BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–2

PLEXY: Efficient Target Prediction for Box C/D snoRNAs

Stephanie Kehr [1]*, Sebastian Bartschat [1], Peter F. Stadler [1–5], 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]Inst. f. 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, Germany
[4]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1,D-04103 Leipzig, Germany
[5]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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'-0-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-RNA target is available, no comparable tool exists for C/D-Box 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 predictions of box C/D snoRNA target sites.

Availability: The source code of PLEXY is freely available at `http://www.bioinf.uni-leipzig.de/Software/PLEXY`

Contact: steffi@bioinf.uni-leipzig.de

1 INTRODUCTION

Box C/D snoRNAs are mainly involved in 2'-0-ribose methylation of specific nucleotides in ribosomal and spliceosomal RNAs (Terns & Terns, 2002). The targeted position is located exactly 5 nucleotides upstream of the 5′ end of the D or D' box. It is determined by sequence-specific hybridization, Fig. 1A. The base-pairing region has a length of 7-20 nts and exhibits a simple structure consisting of stacked base-pairs and a few mismatches only. In particular, bulges are absent (Ni *et al.*, 1997).

Recently, an efficient and reliable tool for predicting the much more complex interactions of H/ACA snoRNAs with their targets has become available (Tafer *et al.*, 2010), which is based on the thermodynamics principles of RNA folding. No comparable approach is currently available for the simple C/D snoRNA-RNA duplexes. `snoTarget` (Bazeley *et al.*, 2008), at present the only computer program devoted to C/D snoRNA target prediction, employs pattern matching to find candidates, which are then ranked by the co-folding 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 20 nt sequence segments upstream of D- and D'- boxes ers extraceted as putative interaction regions. PLEXY then calls the `RNAplex` algorithm to compute stable duplexes of the snoRNA antisense region 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 & Hofacker, 2008). The list of duplexes is then filtered using the rules compiled by (Chen *et al.*, 2007):
(a) the interaction should be at least 7nts long,
(b) no bulges are allowed,
(c) the core duplex region contains at most one mismatch,
(d) the methylated residue forms a Watson-Crick pair.
Finally, the putative target sites are ranked by the computed duplex energies.

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

Accuracy In order to compare the performance of PLEXY and `snoTarget` we used a collection of experimentally verified snoRNA-rRNA interactions of yeast (Lowe & Eddy, 1999) and human (Lestrade & Weber, 2006), and used yeast (Samarsky & Fournier, 1999) and human (Lestrade & Weber, 2006) snoRNA and rRNA sequences. In the yeast dataset, PLEXY correctly predicted all 50 target sites, 49 (98%) being ranked first. In contrast,

*to whom correspondence should be addressed

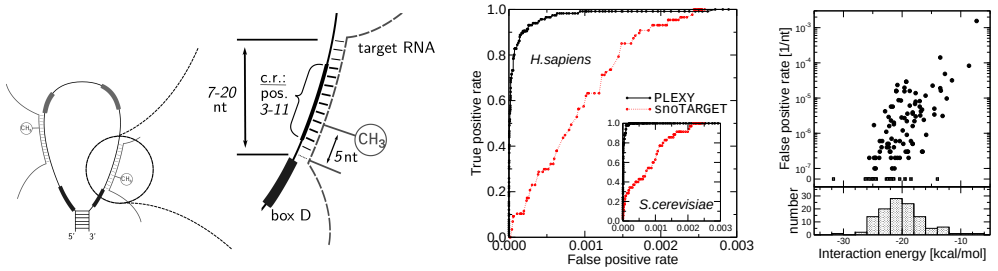

Fig. 1: **A:** Box C/D snoRNA interacting with target RNA. The duplex length varies between 7 and 20 nts, the core duplex region (c.r.) extends from the $3rd$ to $11th$ nt upstream of the D or D'-box. The methylated residue is always the $5th$ nucleotide upstream of the D/D'-box $5'$ end. **B:** ROC-curves 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 nucleotides. Below that the histogram of interaction energies with known targets is shown.

