BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 20 2011, pages 2828–2835
doi:10.1093/bioinformatics/btr495

Structural bioinformatics

Advance Access publication August 27, 2011

Feature-based 3D motif filtering for ribosomal RNA

Ying Shen[1], Hau-San Wong[1,*], Shaohong Zhang[1] and Zhiwen Yu[2]

[1]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong and [2]The School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: RNA 3D motifs are recurrent substructures in an RNA subunit and are building blocks of the RNA architecture. They play an important role in binding proteins and consolidating RNA tertiary structures. RNA 3D motif searching consists of two steps: candidate generation and candidate filtering. We proposed a novel method, known as Feature-based RNA Motif Filtering (FRMF), for identifying motifs based on a set of moment invariants and the Earth Mover's Distance in the second step.

Results: A positive set of RNA motifs belonging to six characteristic types, with eight subtypes occurring in HM 50S, is compiled by us. The proposed method is validated on this representative set. FRMF successfully finds most of the positive fragments. Besides the proposed new method and the compiled positive set, we also recognize some new motifs, in particular a π-turn and some non-standard A-minor motifs are found. These newly discovered motifs provide more information about RNA structure conformation.

Availability: Matlab code can be downloaded from www.cs.cityu.edu.hk/~yingshen/FRMF.html

Contact: cshswong@cityu.edu.hk

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

Received on February 13, 2011; revised on August 3, 2011; accepted on August 21, 2011

1 INTRODUCTION

In recent years, non-protein-coding RNAs (ncRNAs), in particular long ncRNAs (>200 nt), are receiving more and more attention (Mercer *et al*., 2009; Soldà *et al*., 2009). Increasing evidences indicate that these RNAs have important regulatory functions, e.g. the regulation of epigenetic processes related to cell differentiation, etc. (Amaral and Mattick, 2008; Mattick *et al*., 2009; Pang *et al*., 2009).

Starting from a nucleotide sequence, RNA will experience a process of hierarchical folding and finally form a unique compact 3D structure to perform certain functions. In the final 3D structure, some nucleotides that are originally remote from each other at the sequence level may be connected by chemical bonds or form tertiary interactions. Those nucleotides undergoing tertiary interactions, together with some related nucleotides, often compose special substructures that occur recurrently in an RNA subunit. These recurrent substructures are called RNA 3D motifs. RNA 3D motifs have various functions, e.g. interacting with proteins and consolidating the RNA tertiary structures (Apostolico *et al*., 2009; Francois *et al*., 2005). By characterizing these motifs, researchers can find a way to explore the possible functions of ncRNAs.

RNA 3D motifs are also important for predicting the possible tertiary structures for an RNA sequence, since they are the building blocks of RNA 3D structure. Understanding RNA structures relies on the ability to identify the component motifs (Laing and Schlick, 2010). Many fragment assembly approaches for RNA tertiary structure prediction, e.g. MC-Fold/MC-Sym (Parisien and Major, 2008), are based on the reconstructed base pairs and/or motifs. Despite the efforts made in tRNA and medium size RNA structure prediction (Jonikas *et al*., 2009), to our knowledge, no significant improvement has been reported in long ncRNAs tertiary structure prediction. If suitable rules can be inferred from RNA 3D motif composition, methods for RNA tertiary structure prediction will be expected to benefit from these rules and may achieve higher accuracy.

The goal of RNA motif searching is to identify the complete set of motifs in an RNA subunit, which have similar structures to the query motif. Query motifs are often observed recurrent substructures reported in previous publications. In early studies, RNA motif searching algorithms are based on the secondary structure of RNA (Reeder *et al*., 2007; Yang *et al*., 2003). These algorithms first produce 2D symbolic representations of the RNA 3D structures. Then, recurrent motifs are manually determined using the 2D representations generated in the first step. Some methods that identify RNA motifs from sequences were also developed. Yao *et al*. (2006) and Rabani *et al*. (2008) find local motifs from sequences based on a probabilistic model, and Michal *et al*. (2007) finds motifs from homologous sequences based on Genetic Programming. However, motifs discovered by the above approaches are actually 2D motifs like hairpins and contiguous base pairs, and 3D motifs based on tertiary interactions are quite difficult to discover through these methods. In order to find RNA 3D motifs, new methods are proposed and can be categorized into two types. The first type of methods is based on geometrical matching (Apostolico *et al*., 2009; Duarte *et al*., 2003; Gendron *et al*., 2001; Huang *et al*., 2005; Sarver *et al*., 2008; Sargsyan and Lim, 2010; Wadley and Pyle, 2004). These methods use the root mean square deviation (RMSD) and other distance metrics to calculate the distance between the two RNA fragments. Among these geometry-based approaches, FR3D (Sarver *et al*., 2008) is an effective method that is mainly used for RNA 3D motif searching. When comparing two structures, it first superimposes the candidate and the query motif in 3D space, and then uses the fitting error and the orientation error to measure the distance between the two fragments. On the other hand, Apostolico *et al*. (2009) and Sargsyan and Lim (2010) use the cosine measure to

*To whom correspondence should be addressed.

measure the discrepancy of two fragments based on their distance histograms. The second type of method is based on graph theory (Harrison *et al.*, 2003; Lescoute *et al.*, 2005; Major *et al.*, 1991; Zhong *et al.*, 2010). RNA 3D structures stored in the database [e.g. PDB (Berman *et al.*, 2000)] only retain the coordinates of discrete atoms. Graph-based methods first reduce the RNA 3D structures to graphs, and then apply subgraph isomorphism to search for motifs in the reduced graphs. When constructing the graphs, nucleotides are regarded as nodes, and edges are added between the two nodes if the corresponding nucleotides are considered to have undergone interactions according to certain rules. Still adopting a graph-based representation, the approach described in Djelloul and Denise (2008) focuses on finding novel recurrent substructures instead of similar substructures based on a query motif.

