# RNA structural motif recognition based on least-squares distance

**YING SHEN,[1] HAU-SAN WONG,[2,4] SHAOHONG ZHANG,[3] and LIN ZHANG[1]**

[1]School of Software Engineering, Tongji University, Shanghai 200092, China
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[3]Department of Computer Science, Guangzhou University, Guangzhou 510006, China

## ABSTRACT

RNA structural motifs are recurrent structural elements occurring in RNA molecules. RNA structural motif recognition aims to find RNA substructures that are similar to a query motif, and it is important for RNA structure analysis and RNA function prediction. In view of this, we propose a new method known as RNA Structural Motif Recognition based on Least-Squares distance (LS-RSMR) to effectively recognize RNA structural motifs. A test set consisting of five types of RNA structural motifs occurring in *Escherichia coli* ribosomal RNA is compiled by us. Experiments are conducted for recognizing these five types of motifs. The experimental results fully reveal the superiority of the proposed LS-RSMR compared with four other state-of-the-art methods.

Keywords: RNA structural motif; RNA motif recognition

## INTRODUCTION

Long noncoding RNA (ncRNA), which consists of >200 nt and has little or no protein-coding capability, is a remarkable class of RNAs. They are important in various biological processes (Dinger et al. 2008; Guttman et al. 2009, 2011; Mattick et al. 2009; Pang et al. 2009). In their tertiary structures, some of the nucleotides will form special substructures, which have consensus structural patterns and occur repetitively in RNA molecules. These substructures are called RNA structural motifs. They are building blocks of different RNA architectures and play an important role in binding proteins and consolidating RNA tertiary structures (Moore 1999; Leontis et al. 2002; François et al. 2005; Hendrix et al. 2005; Leontis et al. 2006).

Many types of RNA structural motifs have been identified: tetraloop (Woese et al. 1990), sarcin/ricin loop (Szewczak et al. 1993), kink-turn (Klein et al. 2001), π-turn (Wadley and Pyle 2004), and C-loop (Leontis and Westhof 2003), to name a few. Recently, several new types of motifs have been identified: G-ribo (Steinberg and Boutorine 2007), UA_handle (Jaeger et al. 2008), and A-wedge (Gagnon and Steinberg 2010). Certain special types of RNA structural motifs are also regarded as tertiary interactions, such as A-minor motif (Nissen et al. 2001), ribose zipper (Cate et al. 1996), and kissing hairpin (Chang and Tinoco 1994).

Using a known RNA motif as the query, RNA structural motif recognition aims to find all of its occurrences in an RNA molecule (i.e., the search space). There are two prevalent ways to search for RNA structural motifs. The first class of methods is to find motifs based on their geometric features. NASSAM (Harrison et al. 2003) represents RNA motifs and the search space by graphs which are constructed using the distances between atoms of key nucleotides. Then the Ullman algorithm is used to search for RNA motifs on the graph. PRIMOS (Duarte et al. 2003), COMPADRES (Wadley and Pyle 2004), and AMIGOS II (Wadley et al. 2007) characterize RNA structures and motifs using two pseudotorsion angles of the backbones. The distance between two RNA substructures can be calculated based on the distances between their pseudotorsion angles. Recently, FASTR3D (Lai et al. 2009), an online server for RNA 3D structure search, also adopts this representation scheme as well as the same distance measure. Different from these methods, FR3D (Sarver et al. 2008), one of the prevalent geometry-based methods, integrates geometry search and symbolic search. When performing geometry search, FR3D uses the base centers of the nucleotides to represent RNA structures and motifs. It measures the fitting error and the orientation error between the query motif and the candidate to compute their overall discrepancy. Additionally, Apostolico et al. (2009) and Sargsyan and Lim (2010) use shape histogram as the signature of RNA motifs. The distance between two RNA substructures is measured by the cosine distance between their shape histograms. The other trend in RNA structural motif recognition is to use an RNA 2D diagram of base pair isostericity to search for motifs. The rationale of RNAMotifScan (Zhong et al. 2010) is based on the assumption that RNA substructures that have isosteric edges to the query

motif should have similar shapes. Therefore, motifs that are not conserved in sequence can also be found.

Despite the great efforts spent in RNA structural motif recognition, there remain several problems for further investigation. The first problem is how to find a suitable representation scheme for RNA structures and query motifs. In geometry-based methods, RNA substructures are represented by sets of atoms or pseudoatoms. Most atoms have fixed positions, and the distances between them will not change dramatically. However, RNA motifs may have some free bases that are allowed to rotate along the backbone. The movement of these atoms will bring in local deformation to the query motif. A suitable representation scheme can remove such kind of deformation and preserve the invariance within a motif type to a large extent. Meanwhile, it should have enough discriminating power to separate those positive instances from the negative ones. The second problem is how to generate candidates. If we simply compare the query motif with all RNA fragments of the same length, the time cost will become unacceptable. Therefore, a candidate generation module that produces a reasonable number of candidates becomes necessary.

In this paper, we propose a new method, namely RNA Structural Motif Recognition based on Least-Squares distance (LS-RSMR), which aims to solve the aforementioned problems. LS-RSMR divides the process of RNA structural motif recognition into two steps. First, a moderate-sized set of candidates is generated. Then, these candidates are filtered according to their distances to the query motif. Specifically, for each candidate, LS-RSMR will compute a least-squares distance value by superimposing the candidate on the query motif. Those candidates with smaller distances will be regarded as positive instances.

## RESULTS

In the experiments of evaluating the performance of the proposed LS-RSMR, we first construct a query set and a test set using motifs from the HM50S ribosomal RNA (PDB: 1S72) and the *Escherichia coli* ribosomal RNA (PDB: 2QBE), respectively. Query motifs are selected from the query set, and performance is evaluated on the test set. Meanwhile, we also evaluate four other state-of-the-art methods on the same data set and compare their performance with that of LS-RSMR. The experimental results confirm the effectiveness of our LS-RSMR method.

