Design of an RNA structural motif database

Dongrong Wen and Jason T.L. Wang*

Bioinformatics Program and Department of Computer Science,
New Jersey Institute of Technology,
Newark, NJ 07102, USA
E-mail: dw39@njit.edu
E-mail: wangj@njit.edu
*Corresponding author

Abstract: In this paper we present the design and implementation of an RNA structural motif database, called RmotifDB. The structural motifs stored in RmotifDB come from three sources:

- collected manually from the biomedical literature
- submitted by scientists around the world
- discovered by a wide variety of motif mining methods.

We present here a motif mining method in detail. We also describe the interface and search mechanisms provided by RmotifDB and report its current status. The RmotifDB system is fully operational and accessible on the web at http://datalab.njit.edu/bioinfo/.

Keywords: RNA; untranslated region; structural motif; database.

Reference to this paper should be made as follows: Wen, D. and Wang, J.T.L. (2009) 'Design of an RNA structural motif database', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.32–41.

Biographical notes: Dongrong Wen is currently a PhD candidate of Computer Science in the College of Computing Sciences at New Jersey Institute of Technology under supervision of Dr. Jason T.L. Wang. He completed his BS Degree in Information Engineering and Computer Science at Feng Chia University, Taichung, Taiwan, and MS Degree in Computer Science from the Courant Institute of Mathematical Sciences, New York University. His research interests include bioinformatics, databases and data mining.

Jason T.L. Wang received the BS Degree in Mathematics from National Taiwan University, Taipei, Taiwan, and the PhD Degree in Computer Science from the Courant Institute of Mathematical Sciences, New York University, in 1991. He is a full Professor of Computer Science in the College of Computing Sciences at New Jersey Institute of Technology and Director of the University's Data and Knowledge Engineering Laboratory. His research interests include data mining and databases, bioinformatics, and cyber infrastructure. He has published over 120 refereed journal and conference papers as well as five books in these areas.

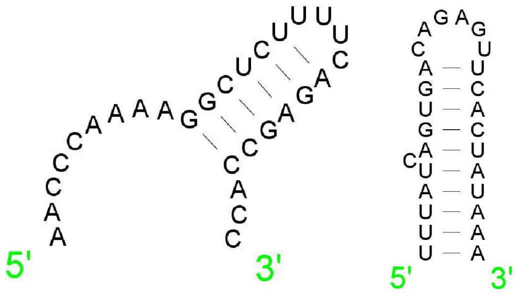
1 Introduction

Post-transcriptional control is one of the mechanisms that regulate gene expression in eukaryotic cells. RNA elements residing in the Untranslated Regions (UTRs) of mRNAs have been shown to play various roles in post-transcriptional control, including mRNA localisation, translation, and mRNA stability (Keene and Tenenbaum, 2002; Kuersten and Goodwin, 2003; Mignone et al., 2002; Wilkie et al., 2003). RNA elements in UTRs can be roughly divided into two groups: elements whose functions are primarily attributable to their sequences and elements whose functions are attributable to their secondary or tertiary structures. For simplicity, they are called sequence elements and structure elements respectively. Well-known sequence elements include AU-Rich Elements (AREs), some of which contain one or several tandem AUUUA sequences and are involved in regulating mRNA stability (Bakheet et al., 2001; Chen and Shyu, 1995; Wilusz and Wilusz, 2004), and miRNA target sequences, which are partially complementary to cognate miRNA sequences and are involved in regulating translation or mRNA stability (Bartel, 2004; John et al., 2004; Lewis et al., 2003).

Among all structure elements in UTRs, the histone 3’-UTR stem-loop structure (HSL3) and the Iron Response Element (IRE) have been most extensively studied (Marzluff and Duronio, 2002; Pesole et al., 2002). Both sequence and structure are important for the functions of structure elements. HSL3 is a stem-loop structure of about 25 nucleotides (nt) that exists in 3’-UTRs of most histone genes. Figure 1(a) shows the graphical representation of HSL3 drawn by XRNA (http://rna.ucsc.edu/rnacenter/xrna /xrna.html). The HSL3 structure is critical for both termination of their transcription and stability of mRNAs. These functions are exerted by the Stem-Loop Binding Protein (SLBP) that interacts with HSL3.

IRE is a stem-loop structure of ~30 nt with a bulge or a small internal loop in the stem. Figure 1(b) shows the graphical representation of IRE drawn by XRNA. IREs have been found in both 5’-UTRs and 3’-UTRs of mRNAs whose products are involved in iron homeostasis in higher eukaryotic species. IREs bind to the Iron Regulatory Proteins (IRPs) of these species, which control translation and stability of IRE-containing mRNAs.

