNAz` hits		
Class	Number	Fraction	Number	Fraction	Enr.
I	2036	0.028	883	0.012	2.3
II	1193	0.017	651	0.009	1.9
III	1949	0.027	1348	0.019	1.4
IV	2087	0.029	1861	0.026	1.1

snoRNAs, which is in line with other reports that snoRNAs may play a role in mRNA modification (Kishore and Stamm, 2006).

Consistent with published data, we identify e.g. the interaction between *mir-196a* and its target mRNA *HOXB8*. This miRNA has an exceptionally high complementarity to its target mRNAs compared to other miRNAs (Yekta *et al.*, 2004) and one would therefore expect to find a particularly strong interaction. Nevertheless, we classify this interaction as not significantly more stable than random. This may indicate that microRNA function is not governed by RNA-RNA interaction energy but is dominated by structural constraints imposed by the RISC complex. This view is consistent with the observation that most miR-mRNA interactions are far from exact complementarity (Du and Zamore, 2005). At this point we cannot rule out, however, that the interaction energy model used here is too crude to properly describe individual binding patterns.

We have identified here a large number of evolutionary conserved structured ncRNA candidate genes that interact with mRNAs significantly stronger than random sequences. Almost none of them belong to one of the established ncRNA families. This observation stimulates speculations on the functional role of these transcripts. Given the stable interactions, one might consider siRNA-like functions. Alternatively, it is conceivable that some of these genes act as "modifier RNAs" by influencing mRNA secondary structure (Hackermüller *et al.*, 2005). The fact that these ncRNAs are conserved in sequence and structure may suggest that other co-factors, such as proteins which recognize specific structured binding motifs, are involved in their function. It remains to be demonstrated whether these observed interactions are restricted to conserved structured RNAs or are also common among conserved non-structured RNAs.

6.4 Structured RNAs are not enriched in predicted TFBS

A recent study by Drake *et al.* (2005) demonstrated that evolutionarily conserved noncoding sequences are selectively constrained and thus can be expected to have discernible function(s). These sequences are most often interpreted as *cis*-acting DNA motifs. This class of functional sequence motifs consists in particular of binding sites for proteins involved in transcriptional regulation (Tagle *et al.*, 1988; Davidson, 2001; Butler and Kadonaga, 2002). In order to corroborate the fact that the `RNAz` predictions are indeed functional at the RNA level, we consider the distribution of known transcription factor binding sites with in the `RNAz` candidates.

We consider a subset of 493 vertebrate transcription factor binding site (TFBS) patterns from the `transfac` database (Heinemeyer *et al.*, 1998). These are mapped to the human sequence of every 10th alignment "slice" that scored as "structured RNA" in the mammalian `RNAz` screen by (Washietl *et al.*, 2005a). For comparison, 10% of the negatively scored input alignments as well as shuffled datasets of both the positive and negative sets were used. The mapping was performed with `pwmatch`[4], a re-implementation of the scoring algorithm published by Kel *et al.* (2003), using a cut-off of 0.9. For simplicity we will refer to these hits TFBS in the following, irrespective of whether the detected sequence motif is a functional binding site *in vivo* or not.

We find that TFBS are slightly enriched in true versus shuffled data sets. Furthermore, there is small enrichment of predicted TFBS in conserved noncoding DNA that is not classified as structured RNA (0.24TFBS/nt) compared to the putative ncRNAs (0.20TFBS/nt). Since randomized sequences have only a slightly smaller density of TFBS ($0.18 - 0.19$ TFBS/nt) we conclude this (high) background level is spurious, i.e., that most of the computationally predicted TFBS are not functional. The false discovery rate of the human `RNAz` was estimated on the order of 10% (Washietl *et al.*, 2005a). The data are thus consistent with an increased frequency of TFBS in evolutionarily conserved non-coding DNA, while structured RNAs approximately behave like random background.

[4] The `pwmatch` tool is available from `www.bioinf.uni-leipzig.de/Software/pwmatch`.

