Nucleic Acids Research Advance Access published June 10, 2013

RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins

Hilal Kazan[1],* and Quaid Morris[2,3,4]

[1]Department of Computer Engineering, Antalya International University, Antalya, 07190, Turkey, [2]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada, [3]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada and [4]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5G 1L6, Canada

Received March 17, 2013; Revised April 30, 2013; Accepted May 6, 2013

ABSTRACT

RBPmotif web server (http://www.rnamotif.org) implements tools to identify binding preferences of RNA-binding proteins (RBPs). Given a set of sequences that are known to be bound and unbound by the RBP of interest, RBPmotif provides two types of analysis: (i) *de novo* motif finding when there is no a priori knowledge on RBP's binding preferences and (ii) analysis of structure preferences when there is a previously identified sequence motif for the RBP. *De novo* motif finding is performed with the previously published RNAcontext algorithm that learns discriminative motif models to identify both sequence and structure preferences. The results of this analysis include the inferred binding preferences of the RBP and the added predictive value of incorporating structure preferences. Second type of analysis investigates whether the instances of the previously identified sequence motif are enriched in a particular structure context in bound sequences, relative to its instances in unbound sequences. On completion, the results page shows the comparison of structure contexts of the motif instances between bound and unbound sequences and an assessment of statistical significance of detected preferences. In summary, RBPmotif web server enables the concurrent analysis of sequence and structure preferences of RBPs through a user-friendly interface.

INTRODUCTION

Post-transcriptional regulation is carried out by RNA-binding proteins (RBPs) that bind to specific mRNAs to control their splicing, transport, localization, stability and degradation. Eukaryotic cells encode hundreds of RBPs, but binding specificities of most RBPs remain uncharacterized. Recently developed high-throughput experimental methods promise to rapidly expand our knowledge by identifying the RNA targets of several RBPs (1–3). However, because the locations of the binding sites within the targets are unknown and because RBPs can recognize both sequence and RNA secondary structure elements in their binding sites, identification of RBP binding preferences from these data requires further analysis with computational methods.

Motif models that are originally developed for DNA-binding proteins are commonly used to infer RBP binding preferences from high-throughput binding data. However, because these models consider only the sequence content of the binding sites, they can give inaccurate results when the RBP has a non-trivial preference for RNA secondary structure. For example, Vts1p is a yeast RBP that preferentially binds CNGG sequences located in RNA hairpin loops (4). Detecting Vts1p's sequence specificity can be difficult without consideration of its structural preference [e.g. (5)]. This observation led to the development of RBP-specific motif models that consider RNA secondary structure. Our previously published RNAcontext algorithm (6) is one such model that can query preferences for multiple structure contexts in addition to the sequence preferences. RNAcontext uses a novel representation of RNA secondary structure that takes into account the uncertainty in the secondary structure that an RNA sequence can assume. We showed that RNAcontext can infer the RBP sequence and structure binding preferences accurately by applying it to several experimental binding data. However, the lack of a web server implementation of RNAcontext has limited its use by biologists.

RBPmotif web server provides two types of analysis depending on the current knowledge of binding preferences of the RBP. If there is no a priori knowledge on RBP binding preferences, the user can choose to run

*To whom correspondence should be addressed. Tel: +90 242 245 0271; Fax: +90 242 245 0045; Email: hilal.kazan@antalya.edu.tr

RNAcontext to identify sequence and structure preferences of the RBP. The required inputs for this analysis are the set of bound and unbound sequences, range of lengths of the motif and parameters specifying the representation and prediction procedure of secondary structure. As a result, the user can obtain the predicted sequence and structure preferences and can also assess whether the incorporation of structure preferences has an added predictive value on held-out data. In addition, RBPs with similar sequence preferences to the predicted sequence motif are identified by searching existing databases of RBP binding sites. If there is a previously identified sequence motif for the RBP, the user can choose to apply the second type of analysis to investigate whether the RBP has an additional preference for the structure context of this motif. In addition to the set of input sequences and parameters for secondary structure prediction, the IUPAC representation of the previously identified motif is required for this analysis. As output, comparison of the secondary structure profiles of the instances of this motif between bound and unbound sequences is shown with a bar graph. Also, results of Wilcoxon rank sum test are provided to show the significance of the identified structure preference(s). We have previously used a similar type of analysis to investigate the structure context of LIN28 binding sites identified by CLIP-seq (7).