Figure 1 (a) An example of the HSL3 motif. and (b) an example of the IRE motif (see online version for colours)



(a) (b)

HSL3 and IRE are similar in several aspects: both are small simple RNA structures with less than 40 nt; both exist in UTRs of several genes with related functions; and both bind to cellular proteins and are involved in post-transcriptional gene regulation. The regulations via HSL3 and IRE constitute a distinct mode of gene regulation, whereby expression of several genes can be modulated via a common RNA structure in UTRs.

Functional sequence motifs in genomes have been heavily studied in recent years, particularly for the promoter region and sequences involved in splicing (Blanchette and Tompa, 2002; Boffelli et al., 2003; Fairbrother et al., 2002; Marino-Ramirez et al., 2004; Smith et al., 2005; Xie et al., 2005). In contrast, RNA structure elements have been investigated to a much lesser extent, largely due to the difficulties in predicting correct RNA structures and conducting RNA structure alignments, where huge computing costs are involved. While some success has been achieved using phylogenetic approaches to gain accuracy in RNA structure prediction (Akmaev et al., 2000; Rivas et al., 2001; Washietl and Hofacker, 2004), large-scale mining for conserved structures in eukaryotic UTRs has not been attempted. In addition, current methods for finding common stem-loop structures solely rely on structure similarities (Gorodkin et al., 2001). Gene Ontology information has not been used in the study of RNA structures.

Here we present a database, called RmotifDB, that contains structure elements (or structural motifs) found in 5' and 3' UTRs of eukaryotic mRNAs. The structural motifs are linked with Gene Ontology entries concerning the motifs. A wide variety of motif mining methods are developed. In particular, we present here a histogram-based method for discovering motifs in eukaryotic UTRs. In the following section we describe RmotifDB and its search interface. In Section 3, we describe the histogram-based method in detail. Finally we conclude the paper and point out some directions for future research.

2 The RmotifDB system

RmotifDB is designed for storing the RNA structural motifs found in the UTRs of eukaryotic mRNAs. RmotifDB is a web-based database system which supports retrieval and access of RNA structural motifs from its database. The system allows the user to search RNA structural motifs in an effective and friendly way. RmotifDB is accessible on the web at http://datalab.njit.edu/bioinfo/.

The RNA structural motifs stored in RmotifDB come from three sources:

- collected manually from the biomedical literature
- submitted by scientists around the world
- discovered by a wide variety of motif mining methods.

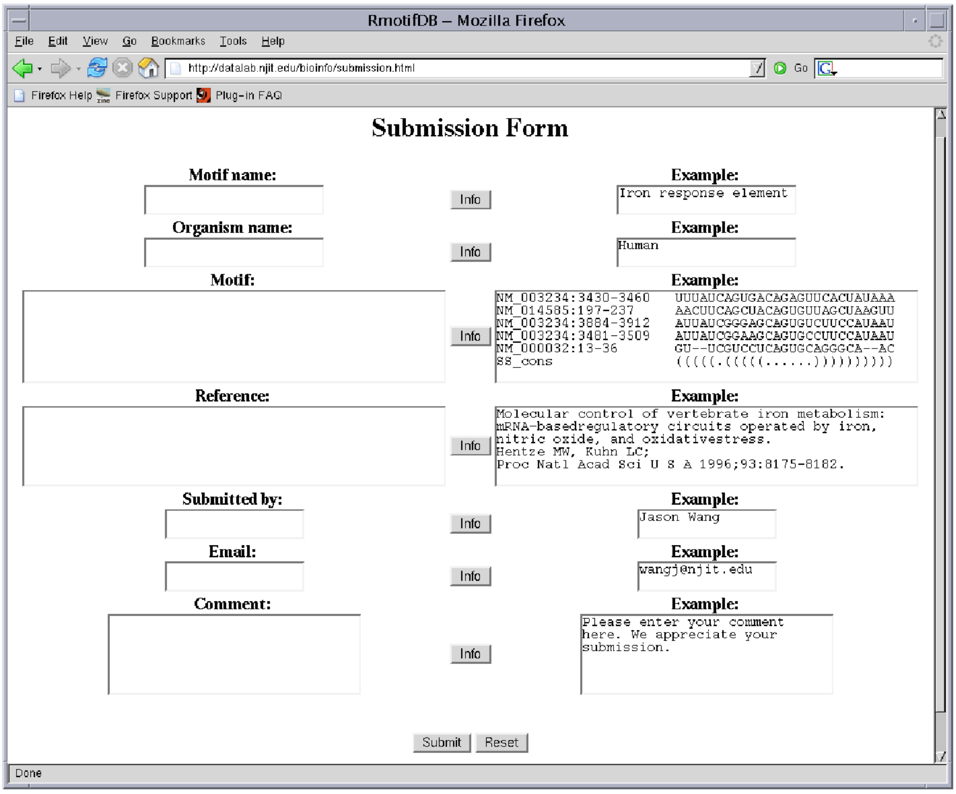
Figure 2 shows the interface where scientists can submit an RNA structural motif.