Although many approaches have been developed for RNA 3D motif searching, the discovery of all positive motifs for known RNA subunits is still a long way off. The RNA motif searching process can be divided into two steps: candidate generation and candidate filtering. In this article, we propose a new approach, which is referred to as Feature-based RNA Motif Filtering (FRMF), to be used in the second step for identifying RNA 3D motifs from the candidates. Our method is based on the geometric features of 3D structures, specifically the moment invariants and the Earth Mover's Distance (EMD), for structure comparison. Moment invariants receive a lot of studies in 2D image analysis and have been extended to 3D space (Flusser *et al.*, 2010; Mamistvalov, 1998). But it is only in recent years that moment invariants are introduced into structural bioinformatics (Sommer *et al.*, 2007). EMD is widely used in pattern recognition to compute feature histogram distance. Compared with other distance metrics, it has the advantage of supporting adaptive binning and partial match (Yu and Herman, 2005).

Through our experiments, we discover a number of new motifs, in particular a new π-turn and non-standard instances of type I A-minor motif. These newly observed motifs are essential for binding proteins and consolidating the RNA tertiary structure. They also lend evidence to the possible variations of π-turn and type I A-minor motif, and provide an insight into RNA 3D structures and functions. We also collect a set of positive RNA 3D motifs of different classes. These motifs are either published in previous papers or discovered by us. We anticipate that this positive set could be useful for the development of further approaches in RNA 3D motif searching.

The rest of this article is organized as follows. Sections 2.1 and 2.2 introduce RNA 3D motifs and the ribosomal RNA subunit HM 50S. Sections 2.3 and 2.4 provide some preliminary knowledge about moment invariants and EMD. In Sections 2.5 and 2.6, FRMF, an approach proposed by us for identifying RNA 3D motifs, will be presented. Section 3 reports our experimental results. Discussion and analysis on the newly discovered motifs is given in Section 4, and we conclude our article in Section 5.

2 MATERIALS AND METHODS

2.1 Classification of RNA 3D motifs

RNA motifs are recurrent substructures occurring independently in RNA subunits. There are many types of RNA motifs, e.g. *hairpin loop*, *sarcin/ricin loop*, *π-turn* and *kink-turn*, etc. Some of them (e.g. kink-turn) can be further categorized into the local type and the composite type. Local kink-turn is composed of two consecutive strands, one of which is the characteristic

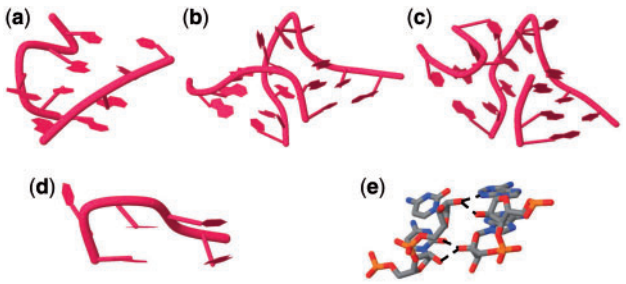

Fig. 1. (**a**) Sarcin/ricin loop; (**b**) local kink-turn; (**c**) composite kink-turn; (**d**) π-turn; (**e**) ribose zipper. Dashed lines represent hydrogen bonds.

Fig. 2. (**a**) Standard tetraloop; (**b**) tetraloop with deletion; (**c**) tetraloop with insertion.

strand (Fig. 1b). Composite kink-turn consists of three strands, but it has a similar characteristic strand to that of the local kink-turn (Fig. 1c). Some types of RNA motifs are called RNA tertiary interaction, because they are formed by several remote nucleotides that have interactions in 3D space. There are seven kinds of tertiary interactions according to the SCOR database (Klosterman *et al.*, 2002). Because there are too many types of motifs, we select some representative types for identification and they are introduced respectively here.

Tetraloop: a tetraloop is a hairpin loop that consists of four consecutive residues and terminates a single RNA helix (Correll and Swinger, 2003; Jaeger *et al.*, 1994). There are three types of tetraloops (Hsiao *et al.*, 2006): standard, with deletion and with insertion (Fig. 2). Standard tetraloop is also called GNRA tetraloop, because GNRA (N is any nucleotide in A, U, G, C; R is A or G) is one of the most common patterns appearing in the standard tetraloops.

Sarcin/ricin loop: the complete sarcin/ricin loop contains nine nucleotides, among which there are five non-canonical base pairs called the core of the sarcin/ricin loop. Sarcin/ricin loop joins two segments of helices. A complete sarcin/ricin loop structure is shown in Figure 1a.

Kink-turn: the kink-turn motif is a two-stranded, helix-internal loop-helix substructure comprising ~15 nt (Klein *et al.*, 2001). The internal loop connects two segments of helices, the orientation of which differs by ~120°. Kink-turn motifs can be further categorized as local kink-turns (consisting of two strands, see Fig. 1b) and composite kink-turns (consisting of three strands, see Fig. 1c).

π-turn: the π-turn consists of five consecutive nucleotides with ~120° change in backbone direction (Fig. 1d) resulting in a pinched strand. On the 5′-side, two consecutive nucleotides stack on a helix (the right two nucleotides in Fig. 1d). On the 3′-side, two discontiguous nucleotides are arranged side-by-side (the middle two nucleotides in Fig. 1d). The nucleotide between them (the left nucleotide in Fig. 1d) extends out to form tertiary interactions with RNAs, proteins or both. π-turn is similar to the kink-turn in structure, but has distinct conformational features (Wadley and Pyle, 2004).

Ribose zipper: ribose zipper (Cate *et al.*, 1996) is a tertiary interaction formed by consecutive hydrogen bonds between the backbone ribose 2′-hydroxyls and bases from two distinct strands (Fig. 1e).