### Query motif set and search space

We have previously constructed a data set which contains six types (including eight subtypes) of RNA structural motifs in HM50S ribosomal RNA (Shen et al. 2011). In the current experiments, five types of motifs (tetraloop, sarcin/ricin loop, $\pi$-turn, k-turn, and ribose zipper) are selected to construct a query set. The instances are listed in Supplemental Data Set 1 in the Supplemental Material. The structure alignment along the

backbone has been performed and shown in Supplemental Figure S1. The rest of the instances of the same type are aligned to the first motif shown on the list. In the previous work, when evaluating a proposed method, researchers usually arbitrarily chose one of the published motifs as the query. However, for each method, different query motifs will result in different performance. In order to perform an impartial comparison, in our experiments, we select the instance that has the smallest sum of square distance to the other positive instances as the query. The distance between two positive instances is computed using the least-squares distance introduced in Material and Methods. For each query, one instance is selected from the query set and used as the query template. The search is conducted only based on the structure of the selected query motif without any other information of the query set.

We choose another ribosomal RNA, *E. coli* ribosomal RNA, as the search space. We construct a test set that contains the same types of motifs as in 1S72. The sequences of these motifs are also listed in the Supplemental Material (see Supplemental Data Set 2). Their structure alignment is shown in Supplemental Figure S2. The numbers of motifs in the query set and test set are shown in Supplemental Table S1.

All the positive instances collected in the query set and test set are RNA fragments having consensus structural patterns. The reason is that those fragments having consensus structural patterns will be considered to be the most important during a search. Although some nonstandard fragments will be deemed to be similar to the query motif by certain methods (such as RNAMotifScan), generally they are less important than the standard ones. In addition, without expertise it is difficult to determine whether they are of the same type as that of the query motif. Therefore, those nonstandard fragments are currently not studied in the query and test sets.

### Methods for performance evaluation

In the following experiments, we search for five types of motifs in 2QBE, and the proposed LS-RSMR is compared with four state-of-the-art methods: FR3D (Sarver et al. 2008), shape histogram used by Apostolico et al. (2009), AMIGOS II (Wadley et al. 2007), and RNAMotifScan (Zhong et al. 2010). The results of FR3D are generated using WebFR3D (Petrov et al. 2011) with only the geometric search function used. Its parameter *discrepancy* was set to 0.9, which is the largest discrepancy value accepted by WebFR3D. Consequently, WebFR3D could return as many search results as possible and achieves its best performance. The method in Apostolico et al. (2009) has two parameters to adjust: thresholds for cosine distance and RMSD. The default values of cosine distance and RMSD used by the authors are as follows: (0.95, 2 Å) for tetraloops, (0.9, 4 Å) for single strand of k-turns, and (0.75, 2 Å) for $\pi$-turns. Based on these default values, we adjusted the two thresholds to allow the method to achieve a better performance. In our experiments, when searching for tetraloops, k-turns, and $\pi$-turns, thresholds for cosine distance and RMSD

were set to (0.95, 2 Å), (0.85, 2 Å), and (0.75, 2 Å), respectively. For the cases of core of sarcin/ricin loops and ribose zippers, the thresholds for cosine distance were both set to 0.9 (RMSD is not considered for the two-loop motifs). When evaluating RNAMotifScan, the threshold (false positive rate) was set to one, which is the largest value of the threshold, so that RNAMotifScan could return the largest number of search results and achieve its best performance.

Each method ranks candidates and lists them according to their distances (e.g., *discrepancy* in FR3D and $d_{LS}$ in LS-RSMR) to the query motif. We plot the ranked lists in a positive/negative space (see Supplemental Fig. S3).

Intuitively, the more positive candidates appear at the top of the output list, the more effective is the method. We can quantify this using the precision at rank $k$ [Precision($k$)] measure as defined below:

$$\text{Precision}(k) = \frac{\#\,\text{of true positives in the first } k \text{ instances}}{k} \quad (1)$$

In order to quantitatively measure their performance, we adopt the concept of Average Precision (AP) (Zhu 2004; Manning et al. 2008).

The AP values for the results in our experiments can be computed using Eq. (2) below:

$$\text{Precision}_{\text{AVG}} = \frac{\sum_{k=1}^{n} P(k) \times \text{label}(k)}{\#\,\text{of positive instances in test set}} \quad (2)$$

where $n$ is the number of ranked candidates, $k$ is the rank in the sequence of candidates, $P(k)$ is the precision at the $k$-th candidate, and label($k$) is 1 if the $k$-th candidate is positive and 0 if the candidate is negative.

## Results

Experiments were performed on a Dell Optiplex 760 (Intel Core2 Duo CPU 3.16 GHz). Results of the five computational methods on searching five types of RNA structural motifs are shown in Supplemental Figure S3. AP values are computed for the five methods and listed in Table 1. The best performance is highlighted in bold.

LS-RSMR, FR3D, AMIGOS II, and Apostolico et al. (2009) are geometry-based methods, which accept the coordinates of the same query motif as inputs. However, RNAMotifScan is based on extracted isostericity edges of structural patterns,

which is quite different from the other methods. Therefore, in our experiments, we use the same query motif for the first four methods and use the query template provided by RNAMotifScan package for RNAMotifScan. If there is no suitable query template, a template will be created by us, such as the case when searching ribose zippers. In some cases (e.g., k-turn search), the package of AMIGOS II also provides the query template. For this case, we will compute its performance by using the query motif and the query template, and the best performance is chosen and compared with other methods.

### Tetraloop

The query motif used by LS-RSMR, FR3D, Apostolico et al. (2009), and AMIGOS II is 9:G90-A93. There is a GNRA tetraloop template in RNAMotifScan package, and we use it as the query for RNAMotifScan.

In this experiment, FR3D achieves the best performance with its AP value equal to 0.971. Our LS-RSMR method is second best with an AP value of 0.905. The less satisfactory performance of LS-RSMR is due to a false positive in the results of LS-RSMR but that is not found in the output list of FR3D: B:G1807-A1810 (see Supplemental Fig. S4a). We align this sequence with the query motif along the backbone (see Supplemental Fig. S4b). Although it is quite similar to the query motif, we currently consider it as a negative instance.