Figure 3 shows the search interface of RmotifDB. The system provides two search options: Query By Sequence (QBS) and Query By Structure (QBR). With QBS, the user enters an RNA sequence in the standard FASTA format and the system matches this query sequence with motifs in the database using either our previously developed RNA alignment algorithm RSmatch (Liu et al., 2005) or Infernal (Eddy, 2002). Since RSmatch accepts, as input data, RNA secondary structures only, the system needs to invoke Vienna

RNA v1.4 (Hofacker, 2003) to fold the query sequence into a structure before a match is performed. With QBR, the user enters an RNA secondary structure represented by the Vienna style Dot Bracket format and the system matches this query structure with motifs in the database using RSmatch. The result is a ranked list of motifs that are approximately contained in the query sequence or the query structure. In addition, the user can search RmotifDB by choosing a Gene ID or RefSeq ID from a pre-defined list of Gene IDs and RefSeq IDs obtained from http://www.ncbi.nlm.nih.gov/RefSeq/ and provided by the RmotifDB system. This pre-defined list contains the IDs of the mRNA sequences used by our motif mining methods to discover the structural motifs stored in RmotifDB. The result of this search is a list of structural motifs found in the input mRNA sequence.

The user can click each motif to see detailed information concerning the motif. Figure 4 shows the result of displaying a motif and its related information. Here the motif is an Iron Response Element (IRE) in human shown in the Stockholm format (Eddy, 2002), which is a multiple sequence alignment with structural annotation in the Vienna style Dot Bracket format. The graphical representation of the motif is shown in the bottom right-hand corner of the window. Also displayed are the Gene Ontology information concerning the motif and relevant articles that publish this motif.

Figure 2 The interface of RmotifDB where scientists can submit an RNA structural motif (see online version for colours)


RmotifDB – Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://datalab.njit.edu/bioinfo/submission.html
Go
Firefox Help
Firefox Support
Plug-in FAQ
Info
Submit
Reset
Done

Submission Form

Motif name:

Example: Iron response element

Organism name:

Example: Human

Motif:

Example:

```
NM_003234:3430-3460   UUUAUCAGUGACAGAGUUCACUAUAAA
NM_014585:197-237     AACUUCAGCUACAGUGUUAGCUAAGUU
NM_003234:3884-3912   AUUAUCGGGAGCAGUGUCUUCCAUAAU
NM_003234:3481-3509   AUUAUCGGAAGCAGUGCCUUCCAUAAU
NM_000032:13-36       GU--UCGUCCUCAGUGCAGGGCA--AC
SS_cons               (((((.((((((......))))))))))
```

Reference:

Example:

Molecular control of vertebrate iron metabolism: mRNA-basedregulatory circuits operated by iron, nitric oxide, and oxidativestress.
Hentze MW, Kuhn LC;
Proc Natl Acad Sci U S A 1996;93:8175-8182.

Submitted by:

Example: Jason Wang

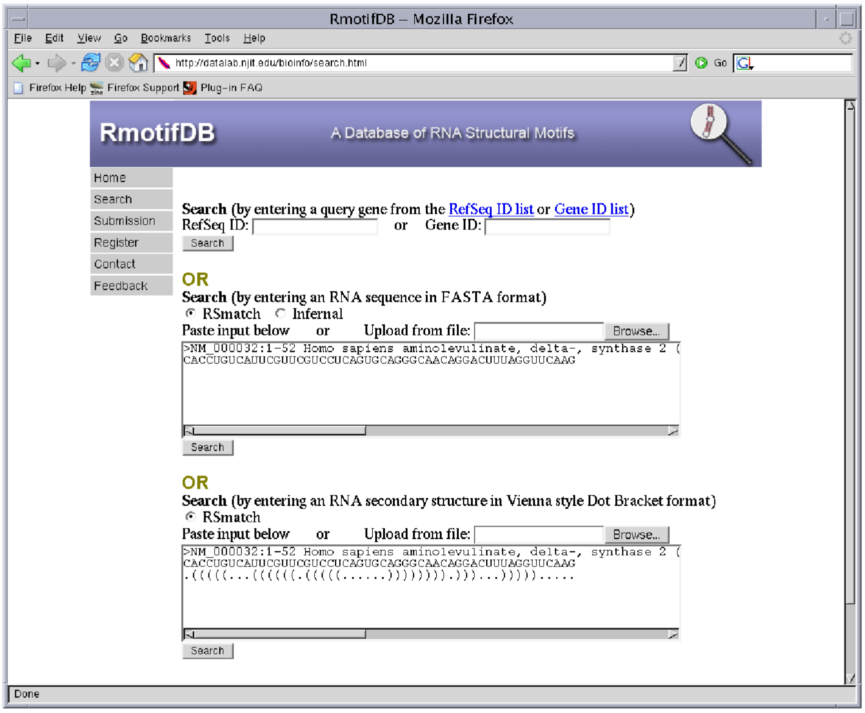
Email:

Example: wangj@njit.edu

Comment:

Example: Please enter your comment here. We appreciate your submission.

Figure 3 The search interface of RmotifDB (see online version for colours)



RmotifDB – Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://datalab.njit.edu/bioinfo/search.html
Go
Firefox Help Firefox Support Plug-in FAQ
RmotifDB
A Database of RNA Structural Motifs
Home
Search
Submission
Register
Contact
Feedback
Done

Search (by entering a query gene from the RefSeq ID list or Gene ID list)
RefSeq ID: ______ or Gene ID: ______
Search

OR

Search (by entering an RNA sequence in FASTA format)
● RSmatch ○ Infernal
Paste input below or Upload from file: ______ Browse...

```
>NM_000032:1-52 Homo sapiens aminolevulinate, delta-, synthase 2 (
CACCUGUCAUUCGUUCGUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG
```

Search

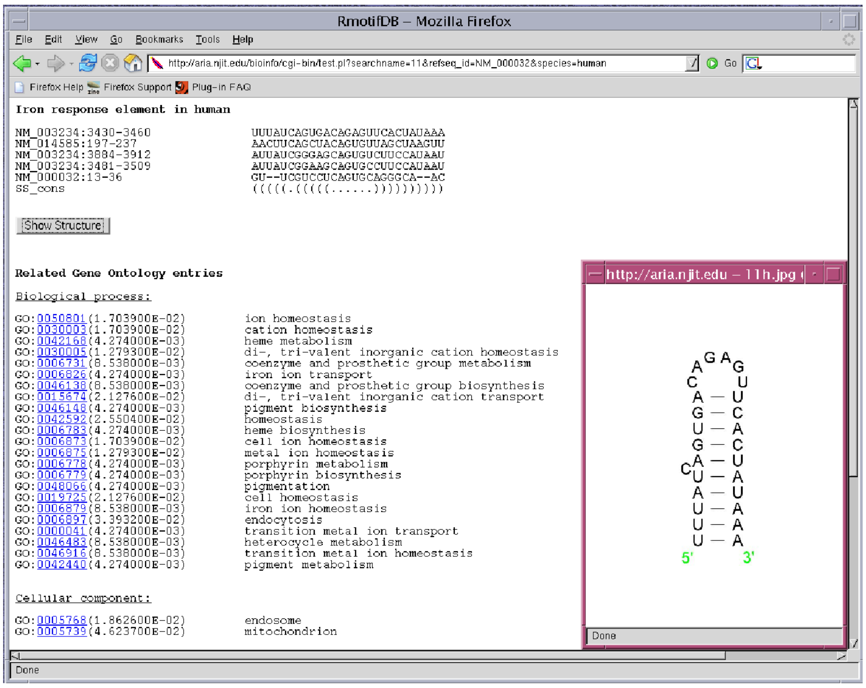
OR

Search (by entering an RNA secondary structure in Vienna style Dot Bracket format)
● RSmatch
Paste input below or Upload from file: ______ Browse...

```
>NM_000032:1-52 Homo sapiens aminolevulinate, delta-, synthase 2 (
CACCUGUCAUUCGUUCGUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG
.((((...((((((.((((((......))))))))).)))...))))).....
```

Search

Figure 4 The output showing a structural motif stored in RmotifDB and related information (see online version for colours)