RBPMOTIF WEB SERVER

We will first describe how we predict secondary structures of input sequences, a step common in both types of analysis. Then, we will explain each type of analysis in detail by explaining the required inputs, implementation details and provided results.

RNA secondary structure prediction

Recent experimental techniques to determine the secondary structures of RNAs have been promising; however, to allow the applicability of RBPmotif server with any set of RNA sequences (i.e. including mRNAs without experimental secondary structure information or designed RNA sequences), we decided to use computational algorithms to predict RNA secondary structure.

A large class of RNA secondary structure prediction methods uses empirically derived thermodynamic parameters to calculate the free energy of a secondary structure. These algorithms often predict the structure with the minimum free energy. However, as thermodynamic parameters have substantial inaccuracies and an RNA sequence can fold into multiple structures during its lifetime, the predicted minimum free energy structure may not be representative of the typical interactions occurring in the structure. As such, a number of methods have considered the distribution of possible structures that an RNA sequence can form. For instance, Sfold (8), the structure prediction method that we used in the original RNAcontext article, draws a representative sample of structures from the Boltzmann ensemble of secondary structures. The secondary structures in this sample are then parsed to calculate, for each position, the distribution over structural contexts such as being paired, being in a hairpin loop, etc. RNAplfold (9) is another ensemble-based method that computes base pair probabilities directly from the Boltzmann equilibrium distribution of all possible structures, rather than a set of structures sampled from this distribution. RNAplfold differs from Sfold also by its local folding strategy where only base pairs within a certain window are considered possible. In this method, mean base pair probabilities are calculated by averaging over all windows that contain the pair. RNAplfold's local folding approach has been shown to produce more accurate results for mRNA sequences when compared with the global approach (i.e. folding the entire mRNA sequence at once, as in Sfold) (10). Also, RNAplfold's running time is much shorter than the global folding approaches, as the number of possible structures increases exponentially with the length of the mRNA. To enable the analysis of long mRNA sequences on RBPmotif server, we chose to use RNAplfold to predict RNA secondary structures. We modified the original source code of RNAplfold to obtain separate probabilities for each position of a sequence to be in a hairpin loop (H), external loop (E), internal loop (I), multiloop (M) or to be paired (P). Figure 1 illustrates how multiple possible structures of an RNA sequence are taken into account to calculate these probabilities. In this toy example, we assume that the RNA sequence can fold into five possible structures with equal free energies. Probability of a base at a specific position to be in a particular structural context, such as being in paired region, is calculated as the proportion of times that base appears in that structural context.

