Non-coding RNAs in Ciona intestinalis

Kristin Missal[a,*], *Dominic Rose*[a], *and Peter F. Stadler*[a,b,c]

[a] *Bioinformatics Group, Dept. of Computer Science, University of Leipzig, Germany*
[b] *Institute for Theoretical Chemistry, University of Vienna, Austria*
[c] *Santa Fe Institute, Santa Fe, New Mexico*

ABSTRACT

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. A computational survey of vertebrate genomes has predicted thousands of previously unknown ncRNAs with evolutionary conserved secondary structures. Extending these comparative studies beyond vertebrates is difficult, however, since most ncRNAs evolve fast at the sequence level while conserving their characteristic secondary structures.

Results: We report on a computational screen of structured ncRNAs in the urochordate lineage based on a comparison of the genomic data from *Ciona intestinalis*, *Ciona savignyi*, and *Oikopleura dioica*. We predict more than 1000 ncRNAs with an evolutionarily conserved RNA secondary structure. Of these, about a quarter is located in introns of known protein coding sequences. Few RNA motifs can be identified as known RNAs, including about 300 tRNAs, some snRNA genes, as well as a few microRNAs and snoRNAs.

Contact: kristin@bioinf.uni-leipzig.de

Data: www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/05-008/

Table 1. Statistics of `RNAz` ncRNA screens.

ncRNAs	CiCs		CiCsOd	
	$p > 0.5$	$p > 0.9$	$p > 0.5$	$p > 0.9$
intronic	830	546	70	62
UTR	65	35	2	1
isolated	1697	1091	161	148
total	3332	2109	329	296
length(nt)	405,758	268,258	42,293	38,449
specificity	97.9%	99.2%	97.6%	99.1%
false positive rate	17.1%	11.4%	6.7%	3.5%

A ncRNA is classified as "isolated" if it is at least 1kb away from the closest known protein coding gene in *Ciona intestinalis*; a ncRNA is classified as "UTR" if it is located within 200nt of a coding region according to the JGI annotation (`http://genome.jgi-psf.org/ciona4/`)
Ci *C. intestinalis*, Cs *C. savignyi*, Od *O. dioica*.

Recently, there has been mounting evidence for the existence of a large number of so far unknown non-coding RNAs: tiling array experiments, for example, demonstrated that a large fraction of animal genomes is transcribed, i.e., that most transcripts do not code for proteins. A computational survey for non-coding RNAs in vertebrate, and in particular mammalian genomes, identified thousands of putative ncRNAs (Washietl *et al.*, 2005b). This contrasts the moderate number of still undiscovered ncRNAs predicted in the much smaller yeast genome (McCutcheon & Eddy, 2003), or in in bacteria (Rivas *et al.*, 2001).

Two rounds of genome duplications shaped the vertebrate genome (Holland *et al.*, 1994) and probably have contributed to the expansion of their ncRNA inventory. Urochordates, the sister group of vertebrates, do not share these genome duplications; hence their ncRNA inventory is of particular interest for comparative purposes. It is not straightforward, however, to simply include urochordate sequences in the ncRNA screen for vertebrates: The large evolutionary distance makes it hard or impossible to obtain the reliable sequence alignments that form the basis for all comparative genomics approaches towards RNA gene finding.

Fortunately, however, the genomes of two ascidians, *Ciona intestinalis* (Dehal & *al.*, 2002) and *Ciona savignyi* have been sequenced, and a third project for the larvacean *Oikopleura dioica* is on the way, providing us with sufficient data and annotation to screen these urochordate genomes for ncRNAs independently of the vertebrate data.

Sequences from *Ciona savignyi* are taken from the website of the Broad Institute, for *Oikopleura dioica* we use the shotgun traces from the NCBI trace repository as well as the genomic sequences surrounding the *Hox* genes (Seo & *al.*, 2004). The JGI gene annotation and the repeat annotation from the UCSC genome browser are used to define non-coding DNA in the *C. intestinalis* genome. Potentially homologous regions are determined by pairwise `blast` alignments, $E < 10^{-3}$. Regions with short distances ($\leq$ 30nt) between them are combined providing the local `blast` alignments are

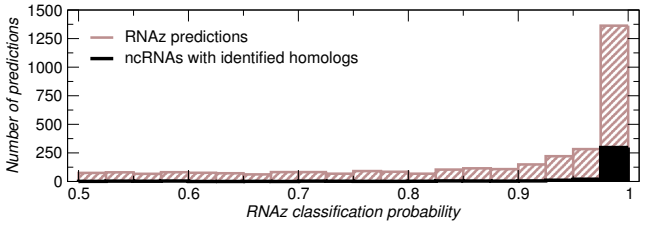

Fig. 1. Distribution of `RNAz` classification probabilities of all 3332 non-coding RNA predictions with $p > 0.5$, amounting to 2.55% of the alignable non-coding DNA. The black bars refer to the 364 ncRNAs with identified homologous RNAs in other species.

Table 2. Annotation of tRNAs and microRNAs.

		genome	$p > 0.5$ CiCs	$p > 0.5$ CiCsOd	$p > 0.9$ CiCs	$p > 0.9$ CiCsOd
tRNA	functional	550	301	266	279	252
	pseudogene	624	2	2	2	2
sensitivity			54.7%	48.3%	50.7%	45.8%
miRNA	candidates		41	0	38	0
	known	7	4	0	4	0

The 7 "known" microRNAs are taken from Legendre *et al.* (2004), one which has no homologous miRNA in the current assembly of the *C. savignyi* genome.

consistent. Global alignments of these regions are then computed using `clustalw`. A multiple alignment is constructed whenever a `blast` hit of the same *C. intestinalis* region is found with both other genomes. Starting with 80,348,523nt of non-protein-coding DNA that are not annotated as repetitive DNA, we obtain pairwise alignments for 12,193,024nt (10.4% of the 116,731,843nt genome of *Ciona intestinalis*) and 393,414nt (0.34%) of conserved non-coding sequence between all three urochordates.