A-minor motif: A-minor motif involves the insertion of adenosines (Fig. 3a) into the minor grooves of RNA helices (Nissen *et al.*, 2001). It has four versions based on the positions of the O2′ and N3 atoms of the adenosines inserted. In type I (Fig. 3b), both the O2′ and the N3 atoms of

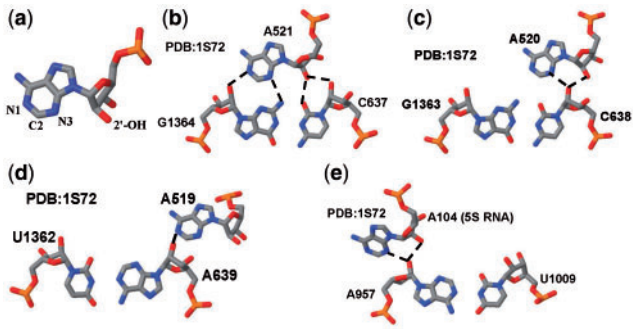

Fig. 3. (**a**) Adenosine; (**b**) A-minor motif, type I; (**c**) A-minor motif, type II; (**d**) A-minor motif, type III; (**e**) A-minor motif, type 0. Dashed lines represent hydrogen bonds connecting the inserted A and the Watson–Crick base pairs.

Table 1. Frequencies of different types of motifs appearing in HM 50S

Motif	Frequency
Tetraloop, standard	19
Tetraloop, with deletion	10
Sarcin/ricin loop, core	11
Kink-turn, local	6
Kink-turn (local and composite)	8
π-turn	7
Ribose zipper	39
A-minor motif, type I	60

the inserted adenosine are inside the minor groove of the helix. The number of hydrogen bonds formed for this type is the maximum among all the four types. Therefore, type I corresponds to the strongest interaction compared with the other three types. In type II (Fig. 3c), the O2′ of the inserted adenosine is outside the near strand O2′ of the Watson–Crick base pair, whereas the N3 is inside. In type III (Fig. 3d), both O2′ and N3 of the inserted adenosine are outside the near strand O2′. The fourth version is type 0 (Fig. 3e), in which the N3 of the inserted adenosine is outside the O2′ of the nucleotide on the far strand.

2.2 *Haloarcula marismortui* 50S ribosomal subunit

We focus on the 50S ribosomal subunit of *Haloarcula marismortui* (PDB ID 1S72) in our experiments. Motifs in HM 50S have been studied in previous publications, and we collect as many as possible the motifs discovered in HM 50S published in previous works in order to construct a positive set of motifs. Specifically, standard tetraloops and tetraloops with deletion are collected from Apostolico *et al*. (2009) and Sargsyan and Lim (2010); sarcin/ricin loops and local kink-turns are collected from Sarver *et al*. (2008); composite kink-turns are collected from Apostolico *et al*. (2009); π-turns are collected from Wadley and Pyle (2004); A-minor motifs and ribose zippers are collected from Xin *et al*. (2008). In addition to the published motifs, this positive set also includes some motifs newly discovered by our proposed FRMF approach. We use this set to validate the performance of different approaches for identifying RNA 3D motifs. The frequencies of different types of motifs appearing in HM 50S are listed in Table 1. The complete set of positive motifs is provided in the Supplementary Material.

2.3 Moment invariants

The geometric moment with order p of a 3D structure is defined as follows:

$$m_{p_1p_2p_3}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^{p_1}y^{p_2}z^{p_3}f(x,y,z)dxdydz \tag{1}$$

where $f(x,y,z)$ is the density function of the 3D structure and $p=p_1+p_2+p_3$. The central geometric moment is defined as:

$$\mu_{p_1p_2p_3}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-x_c)^{p_1}(y-y_c)^{p_2}(z-z_c)^{p_3}f(x,y,z)dxdydz \tag{2}$$

where $x_c=m_{100}/m_{000}$, $y_c=m_{010}/m_{000}$ and $z_c=m_{001}/m_{000}$. (x_c, y_c, z_c) is the center of the 3D structure.

We use moment invariants to describe the characteristics of 3D structures. Moment invariants are quantities insensitive to a particular set of transformations, and they provide adequate discrimination power to distinguish between structures belonging to different classes (Flusser *et al.*, 2010). We list three of these invariants below:

$$I_1=\frac{1}{2}\sqrt{\frac{1}{\pi}}(\mu_{200}+\mu_{020}+\mu_{002}) \tag{3}$$

$$I_2=\frac{5}{4\pi}(\mu_{200}^2+\mu_{020}^2+\mu_{002}^2-\mu_{200}\mu_{020}-\mu_{200}\mu_{002}-\mu_{020}\mu_{002}+3\mu_{110}^2+3\mu_{101}^2+3\mu_{011}^2) \tag{4}$$

$$\begin{aligned}I_3=&\frac{5}{32\pi}\sqrt{\frac{5}{2\pi}}(-2\mu_{200}^3-2\mu_{020}^3-2\mu_{002}^3+3\mu_{200}^2\mu_{020}+3\mu_{200}^2\mu_{002}\\&+3\mu_{020}^2\mu_{200}+3\mu_{020}^2\mu_{002}+3\mu_{002}^2\mu_{200}+3\mu_{002}^2\mu_{020}-12\mu_{200}\mu_{020}\mu_{002}\\&+18\mu_{200}\mu_{011}^2+18\mu_{020}\mu_{101}^2+18\mu_{002}\mu_{110}^2-9\mu_{200}\mu_{110}^2-9\mu_{200}\mu_{101}^2\\&-9\mu_{020}\mu_{110}^2-9\mu_{020}\mu_{011}^2-9\mu_{002}\mu_{101}^2-9\mu_{002}\mu_{011}^2-54\mu_{110}\mu_{101}\mu_{011})\end{aligned} \tag{5}$$

In general, structures in the same class have similar values for all three moment invariants, while structures in different classes have significantly different values for one or more moment invariants. Based on these three second-order moment invariants, we can construct a feature vector $v=[I_1\ I_2\ I_3]$ for a 3D structure. Given two 3D structures S and S', two feature vectors $v=[I_1\ I_2\ I_3]$ and $v'=[I_1'\ I_2'\ I_3']$ can be constructed using Equations (3), (4) and (5). The discrepancy between two structures S and S' can then be measured using Equation (6) below.

$$r(v,v')=\left(\frac{I_1-I_1'}{I_1}\right)^2+\left(\frac{I_2-I_2'}{I_2}\right)^2+\left(\frac{I_3-I_3'}{I_3}\right)^2 \tag{6}$$

From Equation (6), it can be seen that, in our approach, the final discrepancy is defined as the sum of the relative distances from three moment invariants. The reason to choose the relative distance instead of the squared Euclidian distance is that the ranges of the three moment invariants are quite different, thus the relative distance can reduce the effect of different value ranges on the three invariants.

