BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 2 2011, pages 279–280
doi:10.1093/bioinformatics/btq642

Sequence analysis

Advance Access publication November 13, 2010

PLEXY: efficient target prediction for box C/D snoRNAs

Stephanie Kehr[1,*], Sebastian Bartschat[1], Peter F. Stadler[1,2,3,4,5,6] and Hakim Tafer[1]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany, [2]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria, [3]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, [4]RNomics Group, Fraunhofer Institute Cell Therapy and Immunology, Perlickstrasse 1, Leipzig, Germany, [5]Center for non-coding RNAs in Technology and Health (RTH), University of Copenhagen, Copenhagen, Denmark and [6]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Small nucleolar RNAs (snoRNAs) are an abundant class of non-coding RNAs with a wide variety of cellular functions including chemical modification of RNA, telomere maintanance, pre-rRNA processing and regulatory activities in alternative splicing. The main role of box C/D snoRNAs is to determine the targets for 2′-*O*-ribose methylation, which is important for rRNA maturation and splicing regulation of some mRNAs. The targets are still unknown, however, for many 'orphan' snoRNAs. While a fast and efficient target predictor for box H/ACA snoRNAs is available, no comparable tool exists for box C/D snoRNAs, even though they bind to their targets in a much less complex manner.

Results: PLEXY is a dynamic programming algorithm that computes thermodynamically optimal interactions of a box C/D snoRNA with a putative target RNA. Implemented as scanner for large input sequences and equipped with filters on the duplex structure, PLEXY is an efficient and reliable tool for the prediction of box C/D snoRNA target sites.

Availability: The perl script PLEXY is freely available at http://www.bioinf.uni-leipzig.de/Software/PLEXY.

Contact: steffi@bioinf.uni-leipzig.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 9, 2010; revised on October 28, 2010; accepted on November 10, 2010

1 INTRODUCTION

Box C/D snoRNAs are mainly involved in 2′-*O*-ribose methylation of specific nucleotides in ribosomal and spliceosomal RNAs (Terns and Terns, 2002). They are characterized by two sequence motifs, the C-box (RTGATGA) close to the 5′-end, the D-box (CTGA) close to the 3′-end, and in most cases, an additional C′- and D′-box. The targeted position is located exactly 5 nt upstream of the 5′-end of the D- and/or D′-box. It is determined by sequence-specific hybridization (Fig. 1A). The base-pairing region has a length of 7–20 nt and exhibits a simple structure consisting of stacked base pairs and only a few mismatches. In particular, bulges are absent (Ni *et al.*, 1997).

Recently, an efficient and reliable tool for predicting the much more complex interactions of box H/ACA snoRNAs with their targets has become available (Tafer *et al.*, 2010). It is based on the thermodynamic principles of RNA folding. No comparable approach is currently available for the simple box C/D snoRNA-RNA duplexes. snoTarget (Bazeley *et al.*, 2008), at present the only computer program devoted to box C/D snoRNA target prediction, employs pattern matching to find candidates, which are then ranked by the cofolding energy of snoRNA and target as computed by RNAcofold (Bernhart *et al.*, 2006). In contrast, PLEXY directly computes the interaction energies by means of dynamic programming.

2 RESULTS

The PLEXY Algorithm: PLEXY takes a snoRNA sequence with annotated box-motifs and a list of potential target RNAs as input. First, PLEXY extracts the putative antisense elements (ASEs), which are defined as the 20 nt long segments directly upstream of the D/D′-boxes. PLEXY then calls the RNAplex algorithm to compute stable duplexes between the ASE and the putative targets. RNAplex is a fast folding algorithm for unbranched RNA structures that utilizes a linearized energy model to achieve a linear runtime behavior (Tafer and Hofacker, 2008). In contrast to RNAcofold used in snoTarget, RNAplex neglects the internal structure of the interacting RNA sequences and hence is fast enough for genome-wide searches without the need for additional prefiltering steps. The resulting duplexes are then filtered using the rules compiled by Chen *et al.* (2007):

(a) the interaction should be at least 7 nt long,

(b) no bulges are allowed,

(c) the core duplex region contains at most one mismatch, and

(d) the methylated residue forms a Watson–Crick pair.

Finally, putative target sites are ranked by their duplex energies (Supplementary Material S2).

Runtime: the runtime of PLEXY scales linearly with the length of the target sequence. It scans 10^6 nt of target sequences in 19 s on a 2.66 GHz Intel processor (Q9400). This is only four times slower than the pattern search algorithm employed by snoTarget.

Accuracy: to compare the performance of PLEXY and snoTarget, we used a collection of experimentally verified snoRNA–rRNA interactions of yeast (Lowe and Eddy, 1999) and human (Lestrade and Weber, 2006) and used yeast (Samarsky and Fournier, 1999) and human (Lestrade and Weber, 2006) snoRNA

*To whom correspondence should be addressed.

