# Mining RNA Tertiary Motifs with Structure Graphs

Xueyi Wang[*], Jun Huan[†], Jack S. Snoeyink[*] and Wei Wang[*]
[*]*Department of Computer Science, University of North Carolina at Chapel Hill*
*Chapel Hill, NC, 27599-3175, USA*
[†]*Department of Electrical Engineering and Computer Science, University of Kansas*
*Lawrence, KS, 66047-7621, USA*
*Email: [*]{xwang, snoeyink, weiwang}@cs.unc.edu and [†]jhuan@eecs.ku.edu*

## Abstract

*We present a novel application of graph database mining to identify tertiary motifs in RNA structures. In our method, we abstract an RNA molecule as a labeled graph and use a frequent subgraph mining technique to derive tertiary motifs. By applying our technique to ribosome RNA and transfer RNA, we have identified known RNA tertiary motifs such as the ribose zipper and U-turn, plus candidates for novel tertiary motifs. Finally, we suggest an iterative multiple structure alignment algorithm to classify tertiary motifs and generate consensus motifs.*

## 1. Introduction

We present a novel application of graph database mining in the bioinformatics domain: that of identifying tertiary motifs from RNA molecular structures. Our goal is automated motif discovery by (1) modeling RNA structures as graphs and (2) mining a graph database to identify structurally important features (tertiary motifs) from RNA.

RNA plays critical roles in biological systems. RNA molecules play critical roles in biological systems. Recent research shows that RNA can restore and transmit genetic information [1], catalyze chemical reactions [2], and regulate gene expressions [3]. For example, RNA interference (RNAi) is a widely used experimental technique that utilizes short RNA sequences to regulate gene expression in eukaryotic cells, and RNAi-based drugs have become an important target.

An RNA molecule is a linear polymer of nucleotides connected by covalent bonds. Each unit has a backbone (phosphate + sugar) with one of the four nucleotide bases (A, C, G, and U) attached. Like protein, RNA has four levels of structural organization: primary, secondary, tertiary, and quaternary. *Primary structure* is the linear sequence of nucleotides, *Secondary structure* is the collection of pairs of bases in 3D structure, *tertiary structure* is the overall shape of an RNA molecule, and *quaternary structure* is the organization of two or more RNA molecules.

An *RNA motif* is a short fragment of RNA (continuous or noncontinuous) that appears repeatedly in a variety of RNA molecules and plays an important role in biological function [4]. Our algorithms focus on the first half of this definition, "appears repeatedly," because that is a property that can be determined purely from a set of RNA structures, and because natural selection in molecular evolution suggests that motifs with an important role are biased to appear.

Identifying RNA motifs is a step in understanding RNA structures and their function. There are three types of RNA motif: *Sequence motif* is a fragment of RNA sequence. *Secondary motif* reflects RNA base pairing relations, which form the scaffold of RNA structures and serve important biological roles like regulating cellular processes. *Tertiary motif* [5,6,7] reflects spatial interactions between nucleotides and is related to biological function such as stabilizing structure or metal binding. Although tertiary motifs are important for RNA folding and function, current RNA motif identification algorithms focus on finding sequence motifs and secondary motifs, but not tertiary motifs.

We investigate algorithms that represent the 3D structure of RNA molecules as a database of graphs, discover subgraphs (tertiary motifs) with frequent subgraph mining techniques, and build consensus motifs (representatives of subgraphs in same groups) by geometric algorithms.

Our graphs include three types of edges: *backbone edges* that encode connectivity along the primary sequence of an RNA molecule, *base pair edges* that

19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)
0-7695-2868-6/07 $25.00 © 2007 IEEE

encode base pair interaction of nucleotides, and *contact edges* that encode non-local contacts from the tertiary structure of the molecule. Thus we capture aspects of RNA primary, secondary and tertiary structures in the graph.