2.4 EMD

The EMD is a distance measure between two probability distributions (Levina and Bickel, 2001; Rubner *et al*., 1998). Currently, EMD is widely used in computing feature histogram distance, because it judiciously extends a discrepancy measure between individual features to a corresponding measure between histograms, and also supports adaptive binning (Yu and Herman, 2005). In this work, we use EMD to measure the difference between the corresponding interatomic distance histograms of two RNA substructures. Specifically, a distance histogram summarizes the distribution of the interatomic distances in a substructure, which corresponds to a distinctive and robust descriptor for RNA motifs (Apostolico *et al.*, 2009)

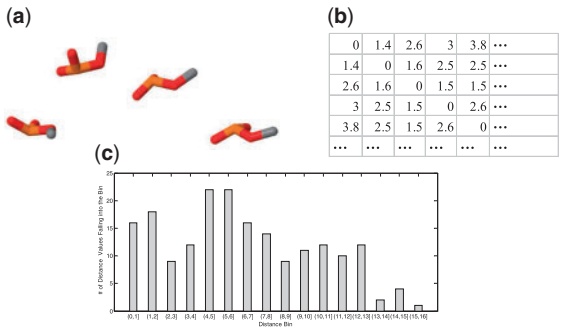

Fig. 4. (**a**) S_1 structure of a standard tetraloop; (**b**) distance matrix for atoms in (a); (**c**) distance histogram for the distance matrix in (b).

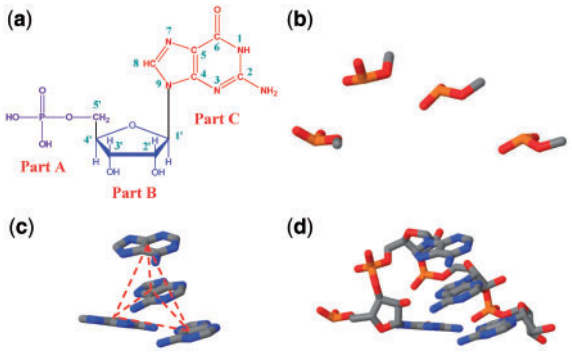

Fig. 5. (**a**) RNA guanine; (**b**) S_1 structure; (**c**) S_2 structure (only contains the center points of the four bases); (**d**) S_3 structure.

and is easy to compute. Given an interatomic distance matrix (Fig. 4b), a distance histogram can be constructed by assigning the matrix entries into a discrete number of bins and counting the frequencies of values falling into the same bin (Fig. 4c). If two distance histograms P and Q both contain m bins, they can be represented as $\{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$ and $\{(q_1, w_{q1}), \ldots, (q_m, w_{qm})\}$, where p_i and q_j are the centers of bin i and bin $j (1 \le i \le m, 1 \le j \le m)$ respectively, and w_{pi}, w_{qj} are the corresponding weights of these bins (in practice, they are the frequencies of distance values falling into these bins) for the two histograms. Given a distance measure between p_i and q_j, which is represented by a matrix $\mathbf{D} = [d_{ij}]$, the distance between P and Q can be computed using the following equation.

$$d_{\mathrm{EMD}}(P,Q) = \min \frac{\sum_{i,j=1}^{m} f_{ij} d_{ij}}{\sum_{i,j=1}^{m} f_{ij}}$$

$$\text{s.t.} \sum_{j=1}^{m} f_{ij} \le w_{pi}, \sum_{i=1}^{m} f_{ij} \le w_{qj} \qquad (7)$$

$$\sum_{i,j=1}^{m} f_{ij} = \min\left(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{m} w_{qj}\right)$$

$$f_{ij} \ge 0, 1 \le i \le m, 1 \le j \le m$$

2.5 Representation of an RNA 3D motif

Before identifying the RNA motifs, an important step is to find their representative structures. That is because, even if two motifs belong to the same category, the positions of some of the bases and ribose sugars can be quite different. Representative structures can characterize the similar motif features in the same class to a significant extent. In previous works, researchers adopt different representative structures for RNA 3D motifs. For example, FR3D uses the centers of all bases comprising the motif. In Apostolico *et al.* (2009) and Sargsyan and Lim (2010), the authors use the phosphates and ribose sugars as the representative structures of the motifs. On the other hand, in our approach, we divide a nucleotide into three components: the phosphate part (part A, containing P, OP1, OP2, C5′ and O5′), ribose sugar (part B, containing C4′, O4′, C3′, O3′, C2′, O2′ and C1′) and the base (part C, containing the other atoms of the residue) (Fig. 5a). We extract three substructures from the original 3D structure of a motif as its representation. Take a standard tetraloop as an example: the first substructure S_1 contains the complete part A of motif nucleotides (Fig. 5b), and the second substructure S_2 contains all the centers of part C of the motif (Fig. 5c). The last substructure S_3 is the motif itself (Fig. 5d).

2.6 Feature-based RNA 3D motif filtering

RNA 3D motif searching can be divided into two steps: candidate generation and candidate filtering. Currently, we only focus on candidate filtering. We now introduce the details of FRMF for filtering candidates.

Given a query motif, FRMF first calculates three feature vectors $v_1 = [I_{11}\ I_{12}\ I_{13}]$, $v_2 = [I_{21}\ I_{22}\ I_{23}]$ and $v_3 = [I_{31}\ I_{32}\ I_{33}]$ using Equations (3), (4) and (5) for the three substructures S_1, S_2, S_3 extracted from the query motif following the method presented in Section 2.5. Given a candidate, FRMF calculates another three feature vectors v_1', v_2' and v_3' for the three substructures S_1', S_2', S_3' extracted from the candidate. Three discrepancy values $\mathrm{disc}_1 = r(v_1, v_1')$, $\mathrm{disc}_2 = r(v_2, v_2')$, $\mathrm{disc}_3 = r(v_3, v_3')$ are then calculated using Equation (6).