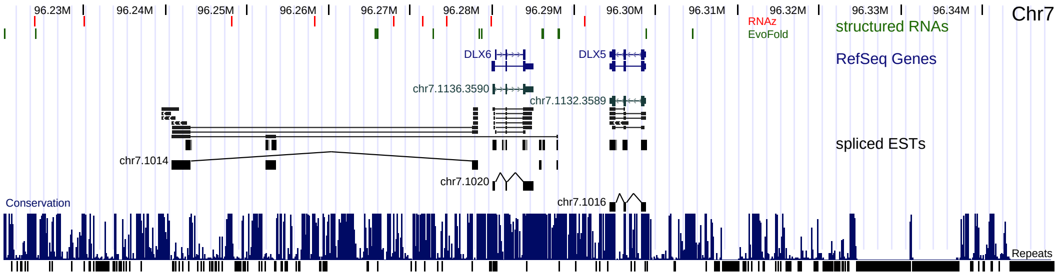

Fig. 9. Chromosomal region of the *Dlx* 5-6 bigene cluster. (UCSC genome browser, hg17 assembly). None of the predicted RNA secondary motives (`RNAz`, `EvoFold`) in this region is located in an exon of the non-coding *Evf-1*.

7 Concluding Remarks: Unstructured RNAs

In procaryotic genomes, the structure of genes, and in particular the promotor and terminator elements are sufficiently well understood that they help to detect non-coding genes independently of RNA structure or comparative sequence information. In eukaryotes, on the other hand, computational approaches to *de novo* ncRNA prediction are at present limited to structured RNAs.

A substantial number of mlncRNAs, including XIST and H19, appear to contain one or more domains with conserved RNA secondary structures, which can be detected (Washietl *et al.*, 2005a). Without additional experimental information such as EST or cDNA data, however, it is not possible at present to reliably predict the structure of such genes from genomic sequence data.

As the example of the non-coding gene *Evf-1* (Kohtz and Fishell, 2004; Faedo *et al.*, 2004) shows, not all well-defined non-coding RNAs are detectable evolutionarily conserved secondary structures, see Fig. 9. This gene is one of the few representative of mlncRNAs that has been studied in some detail. The expression of *Evf-1* depends both on the "Sonic hedgehog" (*shh*) and *Dlx* genes. The molecule exhibits splice variants of similar patterns in human, mouse, and rat. Despite its rather well-defined structure, however, currently available bioinformatics methods are insufficient to detect such unstructured ncRNA genes.

Recent tiling array (Cawley *et al.*, 2004) and cDNA data (Carninci *et al.*, 2005) strongly suggest that ncRNAs genes of this type are the rule rather than the exception. Even in the presence of cDNA and/or EST data it is not an easy task to distinguish genes with short ORFs that code for short peptides from *bona fide* ncRNAs.

The distribution of TFBS discussed in the previous section might provide a starting point. Further investigations with selected sets of TF binding motifs will however be needed to determine whether TFBS frequencies can be used to discern between cis-acting DNA elements and sequence elements that are functional at transcript level.

In this contribution we have attempted to provide an overview of the state of the art in ncRNA annotation. In summary, both the detection of functional RNAs in genomic sequence data and the classification of the candidate sequences is a challenging problem, despite significant recent advances in RNA bioinformatics. Reliably automatic annotation that would be applicable routinely on newly sequences genomes remains elusive beyond those cases that can be handled by sequence homology with known ncRNA genes.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network II" and "non coding RNA", and the German DFG Bioinformatics Initiative BIZ-6/1-2.

References

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

Allen E, Xie Z, Gustafson A, Sung G, Spatafora J, Carrington J, 2004a. Evolution of microrna genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nat Genet 36:1282–1290.

Allen TA, Von Kaenel S, Goodrich JA, Kugel JF, 2004b. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. Nat Struct Mol Biol 11:816–821.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Aravin A, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T, 2003. The small rna profile during drosophila melanogaster development. Dev Cell 5:337–350.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS, 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86.

Backofen R, Will S, 2004. Local sequence-structure motifs in RNA. Journal of Bioinformatics and Computational Biology JBCB 2:681–698.

