RNAstrand: Reading Direction of Structured RNAs in Multiple Sequence Alignments

Kristin Missal [a]*, *Peter F. Stadler* [a,b]

[a] *Bioinformatics Group, Dept. of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[b] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.*

ABSTRACT

Motivation: Genome-wide screens for structured ncRNA genes in mammals, urochordates, and nematodes have predicted thousands of putative ncRNA genes and other structured RNA motifs. A prerequisite for their functional annotation is to determine the reading direction with high precision.

Results: While folding energies of an RNA and its reverse complement are similar, the difference are sufficient at least in conjunction with substitution patterns to discriminate between structured RNAs and their complements. We present here a support vector machine (SVM) that reliably classifies the reading direction of a structured RNA from a multiple sequence alignment.

Software: `RNAstrand` is freely available as a stand-alone tool from `http://www.bioinf.uni-leipzig.de/Software` and included in the latest release of `RNAz`, a part of the `Vienna RNA Package`, from `http://www.bioinf.uni-leipzig.de/RNA`.

Contact: kristin@bioinf.uni-leipzig.de

Genome wide computational screens for structured ncRNA genes in mammals (Washietl *et al.*, 2005b; Pedersen *et al.*, 2006), urochordates (Missal *et al.*, 2005) and nematodes (Missal *et al.*, 2006) resulted in tens of thousands putative structured ncRNAs. Functional and structural annotation of these predictions thus becomes a pressing problem. Evidence for evolutionary conserved RNA structure alone usually does not distinguish very well between the two possible reading directions. This information, however, is crucial already for the most basic annotation information. Direction information is needed e.g. to determine whether a conserved RNA motif is intronic, intergenic, or located within a coding sequence or an untranslated exon. The `RNAstrand` tool is designed specifically to predict the reading direction of a multiple sequence alignment *under the assumption* that the alignment contains an evolutionary conserved RNA secondary structure.

Our task at hand is a conceptually simple two class prediction problem for which we employ a support vector machine (SVM) (Cristianini & Shawe-Taylor, 2000). The basic idea is to devise descriptors that utilize both the small asymmetry in the energy rules (Mathews *et al.*, 1999) and the asymmetric effect of GU pairs. Suppose a particular pair of alignment columns exhibits a GC→GU substitution in one reading direction; this preserves base pairing and hence is consistent with a conserved structure. The reverse complement of the same alignment, however, displays a GC→AC substitution which is inconsistent with a conserved base pair. The patterns of structure conservation thus differ between the reading directions. Note compensatory mutations, such as GC→AU do not provide strand-specific information.

Thermodynamic stability is conveniently quantified by the mean of the energy z-scores of the individual sequences contained in the alignment. We use the same SVM-regression procedure as `RNAz` (Washietl *et al.*, 2005a) to estimate the z-scores from the sequence composition. In contrast, structural conservation is captured by the energy of the consensus fold computed by `RNAalifold` (Hofacker *et al.*, 2002). This program computes the most stable secondary structures that can be formed *simultaneously* by a set of aligned sequences. The `RNAz` program uses the ratio of these two quantities as a "structure conservation index" (SCI) to quantify structural conservation. The energetic differences between strands are captured by the differences of z-scores and mean folding energies, while differences in structure conservation are described by the differences in SCI and consensus energies. Since the relevance of the differences depends also on the sequence similarity of the aligned sequences, we use the average pairwise sequence identity and the length of the alignment as additional descriptors.

We use `libsvm 2.8` (Chang & Lin, 2001) with RBF kernel, $\gamma = 2$, probability estimates, descriptor vectors scaled linearly to the interval $[-1, +1]$ before training, and default settings as listed

Table 1. Evaluation of `RNAstrand`.

		$c = 0$		$c = 0.5$		$c = 0.9$	
ncRNA type	N	S	RNAz	S	1-S-u	S	u
rRNA	947	0.97	[0.96]	0.95	0.01	0.90	0.08
tRNA	213	0.89	[0.48]	0.75	0.06	0.64	0.33
miRNA	2385	0.99	[0.14]	0.98	0.00	0.93	0.06
snoRNA	1403	0.97	[0.90]	0.95	0.01	0.91	0.08
spliceosomal RNA	2771	0.94	[0.84]	0.90	0.04	0.80	0.18
euk. SRP RNA	2364	0.99	[0.85]	0.99	0.00	0.99	0.00
nucl. RNaseP	141	0.94	[0.90]	0.94	0.05	0.92	0.04

N . . . number of alignments in the test sets, S . . . sensitivity, u . . . fraction of undecided cases, $1 - S - u$ . . . fraction of misclassifi ed cases. For comparison we give in brackets the sensitivities of naive strand prediction using the difference of the RNAz ncRNA classifi cation probabilities for the two reading directions.

