Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Genome-wide Search for Coaxial Helical Stacking Motifs

Kevin Byron, Jason T. L. Wang, Dongrong Wen

Bioinformatics Program and Department of Computer Science
New Jersey Institute of Technology
Newark, New Jersey 07102, USA
{byron, wangj, dw39}@njit.edu

***Abstract*—Motif finding in DNA, RNA and proteins plays an important role in life science research. In this paper, we present a computational approach to searching for RNA tertiary motifs in genomic sequences. Specifically, we describe a method, named CSminer, and show, as a case study, the application of CSminer to genome-wide search for coaxial helical stackings in RNA 3-way junctions. A coaxial helical stacking motif occurs in an RNA 3-way junction where two separate helical elements form a pseudocontiguous helix and provide thermodynamic stability to the RNA molecule as a whole. Experimental results demonstrate the effectiveness of our approach.**

***Keywords*- coaxial helical stacking, genome-wide motif finding, RNA junction**

I. INTRODUCTION

Motif finding in DNA, RNA and proteins plays an important role in life science research. Here, we present a method, named CSminer (i.e. Coaxial helical Stacking miner), for finding coaxial helical stackings in genomes. A coaxial helical stacking occurs in an RNA tertiary structure where two separate helical elements form a pseudocontiguous helix [1]. Coaxial helical stacking motifs occur in several large RNA structures, including tRNA [2], pseudoknots [3], group II intron [4] and large ribosomal subunits [5][6][7]. Coaxial helical stackings provide thermodynamic stability to the molecule as a whole [8][9], and reduce the separation between loop regions within junctions [10]. Moreover, coaxial helical stacking interactions form cooperatively with long-range interactions in many RNAs [11] and are thus essential features that distinguish different junction topologies.

Research to unravel the mysteries of (non-coding) RNA is exciting. An unexpected preliminary result of the human ENCODE project indicates that whereas protein-coding sequences (i.e. coding RNA) occupy less than 2% of the human genome, close to 93% of the genome is transcribed into non-coding RNA [12]. The "RNA World" hypothesis proposes that life based on RNA pre-dates the current world of life based on DNA, RNA and proteins [13]. Specialized RNA literature continually emerges [14]. The function of RNA is believed to be closely associated with its 3D structure, which, by virtue of canonical Watson-Crick base pairings (i.e. AU, GC) and wobble base pairing (i.e. GU), is largely determined by its secondary structure [15][16][17]. Many secondary structure prediction tools are available. One of the more highly regarded of these tools is Infernal [18] which has been, and continues to be, frequently cited [19] [20]. Infernal applies stochastic context-free grammar methodology to efficiently predict (non-coding) RNA secondary structures in genome-wide searches [21][22][23]. Databases detailing the 3D structure and features of RNA continue to grow [24][25]. Special interest is paid to RNA junctions [26][27] in which there are one or more coaxial helical stackings [28][29]. Statistical analysis approaches, in particular, ensemble-based approaches, have been successful in non-life science applications [30][31]. Recently, these ensemble-based approaches have been successful in the field of bioinformatics [32][33][34][35][36]. We apply an ensemble-based approach, namely random forests, to predict the existence of a coaxial helical stacking in RNA junctions [1]. In this paper, we extend the functionality of Infernal to create a tool, named CSminer, which can efficiently predict the existence of coaxial helical stackings in genomes. This is accomplished by invoking a random forests classifier within Infernal and filtering Infernal results appropriately. Changes to the Infernal source code are available from the authors upon request.

II. MATERIAL AND METHODS

A. RNA 3-Way Junctions

For this work, we selected samples from known RNA junctions. There are multiple ways for an RNA junction to exist [37]. As a case study, we focus on 3-way junctions here. In [1], we studied 110 distinct RNA 3-way junctions confirmed in available crystal structures. Each 3-way junction contains a multi-branch loop (i.e. MBL) with three helices. Each of these 110 unique junctions is verified in one of 32 crystal structure molecules in PDB [24]. The majority, 75%, of these 110 3-way junctions are found in the relatively complex ribosome subunit molecules, i.e. 51% in 23S rRNA, 20% in 16S rRNA and 4% in 5S rRNA. There is no dominant topological configuration among these 110 3-way junctions in that 47% are categorized as family type C, 35% as family type A and the remaining 18% as family type B [1]. For each of these 110 3-way junctions, the coaxial helical stacking status is known, and the status is one of these four possibilities: H1H2, H1H3, H2H3 or none, where HxHy indicates that helix Hx shares a common axis with helix Hy.

