Prediction of Structured Non-Coding RNAs in the Genomes of the Nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*

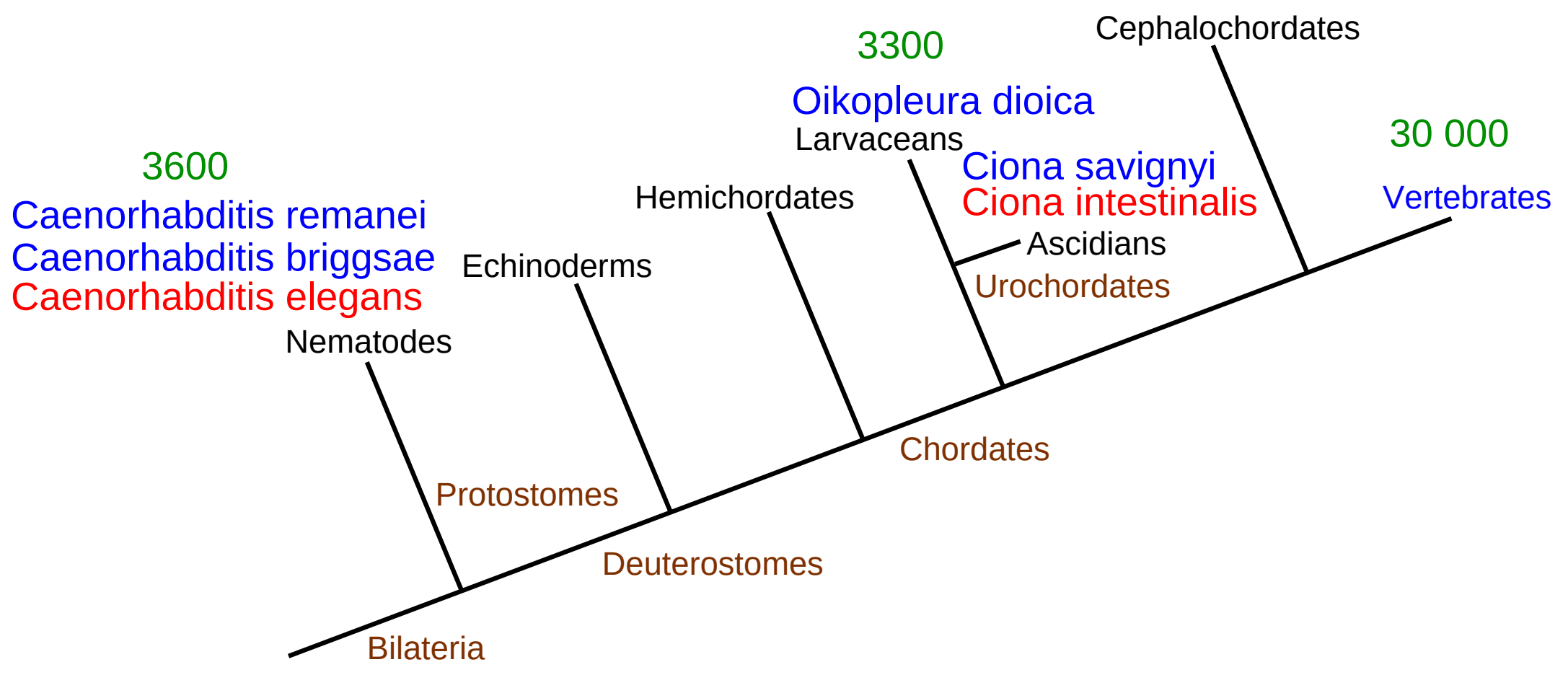
Kristin Missal*[a], Xiaopeng Zhu[b], Dominic Rose[a], Wei Deng*[b],
Geir Skogerbø[b], Runsheng Chen[b,c,d], and Peter F. Stadler[a,e,f]

[a] Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany
Email: {kristin,dominic,studla}@bioinf.uni-leipzig.de
WWW: http://www.bioinf.uni-leipzig.de
[b] Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, China
[c] Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, China
[d] Chinese National Human Genome Center, China
[e] Department of Theoretical Chemistry, University of Vienna, Austria
[f] The Santa Fe Institute, USA

Motivation

The analysis of animal genomes showed that only a minute part of their DNA codes for proteins. Recent experimental results agree, however, that a large fraction of these genomes is transcribed and hence is probably functional at the RNA level [4]. A computational survey of vertebrate genomes has predicted thousands of previously unknown non-coding RNAs (ncRNAs) with evolutionary conserved secondary structures [7]. An extension of these comparative studies beyond vertebrates is difficult, however, since most non-coding RNAs evolve relatively fast at the sequence level while conserving their characteristic secondary structures.