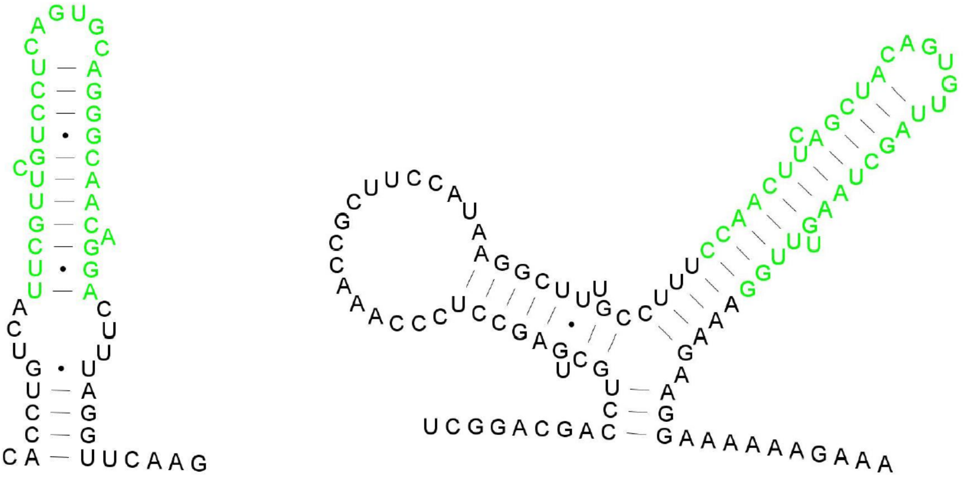
RmotifDB – Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://aria.njit.edu/bioinfo/cgi-bin/test.pl?searchname=11&refseq_id=NM_000032&species=human
Go
Firefox Help Firefox Support Plug-in FAQ
Show Structure
http://aria.njit.edu – 11h.jpg
A G A G
C U
A — U
G — C
U — A
G — C
CA — U
U — A
A — U
U — A
U — A
U — A
5' 3'
Done
Done

```
Iron response element in human

NM_003234:3430-3460              UUUAUCAGUGACAGAGUUCACUAUAAA
NM_014585:197-237                AACUUCAGCUACAGUGUUAGCUAAGUU
NM_003234:3884-3912              AUUAUCGGGAGCAGUGUCUUCCAUAAU
NM_003234:3481-3509              AUUAUCGGAAGCAGUGCCUUCCAUAAU
NM_000032:13-36                  GU--UCGUCCUCAGUGCAGGGCA--AC
SS_cons                          (((((.((((((......))))))))))


Related Gene Ontology entries

Biological process:

GO:0050801(1.703900E-02)       ion homeostasis
GO:0030003(1.703900E-02)       cation homeostasis
GO:0042168(4.274000E-03)       heme metabolism
GO:0030005(1.279300E-02)       di-, tri-valent inorganic cation homeostasis
GO:0006731(8.538000E-03)       coenzyme and prosthetic group metabolism
GO:0006826(4.274000E-03)       iron ion transport
GO:0046138(8.538000E-03)       coenzyme and prosthetic group biosynthesis
GO:0015674(2.127600E-02)       di-, tri-valent inorganic cation transport
GO:0046148(4.274000E-03)       pigment biosynthesis
GO:0042592(2.550400E-02)       homeostasis
GO:0006783(4.274000E-03)       heme biosynthesis
GO:0006873(1.703900E-02)       cell ion homeostasis
GO:0006875(1.279300E-02)       metal ion homeostasis
GO:0006778(4.274000E-03)       porphyrin metabolism
GO:0006779(4.274000E-03)       porphyrin biosynthesis
GO:0048066(4.274000E-03)       pigmentation
GO:0019725(2.127600E-02)       cell homeostasis
GO:0006879(8.538000E-03)       iron ion homeostasis
GO:0006897(3.393200E-02)       endocytosis
GO:0000041(4.274000E-03)       transition metal ion transport
GO:0046483(8.538000E-03)       heterocycle metabolism
GO:0046916(8.538000E-03)       transition metal ion homeostasis
GO:0042440(4.274000E-03)       pigment metabolism


Cellular component:

GO:0005768(1.862600E-02)       endosome
GO:0005739(4.623700E-02)       mitochondrion
```

3 A motif mining method

We have developed several structural motif mining methods based on different RNA representation models. For example, in Chang et al. (1998), Wang et al. (1996, 1998) we represented an RNA secondary structure using an ordered labelled tree and designed a tree matching algorithm to find motifs in multiple RNA secondary structures. More recently we developed a loop model for representing RNA secondary structures. Based on this loop model we designed a dynamic programming algorithm, called RSmatch (Liu et al., 2005), for comparing two RNA secondary structures. The time complexity of RSmatch is $O(mn)$, where m and n are the size of the two compared structures respectively. Figure 5 shows the alignment of two RNA secondary structures produced by RSmatch and drawn by XRNA.