### Core of sarcin/ricin loop

AMIGOS II can only search motifs consisting of a single strand. Therefore, when searching sarcin/ricin loops, only the other four methods are compared. The query motif for LS-RSMR, FR3D, and Apostolico et al. (2009) is 0:G225-A227/A212-A215. The sarcin/ricin loop template in the RNAMotifScan package is used as the query for RNAMotifScan.

In this experiment, LS-RSMR and FR3D successfully find all nine sarcin/ricin loops and achieve the best performance. RNAMotifScan only finds seven loops. Apostolico et al. (2009) identifies six sarcin/ricin loops with a false positive among them.

### π-turn

When searching π-turns, we only compare four methods, excepting RNAMotifScan. This is because RNAMotifScan uses a base pair isostericity diagram as input. The user should first define isostericity of base pairs in the query motif. First, π-turns consist of only one strand without base pairs inside. Second, the isostericity of base pairs formed between nucleotides in π-turns and other residues varies significantly. It is difficult to construct a consensus base pair isostericity matrix for π-turns. Therefore, only the other four methods are used to search π-turns.

**TABLE 1.** AP values of five computational methods on searching five types of RNA structural motifs

| Motif | LS-RSMR | FR3D | Apostolico et al. (2009) | AMIGOS II | RNAMotifScan |
|---|---|---|---|---|---|
| Tetraloop | 0.905 | **0.971** | 0.750 | 0.879 | 0.431 |
| Sarcin/ricin-loop | **1** | **1** | 0.632 | NA | 0.859 |
| π-turn | **1** | 0.662 | **1** | 0.324 | NA |
| k-turn | **1** | **1** | **1** | 0.667 | 0.833 |
| Ribose zipper | **0.914** | 0.821 | 0.644 | NA | 0.001 |

The query motif used by LS-RSMR, FR3D, Apostolico et al. (2009), and AMIGOS II is 0:G1873-G1877. LS-RSMR and Apostolico et al. (2009) successfully find all seven positive instances with AP values equal to one. FR3D and AMIGOS II only find five and three of them, respectively.

### K-turn

We use the characteristic strand to search k-turns so that both local and composite k-turns will be found. The query motif used by LS-RSMR, FR3D, Apostolico et al. (2009), and AMIGOS II is 0:C1146-A1154. The k-turn template in the AMIGOS II package is also used as the query for AMIGOS II. The template in RNAMotifScan package is used as the query for RNAMotifScan.

LS-RSMR, FR3D, and Apostolico et al. (2009) find all three k-turns with the highest performance. RNAMotifScan only finds two of them. When using the query motif and the query template, AMIGOS II achieves the same performance—it only finds two k-turns in both cases.

### Ribose zipper

AMIGOS II cannot be used for searching ribose zipper because ribose zipper consists of two strands. We compare the performance of the other four methods. The query motif used by LS-RSMR, FR3D, and Apostolico et al. (2009) is 0: A520-A521/0:C637-C638. Because there is no suitable query template in the RNAMotifScan package, we construct a query template for RNAMotifScan: iso: AA … CC; edge: SS … SS; struc: ((…)).

LS-RSMR achieves the best performance with its AP value equal to 0.914, ∼10% higher than the second best method, FR3D. RNAMotifScan only returns four positive instances. Considering that there are 44 ribose zippers in 2QBE, the performance of RNAMotifScan is not satisfactory.

Results of the five computational methods show that LS-RSMR identifies more positive instances of all five types of motifs. It achieves a good performance in most cases, and experimental results demonstrate the effectiveness of our LS-RSMR method.

## NEW MOTIF DISCOVERY

In the above experiments, we successfully use LS-RSMR to recognize RNA motifs. Next, we will show the effectiveness of LS-RSMR in extracting new motifs.

The RNA motif discovery problem aims to find new structural patterns occurring in RNA molecules. In order to find new motifs, we propose a method based on LS-RSMR. Specifically, we first generate a set of 2-nucleotide (nt) substructures, each of which consists of a pair of neighboring nucleotides along the chain of the RNA molecule. For each substructure, we add a nucleotide whose distance is smaller than 4 Å to either nucleotide in the 2-nt substructure. In this way, a set of 3-nt substructures can be obtained. Using LS-RSMR, we perform a search using each 3-nt substructure. Setting a stringent threshold, if there are more than two results returned, we will keep the corresponding substructure and refer to it as a 3-nt seed. After finding all 3-nt seeds, we can construct a 3-nt seed set, which may contain potential 3-nt motifs according to the definition of RNA motifs. Based on this set, we can add a new nucleotide to each of its elements and obtain a 4-nt substructure set. Similarly, using LS-RSMR, we shall check each 4-nt substructure and determine whether or not they are 4-nt seeds.

In this paper, we describe two newly discovered motifs to show the effectiveness of LS-RSMR in new motif extraction.

The first new motif is a tertiary interaction composed of three nucleotides. In this motif, two nucleotides (nt 1 and nt 2 in Fig. 1A) form a consecutive strand, which is part of a helix. The third (nt 3) has interactions with nt 1 and nt 2. Specifically, the base of nt 3 interacts with the 2′-OH of nt 1. The 2′-OH of nt 3 interacts with the base and sugar of nt 2. Six occurrences are shown in Figure 1A–F. This type of motif brings two RNA strands close to each other and a groove, where proteins can be embedded, is formed as a consequence. This unique structure allows proteins and RNA to function together. Figure 1G and H reveal two typical bound structures of this motif with ribosomal proteins.

