BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii77–ii78
doi:10.1093/bioinformatics/bti1113

Genes and Genomes

Non-coding RNAs in *Ciona intestinalis*

Kristin Missal[1,*], Dominic Rose[1] and Peter F. Stadler[1,2,3]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany,
[2]Institute for Theoretical Chemistry, University of Vienna, Austria and [3]Santa Fe Institute, Santa Fe, NM, USA

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 are transcribed and hence are probably functional at the RNA level. A computational survey of vertebrate genomes has predicted thousands of previously unknown ncRNAs with evolutionarily conserved secondary structures. Extending these comparative studies beyond vertebrates is difficult, however, since most ncRNAs evolve quickly 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 >1000 ncRNAs with an evolutionarily conserved RNA secondary structure. Of these, about a quarter are located in introns of known protein coding sequences. A few RNA motifs can be identified as known RNAs, including ~300 tRNAs, some 100 snRNA genes and a few microRNAs and snoRNAs.

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

Contact: kristin@bioinf.uni-leipzig.de

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, have demonstrated that a large fraction of animal genomes are 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, http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/05-005.pdf). This contrasts with the moderate number of still undiscovered ncRNAs predicted in the much smaller yeast genome (McCutcheon and Eddy, 2003) or in bacteria (Rivas *et al*., 2001).

Two rounds of genome duplications shaped the vertebrate genome (Holland *et al*., 1994) and probably contributed to the expansion of the 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 *et al*., 2002) and *Ciona savignyi*, have been sequenced, and a third project for the larvacean *Oikopleura dioica* is underway, providing us with sufficient data and annotation to screen these urochordate genomes for ncRNAs independently of the vertebrate data.

Sequences from *C.savignyi* are taken from the website of the Broad Institute; for *O.dioica* we use the shotgun traces from the NCBI trace repository as well as the genomic sequences surrounding the *Hox* genes (Seo *et 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$30 nt) between them are combined provided the local Blast alignments are 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 80 348 523 nt of non-protein-coding DNA that are not annotated as repetitive DNA, we obtain pairwise alignments for 12 193 024 nt (10.4% of the 116 731 843 nt genome of *C.intestinalis*) and 393 414 nt (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 using 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 the 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. Table 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

[*]To whom correspondence should be addressed.

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

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

Table 2. Annotation of tRNAs and microRNAs

	Genome	$P > 0.5$		$P > 0.9$	
		CiCs	CiCsOd	CiCs	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 seven 'known' microRNAs are taken from Legendre *et al*. (2004), one of which has no homologous miRNA in the current assembly of the *C.savignyi* genome.

shuffled alignments that are not classified as structured RNA at given probability level P) of >0.97 ($P > 0.5$) and 0.99 ($P > 0.9$). 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 <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 and Eddy, 1997) (Table 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) and U14 (3). Using a lower Blast cutoff, $E < 10^{-3}$, a tentative annotation for >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 *et 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 cut-off of the energy Z-score at $Z < -3$ and a consensus structure that forms a single hairpin are sufficient to identify microRNAs with >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.

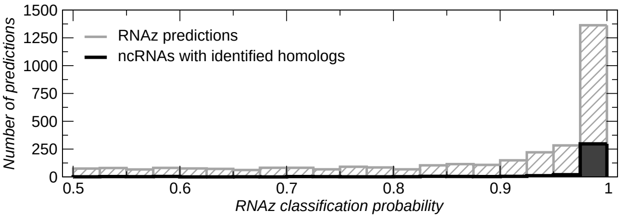

Fig. 1. Distribution of RNAz classification probabilities of all 3332 non-coding RNA predictions with $P > 0.5$, amounting to 2.55% of the alignable non-coding DNA. The black bar refers to the 364 ncRNAs with identified homologous RNAs in other species.

The overwhelming majority of the predicted structurally conserved RNA is located either in introns (approximately one-quarter, despite the compact genome) or relatively far away (>1 kb) 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.

ACKNOWLEDGEMENTS

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

Conflict of Interest: none declared.

REFERENCES

Bonnet,E. *et al*. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

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

Holland,P.W.H. *et al*. (1994) Gene duplication and the origins of vertebrate development. *Dev. Suppl.*, **1994**, 125–133.

Legendre,M. *et al*. (2004) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **21**, 841–845.

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

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

Rivas,E. *et al*. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.

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

Washietl,S. *et al*. (2005a) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005b) Thousands of noncoding RNAs with conserved structure in mammalian genomes.