Figure 5 Alignment of two RNA secondary structures where the local matches found by RSmatch are highlighted with the (light) green colour (see online version for colours)



We describe below a histogram-based scoring method to uncover novel conserved RNA stem-loops in eukaryotic UTRs using the RSmatch tool. This method is an upward extension of our previously developed histogram-based algorithm for DNA sequence classification (Wang et al., 1999a). Given a set of RNA secondary structures, the method uses RSmatch to perform pairwise alignments by comparing two RNA structures at a time in the set. Given an optimal local alignment between two structures A and B found by RSmatch, the set of bases in the aligned region of A is denoted by $Q_A = \{A_i, A_{i+1}, \dots, A_j\}$ where A_i (A_j, respectively) is the 5'-most (3'-most, respectively) nucleotide not aligned to a gap. The set of bases in the aligned region of B is denoted by $Q_B = \{B_m, B_{m+1}, \dots, B_n\}$ where B_m (B_n, respectively) is the 5'-most (3'-most, respectively) nucleotide not aligned to a gap. Each nucleotide $A_k \in Q_A$ that is not aligned to a gap scores $| j - \mathrm{i} + 1 |$ points. All the other bases in the structure A receive 0 point. Thus, the larger the aligned region between A and B, the higher score each base in the region has. When aligning the structure A with another structure C, some bases in Q_A may receive non-zero points and hence the scores of those bases are accumulated. Thus, the bases in a conserved RNA motif will have high scores.

To validate our approach, we conducted experiments to evaluate the effectiveness of this scoring method. The conserved stem-loops we considered were IRE motifs, which contained about 30 nucleotides, located in the 5'-UTRs or 3'-UTRs of mRNAs coding for proteins involved in cellular iron metabolism. The test dataset was prepared as follows. By searching human RefSeq mRNA sequences from the National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov/RefSeq/, we obtained several mRNA sequences, within each of which at least one IRE motif is known to exist. We then extracted the sequences' UTR regions as indicated by RefSeq's GenBank annotation and used PatSearch (Grillo et al., 2003) to locate the IRE sequences. Each IRE sequence was then extended from both ends to obtain a 100 nt sequence. These sequences were mixed with several 'noisy' sequences with the same length. All the resulting sequences were then folded by the Vienna RNA package (Hofacker, 2003) using the 'RNAsubopt' function with setting '–e 0'. This setting can yield multiple RNA structures with the same free energy for any given RNA sequence.

Figure 6 shows the score histograms for three tested RNA structures. It was observed that clusters of bases with high scores correspond to the IRE motifs in the RNA structures. Similar clusters of bases with high scores corresponding to the IRE motifs were observed in the other IRE-containing RNA structures, but not in the 'noisy' structures. This result indicates that our histogram-based scoring method is able to detect biologically significant motifs in multiple RNA structures.