The second new motif is composed of four consecutive nucleotides. The start and the end nucleotide (nt 1 and nt 4 in Fig. 2A) are part of a helix, and the other two nucleotides (nt
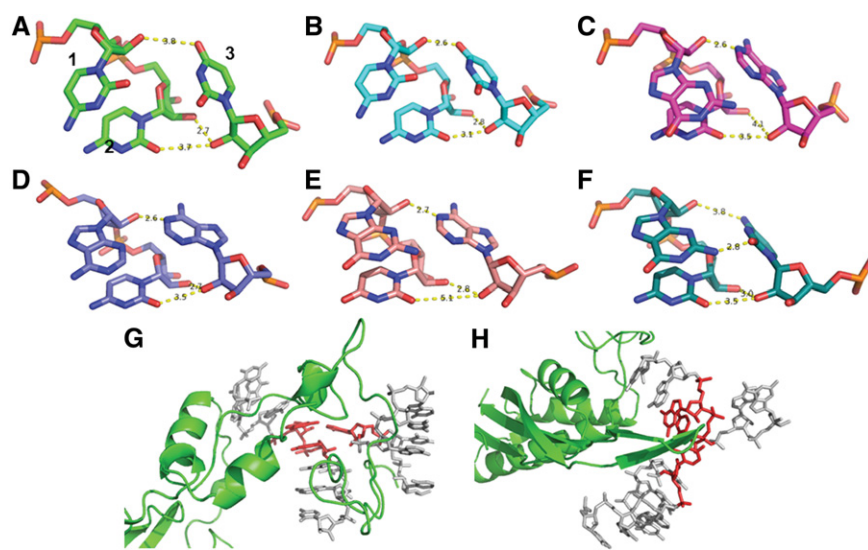


**FIGURE 1.** The structure of the first new motif. (*A*) 0:C2318 C2319 U2322; (*B*) 0:C1403 C1404 U1408; (*C*) 0:G2365 C2366 A2370; (*D*) 0:A1242 C1243 A1247; (*E*) 0:G1849 U1850 A1941; (*F*) 0: G1373 C1374 C1431; (*G*) the structure of *E* and protein L2; (*H*) the structure of *F* and protein L22.
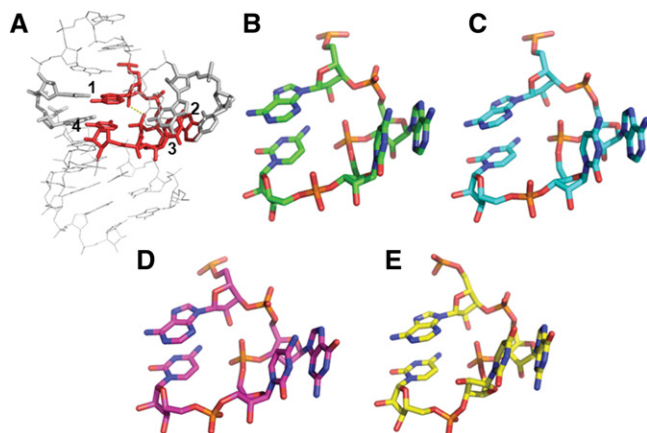
**FIGURE 2.** (*A*) The formation of the second new motif (0:A339-C342) in the HM50S subunit. Numbering is from 5′ to 3′. (*B*) 0:A339-C342; (*C*) 0:A1392-C1395; (*D*) 0:A1448-C1451; (*E*) 0:A2074-C2077.

2 and nt 3) form a bulge, which extrudes outside. The bases of nt 2 and nt 3 are parallel and participate in a strand of another helix (see Fig. 2A).

The sequence of this motif is quite conserved. We identified four occurrences, the structures of which are shown in Figure 2B–E. The sequence pattern of the four occurrences can be summarized as ARYC, where R represents purine and Y represents pyrimidine. We found that the nucleotide at the start position of the 5′-side is always an adenine and at the end position it is always a cytosine. The bulge extruding out contains a purine and a pyrimidine. The shape of this motif is maintained by the hydrogen bond formed between the 2′OH of nt 1 and the phosphate group of nt 3 (see Fig. 2A, the hydrogen bond is shown as the yellow dashed line).

Using these two new types of motifs as queries, we conducted two searches across 550 RNA molecules from a non-redundant list (http://rna.bgsu.edu/rna3dhub/nrlist/release/1.3/all), which includes ribosomal RNAs, ribozymes, and riboswitches. Using two stringent cutoff values ($d_0 = 5$ for the first new motif and $d_0 = 10$ for the second motif), we identified 45 instances of the first new motif in 17 RNA molecules and 55 instances of the second motif in 31 RNA molecules (see Supplemental Table S3, S4). We find that these instances possess consensus structural patterns, and their structures are conserved across species and different types of RNAs.

These newly discovered motifs indicate that LS-RSMR is effective to find similar substructures according to the given template from different types of RNA molecules. It is effective not only for the known RNA motifs, but also for unknown substructures.

## DISCUSSION

In RNA motif recognition, LS-RSMR, when compared with FR3D, produces fewer candidates. For example, when search-

ing ribose zipper in 2QBE, LS-RSMR returns 399 candidates, whereas FR3D returns 12,323 candidates. Meanwhile, in the candidate set of LS-RSMR, there are more positive instances than that of FR3D. A positive candidate of ribose zipper, B: G2607-G2608/U1782-A1783 (see Fig. 3), is found by LS-RSMR but does not appear in the candidate list of FR3D.

When using FR3D to search RNA motifs, if there is no continuity information, FR3D will take a much longer time (>10 min) to generate candidates and return search results. However, without any continuity constraint, LS-RSMR can generate candidates within 1 min. In addition, our representation scheme utilizes information from more atoms of the motifs than FR3D, which increases the discriminating power.

LS-RSMR has only one parameter: a cutoff threshold $d_0$. Users can easily filter candidates by setting a suitable cutoff value. Other methods, such as AMIGOS II and Apostolico et al. (2009), have more than one parameter that needs to be set.

LS-RSMR recognizes RNA motifs based on the degree of structural similarity. Compared with RNAMotifScan, it does not require extra knowledge on base pair isostericity. The query can be a single strand in which there can even be no base pair. In addition, when searching motifs that have no consensus base pair isostericity, e.g., π-turns, ribose zipper, its effectiveness is further highlighted.

LS-RSMR can search both local and composite motifs. In the candidate generation process, nucleotides are added to the partial candidates as long as they have interaction with the existing nucleotides. It does not matter whether they are in the same strand of the partial candidates or not.