Following [1], a 3-way junction is described by three RNA subsequences. For each subsequence, base coordinates and base values (i.e. A, C, G, U) are known. The starting and ending coordinates of each subsequence indicate the 5'

and 3' ends of the subsequence respectively. The 3-way junction formed by these three subsequences includes unpaired bases of the MBL, terminal base pairs of the three helices and the penultimate (i.e. "next-to-last") base pairs of the three helices, as follows. The 5' end of the first subsequence is the 5' base of the penultimate base pair of helix H1. The 3' end of the first subsequence is the 5' base of the penultimate base pair of helix H2. Similarly, the 5' end of the second subsequence is the 3' base of the penultimate base pair of helix H2 and the 3' end of the second subsequence is the 5' base of the penultimate base pair of helix H3. It follows that the 5' end of the third subsequence is the 3' base of the penultimate base pair of helix H3 and the 3' end of the third subsequence is the 3' base of the penultimate base pair of helix H1.

The length of each subsequence is at least 4. The first two bases of each subsequence are part of one helix and the last two bases of that subsequence are part of the next sequential helix. There are zero or more unpaired bases between the two helices that share a subsequence. Unpaired bases of each subsequence are referred to as part of the "loop regions" of the MBL and are used to help determine the coaxial helical stacking status of the 3-way junction as described later. As an example, we illustrate in Figures 1, 2, 3 and 4 a 3-way junction in PDB molecule 2J00, i.e. "Structure of the 70S ribosome complexed with mRNA and tRNA." This 3-way junction has a coaxial helical stacking identified as H1H2, i.e. helices H1 and H2 share a common axis. The RNA segment from position 1072 through 1103 is shown graphically in Figures 1, 3 and 4. Figure 1, obtained using RNAview [38], illustrates helices H1 and H2 aligned with a common axis. In addition to the canonical Watson-Crick base pairings (i.e. AU, GC) and wobble base pairings (i.e. GU), Figure 1 also illustrates tertiary interactions between bases. The primary sequence of RNA chain A obtained from PDB with highlighted 3-way junction subsequences is shown in Figure 2. The 2D structure plot for this RNA segment from position 1072 through 1103 is shown in Figure 3, obtained using S2S [39] and VARNA [40].

In Figure 3, the 3-way junction is enclosed within a red dotted line. The first subsequence of the 3-way junction starts at position 1072 (5'), ends at position 1075 (3') and consists of the bases GUGC. The second subsequence of the 3-way junction starts at position 1082 (5'), ends at position 1088 (3') and consists of the bases GUGUUGG. Finally, the third subsequence of the 3-way junction starts at position 1097 (5'), ends at position 1103 (3') and consists of the bases CCGCAAC. Unpaired bases in the MBL are those bases not part of the terminal base pairs of the three helices.

Figure 4, obtained using Jmol [41], presents a 3D representation of the same PDB 2J00 RNA molecule. This representation is based on the crystal structure 3D coordinates of the 686 atoms constituting this RNA molecule. In this illustration, helix H1 is colored red, helix H2 is colored yellow and helix H3 is colored blue. The coaxial helical stacking of H1 and H2 is apparent in this illustration. In addition, the Jmol software package allows the user to view a 3D visual rotation of the figure. By viewing the rotating figure from virtually every angle, the coaxial helical stacking becomes even more apparent.