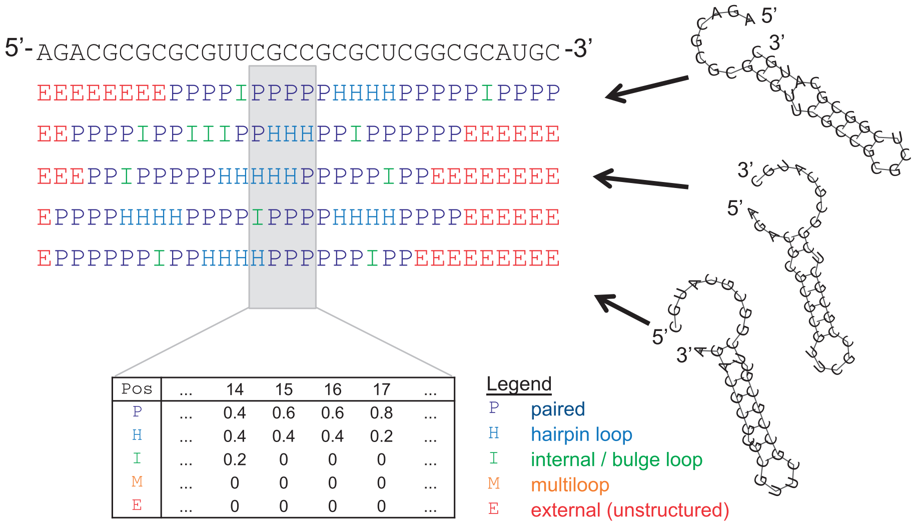
De novo motif discovery with RNAcontext

This part of the web server implements the RNAcontext algorithm for *de novo* discovery of binding preferences when there is no a priori knowledge on the binding preferences of the RBP of interest. RNAcontext infers the sequence and structure preferences of RBPs by maximizing the agreement between provided input labels (i.e. bound or unbound) and RNAcontext predicted scores.

Inputs and analysis

The user is required to provide the following inputs:

- two sets of sequences: (i) a set of sequences that are likely to contain binding sites (i.e. bound sequences) and (ii) a set of sequences that are unlikely to contain binding sites (i.e. unbound sequences). These sequences can be input by either pasting them to text boxes or by uploading as FASTA files.
- range of lengths of the binding site (allowed range is 4–12 nts)
- type of the secondary structure representation (allowed options are PU, PLE, PHTE and PHIME)
- RNAplfold parameters to specify the prediction of RNA secondary structure


5'-AGACGCGCGCGUUCGCCGCGCUCGGCGCAUGC-3'
EEEEEEEEPPPPIPPPPPHHHHPPPPPIPPPP
EEPPPPIPPIIIPPHHHPPIPPPPPPEEEEEE
EEEPPIPPPPPHHHHHPPPPIPPEEEEEEEEE
EPPPPHHHHPPPPIPPPPHHHHPPPPEEEEEE
EPPPPPPIPPHHHHPPPPPPIPPEEEEEEEEE
5'
3'
3'
5'
5'
3'
Legend
P paired
H hairpin loop
I internal / bulge loop
M multiloop
E external (unstructured)

Pos	...	14	15	16	17	...
P	...	0.4	0.6	0.6	0.8	...
H	...	0.4	0.4	0.4	0.2	...
I	...	0.2	0	0	0	...
M	...	0	0	0	0	...
E	...	0	0	0	0	...

Figure 1. Calculation of secondary structure profiles. In this toy example, the RNA sequence is assumed to fold into five equally probable secondary structures. These structures can be represented by annotating each base according to the secondary structure element (e.g. paired, hairpin loop) that it participates in. The distribution of annotation for each base can then be calculated by recording the proportion of times that the base appears in the particular structure context. An example calculation is shown for bases in positions 14–17.

In addition to the set of bound and unbound sequences, users are required to provide the range of lengths of the hypothesized binding site. The length of the motif can range between 4 and 12 nts. Remaining inputs are related to the prediction of secondary structure. Users can choose between four different types of secondary structure representation: PU, PLE, PHTE and PHIME. If the 'PU' option is selected as the type of the secondary structure representation, the resulting profile matrix consists of two values for each position: the probabilities of the base to be paired and to be unpaired. 'PLE' option considers three structural contexts: paired (P), the union of hairpin, internal and multiloops (L) and external loop (E). 'PHTE' option considers four structural contexts: paired (P), hairpin loop (H), internal and multiloop (T) and external loop (E). If the 'PHIME' option is selected instead, the resulting profile matrix consists of five values for each position: the probabilities of the base to be paired (P), to be in a hairpin loop (H), in an internal loop (I), in a multiloop (M) and in an external region (E).