Figure 6 Diagrams illustrating the effectiveness of the proposed scoring method

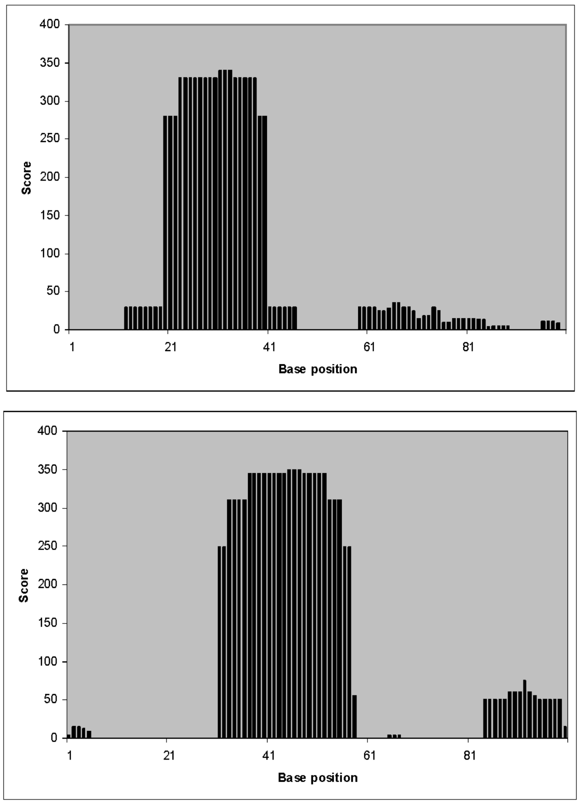
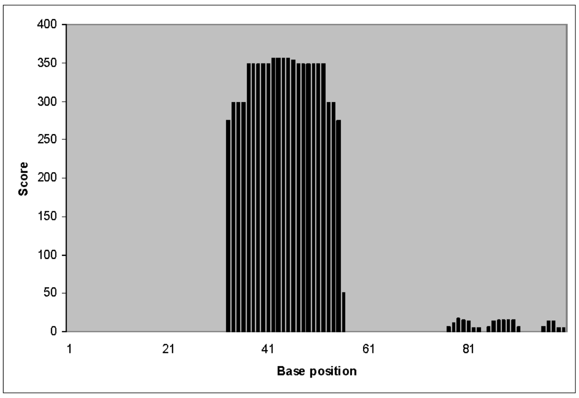


Figure 6 Diagrams illustrating the effectiveness of the proposed scoring method (continued)



4 Conclusion

We presented in the paper an RNA structural motif database, called RmotifDB, and described some features of RmotifDB. We also developed a motif mining method capable of discovering structural motifs in eukaryotic mRNAs. The system presented here is part of a long-term project (Khaladkar et al., 2006; Wang and Wu, 2006) which aims to build a cyber infrastructure for RNA data analysis and mining (http://bioinformatics.njit.edu/rna/). Data mining in bioinformatics has emerged as an important discipline at the interface of information technology and molecular biology (Wang et al., 1999b, 2003, 2005). Our cyber infrastructure will contribute to the field of data mining and bioinformatics in general, and RNA informatics in particular. In future work we plan to develop new data mining methods and tools for RNA structural alignment, classification, clustering and motif discovery with applications to various organisms.

Acknowledgement

We are grateful to Dr. Bin Tian of New Jersey Medical School and Dr. Jianghui Liu for helpful discussions while preparing this paper.

References

Akmaev, V.R., Kelley, S.T. and Stormo, G.D. (2000) 'Phylogenetically enhanced statistical tools for RNA structure prediction', *Bioinformatics*, Vol. 16, No. 6, pp.501–512.

Bakheet, T., Frevel, M., Williams, B.R., Greer, W. and Khabar, K.S. (2001) 'ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins', *Nucleic Acids Research*, Vol. 29, No. 1, pp.246–254.

Bartel, D.P. (2004) 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell*, Vol. 116, No. 2, pp.281–297.

Blanchette, M. and Tompa, M. (2002) 'Discovery of regulatory elements by a computational method for phylogenetic footprinting', *Genome Research*, Vol. 12, No. 5, pp.739–748.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. and Rubin, E.M. (2003) 'Phylogenetic shadowing of primate sequences to find functional regions of the human genome', *Science*, Vol. 299, No. 5611, pp.1391–1394.

Chang, C-Y., Wang, J.T.L. and Chang, R.K. (1998) 'Scientific data mining: a case study', *International Journal of Software Engineering and Knowledge Engineering*, Vol. 8, No. 1, pp.77–96.

Chen, C.Y. and Shyu, A.B. (1995) 'AU-rich elements: characterization and importance in mRNA degradation', *Trends in Biochemical Sciences*, Vol. 20, No. 11, pp.465–470.

Eddy, S.R. (2002) 'A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure', *BMC Bioinformatics*, Vol. 3, No. 18.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A. and Burge, C.B. (2002) 'Predictive identification of exonic splicing enhancers in human genes', *Science*, Vol. 297, No. 5583, pp.1007–1013.

Gorodkin, J., Stricklin, S.L. and Stormo, G.D. (2001) 'Discovering common stem-loop motifs in unaligned RNA sequences', *Nucleic Acids Research*, Vol. 29, No. 10, pp.2135–2144.

Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. and Pesole, G. (2003) 'PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences', *Nucleic Acids Research*, Vol. 31, No. 13, pp.3608–3612.

Hofacker, I.L. (2003) 'Vienna RNA secondary structure server', *Nucleic Acids Research*, Vol. 31, No. 13, pp.3429–3431.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. and Marks, D.S. (2004) 'Human MicroRNA targets', *PLoS Biology*, Vol. 2, No. 11, p.e363.