Next, FRMF will compute the EMD value between S_1 and S_1' based on their distance histograms. To obtain the distance histograms, FRMF first constructs two Euclidean distance matrices for atoms in S_1 and S_1', respectively (Fig. 4a and b). The distance values in the matrices are assigned to several bins as shown in Figure 4c and two histograms h_1 and h_1' are constructed for S_1 and S_1', respectively. The distance d_{EMD} between h_1 and h_1' is calculated using Equation (7).

After the previous steps, four discrepancy values: disc_1, disc_2, disc_3 and d_{EMD} are obtained. Three thresholds t_1, t_2, t_3 are set for (disc_1, disc_2, disc_3), respectively. If the values of disc_1, disc_2 and disc_3 are all smaller than the corresponding thresholds, the candidate is regarded as belonging to the same class as that of the query motif. In the end, all the positive candidates are ranked according to their d_{EMD} values.

3 RESULTS

In the following subsections, we present the results on searching different types of motifs. We also list the results obtained from FR3D and the method used in Apostolico *et al.* (2009) for comparison. RNA fragments are represented in the form of 'sequence_ID chain_ID'. For example, a standard tetraloop, G2412 A2413 A2414 A2415 0000, consists of a guanine and three adenines. Their sequence IDs are 2412, 2413, 2414 and 2415, respectively. All four nucleotides are located on chain 0, and therefore, there are four zeros at the end. FR3D candidate generation module is used to generate possible candidates. There will be $n(n-1)\ldots(n-m+1)$ (i.e. approximately n^m) candidates for a search (m is the size of the query motif and n is the size of the RNA structure). Sarver *et al.* (2008) has shown that possible candidates should satisfy a screening criterion, which can be validated rapidly. In the candidate generation step, the screening algorithm of FR3D can reduce the original size of the candidate set to a few hundred thousand or less. Then the three

approaches are evaluated on the reduced candidate set to identify motifs.

The thresholds for parameters used by FR3D and FRMF are listed in Table 2 (more details about settings are shown in Supplementary Figs S1–S8). Apostolico *et al*. (2009) uses RMSD and the cosine measure to filter candidates. For this method, the same thresholds as in the original paper are adopted for searching tetraloops, kink-turns, π-turns and sarcin/ricin loops. When searching ribose zippers and A-minor motifs, the threshold for the cosine measure is 0.95 (RMSD cannot be used when searching motifs containing more than two strands). The outputs of the three algorithms are listed in Supplementary Tables S1–S6. Results highlighted in yellow indicate positive motifs, while the negative instances are not highlighted. The first instance in each table is used as the query motif.

In order to assess the accuracy of FRMF, FR3D and the method in Apostolico *et al*. (2009) on searching different motifs, precision–recall (PR) curves are plotted. For the given query motif, the three approaches generate their ranked candidate lists using their respective measures [i.e. *discrepancy* in FR3D, d_{EMD} value in FRMF and the cosine measure in Apostolico *et al*. (2009)]. The candidates are classified as positive instances if their distances are smaller than a threshold. The others are classified as negative instances. For a certain threshold, precision and recall can be computed using Equations (8) and (9) below. By adjusting the threshold on the ranked list, precision/recall pairs are computed and shown as a point in the PR space. A PR curve can be obtained by connecting these points (Fig. 6).

$$\mathrm{Precision} = \frac{\mathrm{TruePositive}}{\mathrm{TruePositive}+\mathrm{FalsePositive}} \tag{8}$$

$$\mathrm{Recall} = \frac{\mathrm{TruePositive}}{\mathrm{TruePositive}+\mathrm{FalseNegative}} \tag{9}$$

Area under the PR curve (AUC-PR) is computed for each curve using the approach in Davis and Goadrich (2006). A larger AUC-PR value indicates a better performance.

Table 2. Thresholds for motif filtering

Motif	Guaranteed cutoff	Relaxed cutoff	t_1	t_2	t_3
Tetraloop, standard	0.5	0.5	0.92	2.53	1.24
Tetraloop, with deletion	0.5	0.5	0.67	0.78	1.07
Sarcin/ricin loop, core	0.4	0.4	1.20	1.02	1.54
Kink-turn, local	0.5	0.5	0.82	1.99	0.55
Kink-turn (composite and local)	0.95	0.95	0.70	6.14	2.33
π-turn	0.95	0.95	0.59	1.47	1.18
Ribose zipper	0.5	0.5	0.70	0.99	0.75
A-minor motif, type I	0.5	0.5	Infinity	1.02	1.40

Guaranteed cutoff and relaxed cutoff are parameters used by FR3D. t_1, t_2, and t_3 are parameters used by FRMF.

3.1 Tetraloop

There are three subtypes of tetraloops: standard, with deletion and with insertion. Because the number of insertion cases is too small, they are not considered in the experiments. We construct two lists of the standard tetraloops and tetraloops with deletion, respectively, for identification. There are 19 instances of standard tetraloop and 10 instances of tetraloop with deletion. Search results are listed in Supplementary Tables 1 and 2. The first instances in the two tables are used as the query motifs. The parameters of FR3D used for searching tetraloops are shown in Supplementary Figures 1 and 2. PR curves for the results based on the three approaches are shown in Figure 6a and b. The corresponding AUC-PR values are listed in Table 3.

When searching the standard tetraloops, all the three methods find all the 19 standard tetraloops. However, there are four false positives in the results of FRMF and Apostolico *et al*. (2009) and one false positive in the result of FR3D. FR3D has the largest AUC-PR value of 0.996. FRMF and Apostolico *et al*. (2009) have similar AUC-PR values.

When searching the tetraloops with deletion, all the three methods have some difficulty to discriminate the deletion cases from the standard ones. Compared with the other two methods, FRMF has the largest AUC-PR value of 0.588, while FR3D has the smallest AUC-PR value (only 0.316).