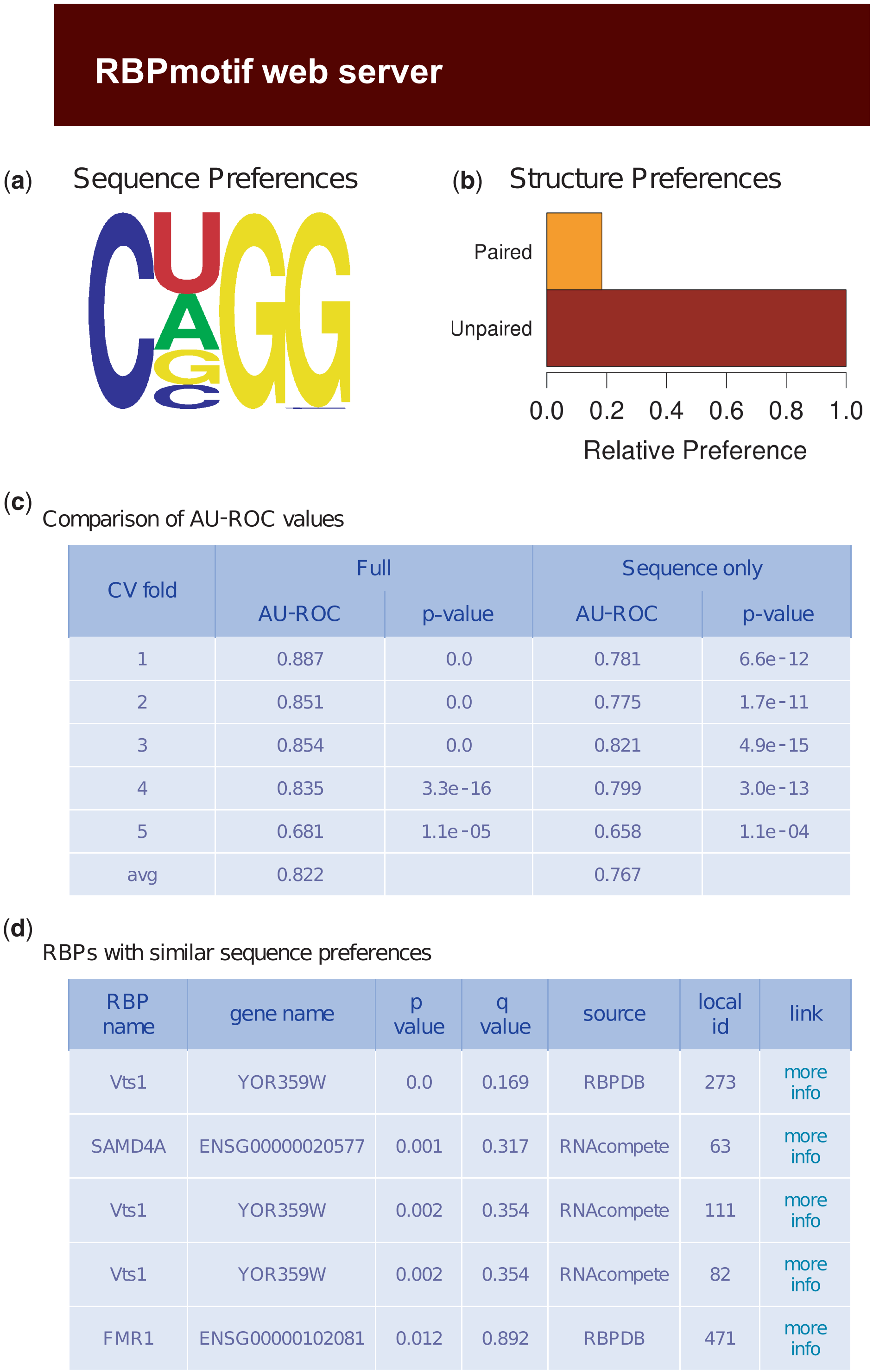
We compute the secondary structure profiles of the input sequences with RNAplfold program. Users can choose between local and global folding techniques. When local folding is chosen, users have to provide two additional arguments of RNAplfold: the length of the local windows (-*W*) and the length of the maximum base pair span (-*L*). When global folding is chosen, we assign -*W* and -*L* parameters equal to the length of the sequence. Lastly, RNAplfold's -*u* option is fixed to 1 so that probabilities of occurrence in structural contexts are determined for each position.