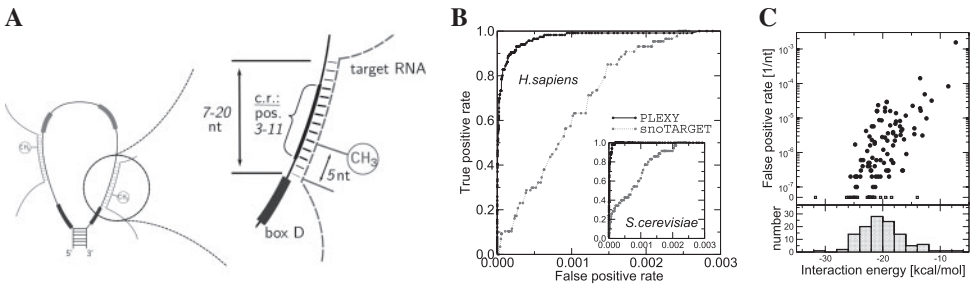

Fig. 1. (**A**) Box C/D snoRNA interacting with target RNA. The duplex length varies between 7 and 20 nt, the core duplex region (c.r.) extends from the 3-rd to 11-th nt upstream of the D or D′-box. The methylated residue is always the 5-th nucleotide upstream of the D/D′-box 5′-end. (**B**) Receiver operating characteristic (ROC) curves of the target predictions by `PLEXY` (solid line) and `snoTarget`(dotted lines) in human and yeast (inset). (**C**) Rate of false positive interaction predictions in genomic DNA as a function of interaction energy with the known target for human snoRNAs. For 24 snoRNAs, no false positive hit is reported in 10^7 nt. Below that the histogram of interaction energies with known targets is shown.

and rRNA sequences. In yeast, `PLEXY` correctly predicted all 50 target sites, 49 (98%) being ranked first. In contrast, `snoTarget` recovered only 37 of 50 (74%) of the methylation sites, and only 20 (40%) achieved the top rank. In human, `PLEXY` found 116 of 118 (98.3%) known rRNA targets, 108 (91.55%) with top rank (Supplementary Material S1). `snoTarget` retrieved 78.88% of these targets and ranked 55.77% of them at the top. The data are summarized as ROC curves in (Fig. 1B). The minimum free energy for the predicted duplexes on the rRNAs averages −20.4 kcal/mol. The energy distribution is shown in Figure 1C.

False positive rate: we tested 116 snoRNAs with known targets on rRNAs against a 10 Mb segment of the human genome (chr12:2M-12M). A duplex is a false positive hit (FP) if its interaction energy is lower than that of the true interaction. For 24 snoRNAs no FP was found, 39 additional snoRNAs had less than one FP per Mb, and more than 80% (98/116) of the snoRNAs had less than one FP per 100 Kb. The false positive rate depends exponentially on the interaction energy (Fig. 1C), hence `PLEXY` cannot reliably predict the few snoRNA–rRNA interactions that have very poor interaction energies.

Targets in mRNAs: in contrast to the majority of the box C/D snoRNAs, the members of the brain-specific *HBII-52* family do not methylate rRNAs or snRNAs but basepair close to an alternative splice junction in the mRNA transcript *5HT-2C*, which codes for the serotonin receptor (Kishore and Stamm, 2006). A target search in a large dataset of (primary) transcripts expressed in brain (covering about $\sim 0.75 \times 10^9$ nt) returned the known target site with median duplex energy of −29.1 kcal/mol for 41 of the 42 members of the snoRNA family, and revealed a second putative target with a median interaction energy of −29.3 kcal/mol in 37 of the 42 snoRNAs. The second region is located in a large intronic region of the neurexin primary transcript (hg18/chr2:50666208-50666189). The example demonstrates that `PLEXY` can be employed for essentially transcriptome-wide target searches.

3 DISCUSSION

Recently, it was discovered that *HBII-52* is also processed into shorter RNAs, so-called psnoRNAs (for processed snoRNAs), that appear to be involded in splicing regulation. The psnoRNAs form RNPs distinct from the 'common' snoRNPs. It is not surprising, therefore, that the psnoRNAs-mediated mode of action follows somewhat different interaction rules, although they involve the same regions of the snoRNA. For instance, psnoRNA–mRNA duplexes appear to have more mismatches than canonical snoRNA–rRNA ones (Kishore *et al.*, 2010). As soon as these recognition parameters are better understood, they could be easily included in the `PLEXY` algorithm by simply adding new filtering rules.

Finally, we remark that the specificity of `PLEXY` can be enhanced by considering evolutionary conservation of the target site. This can be achieved most easily by filtering the predicted putative targets by their sequence conservation or by using the ability of `RNAplex` to compute interactions between multiple sequence alignments.

In summary, `PLEXY` is a computationally efficient tool to predict target sites for box C/D snoRNAs. It is specific enough to reliably identify modification sites on rRNAs and snRNAs. At the same time, it is efficient enough to perform genome-wide searches for potential mRNA targets of orphan snoRNAs.

Funding: European Union under the auspices of the FP-7 QUANTOMICS project.

Conflict of Interest: none declared.

REFERENCES

Bazeley,P.S. *et al.* (2008) snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene*, **408**, 172–179.

Bernhart,S.H. *et al.* (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.

Chen,C.L. *et al.* (2007) Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J. Mol. Biol.*, **369**, 771–783.

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

Kishore,S. *et al.* (2010) The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum. Mol. Genet.*, **19**, 1153–1164.

Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.

Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.

Ni,J. *et al.* (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, **89**, 565–573.

Samarsky,D.A. and Fournier,M.J. (1999) A comprehensive database for the small nucleolar RNAs from Saccharomyces cerevisiae. *Nucleic Acids Res.*, **27**, 161–164.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Tafer,H. *et al.* (2010) RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics*, **26**, 610–616.

Terns,M.P. and Terns,R.M. (2002) Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin. *Gene Expr.*, **10**, 17–39.