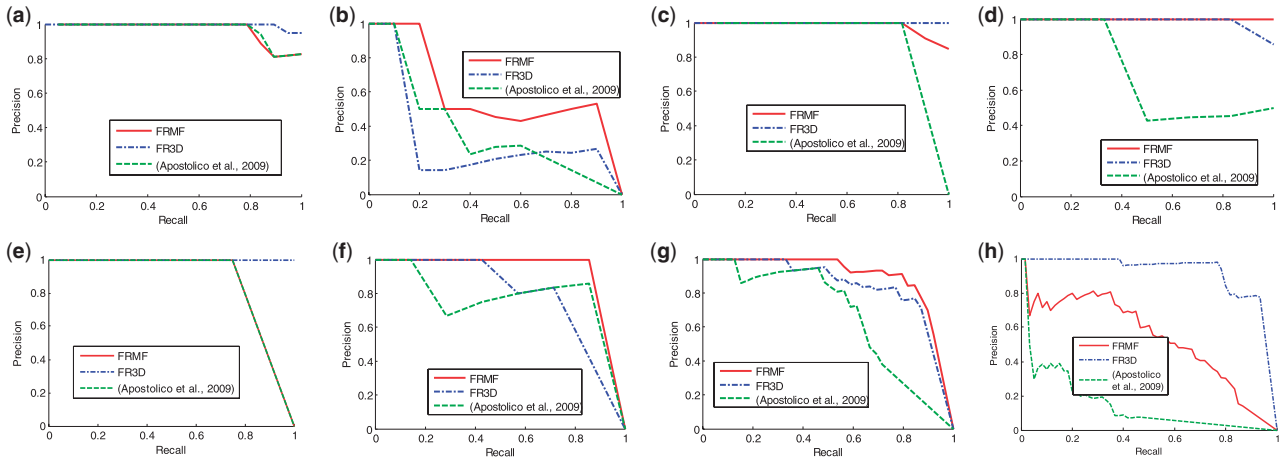

Fig. 6. PR curves for the three methods on searching eight subtypes of RNA motifs. (**a**) Standard tetraloop; (**b**) tetraloop with deletion; (**c**) core of sarcin/ricin loop; (**d**) local kink-turn; (**e**) (composite and local) kink-turn; (**f**) π-turn; (**g**) ribose zipper; (**h**) type I A-minor motif.

Table 3. AUC-PR values of FRMF, FR3D and the method proposed in Apostolico *et al*. (2009) for searching different motifs

Motif	FRMF	FR3D	Apostolico *et al*. (2009)
Tetraloop, standard	0.970	0.996	0.973
Tetraloop, with deletion	0.588	0.316	0.330
Sarcin/ricin loop, core	0.985	1	0.864
Kink-turn, local	1	0.988	0.680
Kink-turn (composite and local)	0.840	0.944	0.840
π-turn	0.929	0.733	0.772
Ribose zipper	0.884	0.808	0.617
A-minor motif, type I	0.528	0.898	0.129

3.2 Sarcin/ricin loop

In HM 50S, there are 11 core of sarcin/ricin loops. Search results of FRMF, FR3D and Apostolico *et al*. (2009) are shown in Supplementary Table S3. FRMF and FR3D both find all 11 positive motifs, but there are two false positives in the FRMF result. Apostolico *et al*. (2009) only finds nine true positives with two false negatives. PR curves for the three methods are shown in Figure 6c. The AUC-PR values are listed in Table 3. FR3D has the largest AUC-PR value 1, and Apostolico *et al*. (2009) has the smallest AUC-PR value of 0.864.

3.3 Kink-turn

There are six local kink-turn motifs appearing in HM 50S. According to Supplementary Table S4, FRMF, FR3D and Apostolico *et al*. (2009) find all six local kink-turn motifs. There is a false positive appearing in the result of FR3D and six false positives in the result of Apostolico *et al*. (2009). FRMF has the largest AUC-PR value 1 and Apostolico *et al*. (2009) has the smallest AUC-PR value of 0.680.

Composite kink-turn motif consists of three distinct strands. Different from the local kink-turn, the characteristic strand of the composite kink-turn is coupled with the two strands and forms two helices. Composite kink-turns cannot be completely found using the local motifs, because there are differences in the complementary strands. To search these composite kink-turns, only the characteristic strand of a local kink-turn is used as the query motif. All kink-turns, including local and composite kink-turns, should be returned as positive instances in the results.

There are eight kink-turns in HM 50S. Supplementary Table S5 shows that FR3D recognizes seven of them with one false negative. FRMF recognizes six positives with one false positive and two false negatives. Apostolico *et al*. (2009) recognizes six of them with two false negatives. FR3D has the largest AUC-PR value of 0.944. FRMF and Apostolico *et al*. (2009) have the same AUC-PR value of 0.84.

3.4 π-turn

In HM 50S, there are six previously discovered π-turns (Apostolico *et al*., 2009; Wadley and Pyle, 2004). From the experiments, a new π-turn 'G269 U270 C271 A272 G273 00000' was identified by FRMF. Therefore, there are seven positive π-turns. Search results are shown in Supplementary Table S6 and PR curves are shown in Fig. 6f. FRMF successfully finds six of them with one false negative. Apostolico *et al*. (2009) also finds six π-turns but with one false positive. FR3D identifies five of them with two false negatives. FRMF has the largest AUC-PR value of 0.929, while FR3D has the smallest AUC-PR value of 0.733.

3.5 Ribose zipper

The ribose zipper motif is composed of four nucleotides. The query motif used for searching ribose zipper consists of A160, A161, C769 and C770 on chain 0. PR curves for the three methods are shown in Figure 6g. The AUC-PR values for the three methods are listed in Table 3. According to the results, FRMF has the largest AUC-PR value of 0.884, while Apostolico *et al*. (2009) has the smallest AUC-PR value of 0.617.

3.6 A-minor motif

A-minor motifs consist of four types: type I, type II, type III and type 0. Because type I A-minor motif is the most common subtype appearing in RNA subunits, we compile an independent list for type I A-minor motifs following the criteria defined in Xin *et al*. (2008).