Once the prediction of secondary structures is completed, the set of bound and unbound sequences together with their associated secondary structure profiles are input to the RNAcontext algorithm to search for motifs of specified range of lengths. The discriminative learning strategy of RNAcontext searches for motifs that are enriched in bound sequences in comparison with unbound sequences. To do this, each k-mer (i.e. subsequence of length k) of the sequence is scored with the model parameters, and these k-mer scores are combined to obtain the score of the entire sequence. The noisy-or function that we used to combine k-mer scores in the original RNAcontext article saturates for long RNA sequences. To avoid this problem, here, we switch to summing the k-mer scores to calculate the score of the entire sequence. To optimize RNAcontext parameters, we use the L-BFGS-B package (11) with three random initializations. The model that gives the minimum training error is displayed on the results page.

Once RNAcontext results are ready, we query the predicted sequence motif to identify similar motifs in RBPDB database (12) and RNAcompete compendium (Ray *et al.* 2013, Nature, in press) using the TomTom tool with Pearson distance metric (13). RBPs that have similar sequence preferences are displayed together with the associated TomTom *P*-values and links to the corresponding entries in the original database.

Outputs

Figure 2 shows an example results page where we input two sets of sequences (500 sequences in each set) that are known to be bound and unbound by Vts1p (1). The sequence parameters inferred by RNAcontext correspond to free energy values. These energy values are converted into probabilities using the Boltzmann distribution, and the resulting Position Frequency Matrix (PFM) is displayed as a motif logo [generated by EnoLOGOS software (14)]. The PFM that is used to plot this logo can be downloaded when the image is clicked (Figure 2a). The secondary structure parameters are displayed as a bar graph where *y*-axis shows the relative affinities to different structural contexts (Figure 2b).



Comparison of AU-ROC values

CV fold	Full		Sequence only	
	AU-ROC	p-value	AU-ROC	p-value
1	0.887	0.0	0.781	6.6e-12
2	0.851	0.0	0.775	1.7e-11
3	0.854	0.0	0.821	4.9e-15
4	0.835	3.3e-16	0.799	3.0e-13
5	0.681	1.1e-05	0.658	1.1e-04
avg	0.822		0.767	

RBPs with similar sequence preferences

RBP name	gene name	p value	q value	source	local id	link
Vts1	YOR359W	0.0	0.169	RBPDB	273	more info
SAMD4A	ENSG00000020577	0.001	0.317	RNAcompete	63	more info
Vts1	YOR359W	0.002	0.354	RNAcompete	111	more info
Vts1	YOR359W	0.002	0.354	RNAcompete	82	more info
FMR1	ENSG00000102081	0.012	0.892	RBPDB	471	more info

Figure 2. Result page of the first type of analysis that involves *de novo* motif finding with RNAcontext. (**a**) Inferred sequence preferences are converted into a PFM and plotted as a motif logos using EnoLOGOS (14). (**b**) Structure parameters are scaled so that the most preferred context gets a value of 1. Resulting relative structure preferences are shown as a bar graph. (**c**) To assess the added predictive value of inferred structure preferences, the AU-ROCs of the full RNAcontext model and a simpler version of it that only includes the sequence preferences on held-out data are compared for each cross-validation run. AU-ROCs and their associated *P*-values are displayed as a table. (**d**) The top 5 RBPs with most similar binding motifs [identified with TomTom (13)] to the predicted sequence motif (shown in a) are displayed as a table. The columns of this table show the name of the RBP, name of the gene, *P*-value, *q*-value, local id and link to the original database entry, respectively.

To asses whether the inferred structure preferences have an added predictive value over the sequence preferences, we use 5-fold cross-validation. We train RNAcontext models on four folds and score the sequences in the other fold in two ways: using the full RNAcontext model or using a simpler model that does not include the structure parameters. We calculate the area under the receiver-operator characteristic curve (AU-ROC) to evaluate the predictions, and compare average AU-ROC values between the two models for each of the cross-validation runs (Figure 2c). The table shown in Figure 2c also includes *P*-values showing the significance of AU-ROCs [*P*-value calculation is based on (15)]. The motifs displayed on the results page are learned from the whole input data, and not a subset of it.

Lastly, RBPs that are found to have similar sequence preferences are displayed in a table (Figure 2d). This table shows the name of the RBP, name of the gene, *P*-value and *q*-value calculated by TomTom and link to the corresponding entry in the original database.