We employ a frequent subgraph mining algorithm [8] to identify the frequently occurring subgraphs (tertiary motifs) in a collection of RNA structure graphs. For each group of subgraphs, we derive a consensus motif by applying a geometric structure alignment algorithm that classifies mirror symmetric subgraphs as right or left handed and performs multiple structure alignment by iteratively finding local optimal solution and converging towards a global minimal. With our alignment algorithm, we show that the aligned tertiary motifs fit well with a 3D Gaussian distribution model.

We demonstrate the overall utility of our algorithm on transfer RNA (tRNA) and ribosome RNA (rRNA). tRNA and rRNA are selected because of their abundance in known RNA structures and the extensive manual study in the SCOR database [9]. SCOR is a comprehensive database for recording RNA secondary and tertiary motifs that classifies RNA information into structural classification, functional classification, and tertiary interaction. By comparing our mined RNA tertiary motifs to the collections of motifs in tertiary interactions in SCOR, we show that our method can find known tertiary motifs, plus novel ones.

## 2. Related Work

Several RNA motif identification algorithms have been developed, with various assumptions. Below, we review some major algorithms, classified into four groups.

The first group of motif identification algorithms involves manual processing to identify tertiary motifs. Klosterman *et al*. [10] described examples of newly found RNA tertiary motifs, including extruded helical single strand, internal loop triples, and U-turns in internal loops. All these tertiary motifs are observed manually, not discovered automatically by tools.

The second group of motif identification algorithms finds sequence motifs only. For example, Morgante *et al*. [11] use a graph representation of sequence and find common non-consecutive motifs for two or more sequences. Rajasekaran *et al*. [12] find common sequence of length $l$ with Hamming distance of $d$ in $t$ sequences of length $n$. Zhao *et al*. [13] find the similar DNA motifs based on a permuted Markov model.

The third group of motif identification algorithms uses simplified representations of RNA structures to find common structural motifs. COMPADRES [14] uses P and C4' atoms to represent a nucleotide, reduces RNA 3D structure to a sequence of dihedrals by continuous P and C4' atoms, and clusters the dihedrals. Huang *et al*. [15] cut an RNA sequence into 6-nt fragments, compare their RMSD values, and cluster into a hierarchy structure by the unweighted pair group method with arithmetic mean (UPGMA). The structural motifs discovered by these two methods are fixed to short consecutive sequences since they use no knowledge of secondary and tertiary interactions.

The fourth group of motif identification algorithms uses structure alignment to derive tertiary motifs. ARTS [16], which stands for alignment of RNA tertiary structures, compares two RNA sub-structures with sizes from two to thousands of nucleotides. It uses a set of base pairs as seed, compares their minimum RMSD every two consecutive base pairs, extends to the whole structures, and scores the matching. RAG [17] represents RNA secondary structure as tree and dual-graph motifs, enumerates all possible motifs, and clusters based on topological characteristics. These methods have difficulty finding tertiary motifs because they do not consider tertiary interactions.

In our previous study, we also applied graph modeling and graph mining for analyzing 3D protein structures [18]. Adapting the same technique to RNA analysis is non-trivial because of the following reasons: (1) Modeling RNA structure is different from that of protein structure: RNA structures are much larger and less stable than protein structures. (2) RNA is composed of 4 residues rather than 20 in proteins, which means that we have smaller set of node labels in RNA graph mining. Our current method is fully automated, fast, and works directly on 3D RNA structures.

## 3. Algorithms

First, we define labeled graphs, which serve as the formal base of our graph representation of RNA molecules, and the data structure used by the frequent subgraph mining algorithm. Second, we discuss constructing graph representations for RNA molecules. Finally, we introduce the novel structure alignment algorithm for building consensus motifs.