The query motif used for searching type I A-minor motif is 'A521 C637 G1364 000'. PR curves for the three methods are shown in Figure 6h. The corresponding AUC-PR values are listed in Table 3. In this experiment, FR3D has the largest AUC-PR value of 0.898. Apostolico *et al*. (2009) has the smallest AUC-PR value (only 0.129). On the other hand, we observe that FRMF is capable of discovering non-standard type I A-minor motifs. There are 22 non-standard type I A-minor motifs discovered that can help researchers to better understand various RNA motif structures.

We have further compared FRMF and FR3D with RNAMotifScan (Zhong *et al*., 2010). In general, this approach is quite different from FRMF and FR3D due to its adoption of a graph representation for the RNA motif, and its requirement of the availability of the secondary structure of the motif. We performed the comparison using the local kink-turns and the core of sarcin/ricin loops, based on which FRMF and FR3D attain the best search result, respectively. These motifs were also the focus of the studies in Zhong *et al*. (2010). For local kink-turns, the AUC-PR value of RNAMotifScan is 1 (all six positive instances were found with no false positives), the same value as that of FRMF. For sarcin/ricin loops, the AUC-PR value is 0.932 which is smaller than that of FRMF and FR3D.

Based on all the results, we can thus conclude that FRMF serves as an important complementary approach to FR3D for searching specific types of RNA motifs. In addition, FRMF is also capable of finding new motifs, as we shall now discuss below.

4 DISCUSSION

When evaluating the three methods, we discover a number of new motifs in the experiments. The first discovery is a new π-turn motif G269-G273, the structure of which is shown in Figure 7a. The RMSD between the backbone of this new motif and the standard π-turn G1873-G1877 is 0.56 Å (The RMSD values of the other five π-turns to G1873-G1877 range from 0.31 Å to 0.55 Å. Structure alignments are shown in Fig. 7c). Beside this new π-turn, we also discovered a motif, U866-G870, which has a structure similar to a π-turn (Fig. 7b). U866-G870 also connects two helices and forms a π shape. Compared with the standard π-turns, the first nucleotide on the 5′-side (U866) is elongated and far away from the neighboring nucleotide A867. The vacancy between U866 and A867 is occupied by a remote nucleotide A776. The RMSD between the backbones of

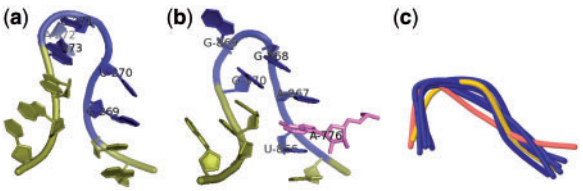

Fig. 7. (**a**) New π-turn G269-G273, chain 0 from HM 50S. The five nucleotides (blue) are flanked by two helical strands (yellow); (**b**) the π-turn-like motif U866-G870. Its five nucleotides are also flanked by two helices, but a remote nucleotide (A776, in violet) is inserted that results in a possible composite motif; (**c**) we aligned the backbones of seven positive π-turns and motif U866–870 using the Jmol software. The new π-turn G269-G273 is shown in yellow and the motif U866-G870 in pink.

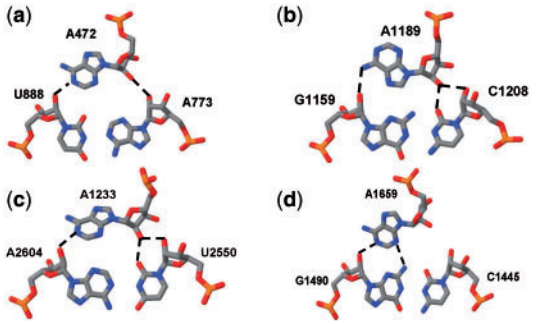

Fig. 8. Non-standard type I A-minor motif. (**a**) A docks into an A–U base pair; (**b**) A docks into a G–C base pair with its base rotated by 180°; (**c**) A docks into an A–U base pair with its base rotated by ∼60°; (**d**) only N1 and N3 of the inserted A form hydrogen bonds with the G–C base pair.

U866-G870 and G1873-G1877 is 0.6 Å (Fig. 7c). U866-G870 can be regarded as a special case of π-turn because U866 is far away from its standard position. It can also be regarded as a possible composite π-turn if U866 is replaced by A776.

Another main discovery is the observation of some instances of non-standard type I A-minor motif. Standard type I A-minor motif consists of four hydrogen bonds (Fig. 3b). However, the examples presented in Figure 8 consist of fewer hydrogen bonds than the standard type, which indicates that these non-standard instances are easier to form, but could be less stable than the standard ones. In Figure 8a, an adenosine is inserted into an A–U pair instead of a G–C pair. No hydrogen bond is formed between N3 of the inserted A and the bases of the A–U pair. In Figure 8b and c, the bases of the inserted A rotate by a certain angle (180° in Fig. 8b and ∼60° in Fig. 8c). After rotation, N3 is far away from the Watson–Crick base pair and fails to form a hydrogen bond. In Figure 8d, the ribose sugar and phosphate of the inserted A are raised from their typical positions. As a result, only N1 and N3 of the inserted A interact with the N2 and 2′-OH of G. Through the experiments, 38 standard instances and 22 non-standard instances are observed. The large number of non-standard cases indicates that type I A-minor motifs are not as stable as previously supposed.

These newly observed motifs lend evidence to the possible variations of π-turns and type I A-minor motifs. They also provide additional information about the RNA tertiary structures and functions.

5 CONCLUSION

In this article, we compile a set of positive RNA 3D motifs occurring in HM 50S. This set includes six characteristic types and eight subtypes of RNA 3D motifs. In addition to the published motifs, this set also contains a new π-turn motif and 22 non-standard instances of type I A-minor motifs discovered by us. FRMF, FR3D and the method used in Apostolico *et al*. (2009) have been evaluated on this positive set. The experimental results suggest that: (i) the method used in Apostolico *et al*. (2009) does not achieve the best result for any of the motifs; (ii) FRMF outperforms FR3D in the case of searching tetraloops with deletion, local kink-turns, π-turns and ribose zippers (AUC-PR value of FRMF is greater by 27, 1, 20 and 8 percentage points, respectively); (iii) FR3D outperforms FRMF in the other cases. As a result, the advantage of FRMF lies in serving as an important complementary approach to FR3D for particular types of RNA motifs and its capability to find new motifs.