## CONCLUSIONS

In this paper, we have proposed a new method, namely LS-RSMR, for RNA structural motif recognition. To quantitatively measure the performance of LS-RSMR, we apply it to search for five types of motifs occurring in 2QBE. We also compare the performance of LS-RSMR with four other state-of-the-art methods. The experimental results show the effectiveness of our method.

In addition to its effectiveness, LS-RSMR has three other advantages. First, only one parameter is required to be adjusted, which makes it easy to use. Second, it is capable of searching both local and composite motifs. Third, to start a search, it only requires the residue sequence numbers of the query
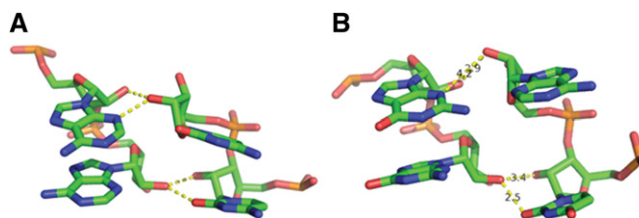


**FIGURE 3.** (*A*) A standard ribose zipper; (*B*) a positive candidate found by LS-RSMR in 2QBE: B:G2607-G2608/U1782-A1783.

nucleotides without any other structural knowledge about the query fragments.

We have also extended LS-RSMR for the discovery of new motifs. Currently, we successfully find two new motifs. These new discoveries further demonstrate the power of LS-RSMR in RNA substructure recognition.

In the near future, we plan to extend LS-RSMR to allow users to set constraints on the relationship between nucleotides in the generated candidates. The idea of symbolic search, which has been adopted by FR3D, will also be considered as our future work.

## MATERIALS AND METHODS

### RNA structural motif

RNA structural motifs are recurrent substructures occurring in RNA molecules. They are building blocks of RNA tertiary structures and are important for RNAs fulfilling their biological functions. Many types of motifs have been discovered, such as tetraloop, sarcin/ricin loop, kink-turn, π-turn, and ribose zipper. They are briefly introduced in the Supplemental Material.

### Representative structure of an RNA motif

Motifs in the same class share a similar structure. In general, their structures are similar but tiny differences exist, especially in the positions of some unrestricted bases. Since RNA motifs are largely invariant but slightly different due to local variations, the problem is how to find a representation scheme to preserve their identity and at the same time to have adequate discriminating power. Reijmers et al. (2001) argues that Cartesian coordinates is the most basic representation for RNA structures and other representations, such as torsion angles, can be deduced from it. Duarte et al. (2003) and Apostolico et al. (2009) use backbones as the representative structures of RNA motifs because backbones of motifs are relatively stable and similar within a class. On the other hand, FR3D (Sarver et al. 2008) utilizes the centroids of the bases as the representation of RNA structures.

After re-examining the structures of motifs, we found that the atoms from the backbone (including the phosphate group and sugar, i.e., P, OP1, OP2, C1′, C2′, C3′, C4′, C5′, O2′, O3′, O4′, and O5′) and the centers of the bases form a stable representative structure (see Fig. 4). Such a representation scheme can balance the impact of backbone and bases on the structural identity of RNA motifs to a certain extent. Since the number of atoms from the backbone is much larger than the base centers, once a candidate matches the query mo-
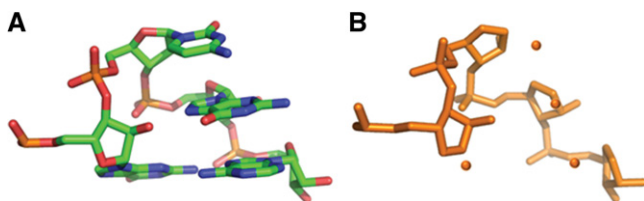


**FIGURE 4.** (*A*) A tetraloop; (*B*) its representative structure.

tif in their backbones, the difference in the base centers will not significantly affect the final matching score. Meanwhile, the positions of the base centers can help to align the candidate with the query motif. An only concern of using this representation scheme is that it may allow fragments that have rotated bases to appear in the results, such as the instance shown in Supplemental Figure S4.

### Least-squares distance between two representative structures

Similar to the approach adopted by FR3D, RNA structural motif recognition is partitioned into two phases by using the proposed LS-RSMR: the candidate generation phase and the candidate filtering phase. In this section, we describe the second phase in which candidates are filtered based on a least-squares distance measure. Specifically, we shall first introduce the least-squares distance measure and the candidate filtering method. The method for candidate generation will be described in "Candidate generation."

Suppose we have a set of candidates. The process of candidate filtering is that, given a query motif, each candidate will be compared with the query. A discrepancy value between each candidate and the query motif can be computed using a suitable distance measure. Candidates with smaller distances will be regarded as positive instances.

Under the representation scheme introduced previously, the RNA substructures, including query motif and candidates, can be treated as rigid objects. The distance between them can be measured by the sum of Euclidean distances between the corresponding atoms after a suitable linear transformation. The computational details can be found in the Supplemental Material.

In the problem of RNA structural motif recognition, we assume that there is no scale change between the two RNA substructures to be matched. Therefore, the scale factor $s$ in Supplemental Equation S1 is set to 1, instead of using the solution given by Supplemental Equation S2. We can compute a least-squares distance $d_{LS}$ using Supplemental Equations S1 and S2 between each candidate and the query motif. Given a threshold $d_0$, candidates whose $d_{LS}$ values are smaller than $d_0$ will be regarded as positive instances and listed in ascending order of $d_{LS}$. How to set $d_0$ will be discussed at the end of "Candidate generation."

Least-squares distance is also adopted by FR3D and some protein alignment methods (McLachlan 1979). We found that compared with other distance measures, least-squares distance has the best discriminating power in Euclidean space. Therefore, LS-RSMR also adopts this distance measure for RNA substructure comparison. Although both LS-RSMR and FR3D use the similar distance measure, the computed rotation matrices and translation vectors are different. This is because points in the representative structures constructed by the two methods are quite different; and consequently, the transformation parameters are also different. In addition, FR3D also incorporates an orientation error in addition to the least-squares fitting error. As a consequence, the search results of LS-RSMR are different from the results of FR3D.