Analyzing the secondary structure context of a previously identified motif

This analysis is applicable when the sequence specificity of the RBP of interest is already known. Given a set of bound and unbound sequences, RBP's structure preferences can be queried by comparing the secondary structure profiles of the instances of the motif between bound and unbound sequences.

Inputs and analysis

The user is required to provide the following inputs:

- a set of bound and unbound sequences, as explained in the previous section
- type of the secondary structure representation (allowed options are PU, PLE, PHTE and PHIME)
- RNAplfold parameters to specify the prediction of RNA secondary structure
- IUPAC representation of the previously identified sequence motif

The secondary structure profiles of the input sequences are computed with RNAplfold, as described in the previous section. The occurrences of the input motif are found by scanning the bound and unbound sequences with the IUPAC motif. The structure profiles of instances in bound and unbound sequences are represented as two matrices, where rows correspond to motif instances and columns correspond to structural contexts. An entry of the profile matrix shows the average probability of the motif to appear in a particular structural context. This value is calculated by taking the average of single nucleotide probabilities across the positions of the motif. Pairs of columns from the two profile matrices are compared using Wilcoxon's rank sum test (two-sided) with Bonferroni multiple testing correction. In other words, if the RBP has a preference for a particular structure context, the distribution of probabilities for that structure context among the motif instances should be different between bound and unbound sets.

Outputs

Figure 3 shows an example results page where the bound and unbound sequences that we previously prepared for Vts1p are scanned with the motif CUGG. For each structural context, average profile value across the motif instances (across the rows of the profile matrix) is calculated. These values together with the standard error of the mean are displayed with a bar graph. The data used to plot the bar graph can be downloaded when the figure is clicked (Figure 3a). Additionally, the results of Wilcoxon's rank sum test are displayed as a table where the statistically significant differences are shown (Figure 3b).