These alignments were screened with `RNAz` (Washietl *et al.*, 2005a) to detect regions that are also conserved on the level of RNA secondary structure. The `RNAz` algorithm evaluates the thermodynamic stability of RNA secondary structures (relative to an ensemble of shuffled sequences) and quantifies the evidence for stabilizing selection by comparing the energy of a consensus structure with the ground-state energies of the individual structures. The classification is performed by a support vector machine based on the length and sequence divergence of the alignment, the number of aligned sequences, the energy z-score and a structure conservation index. For each global alignment, both possible reading directions are considered, because the classification of `RNAz` is based on thermodynamic stability of the potentially transcribed RNA, which is inherently direction-dependent.

In order to annotate ncRNAs and other structurally conserved RNA motifs in the *C. intestinalis* genome, we merged overlapping regions of the same alignment that were identified as conserved RNA structures into a single conserved structure. We map different alignments to the same genomic location, if they overlap to at least 90% independently of their reading direction and filter the genomic location so that each genomic location is represented in at most one ncRNA candidate. Tab. 1 summarizes the results. Repeating the entire screen with shuffled input alignments demonstrates that `RNAz` has a specificity (measured as the fraction of individual `RNAz` windows of the shuffled alignments that are not classified as structured RNA at given probability level p) of more than 0.97 ($p > 0.5$) and 0.99 ($p > 0.9$), resp. The overall false positive rate, which is defined as the fraction of individual `RNAz` scanning windows classified as ncRNA at probability level p in the shuffled alignments and original alignments, is less then 18%.

Since there is no comprehensive annotation of non-coding RNAs in urochordates we estimate the sensitivity of our screen using tRNA predictions by `tRNAscan-SE` (Lowe & Eddy, 1997), Tab. 2. A comparison (`blast` $E < 10^{-6}$) with the `noncode` database identifies 67 sequences as known ncRNAs, mostly the classical snRNAs U1 (4 loci), U2 (5), U4 (4), U5 (10), U6 (1), and U6atac (1); the signal recognition particle 7SL RNA (4); the snoRNAs U3 (4), U14 (3). Using a lower `blast` cutoff, $E < 10^{-3}$, a tentative annotation for more than 100 additional sequences was obtained, including several snoRNAs and a few microRNAs, among them the "known" miRNAs *mir-92* and *mir-124*. Furthermore, candidates for both RNAseP and RNAse MRP were identified and verified by more detailed comparative sequence analysis. A `blast` comparison with the 5S rRNA of *Halocynthia roretzi*, a closely related urochordate, identified 34 copies of 5S rRNAs in the *C. intestinalis* genome. Other ribosomal RNAs (with the exception of two sequence fragments) cannot be found in our screen because the ribosomal RNA genes are deliberately excluded from the assembly of the *C. intestinalis* genome (Dehal & *al.*, 2002). Overall, we unambiguously annotate 364 non-redundant ncRNA candidates, Fig. 1. In general, however, urochordate ncRNAs are too different from the much better-known vertebrate sequences to be identifiable by sequence comparison alone. MicroRNAs are much more stable than the structures of comparable random sequences (Bonnet *et al.*, 2004). Using microRNA families from the `Rfam` database and screening both the true data and shuffled data, we find that a cutoff of the energy z-score at $z < -3$ and a consensus structure that forms a single hairpin are sufficient to identify microRNAs with more than 90% sensitivity and 95.0% specificity. We find 41 candidates, including the "known" miRNA let-7, in our pairwise alignments. The sequence of *O. dioica*, however, is too distant, so that no identifiable microRNA is contained in the three-species alignments.

The overwhelming majority of the predicted structurally conserved RNA is located either in introns (about 1/4, despite the compact genome) or relatively far away ($>$ 1kb) from any known protein coding gene. We predict that a large fraction of these are indeed ncRNAs. The small number of signals in UTRs of known genes are probably cis-acting regulatory motifs of the corresponding mRNAs.

Acknowledgment. Discussions with Ivo Hofacker, Andrea Tanzer, and Stefan Washietl and financial support by the German *DFG* Bioinformatics Initiative, BIZ-6/1-2, are gratefully acknowledged.

REFERENCES

Bonnet, E., Wuyts, J., Rouzé, P. & Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

Dehal, P. & *al.* (2002). The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.

Holland, P. W. H., Garcia-Fernández, J., Williams, N. A. & Sidow, A. (1994). Gene duplication and the origins of vertebrate development. *Development*, **(Suppl.)**, 125–133.

Legendre, M., Lambert, A. & Gautheret, D. (2004). Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*. DOI: 10.1093/bioinformatics/bti073.

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

McCutcheon, J. P. & Eddy, S. R. (2003). Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.*, **31**, 4119–4128.

Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.

Seo, H.-C. & *al.* (2004). *Hox* cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature*, **431**, 67–71.

Washietl, S., Hofacker, I. L. & Stadler, P. F. (2005a). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.

Washietl, S., Hofacker, I. L. & Stadler, P. F. (2005b). Thousands of noncoding RNAs with conserved structure in mammalian genomes. Under review, http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/05-005.pdf.