### Candidate generation

Suppose the query motif contains $m$ nucleotides, the first stage of RNA structural motif recognition is to generate all possible $m$-nt RNA fragments in the search space. Specifically, given a search space

that contains $n$ nucleotides, the number of all $m$-nt fragments will be $n(n-1), \ldots, (n-m)$, i.e., $O(n^m)$. When $n$ is large, it is inefficient to retain this many candidates and compute their distances to the query motif. In view of this, an efficient candidate generation method is indispensable so that fewer candidates can be produced for filtering in the next phase. Previously, FR3D proposed a fast and efficient screening algorithm for determining a reasonable number of candidates. On the other hand, in LS-RSMR, we adopt a different, yet also efficient, candidate generation scheme for this purpose.

Our candidate generation method consists of two steps. The first step is to reorder the nucleotides in the query motif. The second step is to generate candidates in the search space according to the new order of the query motif.

### Determine a new order of nucleotides in the query motif

The sequence of nucleotides in the query motif is given by the user. We should first reorganize the sequence of these nucleotides so that in the second step candidates can be generated in a more efficient way. The pseudocode is shown in Supplemental Table S2. The process of reordering the query motif is as follows. Suppose the query motif contains $m$ nucleotides. We first construct a distance matrix $\mathbf{D} = [d_{ij}]_{i,j=1,\ldots,m}$. Each entry of $\mathbf{D}$, $d_{ij}$, is the smallest distance between atoms in nucleotides $i$ and $j$. Next, we will construct two lists. One of them is $\mathrm{Seq_{new}} = \{n'_i\}_{i=1,\ldots,m}$, which contains the new order of nucleotides in the query. The other is $\mathrm{ClosestPairs} = \{(n'_i, n'_j)\}_{j=2,\ldots,m;i<j}$. $(n'_i, n'_j)$ means $n'_i$ is the closest nucleotide to $n'_j$ in the first $j-1$ nucleotides of $\mathrm{Seq_{new}}$.

Suppose the given sequence of a query motif is $\mathrm{Seq_{old}} = \{n_1,n_2, \ldots, n_m\}$. First, we set $\mathrm{Seq_{new}} = \{n'_1 = n_1\}$. Then, we set $\mathrm{Seq_{remain}} = \{n_1,n_2,\ldots,n_m\}\backslash\mathrm{Seq_{new}}$, where '\' means the exclusion of members of $\mathrm{Seq_{new}}$ from $\{n_1,n_2,\ldots,n_m\}$. For each residue from $\mathrm{Seq_{remain}}$, we can always find a nucleotide $n_s$, which has the smallest minimum distance to one of the members in $\mathrm{Seq_{new}}$. Suppose $n'_i$ is the nucleotide closest to $n_s$ in $\mathrm{Seq_{new}}$, then we set $n'_{k+1} = n_s$ and $\mathrm{Seq_{remain}} = \mathrm{Seq_{remain}}\backslash n_s$, and $(n'_i, n'_{k+1})$ will be added to $\mathrm{ClosestPairs}$. This process is repeated until $\mathrm{Seq_{new}}$ contains all $m$ nucleotides.

### Construct candidates

Based on the two lists: $\mathrm{Seq_{new}} = \{n'_1, n'_2, \ldots, n'_m\}$, and $\mathrm{ClosestPairs} = (n'_i, n'_j)\}_{j=2,\ldots,m;i<j}$, we can now construct the candidates.

First, we select any one of the nucleotides in the search space and denote it by $c_1$. Now a partial candidate containing only one nucleotide has been created. Next, we should find the second nucleotide $c_2$ such that $\{c_1,c_2\}$ is similar to $\{n'_1, n'_2\}$ in $\mathrm{Seq_{new}}$. The similarity between $\{c_1,c_2\}$ and $\{n'_1, n'_2\}$ can be measured by Supplemental Equation S1. If the distance between $\{c_1,c_2\}$ and $\{n'_1, n'_2\}$ is smaller than a threshold $d_2$, then $c_2$ will be kept and $\{c_1,c_2\}$ forms a new partial candidate.

The naïve way to choose $c_2$ from the search space is to check all the nucleotides in the search space and see whether its combination with $c_1$ satisfies the similarity constraint. However, this is time consuming considering that there are thousands of nucleotides in the search space. When observing the first two nucleotides in $\mathrm{Seq_{new}}$, it can be found that $n'_1$ is the closest nucleotide to $n'_2$. Therefore, $c_1$ should also be close to $c_2$ if $\{c_1,c_2\}$ and $\{n'_1, n'_2\}$ are similar. This means $c_1$ and $c_2$ are either neighbors along the chain or they have an interaction. If $c_1$ and $c_2$ are neighbors along the chain, their smallest distance is <2 Å. If they have an interaction, their distance should be <4 Å. Based

on this observation, when selecting a suitable $c_2$, we only need to check the neighbors of $c_1$ that are located within a distance of 4 Å (in practice, we relax this distance constraint to 10 Å to avoid missing some nucleotides not in the standard position). This operation greatly reduces the number of nucleotides for checking. It should be noted that our method relies on finding a chain of nearest neighbors. As long as the chain has distances below 10 Å in distinct parts of the query motif, our method will perform well.

Now suppose we have a partial candidate containing $k$ $(k < m)$ nucleotides: $\{c_1,c_2,\ldots,c_k\}$. We will now add a new residue $c_{k+1}$. Similar to the case of $\{c_1,c_2\}$, we first check ClosestPairs to find the pair $(n'_i, n'_{k+1})$. $(n'_i, n'_{k+1})$ means $n'_i$ is the closest nucleotide to $n'_{k+1}$ when considering the first $k$ nucleotides of $\mathrm{Seq_{new}}$. Then, we turn to the $i$-th nucleotide $c_i$ in the partial candidate and find all its neighbors within 10 Å in the search space. By adding each neighbor to $\{c_1,c_2,\ldots,c_k\}$, respectively, we can construct several new partial candidates in the form of $\{c_1,c_2,\ldots,c_{k+1}\}$. We then check the distance $d_{\mathrm{LS}}$ between $\{c_1,c_2,\ldots,c_{k+1}\}$ and $\{n'_1, n'_2, \ldots, n'_{k+1}\}$. If the $d_{\mathrm{LS}}$ value is smaller than the threshold $d_{k+1}$, $\{c_1,c_2,\ldots,c_{k+1}\}$ will be regarded as a new partial candidate. When there are $m$ nucleotides in each candidate, this process will stop. By changing $c_1$ to different nucleotides in the search space, a complete candidate set can be constructed.