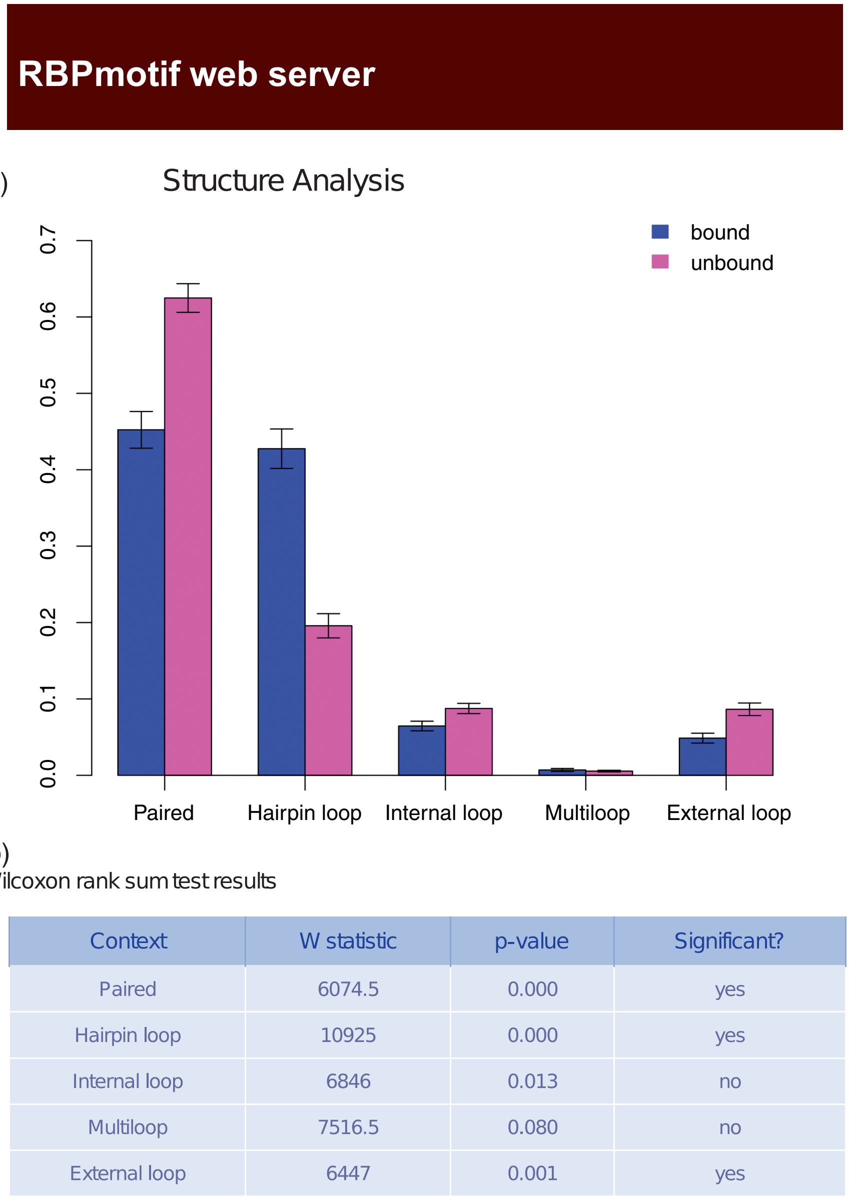
Implementation

The server is implemented in HTML, PHP, Javascript, Python and R. The web service is run on a Red Hat Enterprise Linux 6 with 4× AMD Opteron 6164HE 1.7 GHz 12 core processor and 64 GB memory. To provide results in a reasonable amount of time, we apply a set of limits on the number and length of input sequences. These limits are summarized on the main page. The source code for RNAcontext is also available to download from the help page of RBPmotif web server.

OTHER PROGRAMS FOR IDENTIFYING THE BINDING PREFERENCES OF RBPS

Other methods to identify RBP binding preferences include MEMERIS (16), StructRED (17), CMfinder (18), RNApromo (19), PARalyzer (20) and Aptamotif (21). MEMERIS extends the popular DNA motif-finding algorithm MEME (22) by preferentially searching for single-stranded regions. MEMERIS performs much better than MEME when predicting the binding sites of a number of RBPs. However, MEMERIS lacks a web server and can only assess a single pre-defined structural context. StructRED extends the MatrixREDUCE (23) algorithm to identify stem-loop motifs that explain post-transcriptional events. StructRED lacks a web server implementation and only searches for stem-loop motifs.

Another class of methods that can learn RNA motifs has originated from covariance models (CMs) (24). The applicability of these methods for inferring RBP binding preferences is disputable. For example, CMfinder, also implemented as a web server, is one such model that had promising results in discovering RNA motifs from families of noncoding RNAs. CMfinder preferentially searches for motifs represented as secondary structures and cannot represent the preference of RBPs that bind unpaired, single-stranded RNA, as many RBPs do. Also, the minimum allowed length for a motif is much longer than the typical length of RBP binding sites. Therefore, CMfinder is not suitable for discovering short motifs in a set of long unaligned RNA sequences. RNApromo (available as a web server) is another CM-based method designed to discover local RNA motifs. Like CMfinder, RNApromo is unable to detect a preference for binding unpaired RNA. Aptamotif finds sequence-structure motifs in SELEX (Systematic Evolution of Ligands by EXponential Enrichment)-derived aptamers by adapting the iterative



Wilcoxon rank sum test results

Context	W statistic	p-value	Significant?
Paired	6074.5	0.000	yes
Hairpin loop	10925	0.000	yes
Internal loop	6846	0.013	no
Multiloop	7516.5	0.080	no
External loop	6447	0.001	yes

Figure 3. Result page of the second type of analysis. (**a**)The bar graph compares the mean profile values of motif instances between bound and unbound sequences. The standard errors of the mean are also shown as error bars. (**b**) The table shows the results of Wilcoxon's rank sum test to compare the distribution of structure profiles of motif instances among bound and unbound sequences. The significance threshold for *P*-values is 0.05 after Bonferroni multiple testing correction.

learning procedure of CMs. PARalyzer, not available as a web server, uses a kernel density estimate classifier to identify RBP–RNA interactions sites in photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) (25) data. The interactions sites determined by PARalyzer are further analyzed with a motif-finding algorithm that ignores RNA secondary structure.