Bafna V, Zhang S, 2004. FastR: Fast database search tool for non-coding RNA. Proc IEEE Comp Systems Bioinformatics Conference .

Baskerville S, Bartel DP, 2005. Microarray profiling of micrornas reveals frequent coexpression with neighboring mirnas and host genes. RNA 11:241–247.

Belisova A, Semrad K, Mayer O, Kocian G, Waigmann E, Schroeder R, Steiner G, 2005. Rna chaperone activity of protein components of human ro rnps. RNA 11:1084–1094.

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

Berezikov E, Guryev V, van de Belt J, Wienholds E, Ronald Plasterk HA, 2005. Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120:21–24.

Berezikov E, Plasterk RH, 2005. Camels and zebrafish, viruses and cancer: a microrna update. Hum Mol Genet 14 Spec No. 2:183–190.

Bernhart S, Hofacker IL, Stadler PF, 2006. Local rna base pairing probabilities in large sequences. Bioinformatics 22:614–615.

Berteaux N, Lottin S, Adriaenssens E, Van Coppenolle F, Van Coppennolle F, Leroy X, Coll J, Dugimont T, Curgy JJ, 2004. Hormonal regulation of h19 gene expression in prostate epithelial cells. J Endocrinol 183:69–78.

Berteaux N, Lottin S, Monté D, Pinte S, Quatannens B, Coll J, Hondermarck H, Curgy JJ, Dugimont T, Adriaenssens E, 2005. H19 mrna-like noncoding rna promotes breast cancer cell proliferation through positive control by e2f1. J Biol Chem 280:29625–29636.

Bertone P, Stoc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M, 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242–2246.

Bompfünewerer AF, Flamm C, Fried C, Fritzsch G, Hofacker IL, Lehmann J, Missal K, Mosig A, Müller B, Prohaska SJ, Stadler BMR, Stadler PF, Tanzer A, Washietl S, Witwer C, 2005. Evolutionary patterns of non-coding rnas. Th Biosci 123:301–369.

Bonnet E, Wuyts J, Rouzé P, Van de Peer Y, 2004. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. Proc Natl Acad Sci USA 101:11511–11516.

Butler J, Kadonaga J, 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev 16:2583–2592.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)*, 2005. The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR, 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116:499–509.

Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD, 2005a. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 21:326–329. URL `http://www.hubmed.org/display.cgi?uids=15922830`.

Chen PY, Manninga H, Slanchev K, Chien M, Russo JJ, Ju J, Sheridan R, John B, Marks DS, Gaidatzis D, Sander C, Zavolan M, Tuschl T, 2005b. The developmental mirna profiles of zebrafish as determined by small rna cloning. Genes Dev 19:1288–1293.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR, 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154.

Childs JL, Disney MD, Turner DH, 2002. Oligonucleotide directed misfolding of RNA inhibits *Candida albicans* group I intron splicing. Proc Natl Acad Sci USA 99:11091–11096.

Collins LJ, Macke TJ, Penny D, 2004. Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif. J Integ Bioinf #6:15p. `http://journal.imbio.de/`.

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

David L, Huber W, Granovskaia Marina Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM, 2006. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA 103:5320–5325.

Davidson E, 2001. Genomic Regulatory Systems. San Diego: Academic Press.

Deininger P, Batzer M, 2002. Mammalian retroelements. Genome Res 12:1455–1465.

Deng W, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, He Housheng Cai L, Sun H, Liu C, Li BL, Bai B, Wang J, Cui Y, Jai D, Wang Y, Du D, Chen R, 2006. Organisation of the *Caenorhabditis elegans* small noncoding transcriptome: genomic features, biogenesis and expression. Genome Res 16:30–36.

Dezulian T, Remmert M, Palatnik JF, Weigel D, Huson DH, 2006. Identification of plant microRNA homologs. Bioinformatics 22:359–360.

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

Dimitrov RA, Zuker M, 2004. Prediction of hybridization and melting for double-stranded nucleic acids. Biophys J 87:215–226.

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

Drake JAD, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Raymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN, 2005. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genetics 38:223–227.