In the previous process, there is a threshold $d_k$ $(k = 2, \ldots, m-1)$ used for filtering partial candidates with $k$ nucleotides. Next we shall explain the relationship between $d_k$ and $d_0$ as well as how to set their values. Suppose the query motif is $\{q_1,q_2,\ldots,q_m\}$ (which has been reordered using the algorithm in Supplemental Table S2). The least-squares distance $d_{\mathrm{LS}}$ of a candidate $\{c_1,c_2,\ldots,c_m\}$ to the query is supposed to be smaller than $d_0$. The corresponding rotation matrix and translation vector for $m$ nucleotides is denoted by $R_1$ and $t_1$, respectively. When only the first $k$ nucleotides of the query motif and the candidate are considered, the optimal rotation matrix and translation vector are denoted by $R_2$ and $t_2$. It is obvious that

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{l} &||q_{ij} - [R_2(c_{ij}) + t_2]||^2 \\
&\leq \sum_{i=1}^{k}\sum_{j=1}^{l} ||q_{ij} - [R_1(c_{ij}) + t_1]||^2 \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{l} ||q_{ij} - [R_1(c_{ij}) + t_1]||^2 \leq d_0
\end{aligned}
\tag{3}
$$

where $q_{ij}$ and $c_{ij}$ are the $j$-th points in $q_i$ and $c_i$, and $l$ is the number of points in $q_i$ $(c_i)$.

Equation 3 means that, if the least-squares distance of a candidate is smaller than $d_0$, all its $k$-nt substructures should also have a least-squares distance smaller than $d_0$ when compared with the corresponding $k$-nt subsequence of the query motif. If a $k$-nt RNA fragment has a least-squares distance larger than $d_0$, all $m$-nt candidates starting with it will not be regarded as positive instances in the filtering process. This means that $d_0$ is the upper bound of $d_k$.

In Figure 5, we have plotted the relationship between $k$ and $d_{\mathrm{LS}}(k) = d_{\mathrm{LS}}^{(lk)}$ (under the assumption that there are $l$ representative points in each nucleotide and $n$ is set to $lk$ in Supplemental Equation S1) based on positive instances of tetraloop, sarcin/ricin loop, $\pi$-turn, k-turn, and ribose zipper in two RNA subunits (PDB: 1S72 and 2QBE). We choose one of the positive instances in each type
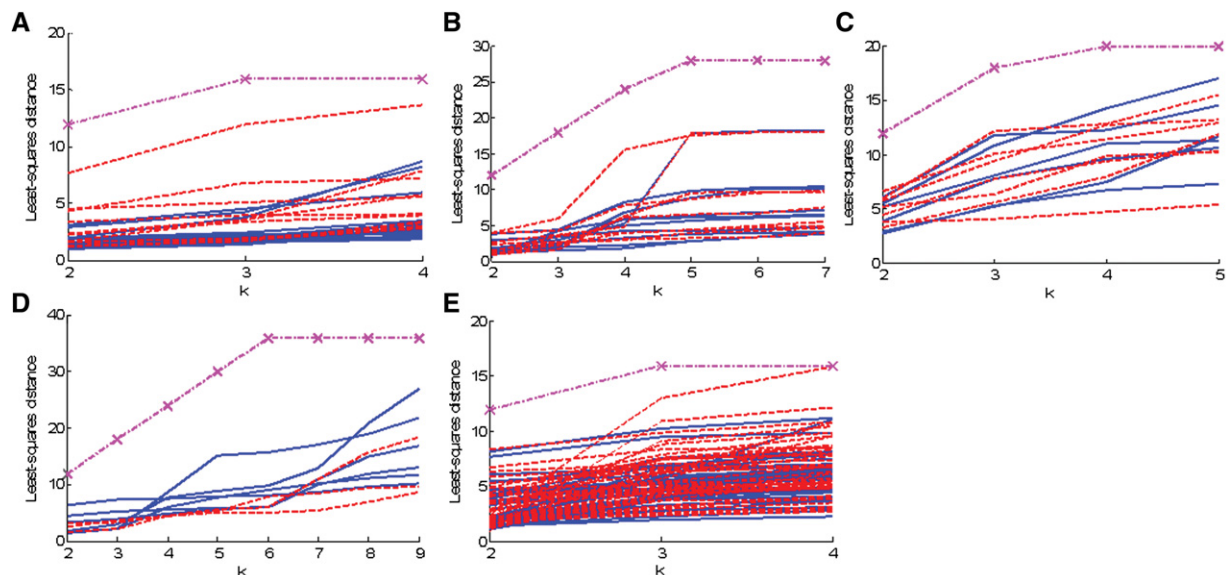
**FIGURE 5.** Relationship between least-squares distance $d_{LS}(k)$ and the number of nucleotides $k$ in the partial positive instances and the query. Dashed lines correspond to motifs from 1S72 and solid lines correspond to motifs from 2QBE. The purple lines with "×" markers show the filtering constraints of $\min(d_0, 6k)$. It is shown that distances of all positive instances to the query are smaller than the filtering constraints. (*A*) Tetraloop; (*B*) core of sarcin/ricin loop; (*C*) π-turn; (*D*) k-turn; (*E*) ribose zipper.

as the query and compute the distances of the other positive instances to the query. The dashed lines correspond to instances from 1S72 and solid lines for instances from 2QBE. It is obvious that $d_{LS}(k)$ always increases with $k$. The largest value of $d_0$ can be estimated as $4m$ ($m$ is the number of nucleotides in the query). For motifs with moderate length (ranging from ∼3 to 10 nt), the estimated value is large enough to include all the positive instances and guarantees the computational efficiency simultaneously. However, for some large motifs (>10 nt), the estimated value may be a bit high, which may lead to heavy computation. Instead, for these cases, we suggest using a value smaller than $4m$. The adjustment of $d_0$ will not affect the order of search results but the number of instances shown in the results. A smaller $d_0$ corresponds to a smaller number of search results and vice versa. Therefore, when the length of motif is larger than 10 nt, we suggest that users adopt a smaller value of $d_0$. If the returned instances are all positive, $d_0$ can be increased until enough negative instances appear at the end of the list, which means that most of the positives instances are found.