snoTarget recovered only 37 of 50 (74%) of the methylation sites, only 20 (40%) achieving the top rank. In human, PLEXY finds 116 out of 118 (98.3%) of the known rRNA targets, 108 (91.55%) with top rank. snoTarget retrieves 78.88% of these targets and ranks 55.77% of them at the top of the list. The data are summarized as ROC-curves in Figure 1B. The minimum free energy for the predicted duplexes on the ribosomal RNAs averages -20.4 [kcal/mol]. The energy distribution is shown in Figure 1C.

False Positive Rate We tested 117 snoRNAs with known targets on rRNAs or snRNAs against a 10 Mb segment of the human genome. A duplex is a false positive if its interaction energy is lower than that of the true interaction. For 24 snoRNAs we found no false positive hit, for 39 additional snoRNAs, there is less than one false positive per megabase, and more than 80% (98/117) of the snoRNAs have less than 1 false positive in 100kb. The false positive rate depends exponentially on the interaction energy, Figure 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 brain-specific *HBII-52* family do not methylate rRNAs or snRNAs but guide modifications close to an alternative splice junction in the mRNA transcript *5HT-2C*, which codes for the seretonin receptor (Kishore & Stamm, 2006). A search of a large dataset of (primary) transcripts expressed in brain (covering abour $\sim 0.75 \times 10^9$ nt) returned the known target site with a 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 2nd region is located in an large intronic region. 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 psnoRNA form RNPs distinct from the "common" snoRNPs. It is not surprising, therefore, that the interactions guiding methylation and the psnoRNAs-mediated mode of action follow somewhat different rules, although they involve the same regions of the snoRNA. For instance, psnoRNA-mRNA duplexes appear to have more mismatches than canonical snoRNA-rRNA interactions (Kishore *et al.*, 2010). As soon as these recognition parameters are better understood they can be easily included in the PLEXY algorithm by simply adding rules to select approriate duplexes from the RNAplex output.

Finally, we remark that the specificity of PLEXY can be enhanced by considering evolutionary conservation of the target site. This is achieved most easily by filtering the predicted putative targets by their sequence conservation. Alternatively, the capability of RNAplex to compute interactions regions between multiple sequence alignements could by employed.

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 ribosomal and spliceosomal RNAs. 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.

REFERENCES

Bazeley, P. S., Shepelev, V., Talebizadeh, Z., Butler, M. G., Fedorova, L., Filatov, V. & Fedorov, A. (2008) snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene*, **408** (1-2), 172–9.

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F. & Hofacker, I. L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3 [epub].

Chen, C. L., Perasso, R., Qu, L. H. & Amar, L. (2007) Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J Mol Biol*, **369** (3), 771–83.

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P. J., Stefan, M., Beach, C., Nicholls, R. D., Zavolan, M. & Stamm, S. (2010) The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum Mol Genet*, **19** (7), 1153–64.

Kishore, S. & Stamm, S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311** (5758), 230–2.

Lestrade, L. & Weber, M. J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, **34** (Database issue), D158–62.

Lowe, T. M. & Eddy, S. R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283** (5405), 1168–71.

Ni, J., Tien, A. L. & Fournier, M. J. (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, **89**, 565–573.

Samarsky, D. A. & Fournier, M. J. (1999) A comprehensive database for the small nucleolar RNAs from Saccharomyces cerevisiae. *Nucleic Acids Res*, **27** (1), 161–4.

Tafer, H. & Hofacker, I. L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24** (22), 2657–63.

Tafer, H., Kehr, S., Hertel, J., Hofacker, I. L. & Stadler, P. F. (2010) RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics*, **26** (5), 610–6.

Terns, M. P. & Terns, R. M. (2002) Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin. *Gene Expr*, **10** (1-2), 17–39.