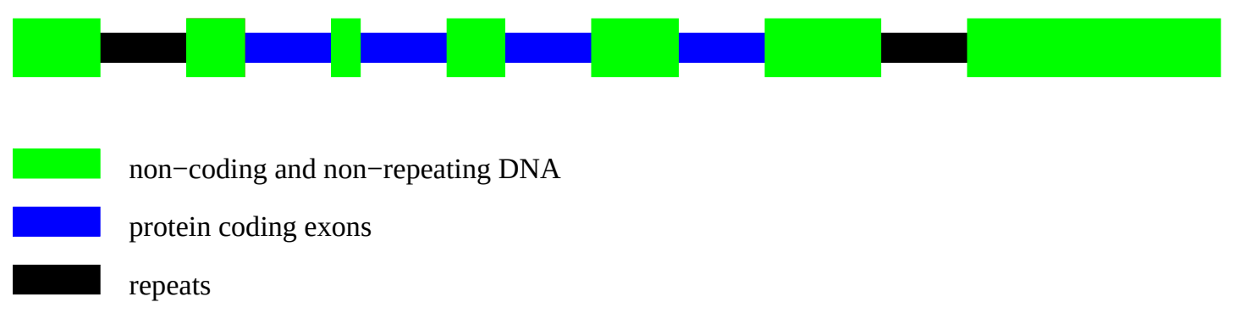
Hence, independent screens in invertebrates are necessary. A first ncRNA prediction approach amog urochordates revealed some thousand putative structured RNAs [5]. Here we extend the phylogenetic range of systematic surveys for ncRNAs to the nematodes *C. elegans* and *C. briggsae*.



Phylogenetic classification of the nematodes *C. elegans* and *C. briggsae*; green numbers represent the amount of predicted ncRNA candidates.

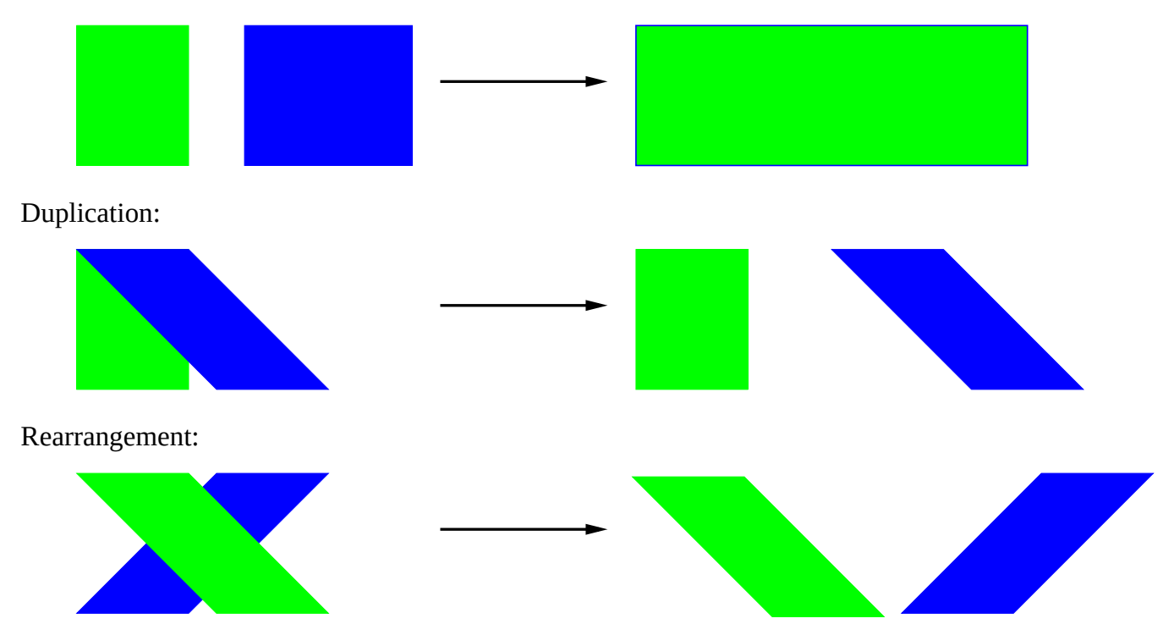
Methods

The sequences of *C. elegans* are taken from the website of the Sanger Institute in version `WS120` of March 2004, for which a gene and repeat annotation exists at the `UCSC genome browser`. Sequences of *C. briggsae* are used in version `cb25.agp8` of July 2002. The gene and repeat annotation of the UCSC genome browser are used to define non-coding DNA in the *C. elegans* genome:



Contiguous regions except protein-coding and repetitive elements define putative nc DNA.