In Figure 5, sometimes the estimated value of $d_0$ is much larger than $d_{LS}(k)$, especially when $k$ is small. If we use $d_0$ to filter small partial candidates, most of them will be kept at the beginning of the candidate generation step, which consequently affects the speed of the method. In order to achieve a higher efficiency, we set the value of $d_k$ as the smaller value between $d_0$ and $6k$ so that fewer small partial candidates will be kept.

## LS-RSMR

Given a query motif and a cutoff threshold $d_0$, LS-RSMR first determines the new sequence order of the query using the algorithm in Supplemental Table S2. Then, the algorithm will generate candidates according to this new query motif sequence order. After that, each candidate will be superimposed on the query motif and a least-squares distance will be computed using Supplemental

Equation S1. Candidates whose least-squares distance values are less than $d_0$ will be retained as the positive instances.

## SUPPLEMENTAL MATERIAL

Supplemental material is available for this article, and codes of LS-RSMR are available at http://sse.tongji.edu.cn/yingshen/RSMR/LS_RSMR.html.

## ACKNOWLEDGMENTS

## REFERENCES

Apostolico A, Ciriello G, Guerra C, Heitsch CE, Hsiao C, Williams LD. 2009. Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res* **37:** e29.

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principle of RNA packing. *Science* **273:** 1678–1685.

Chang K, Tinoco I. 1994. Characterization of a "kissing" hairpin complex derived from the human immunodeficiency virus genome. *Proc Natl Acad Sci* **91:** 8705–8709.

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, et al. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18:** 1433–1445.

Duarte CM, Wadley LM, Pyle AM. 2003. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res* **31:** 4755–4761.

François B, Russell RJ, Murray JB, Aboul-ela F, Masquida B, Vicens Q, Westhof E. 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.

Gagnon MG, Steinberg SV. 2010. The adenosine wedge: A new structural motif in ribosomal RNA. *RNA* **16:** 375–381.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 259–300.

Harrison AM, South DR, Willett P, Artymiuk PJ. 2003. Representation, searching discovery of patterns of bases in complex RNA structures. *J Comput Aided Mol Des* **17:** 537–549.

Hendrix DK, Brenner SE, Holbrook SR. 2005. RNA structural motifs: Building blocks of a modular biomolecule. *Q Rev Biophys* **38:** 221–243.

Jaeger L, Verzemnieks EJ, Geary C. 2008. The UA_handle: A versatile submotif in stable RNA architectures. *Nucleic Acids Res* **37:** 215–230.

Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: A new RNA secondary structure motif. *EMBO J* **20:** 4214–4221.

Lai C, Tsai MY, Liu YC, Wang CW, Chen KT, Lu CL. 2009. FASTR3D: A fast and accurate search tool for similar RNA 3D structures. *Nucleic Acids Res* **37:** W287–W295.

Leontis NB, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol* **16:** 300–308.

Leontis NB, Stombaugh J, Westhof E. 2002. The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **30:** 3497–3531.

Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16:** 279–287.

Manning CD, Raghavan P, Schütze H. 2008. *Introduction to information retrieval.* Cambridge University Press, New York.

Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF. 2009. RNA regulation of epigenetic processes. *BioEssays* **31:** 51–59.

McLachlan AD. 1979. Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol* **128:** 49–79.

Moore PB. 1999. Structural motifs in RNA. *Annu Rev Biochem* **68:** 287–300.

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. 2001. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci* **98:** 4899–4903.

Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W, Mattick JS. 2009. Genome-wide identification of long noncoding RNAs in CD8[+] T cells. *J Immunol* **182:** 7338–7348.

Petrov AI, Zirbel CL, Leontis NB. 2011. WebFR3D—a server for finding, aligning and analyzing recurrent RNA 3D motifs. *Nucleic Acids Res* **39:** W50–W55.

Reijmers TH, Wehrens R, Buydens LM. 2001. The influence of different structure representations on the clustering of an RNA nucleotides data set. *J Chem Inf Comput Sci* **41:** 1388–1394.

Sargsyan K, Lim C. 2010. Arrangement of 3D structural motifs in ribosomal RNA. *Nucleic Acids Res* **38:** 3512–3522.

Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. 2008. FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol* **56:** 215–252.

Shen Y, Wong HS, Zhang S, Yu Z. 2011. Feature-based 3D motif filtering in ribosomal RNA. *Bioinformatics* **27:** 2828–2835.

Steinberg SV, Boutorine YI. 2007. G-ribo: A new structural motif in ribosomal RNA. *RNA* **13:** 549–554.

Szewczak AA, Moore PB, Chang YL, Wool IG. 1993. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc Natl Acad Sci* **90:** 9581–9585.

Wadley LM, Pyle AM. 2004. The identification of novel RNA structural motifs using COMPADRES: An automated approach to structural discovery. *Nucleic Acids Res* **32:** 6650–6659.

Wadley LM, Keating KS, Duarte CM, Pyle AM. 2007. Evaluating and learning from RNA pseudotorsional space: Quantitative validation of a reduced representation for RNA structure. *J Mol Biol* **372:** 942–957.

Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci* **87:** 8467–8471.

Zhong C, Tang H, Zhang S. 2010. RNAMotifScan: Automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res* **38:** e176.

Zhu M. 2004. Recall, precision and average precision. Technical report, University of Waterloo, Waterloo, Ontario.