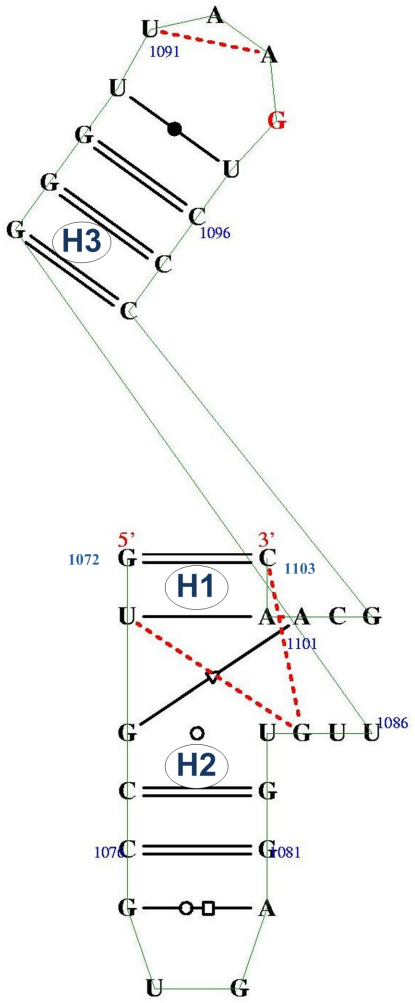

Figure 1. 2D (including tertiary interactions) illustration of bases 1072 through 1103 of RNA chain A from PDB ID 2J00. The three helical stacks are labeled H1, H2 and H3 respectively. In this illustration, helices H1 and H2 are seen to be coaxially stacked, i.e. sharing a common axis.

GUGCCGUGAGGUGUUGGGUUAAGUCCCGCAAC

Figure 2. Primary sequence of RNA chain A from PDB ID 2J00 illustrated in Figures 1, 3 and 4. Highlighted in yellow are the three subsequences that constitute the 3-way junction.

B. Feature Selection

A coaxial helical stacking motif in an RNA 3-way junction can be predicted by a random forests classifier that has been trained using certain specifically chosen "features" readily available in the secondary structure of known RNA 3-way junctions, i.e. the 110 element dataset described above. Selecting appropriate features for motif prediction is

one of the most fundamental challenges in bioinformatics, pattern recognition and machine learning. Features selected for this work are based on three principles [1]. First, a short loop region in a MBL, i.e. the region between adjacent helices, is more likely to be associated with a coaxial helical stacking. For this reason, the sizes of the three loop regions (i.e. the numbers of unpaired nucleotides in the three loop regions) of a 3-way junction are selected as features as well as the manner in which these three sizes relate to one another, e.g. the minimum of the three sizes. Second, it is known that consecutive unpaired adenine bases tend to interact via hydrogen bonding with the minor groove of a neighboring helix. This common interaction, known as A-minor motif, stabilizes contacts between RNA helices. In fact, the A-minor motif is the most common tertiary interaction in the large ribosomal subunits. For this reason,

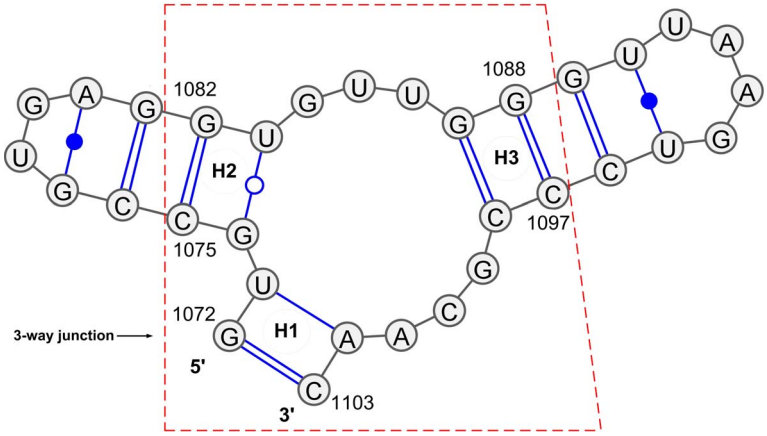

Figure 3. 2D plot produced by VARNA [40] using the primary sequence of bases 1072 through 1103 of PDB ID 2J00 in CT format provided by S2S [39]. 3-way junction is enclosed by a dotted red line. The three helical stacks are labeled H1, H2 and H3 respectively. In this illustration, helices H1 and H2 are not seen to be coaxially stacked, i.e. sharing a common axis, as compared with Fig. 1.