We identify conserved non-coding DNA regions between *C. elegans* and *C. briggsae* by `blast` alignments ($E < 10^{-3}$). Hits with short distance between are combined considering consistence checks:



Global alignments of the resulting regions are computed using `clustalw`. They are screened with `RNAz` [8] to detect regions that are also conserved at the secondary structure level. The `RNAz` algorithm evaluates thermodynamic stability and the evolutionary conservation of secondary structure. Evolutionary conserved secondary structure indicates functional significance and a z-score of thermodynamic stability relative to an ensemble of shuffled sequences evaluates if the potentially transcribed RNA is more stable than by chance. For each global alignment, both possible reading directions are considered, because calculating thermodynamic energy is direction dependent.

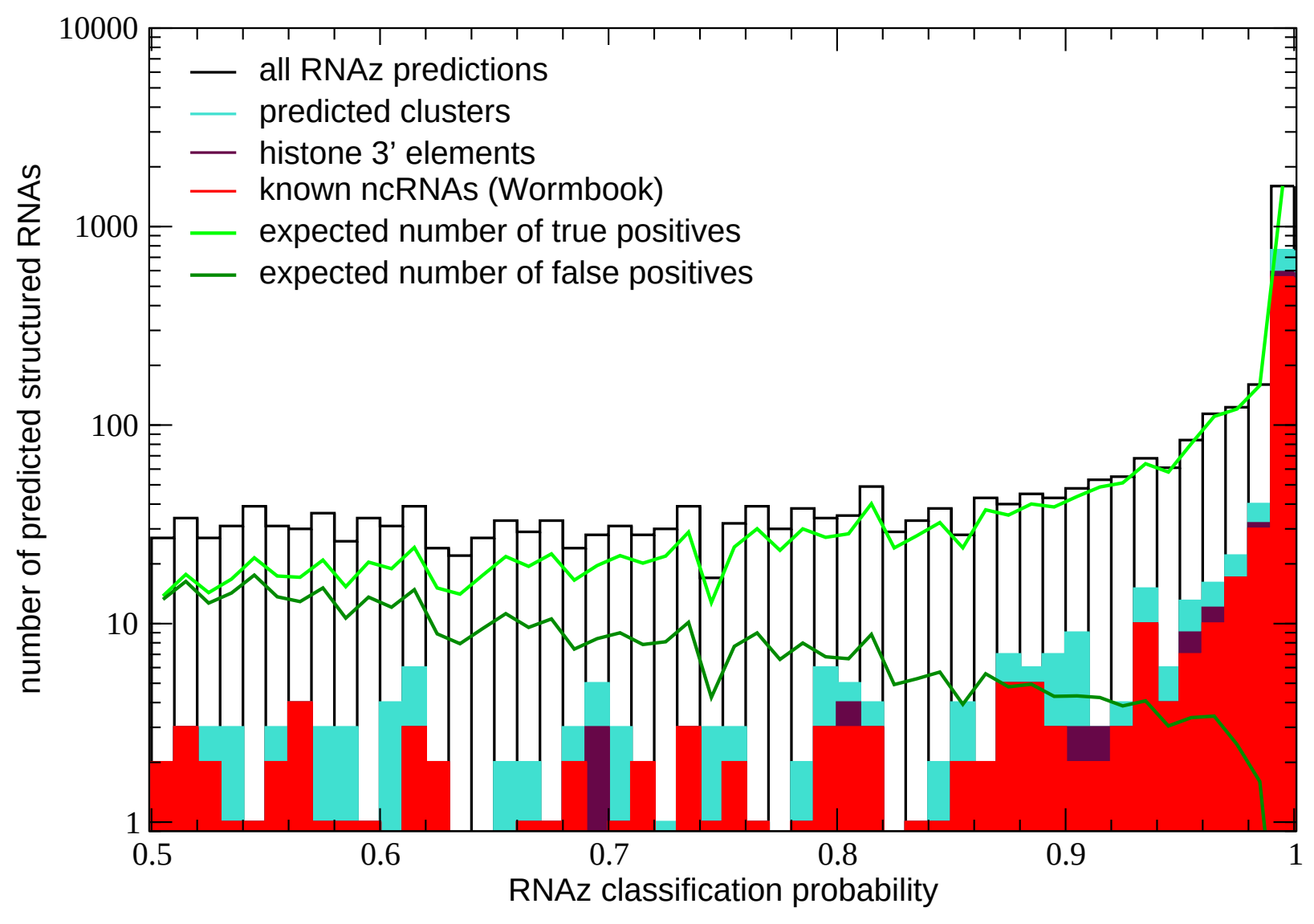
Upcoming statistical values describe the number of the genomic loci in *C. elegans.*