Du T, Zamore PD, 2005. microprimer: the biogenesis and function of microrna. Development 132:4645–4652.

Eddy SR, 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3:18.

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

Elbashir S. Lendeckel W, Tuschl T, 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev 15:188–200.

Erdmann V, Szymanski M, Hochberg A, de Groot N, Barciszewski J, 1999. Collection of mrna-like non-coding rnas. Nucleic Acids Res 27:192–195.

Erdmann V, Szymanski M, Hochberg A, Groot N, Barciszewski J, 2000. Non-coding, mrna-like rnas database y2k. Nucleic Acids Res 28:197–200.

Faedo A, C. QJ, Stoney P, Long JE, Dye C, Zollo M, Rubenstein J, Price D, Bulfone A, 2004. Identification and characterization of a novel transcript down-regulated in dlx1/dlx2 and up-regulated in pax6 mutant telencephalon. Dev Dyn 231:614–620.

Farris AD, Koelsch G, Pruijn GJ, van Venrooij WJ, Harley JB, 1999. Conserved features of y rnas revealed by automated phylogenetic secondary structure analysis. Nucleic Acids Res 27:1070–1078.

Fedorov A, Stombaugh J, Harr MW, Yu S, Nasalean L, Shepelev V, 2005. Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. Nucleic Acids Res 33:4578–4583.

Flamm C, Fontana W, Hofacker I, Schuster P, 2000. RNA folding kinetics at elementary step resolution. RNA 6:325–338.

French PJ, Bliss TV, O'Connor V, 2001. Ntab, a novel non-coding rna abundantly expressed in rat brain. Neuroscience 108:207–215.

Gajendran K, Gonzales MD, Farmer A, Archuleta E, Win J, Waugh ME, Kamoun S, 2006. Phytophthora functional genomics database (PFGD): functional genomics of phytophthora-plant interactions. Nucleic Acids Res 34:465–470.

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

Gautheret D, Lambert A, 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 313:1003–1011.

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

Gogolevskaya I, Koval A, Kramerov D, 2005. Evolutionary history of 4.5sh rna. Mol Biol Evol 22:1546–1554.

Gogolevskaya I, Kramerov D, 2002. Evolutionary history of 4.5si rna and indication that it is functional. J Mol Evol 54:354–364.

Gopinath SC, Matsugami A, Katahira M, Kumar PK, 2005. Human vault-associated non-coding rnas bind to mitoxantrone, a chemotherapeutic compound. Nucleic Acids Res 33:4874–4881.

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

Gottesman S, 2004. The small RNA regulators of *Escherichia coli*: roles and mechanisms. Annu Rev Microbiol 58:303–328.

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

Griffiths-Jones S, 2004. The microRNA Registry. Nucl Acids Res 32:D109–D111.

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

Grundhoff A, Sullivan CS, Ganem D, 2006. A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. RNA [epub] doi:10.1261/rna.2326106.

Hackermüller J, Meisner NC, Auer M, Jaritz M, Stadler PF, 2005. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: A quantitative model. Gene 345:3–12.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA, 1998. Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. Nucl Acids Res 26:364–370.

Henras AK, Dez C, Henry Y, 2004. RNA structure and function in C/D and H/ACA s(no)RNAs. Curr Op Struct Biol 14:335–343.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, The Students of Bioinformatics Computer Labs 2004 and 2005, 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25 [epub].

Hertel J, Stadler PF, 2006. Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data. ISMB Submitted.

Hofacker IL, Bernhart SHF, Stadler PF, 2004a. Alignment of RNA base pairing probability matrices. Bioinformatics 20:2222–2227.

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

Hofacker IL, Priwitzer B, Stadler PF, 2004b. Prediction of locally stable rna secondary structures for genome-wide surveys. Bioinformatics 20:191–198.

Hull Havgaard JH, Lyngsø R, Stormo GD, Gorodkin J, 2005. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 21:1815–1824.

Hüttenhofer A, Schattner P, Polacek N, 2005. Non-coding RNAs: hope or hype? Trends Genet 21:289–297.