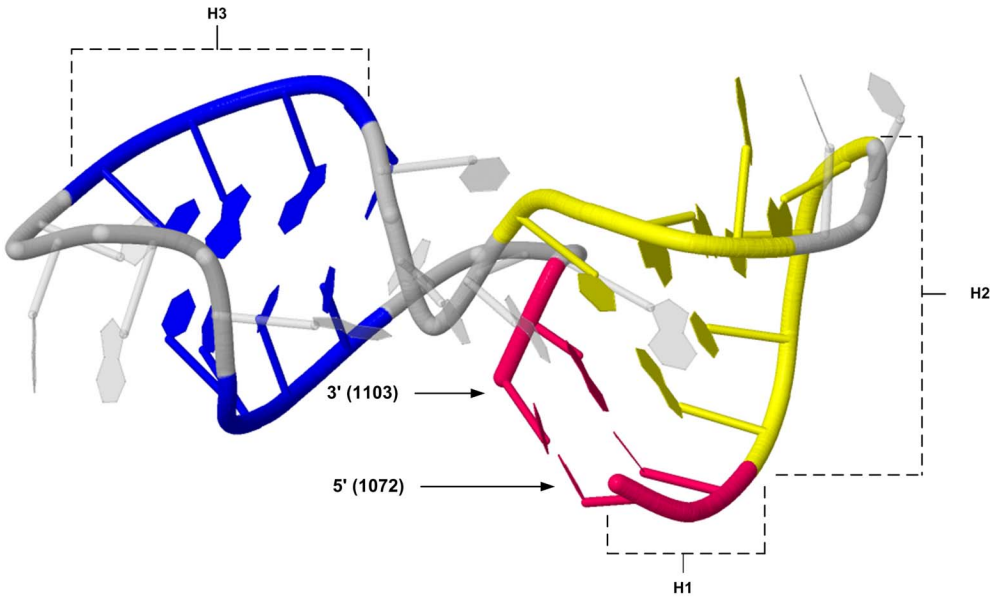

Figure 4. 3D plot produced by Jmol [41] using 3D coordinates of bases 1072 through 1103 of PDB ID 2J00. Helix H1 is colored red, helix H2 is colored yellow and helix H3 is colored blue. In this illustration, helices H1 and H2 are seen to be coaxially stacked, i.e. sharing a common axis.

information about consecutive unpaired adenine bases is selected to be used as features. Third, thermodynamic free-energy associated with the base pairs at the helix termini and the loop regions between adjacent helices is obtained and used as features. It is known that as thermodynamic free-energy declines in a conformation, stability increases. Totally, we selected 15 features used for coaxial helical stacking prediction.

C. The CSminer Approach

The CSminer program combines the trained random forests classifier described above with Infernal [18], which is a widely used tool capable of performing genome-wide search for non-coding RNAs. Using S2S [39], we obtained secondary structures for the 110 RNA 3-way junctions in our dataset. We clustered these 110 secondary structures using RNAforester [42]. We then selected three 3-way junctions whose secondary structures were in a common cluster produced by RNAforester. These three 3-way junctions had similar secondary structures and known coaxial helical stackings. The three 3-way junctions belonged to PDB molecules with identifiers 1NKW, 2AW4 and 1S72 respectively. Since a Stockholm alignment is required to create an Infernal covariance model, we manually constructed the appropriate Stockholm alignment of the three 3-way junctions using S2S and a text editor (Figure 5). The Stockholm alignment is a multiple alignment of RNA sequences together with the consensus secondary structure of the sequences. The secondary structure is shown in dot-bracket notation in Figure 5, in which dots represent bases and brackets represent base pairs. We created a covariance model from the constructed Stockholm alignment using Infernal's CMbuild utility [18]. We modified the source code in Infernal's CMsearch utility to execute the trained random forests classifier whenever an RNA secondary structure similar to our covariance model was detected during genome-wide searches performed by CMsearch. The resulting program is named CSminer.