Results

We detect 3672 structured RNA motifs, of which only 678 are known ncRNAs or clear homologs of known *C. elegans* ncRNAs. Most of these signals are located in introns or at a distance from known protein-coding genes.

Genomic context	`blast` alignments length	Number of ncRNA candidates $p_c = 0.5$	$p_c = 0.9$
intronic	597,128	1235	891
5'UTR	116,193	119	65
3'UTR	128,766	130	69
intergenic	810,989	1221	726
total		3672	2366
length(nt)	13,567,851	432,536	291,499

Statistics of the `RNAz` ncRNA screen for *C. elegans* and *C. briggsae.* NcRNAs are slightly enriched in introns, while UTR elements are rare; 54 ncRNAs are annotated as 5'UTR as well as 3'UTR, which might be regulatory elements for polycistronic transcripts [1].



Distribution of classification probabilities p among `RNAz` predictions. Colors indicate the fractions of known ncRNAs, predicted histone elements, and predicted families with two or more homologous in each histogram bar.

				`RNAz`					
				$p_c = 0.5$			$p_c = 0.9$		
	N_g	N_a	s_g	N	s_g	s_a	N	s_g	s_a
tRNA (functional)	591	584	0.98	509	0.86	[0.87]	465	0.78	[0.79]
tRNA (pseudogene)	1072	70		50			44		
miRNA	117	40	0.34	34	0.29	[0.85]	34	0.29	[0.85]
snoRNA	31	26	0.84	13	0.41	[0.50]	9	0.29	[0.35]
snRNA (spliceosomal)	72	72	1.00	54	0.75	[0.75]	47	0.65	[0.65]
snRNA (spliced leader)	30	26	0.87	26	0.87	[1.00]	26	0.87	[1.00]
rRNA	22	20	0.9	5	0.22	[0.25]	4	0.18	[0.2]

The sensitivity of `RNAz`-detected ncRNAs is based on known ncRNA annotations from the Wormbook [6]. We compare the numbers of genes known in the genome (N_g) and those contained in our input alignments (N_a) with those classified as structured RNAs by `RNAz` (N) at two different classification probability levels. In addition, sensitivities are listed as fraction s_g of known genomic sequences, and as fraction s_a of known sequences contained in the input alignments (given in brackets).

Type				`RNAz`					
				$p_c = 0.5$			$p_c = 0.9$		
	N_g	N_a	s_g	N	s_g	s_a	N	s_g	s_a
in Wormbook	97	90	0.93	63	0.64	[0.70]	55	0.56	[0.61]
H/ACA snoRNA	41	31	0.76	11	0.26	[0.35]	9	0.21	[0.29]
CD snoRNA	28	19	0.68	3	0.10	[0.15]	2	0.07	[0.10]
sb RNA	9	3	0.33	2	0.22	[0.66]	2	0.22	[0.66]
snl RNA	8	3	0.38	3	0.37	[1.00]	2	0.25	[0.66]
unknown	14	14	1.00	4	0.28	[0.28]	2	0.14	[0.14]
all novel	101	70	0.69	23	0.23	[0.33]	17	0.17	[0.24]
Total	198	160	0.81	86	0.43	[0.53]	72	0.36	[0.45]