*to whom correspondence should be addressed

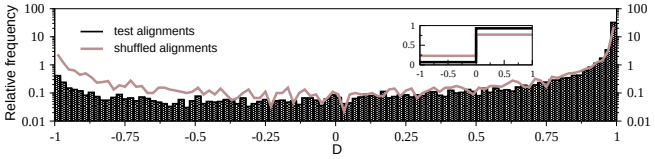

Fig. 1. Distribution of RNAstrand score D for RNAz-positive alignments and shuffled controls, excluding families with $S < 0.1$.

in the README file for all other parameters. Alignments for training were taken from the same sources as in (Washietl *et al.*, 2005a) including representatives for rRNAs, spliceosomal RNAs, tRNAs, miRNAs, small nucleolar RNAs, RNaseP and SRP RNA. Sequence similarity in this dataset ranges from 47% to 99% mean pairwise identity in alignments of $40nt$ to $400nt$ length and containing 2 up to 6 sequences. A total of 5906 alignments, approximately equally representing these ncRNA families, were used after removing 924 alignments that were not recognized as structured RNA by RNAz in the correct reading direction. Half of this dataset was used as positive training data, while the reverse complement of the other half was used as negative training set. The SVM returns an estimated class probability p which we convert into a score $D = 2p - 1$, so that $D \approx +1$ means "RNA in reading direction of input alignment" while $D \approx -1$ means "RNA is reverse complement of input alignment".

Classification performance is evaluated using 39988 alignments of 227 of the 503 ncRNA families from RFAM (version 7.0). The remaining families were excluded because they contained too few (67), too short (5), too long (4) or only too divergent (200) sequences. The test data was created in a similar way as in (Washietl *et al.*, 2005a). For each ncRNA family at most 500 alignments were randomly constructed each for 2 up to 6 sequences. Identical alignments, alignments with a pairwise identity less than 60% and alignments not recognized as RNA in both reading directions were removed. The alignments from each family were split into two distinct subsets, one used as positive test cases while the reverse complement of the other were used as negative test set.

Table 1 lists the classification rates for different threshold values c, i.e., classifying the RNA as "plus strand" for $D > c$ and as "minus strand" for $D < -c$, while $-c \leq D \leq c$ is interpreted as "undecided". The results highlight that our classification task has an intrinsic symmetry: sensitivity and specificity in fact coincide for an unbiased sample since the fraction S of correct classifications must be the same in both reading directions. We observe only a negligible loss of sensitivity when c is increased from 0 to 0.9. The distribution of D (Fig. 1) demonstrates that the majority of alignments are classified correctly with high probability.

A naive way to determine the likely reading direction is to score an alignment and its reverse complement using RNAz, evofold, or another tool for recognizing structured RNAs. This approach was taken e.g. in (Missal *et al.*, 2005, 2006; Washietl *et al.*, 2005b). A manual inspection of the data, however, showed that this approach is problematic in particular in those cases where RNAz scores are similar. Table 2 shows that the reading direction is classified correctly in the majority of the test alignments by both approaches. However, the misclassification rate of the naive approach is almost eight times higher than that of RNAstrand.

Table 2. Comparison of classification accuracies.

	Naive RNAz-based classification	
RNAstrand	correct	incorrect
correct	24510	9122
incorrect	1127	1538

The distribution of RNAstrand scores D depends on whether the alignment is classified as structured RNA or not, Fig. 1. If so, RNAstrand predicts predominately the correct reading direction. The RNAz-negative set was constructed using the shuffling procedure described in (Washietl *et al.*, 2005a), which is designed to destroy the secondary structure signal but at the same time retain as much of the alignment structure as possible. The distribution of RNAstrand scores is non-random, frequently retaining the classification of the original alignments. This indicates that the the shuffling procedure preserves a "shadow" of the structural information, which in turn implies that the false discovery rates of RNAz screens are likely to be over-estimated since the shuffled controls still contain some structural information.

In summary, RNAstrand provides a significant improvement in determining the probably reading direction of a predicted structured RNA motif by reducing the misclassification rate almost eight-fold compared to more naive approaches. This in particular increases the confidence with which we can discriminate predicted RNA motifs located in introns or non-coding UTRs from putative anti-sense RNAs.

Acknowledgement. Helpful comments by Ivo L. Hofacker and Stefan Washietl as well as financial support by the Bioinformatics Initiative (BIZ-6/1-2) of the *DFG* is gratefully acknowledged.

REFERENCES

Chang, C.-C. & Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cristianini, N. & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines. Cambridge University Press, Cambridge UK.

Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Mathews, D. H., Sabina, J., Zuker, M. & Turner, H. (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Missal, K., Rose, D. & Stadler, P. F. (2005). Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, **21 S2**, i77–i78.

Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbø, G., Chen, R. & Stadler, P. F. (2006). Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. *J. Exp. Zool. B: Mol. Dev. Evol.*. In press.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W. & Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. Preprint.

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. *Nature Biotech*, **23**, 1383–1390.