In this work, we have adopted the simplified criterion that, if the distance between the candidate and the query motif is lower than a threshold, the candidate is considered to be of the same type as the query motif. While this notion may not necessarily be the standard one used in motif searching, its adoption makes it easier to compare the performance and rankings of different approaches. In practice, since the measures disc_1, disc_2 and disc_3 are decoupled, we can set the corresponding thresholds t_1, t_2 and t_3 separately by selecting from a suitable transition region between a group of small and more cohesive discrepancy values, and a group of comparatively large and less cohesive values based on their overall distribution. Another possible way is that the thresholds can be determined through an adaptive example-based learning approach based on a set of representative samples for different types of motifs.

In addition, parameters used by FRMF for searching different motifs are not standardized. In future, we shall focus on normalizing the discrepancies of the candidates from different types of motifs to the same range to facilitate the selection of a suitable set of parameters.

Finally, distance histograms, apart from its use in the candidate filtering step, could also serve as a computationally efficient approach during the candidate generation step to distinguish between substructures with grossly different shapes. We shall investigate this approach for candidate generation in our future work.

Funding: City University of Hong Kong (Project No. 7008044); National Natural Science Foundation of China (Project No. 61003174); Natural Science Foundation of Guangdong Province, China (Project No. 10451064101004233); Fundamental Research Funds for the Central Universities (Project No. 2009ZM0255); Doctoral Fund of Ministry of Education of China (Project No. 20100172120031).

Conflict of Interest: none declared.

REFERENCES

Amaral,P.P. and Mattick,J.S. (2008) Noncoding RNA in development. *Mamm. Genome*, **19**, 454–492.

Apostolico,A. *et al*. (2009) Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res.*, **37**, e29.
Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
Cate,J.H. *et al*. (1996) Crystal structure of a group I ribozyme domain: principle of RNA packing. *Science*, **273**, 1678–1685.
Correll,C.C. and Swinger,K. (2003) Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA*, **9**, 355–363.
Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. *Proc. ICML*, **148**, 233–240.
Djelloul,M. and Denise,A. (2008) Automated motif extraction and classification in RNA tertiary structures. *RNA*, **14**, 2489–2497.
Duarte,C.M. *et al*. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.
Flusser,J. *et al*. (2010) *Moments and Moment Invariants in Pattern Recognition.* John Wiley & Sons, UK.
Francois,B. *et al*. (2005) Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. *Nucleic Acids Res.*, **33**, 5677–5690.
Gendron,P. *et al*. (2001) Quantitative analysis of nucleic acid three-dimensional structure. *J. Mol. Biol*., **308**, 919–936.
Harrison,A.M. *et al*. (2003) Representation, searching discovery of patterns of bases in complex RNA structures. *J. Comput. Aided Mol. Des.*, **17**, 537–549.
Hsiao,C. *et al*. (2006) Single nucleotide RNA choreography. *Nucleic Acids Res.*, **34**, 1481–1491.
Huang,H.C. *et al*. (2005) The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA*, **11**, 421–423.
Jaeger,L. *et al*. (1994) Involvement of a GNRA tetraloop in long-range tertiary interactions. *J. Mol. Biol.*, **236**, 1271–1276.
Jonikas,M.A. *et al*. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
Klein,D.J. *et al*. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.
Klosterman,P.S. *et al*. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.*, **30**, 392–394.
Laing,C. and Schlick,T. (2010) Computational approaches to RNA 3D modeling. *J. Phys. Condens. Matter*, **22**, 283101.
Lescoute,A. *et al*. (2005) Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
Levina,E. and Bickel,P. (2001) The Earth Mover's Distance is the Mallows Distance: some insights from statistics. *Proc. ICCV*, 251–256.
Major,F. *et al*. (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, **253**, 1255–1260.
Mamistvalov,A.G. (1998) N-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids. *IEEE Trans. Pattern Anal. Mach. Intell*, **20**, 819–831.
Mattick,J.S. *et al*. (2009) RNA regulation of epigenetic processes. *BioEssays*, **31**, 51–59.
Mercer,T.R. *et al*. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
Michal,S. *et al*. (2007) Finding a common motif of RNA sequences using Genetic Programming: the GeRNAMo system. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **4**, 596–610.
Nissen,P. *et al*. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Nat. Acad. Sci. USA*, **98**, 4899–4903.
Pang,K.C. *et al*. (2009) Genome-wide identification of long noncoding RNAs in $CD8^+$ T cells. *J. Immunol.*, **182**, 7338–7348.
Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
Rabani,M. *et al*. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA*, **39**, 14885–14890.
Reeder,J. *et al*. (2007) Locomotif: from graphical motif description to RNA motif search. *Bioinformatics*, **23**, i392–i400.
Rubner,Y. *et al*. (1998) A metric for distributions with applications to image databases. *Proc. ICCV*, 59–66.
Sargsyan,K. and Lim,C. (2010) Arrangement of 3D structural motifs in ribosomal RNA. *Nucleic Acids Res.*, **38**, 3512–3522.
Sarver,M. *et al*. (2008) FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.
Soldà,G. *et al*. (2009) An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. *Brief. Bioinformatics*, **10**, 475–489.
Sommer,I. *et al*. (2007) Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, **23**, 3139–3146.
Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.
Xin,Y. *et al*. (2008) Annotation of tertiary interactions in RNA structures reveals variations and correlation. *RNA*, **14**, 2465–2477.
Yang,H. *et al*. (2003) Tools for the automatic identification and classification RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
Yao,Z. *et al*. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
Yu,Z. and Herman,G. (2005) On the Earth Mover's Distance as a histogram similarity metric for image retrieval. *Proc. ICME*, 686–689.
Zhong,C. *et al*. (2010) RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res*., **38**, e176.