Keene, J.D. and Tenenbaum, S.A. (2002) 'Eukaryotic mRNPs may represent posttranscriptional operons', *Molecular Cell*, Vol. 9, No. 6, pp.1161–1167.

Khaladkar, M., Bellofatto, V., Wang, J.T.L., Tian, B. and Zhang, K. (2006) 'RADAR: an interactive web-based toolkit for RNA data analysis and research', *Proceedings of the 6th IEEE Symposium on Bioinformatics and Bioengineering*, Arlington, Virginia, USA, pp.209–212.

Kuersten, S. and Goodwin, E.B. (2003) 'The power of 3' UTR: translational control and development', *Nature Reviews Genetics*, Vol. 4, No. 8, pp.626–637.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) 'Prediction of mammalian micro RNA targets', *Cell*, Vol. 115, No. 7, pp.787–798.

Liu, J., Wang, J.T.L., Hu, J. and Tian, B. (2005) 'A method for aligning RNA secondary structures and its application to RNA motif detection', *BMC Bioinformatics*, Vol. 6, No. 89.

Marino-Ramirez, L., Spouge, J.L., Kanga, G.C. and Landsman, D. (2004) 'Statistical analysis of over-represented words in human promoter sequences', *Nucleic Acids Research*, Vol. 32, No. 3, pp.949–958.

Marzluff, W.F. and Duronio, R.J. (2002) 'Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences', *Current Opinion in Cell Biology*, Vol. 14, No. 6, pp.692–699.

Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) 'Untranslated regions of mRNAs', *Genome Biology*, Vol. 3, No. 3, reviews 0004.1–0004.10.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C. and Saccone, C. (2002) 'UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs', *Nucleic Acids Research*, Vol. 30, No. 1, pp.335–340.

Rivas, E., Klein, R.J., Jones, T.A. and Eddy, S.R. (2001) 'Computational identification of noncoding RNAs in E. coli by comparative genomics', *Current Biology*, Vol. 11, No. 17, pp.1369–1373.

Smith, A.D., Sumazin, P. and Zhang, M.Q. (2005) 'Identifying tissue-selective transcription factor binding sites in vertebrate promoters', *Proceedings of the National Academy of Sciences USA*, Vol. 102, No. 5, pp.1560–1565.

Wang, J.T.L. and Wu, X. (2006) 'Kernel design for RNA classification using support vector machines', *International Journal of Data Mining and Bioinformatics*, Vol. 1, No. 1, pp.57–76.

Wang, J.T.L., Rozen, S., Shapiro, B.A., Shasha, D., Wang, Z. and Yin, M. (1999a) 'New techniques for DNA sequence classification', *Journal of Computational Biology*, Vol. 6, No. 2, pp.209–218.

Wang, J.T.L., Shapiro, B.A. and Shasha, D. (Eds.) (1999b) *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications*, Oxford University Press, New York.

Wang, J.T.L., Shapiro, B.A., Shasha, D., Zhang, K. and Chang, C-Y. (1996) 'Automated discovery of active motifs in multiple RNA secondary structures', *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, pp.70–75.

Wang, J.T.L., Shapiro, B.A., Shasha, D., Zhang, K. and Currey, K.M. (1998) 'An algorithm for finding the largest approximately common substructures of two trees', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp.889–895.

Wang, J.T.L., Wu, C.H. and Wang, P.P. (Eds.) (2003) *Computational Biology and Genome Informatics*, World Scientific Publishing Company, Singapore.

Wang, J.T.L., Zaki, M.J., Toivonen, H.T.T. and Shasha, D. (Eds.) (2005) *Data Mining in Bioinformatics*, Springer, London, New York.

Washietl, S. and Hofacker, I.L. (2004) 'Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics', *Journal of Molecular Biology*, Vol. 342, No. 1, pp.19–30.

Wilkie, G.S., Dickson, K.S. and Gray, N.K. (2003) 'Regulation of mRNA translation by 5' and 3'-UTR-binding factors', *Trends in Biochemical Sciences*, Vol. 28, No. 4, pp.182–188.

Wilusz, C.J. and Wilusz, J. (2004) 'Bringing the role of mRNA decay in the control of gene expression into focus', *Trends in Genetics*, Vol. 20, No. 10, pp.491–497.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) 'Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals', *Nature*, Vol. 434, No. 7031, pp.338–345.