The trained random forests classifier is capable of predicting the type of coaxial helical stacking in an RNA 3-way junction based upon the secondary structure detected by CMsearch. The random forests classifier is comprised of numerous classification and regression trees (CARTs) [43], each of which is formed by a small random subset of 4 (i.e. the square root) of the 15 features. Each CART is capable of contributing a "better than random opinion" about the coaxial helical stacking classification of an unknown or unlabeled input. By consolidating all opinions from all CARTs, i.e. by tallying all "votes", the random forests classifier is able to predict the coaxial helical stacking status of the RNA 3-way junction with high accuracy.

III. RESULTS

We applied CSminer to the complete genome of *T. Thermophilus*, i.e. GenBank ID CP0027777.1, obtained from the NCBI GenBank database. The CSminer search was performed on this complete genome, and motifs were detected between positions 14,310 and 14,384 on the plus strand of the genome. Figure 6 illustrates the output of CSminer. The output produced by CSminer is restricted to only those results determined to contain a 3-way junction. Furthermore, in each case of a 3-way junction, a "Coaxial Helical Stacking Status" is reported.

Figure 6 shows that in the genome of *T. Thermophilu*, there is evidence of a 3-way junction. Furthermore, this 3-way junction is predicted, by our random forests classifier, to contain a coaxial helical stacking. The coaxial helical stacking is of type H1H2 (i.e. helix H1 and helix H2 are aligned with a common axis).

This CSminer search result is validated as follows. Based on Blast [44] and manual analyses, we know *T. Thermophilus* is related to PDB molecule 2J01. Specifically, we downloaded the chain A nucleotide FASTA sequence from PDB for the 2J01 structure. Using NCBI Blast [44], we located this downloaded FASTA sequence in the whole genome of *T. Thermophilus*, i.e. GenBank ID CP002777.1, from position 14,310 through 14,384 on the plus strand. These positions are consistent with those outputted by CSminer where the motifs were detected (see Figure 6). Furthermore, based on the analysis in [1], this region of the 2J01 structure contains a 3-way junction with a coaxial helical stacking of type H1H2, which are exactly what CSminer reports. Notice that the 2J01 structure is not among the three molecules with PDB identifiers 1NKW, 2AW4 and 1S72 respectively [24] used to build the covariance model employed in CMsearch.

```
# STOCKHOLM 1.0

1NKW_52        CU--CCCGGAAGACCACCGGGUUAAGAGGCCAGG---CGUGCAC-----GCAUAGCAAUGUGU-----UCAGCG---GAC
1S72_53        UC--CCGCGUACAAGACGCGGUCGAUAGACUCG-GGGUGUG---CGCGUCGAGGUA--ACGAGACGUUA---AGCCCA-C
2AW4_54        GGAACGUUGAAGACGACGACGUUGAUAGGCCGGGUG-UGUA---AG---CGCAGCG--AUGCGUU---G---AGCU-AAC
#=GC SS_cons   ((..(((((.......)))))).....(((((((((....((..........((((((....))))))..........)).....)

1NKW_52        UGGUGCUCAUC--AG
1S72_53        GAGCACUAACA--GA
2AW4_54        CGGUACUAAUGAACC
#=GC SS_cons   ))))..)))......))
//
```

Figure 5. Stockholm alignment of RNA molecules from three organisms recorded in PDB with identifiers 1NKW, 2AW4 and 1S72 respectively.

```
CM: coax_model10-1
>gi|333965676|gb|CP002777.1|

  Plus strand results:

 Query = 1 - 75, Target = 14310 - 14384
 Score = 64.97, GC =  71

 Coax status = H1H2

           ((<<<<<_______>>>>>,,,,<<<<<<<<<<--<<----------__>>>>>>>-->>---
         1 ccCcccGAAgAcgACgggGUUGAuAGgCCgGGuGcgUGaGcgcgauaucgcgauacgCuA 60
           :CCC::GAAGAC AC::GGU GAU:GGCCGGG+G::U+AGCGC::  ::GCG+U ::C+
     14310 ACCCGGGAAGACCACCCGGUGGAUGGGCCGGGGGUGUAAGCGCCGCGAGGCGUUGAGCCG 14369

           ->>>>>->>>,,,))
        61 ACcGGuACUaAUagg 75
           ACCGGU C:AAU G:
     14370 ACCGGUCCCAAUCGU 14384
```