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

Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, Kageyama Y, 2005. Identification and expression analysis of putative mrna-like noncoding rna in drosophila. Genes Cells 10:1163–1173.

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

Jones-Rhoades MW, Bartel DP, Bartel B, 2006. Micrornas and their regulatory roles in plants. Annu Rev Plant Biol .

Jones-Roades MW, Bartel DP, 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 14:787–799.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR, 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14:331–342.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk A, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C, RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium, 2005. Antisense transcription in the mammalian transcriptome. Science 309:1564–1566.

Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, E. W, 2003. `MATCHTM`: a tool for searching transcription factor binding sites in DNA sequences. Nucl Acids Res 31:3576–3579.

Kishore S, Stamm S, 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science 311:230–232.

Klein RJ, Eddy SR, 2003. RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics 4:1471–2105.

Kohtz J, Fishell G, 2004. Developmental regulation of EVF-1, a novel noncoding RNA transcribed upstream of the mouse Dlx6 gene. Gene Expr Patterns 4:407–412.

Kramerov D, Vassetzky N, 2005. Short retroposons in eukaryotic genomes.

Int Rev Cytol 247:165–221.

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

Laslett D, Canback B, Andersson S, 2002. `BRUCE`: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. Nucl Acids Res 30:3449–3453.

Le SY, Chen JH, Currey K, Maizel J, 1988. A program for predicting significant RNA secondary structures. CABIOS 4:153–159.

Legendre M, Lambert A, Gautheret D, 2005. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 21:841–845.

Lestrade L, Weber MJ, 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucl Acids Res 34:D158–D162.

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

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

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

Lowe T, Eddy S, 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 25:955–964.

Lowe TM, Eddy SR, 1999. A computational screen for methylation guide snoRNAs in yeast. Science 19:1168–1171.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R, 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucl Acids Res 29:4724–4735.

Makarova JA, Kramerov DA, 2005. Noncoding rna of u87 host gene is associated with ribosomes and is relatively resistant to nonsense-mediated decay. Gene 363:51–60.

Maraia RJ, Sasaki-Tozawa N, Driscoll CT, Green ED, Darlington GJ, 1994. The human y4 small cytoplasmic rna gene is controlled by upstream elements and resides on chromosome 7 with all other hy scrna genes. Nucleic Acids Res 22:3045–3052.

Mathews DH, Turner DH, 2002. `Dynalign`: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317:191–203.

McCaskill JS, 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105–19.

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

Meisner NC, Hackermüller J, Uhl V, Aszódi A, Jaritz M, Auer M, 2004. mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation. Chembiochem 5:1432–1447.

Meyer IM, Miklós I, 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucl

Acids Res 33:6338–6348.
Missal K, Rose D, Stadler PF, 2005. Non-coding RNAs in *Ciona intestinalis*. Bioinformatics 21 S2:i77–i78.
Missal K, Stadler PF, 2006. xxx .
Missal K, Zhu X, Rose D, Deng W, Skogerbø G, Chen R, Stadler PF, 2006. Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. J Exp Zool Mol Dev Evol In press.
Mossink MH, van Zon A, Scheper RJ, Sonneveld P, Wiemer EA, 2003. Vaults: a ribonucleoprotein particle involved in drug resistance? Oncogene 22:7458–7467.
Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT, 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res 33:3570–3581.
Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z, 2003. The microRNA world: small is mighty. Trends Biochem Sci 28:534–540.
Nulf CJ, Corey D, 2004. Intracellular inhibition of hepatitis C virus (HCV) internal ribosomal entry site (IRES)-dependent translation by peptide nucleic acids (PNAs) and locked nucleic acids (LNAs). Nucl Acids Res 32:3792–3798.
Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al.*, 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573.
Pang KC, Stephen S, Engström PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS, 2005. Rnadb–a comprehensive mammalian noncoding rna database. Nucleic Acids Res 33:125–130.
Paulus M, Haslbeck M, Watzele M, 2004. RNA stem-loop enhanced expression of previously non-expressible genes. Nucl Acids Res 32:9/e78. Doi 10.1093/nar/gnh076.
Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D, 2006. Identification and classification of conserved RNA secondary structures in the human genome Preprint.
Pelczar P, Filipowicz W, 1998. The host gene for intronic u17 small nucleolar rnas in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family. Mol Cell Biol 18:4509–4518.
Piccinelli P, Rosenblad MA, Samuelsson T, 2005. Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res 33:4485–4495.
Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, Grimmond SM, Hume DA, Hayashizaki Y, Mattick JS, 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16:11–19.
Regalia M, Rosenblad MA, Samuelson T, 2002. Prediction of signal recognition particle RNA genes. Nucl Acids Res 30:3368–3377.

Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R, 2004. Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517.

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

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

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

Rodriguez A, Griffiths-Jones S, Ashurst J, Bradley A, 2004. Identification of mammalian microrna host genes and transcription units. Genome Res 14:1902–1910.

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

Sankoff D, 1985. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. SIAM J Appl Math 45:810–825.

Sawata M, Takeuchi H, Kubo T, 2004. Identification and analysis of the minimal promoter activity of a novel noncoding nuclear rna gene, ancr-1, from the honeybee (apis mellifera l.). RNA 10:1047–1058.

Schattner P, Decatur WA, Davis CA, Ares Jr M, Fournier MJ, Lowe TM, 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. Nucleic Acids Res 32:4281–4296.

Sethupathy P, Corda B, Hatziegeorgiou AG, 2006. `TarBase`: A comprehensive database of experimentally supported animal microRNA targets. RNA 12:192–197.

Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M, 2005. Identification of clustered microRNAs using an *ab initio* prediction method. BMC Bioinformatics 6:267 [epub].

Shendure J, Church GM, 2002. Computational discovery of sense-antisense transcription in the human and mouse genome. Genome Biol 3:1–14.

Smalheiser N, Torvik V, 2005. Mammalian micrornas derived from genomic repeats. Trends Genet 21:322–326.

Steigele S, Nieselt K, 2005. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. Nucl Acids Res 33:5034–5044.

Steigele S, Stadler PF, Nieselt K, 2006. Computational prediction and annotation of structured RNAs in yeasts. RECOMB poster submission.

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

Stuart K, Allen TE, Heidmann S, Seiwert SD, 1997. RNA editing in kinetoplastid protozoa. Microbiol Mol Biol Rev 61:105–120.

Sullivan CS, Ganem D, 2005. Micrornas and viral infection. Mol Cell 20:3–7.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT, 1988.

Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439–455.

Tanzer A, Amemiya CT, Kim CB, Stadler PF, 2005. Evolution of microRNAs located within hox gene clusters. J Exp Zoolog B Mol Dev Evol 304:75–85.

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

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

Torarinsson E, Sawera M, Havgaard J, Fredholm M, Gorodkin J, 2006. Thousands of corresponding human an mouse genomic regions unalignable in primary sequece contain common rna structure. Genome Research, submitted.

Uzilov AV, Keegan JM, Mathews DH, 2006. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 7:173 [epub].

van Zon A, Mossink MH, Scheper RJ, Sonneveld P, Wiemer EA, 2003. The vault complex. Cell Mol Life Sci 60:1828–1837.

Washietl S, Hofacker IL, 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 342:19–39.

Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF, 2005a. Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. Nature Biotech 23:1383–1390.

Washietl S, Hofacker IL, Stadler PF, 2005b. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102:2454–2459.

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

Will S, Missal K, Stadler PF, Backofen R, 2006. Structured-based clustering of ncrna candidates. Manuscript in preparation.

Yekta S, Shih Ih, Bartel DP, 2004. MircoRNA-directed cleavage of *HoxB8* mRNA. Science 304:594–596.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, Nemzer S, Pinner E, Walach S, Bernstein J, Savitsky K, Rotman G, 2003. Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21:379–386.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D, 2002. The single, ancient origin of chromist plastids. Proc Natl Acad Sci USA 99:15507–15512.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK, 2006. Combining multi-species genomic data for microRNA identification using a na ıve Bayes classifier. Bioinformatics [epub].