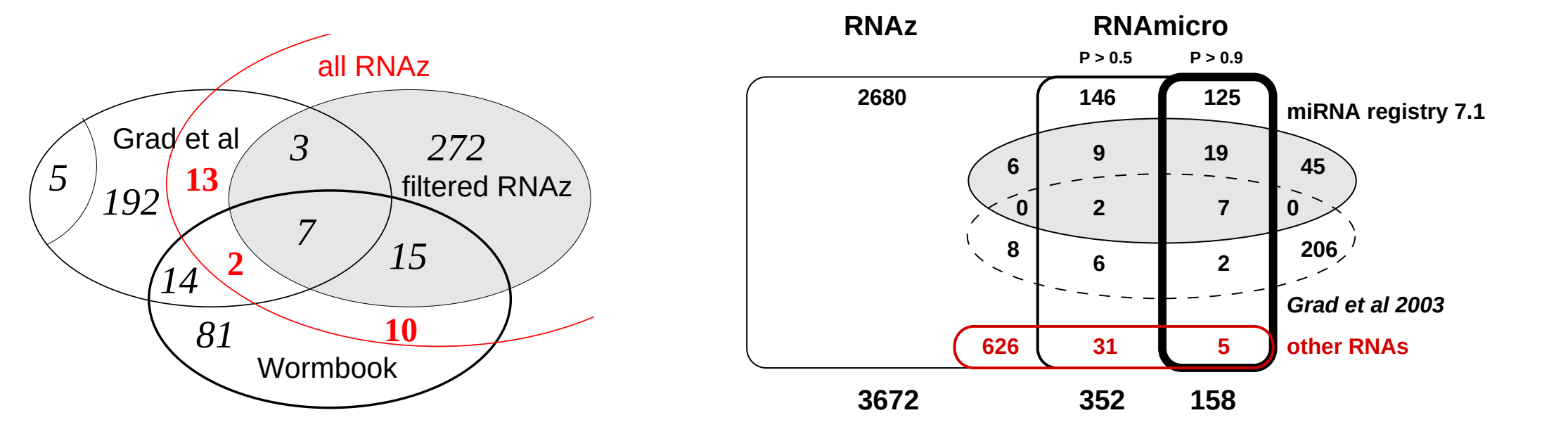
Comparison of the `RNAz` results with experimentally validated ncRNAs [2]. Columns have the same meaning as above.

Annotation

Deng *et al.* identified three putative RNA-specific promotor sequences, denoted by UM1, UM2 and UM3. They form sequence-bulge RNAs and are associated with our ncRNAs. UM1 (90 hits) covers snRNA loci and includes the *C. elegans* proximal sequence element (PSE), UM2 (413 hits) was mainly found upstream of snoRNA genes. However, it is similar to the internal tRNA promotor and thus comprises tRNA loci. UM3 (7 hits) covers the U6 snRNA, RNAse P and 5 functionally unassigned loci.

Furthermore, Deng *et al.* identified a class of snRNA-like ncRNAs characterized by a recognizable SMN-binding site. We use `RNAbob` to search for the sequence motif `AUUUUUG` followed by a hairpin of rather variable stem and loop length, a common generalization of SMN binding sites in known snRNAs. We require that the pattern corequisitely occurs in aligned positions of *C. elegans* and *C. briggsae* ncRNA candidates. This procedure recovers 122 loci of which more than 60 are plausible snRNA candidates (among others we count 9 U1, 19 U2, 5 U4, and 12 U5 loci).

Possible novel microRNA precursors are either identified by manual filtering of the `RNAz`-based predictions or by running `RNAmicro`[3] on the input alignments. `RNAmicro` works in spirit of `RNAz`, but especially is trained to detect microRNA precursors.



Comparison of microRNA candidates manually derived from `RNAz` candidate set (stem-loop structure and z-score<=-3.0 required, left figure) and automatically detected by `RNAmicro` (right figure).

References

[1] T. Blumenthal. Operons in eukaryotes. *Brief Funct Genomic Proteomic*, 3(3):199–211, Nov 2004.

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

[3] J. Hertel and P. F. Stadler. Hairpins in a haystack: Recognizing microrna precursors in comparative genomics data. *submitted*, 2006.

[4] J. S. Mattick. RNA regulation: a new genetics? *Nat Rev Genet*, 5(4):316–323, Apr 2004.

[5] K. Missal, D. Rose, and P. F. Stadler. Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, 21 Suppl 2:ii77–ii78, Sep 2005.

[6] S. L. Stricklin, S. Griffiths-Jones, and S. R. Eddy. C. elegans noncoding RNA genes. *WormBook*, doi/10.1895/wormbook.1.7.1, 2005. http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html.

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

[8] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, 102:2454–2459., 2005.

Printed by Universitätsrechenzentrum Leipzig