CONCLUSION

Most studies investigating the binding preferences of RBPs have ignored the secondary structure preferences of RBPs owing to lack of computational tools. RBPmotif fills this gap by providing user-friendly tools to infer binding preferences of RBPs. Users can either run the *de novo* motif discovery algorithm RNAcontext to identify sequence and structure preferences of the RBP or, as an alternative, they can analyze the structure context of a previously identified sequence motif (when available) to investigate RBP's structure preferences. RBPmotif uses a representation of secondary structure that can detect preferences to multiple structure contexts. Also, the provided results include an assessment of the added predictive value of inferred structure preferences. We think that RBPmotif will be a useful tool for researchers investigating RBP binding specificities.

ACKNOWLEDGEMENTS

The authors thank Harold Rodriguez for setting up the server, Kate Cook for helping with accessing the RBPDB database, Michelle Tran for contributions to an earlier version of the web server, and Wei Jiao for implementing the modification to the original RNAplfold source code.

FUNDING

Canadian Institute of Health Research operating grant [MOP-125894 to Q.M.] and start-up grants (to H.K.). Funding for open access charge: Canadian Institute of Health Research operating grant [MOP-125894 to Q.M.].

Conflict of interest statement. None declared.

REFERENCES

1. Ray,D., Kazan,H., Chan,E., Castillo,L., Chaudhry,S., Talukder,S., Blencowe,B., Morris,Q. and Hughes,T. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.
2. Zhao,J., Ohsumi,T., Kung,J., Ogawa,Y., Grau,D., Sarma,K., Song,J., Kingston,R., Borowsky,M. and Lee,J. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.
3. Ule,J., Jensen,K., Mele,A. and Darnell,R. (2005) CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, **37**, 376–386.
4. Aviv,T., Lin,Z., Ben-Ari,G., Smibert,C. and Sicheri,F. (2006) Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1. *Nat. Struct. Mol. Biol.*, **13**, 168–176.
5. Hogan,D., Riordan,D., Gerber,A., Herschlag,D. and Brown,P. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.*, **6**, e255.
6. Kazan,H., Ray,D., Chan,E., Hughes,T. and Morris,Q. (2010) RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. *PLoS Comput. Biol.*, **6**, e1000832.
7. Wilbert,M., Huelga,S., Kapeli,K., Stark,T., Liang,T., Chen,S., Yan,B., Nathanson,J., Hutt,K., Lovci,M. *et al.* (2012) LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance. *Mol. Cell*, **48**, 195–206.
8. Ding,Y. and Lawrence,C. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, **31**, 7280–7301.
9. Bernhart,S., Hofacker,I. and Stadler,P. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
10. Lange,S., Maticzka,D., Mohl,M., Gagnon,J., Brown,C. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.
11. Byrd,R., Lu,P. and Nocedal,J. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comp.*, **16**, 1190–1208.
12. Cook,K., Kazan,H., Zuberi,K., Morris,Q. and Hughes,T. (2010) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.
13. Gupta,S., Stamatoyannopolous,J., Bailey,T. and Noble,W. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
14. Workman,C., Yin,Y., Corcoran,D., Ideker,T., Stormo,G. and Benos,P. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.
15. Hanley,J. and McNeil,B. (1982) The meaning and use of a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.
16. Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.
17. Foat,B. and Stormo,G. (2009) Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol. Syst. Biol.*, **5**, 268.
18. Yao,Z., Weinberg,Z. and Ruzzo,W. (2006) CMfinder: A covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
19. Rabani,M., Kersetz,M. and Segal,E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA*, **105**, 14885–14890.
20. Corcoran,D.L., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R.L., Keene,J.D. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.
21. Hoinka,J., Zotenko,E., Friedman,A., Sauna,Z. and Przytycka,T. (2012) Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, **28**, i215–i223.
22. Bailey,T., Williams,N., Misleh,C. and Li,W. (2006) MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
23. Foat,B., Morozov,A. and Bussemaker,H. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.
24. Eddy,S. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
25. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A.C., Munschauer,M. *et al.* (2010) PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp.*, **41**, pii:2034.