Figure 6. CSminer's search result from the whole genome of *T. Thermophilus*. Nucleotides shown are from positions 14310 through 14384, i.e. for a length of 75, on the plus strand. Notice the coaxial helical stacking status of "H1H2" indicating that the evidence of a 3-way junction was located and that this 3-way junction is predicted to contain a coaxial helical stacking motif.

In a separate test, a different set of similar 3-way junctions was selected from a different RNAforester cluster. An Infernal covariance model was created from a Stockholm alignment of this new set of 3-way junctions. As expected, different search results were received. Infernal search results are similar to the 3-way junctions that comprise the Infernal covariance model.

IV. CONCLUSION

CSminer combines the strengths of the genome-wide search tool, Infernal, with an ensemble-based random forests classifier to provide an effective predictive instrument. Among the growing number of ensemble-based methodologies, the random forests method is among the most accurate. This allows us to add significant additional functionality to Infernal. Efficient prediction of the presence of coaxial helical stacking motifs in genomes will help to further unravel the mysteries of non-coding RNA. Much remains unknown in this exciting research area. Our conclusion is that genome-wide searching for coaxial helical stacking RNA motifs is feasible and cost effective.

We are currently conducting more extensive CSminer genome searches by using known higher order junctions (i.e. 4-way junctions, 5-way junctions, etc.) to further demonstrate the feasibility of this novel approach. We are also investigating the use of various ensemble-based methodologies, including random forests, to find RNA tertiary motifs which include pseudoknot interactions and A-minors [11]. We are interested in evaluating the effectiveness and efficiency of combining a variety of different machine learning approaches to stubborn RNA motif finding problems.

ACKNOWLEDGMENT

We thank Drs. Miguel Cervantes-Cervantes, Christian Laing, Yongkyu Park and Tamar Schlick for useful comments and suggestions.

REFERENCES

[1] Laing C, Wen D, Wang JTL, Schlick T. Predicting coaxial helical stacking in RNA junctions. Nucleic Acids Res. 2012;40:487-498.

[2] Kim SH, Sussman JL, Suddath FL, et al. The general structure of transfer RNA molecules. Proc Natl Acad Sci USA. 1974;71:4970-4974.

[3] Aalberts DP, Hodas NO. Asymmetry in RNA pseudoknots: observation and theory. Nucleic Acids Res. 2005;33:2210-2214.

[4] Toor N, Keating KS, Taylor SD, Pyle AM. Crystal structure of a self-spliced group II intron. Science. 2008;320:77-82.

[5] Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science. 2000; 289:905-920.

[6] Schluenzen F, Tocilj A, Zarivach R, et al. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. Cell. 2000; 102:615-623.

[7] Wimberly BT, Brodersen DE, Clemons WM Jr, et al. Structure of the 30S ribosomal subunit. Nature. 2000;407:327-339.

[8] Kim J, Walter AE, Turner DH. Thermodynamics of coaxially stacked helixes with GA and CC mismatches. Biochemistry. 1996;35:13753-13761.

[9] Walter AE, Turner DH, Kim J, et al. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci USA. 1994;91:9218-9222.

[10] Aalberts DP, Nandagopal N. A two-length-scale polymer theory for RNA loop free energies and helix stacking. RNA. 2010;16:1350-1355.

[11] Xin Y, Laing C, Leontis NB, Schlick T. Annotation of tertiary interactions in RNA structures reveals variations and correlations. RNA. 2008;14:2465-2477.

[12] Valadkhan S, Nilsen TW. Reprogramming of the non-coding transcriptome during brain development. Journal of Biology. 2010;9:5.

[13] Gilbert W. Origin of life: the RNA world. Nature. 1986;319:618.

[14] MicroRNA. ISSN:2211-5366 (Print) ISSN: 2211-5374 (Online) Volume 1, 2 Issues, 2012.

[15] Pyle AM, Shakked Z. The ever-growing complexity of nucleic acids: from small DNA and RNA motifs to large molecular assemblies and machines (Editorial overview). Current Opinion in Structural Biology. 2011;21:293-295.

[16] Leontis NB, Lescoute A, Westhof E. The building blocks and motifs of RNA architecture. Current Opinion in Structural Biology. 2006;16:279-287.

[17] Reiter NJ, Chan CW, Mondrago A. Emerging structural themes in large RNA molecules. Current Opinion in Structural Biology. 2011;21:319-326.

[18] Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25(10):1335-1337.

[19] Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31:439-441.

[20] Silveira AC, Robertson KL, Lin B, et al. Identification of non-coding RNAs in environmental vibrios. Microbiology. 2010;156:2452-2458.

[21] Wong TK, Lam TW, Sung WK, Yiu SM. Adjacent nucleotide dependence in ncRNA and order-1 SCFG for ncRNA identification. PLoS One. 2010;5(9).

[22] Sun Y, Buhler J, Yuan C. Designing filters for fast known ncRNA identification. IEEE/ACM Trans Comput Biol Bioinform. 2011.

[23] Byron K, Cervantes-Cervantes M, Wang JTL, Lin WC, Park Y. Mining roX1 RNA in *Drosophila* genomes using covariance models. Intl J Comp Bioscience. 2010;1(1):22-32.

[24] Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235-242.

[25] Berman HM, Olson WK, Beveridge DL, et al. The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J. 1992;63:751-759.

[26] Diamond JM, Turner DH, Mathews DH. Thermodynamics of three-way multibranch loops in RNA. Biochemistry. 2001;40(23):6971-6981.

[27] Liu B, Diamond JM, Mathews DH, Turner DH. Fluorescence competition and optical melting measurements of RNA three-way multibranch loops provide a revised model for thermodynamic parameters. Biochemistry. 2011;50(5):640-653.

[28] Laing C, Jung S, Iqbal A, Schlick T. Tertiary motifs revealed in analyses of higher-order RNA junctions. J Mol Biol. 2009;393(1):67-82.

[29] Laing C, Schlick T. Analysis of four-way junctions in RNA structures. J Mol Biol. 2009;390(3):547-559.

[30] Yang BS, Di X, Han T. Random forests classifier for machine fault diagnosis. Journal of Mechanical Science and Technology. 2008;22(9): 1716-1725.

[31] Liu G, Liu J, Cui X, Cai L. Sequence-dependent prediction of recombination hotspots in *Saccharomyces cerevisiae*. J Theor Biol. 2011;293C:49-54.

[32] Wang XF, Chen Z, Wang C, Yan RX, Zhang Z, Song J. Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. PLoS One. 2011;6(10):e26767.

[33] Calle ML, Urrea V, Boulesteix AL, Malats N. AUC-RF: a new strategy for genomic profiling with random forest. Hum Hered. 2011;72(2):121-132.

[34] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonga A. Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes. 2011;4:299.

[35] Statnikov1 A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics. 2008;9:319.

[36] Xiao J, Tang X, Li Y, Fang Z, Ma D, He Y, Li M. Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure. BMC Bioinformatics. 2011;12:165.

[37] Lescoute A, Westhof E. Topology of three-way junctions in folded RNAs. RNA. 2006;12(1):83-93.

[38] Yang H, Jossinet F, Leontis N, et al. Tools for the automatic identification and classification of RNA base pairs. Nucleic Acids Res. 2003;31(13):3450-3460.

[39] Jossinet F, Westhof E. Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics. 2005;21(15):3320-3321.

[40] Darty K, Denise A, Ponty Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009;25(15):1974-1975.

[41] Jmol: an open-source Java viewer for chemical structures in 3D. Available at http://www.jmol.org/.

[42] Hochsmann M, Toller T, Giegerich R, Kurtz S. Local similarity in RNA secondary structures. Proc. IEEE Computational Systems Bioinformatics Conf. 2003:159-168.

[43] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, CA:Wadsworth, 1984.

[44] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403-410.