### 3.1. Labeled graphs and frequent subgraph mining algorithms

A labeled graph $G$ is a quadruple $G = (V, E, \Sigma, \lambda)$. $V$ is a set of nodes, $E \subseteq V \times V$ is a set of undirected edges joining distinct nodes, $\Sigma$ is a set of node labels and

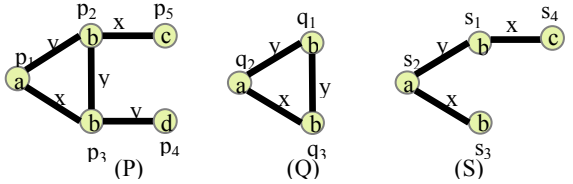



**Figure 1. A database *GD* of three labeled graphs**

edge labels, and the labeling function $\lambda$ defining the mappings from nodes and edges to their labels: $V \cup E \rightarrow \Sigma$. The *size* of a graph $G$ is the cardinality of its node set $V$. A *graph database GD* is simply a group of labeled graphs. Figure 1 shows a graph database with three labeled graphs. The labels of nodes and edges are specified within the nodes and along the edges for each graph.

From graph theory, we formalize the search for tertiary motifs as the search for commonly occurring subgraphs in a group of graphs. A fundamental part of our frequent subgraph mining algorithm is to decide whether a subgraph $G$ occurs in another graph $G_o$. To make this more precise, we define that a graph $G = (V, E, \Sigma, \lambda)$ is *subgraph isomorphic* to $G_o = (V_o, E_o, \Sigma_o, \lambda_o)$ if there exists a one-one mapping $f: V \rightarrow V'$ such that:

$\forall\ u \in V, \lambda(u) = \lambda_o(f(u))$,

$\forall\ u, v \in V, (u, v) \in E \Rightarrow (f(u), f(v)) \in E_o$,

$\forall\ (u, v) \in E,\ \lambda(u, v) = \lambda_o(f(u), f(v))$.

The one-one mapping $f$ is defined as a subgraph isomorphism from $G$ to $G_o$. Figure 1 shows a subgraph isomorphism $f$: $q_1 \rightarrow p_2$, $q_2 \rightarrow p_1$, and $q_3 \rightarrow p_3$ from graph $Q$ to $P$, hence we say that graph $Q$ occurs in $P$ through the subgraph isomorphism $f$. Luke *et al.* [18] show an example of using labeled graphs in protein structures.

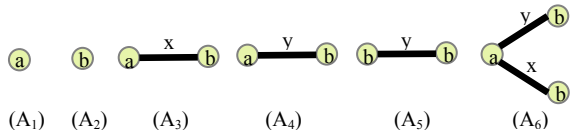



**Figure 2. All frequent connected subgraphs from *G* in Figure 1 with support threshold $\sigma$ = 100%**

Given a graph database $GD$, which contains a set of graphs, the *support* of a subgraph $G$ is the fraction of graphs in $GD$ in which $G$ occurs. Given a threshold $0 \leq \sigma \leq 1$, we define $G$ to be *frequent* if its support is at least $\sigma$. The goal of frequent subgraph mining is to identify all frequent subgraphs from a graph database $GD$ with support threshold $\sigma$. Figure 2 shows all six frequent connected subgraphs with $\sigma$=1 from the three graphs of Figure 1.

We use the Fast Frequent Subgraph Mining algorithm (FFSM) (available at http://www.cs.unc.edu/~huan/FFSM.html), which is competitive or outperforms other state-of-art subgraph mining algorithms [8].

### 3.2. Graph Modeling of RNA Molecules

In our graph representation, each node represents one nucleotide and each edge represents the connection for two nucleotides. We generate RNA graphs from RNA structures in the following way:

Totally there are four different nucleotides in RNA molecule with the same backbone but different bases – A, C, G, and U. In graph representation, each node corresponds to a nucleotide and is labeled either with purine (A and G) or pyrimidine (C and U). We reduce the alphabet to two because these nucleotides do not have significant structural differences, and it is common that mutated and wild-type RNAs have the same motif with different nucleotides [4]. We have tried the alphabet of all four symbols, but we find very few tertiary motifs.

We generate three types of edges to represent RNA primary, secondary, and tertiary structures in the following priority order:

a *backbone edge* connects two contiguous nucleotides,

a *base pair edge* connects two nucleotides recorded as a base pair in the NDB [19],

a *contact edge* connects spatial neighboring nucleotides within 8Å.

Backbone and base pair edges are labeled by their types. For each nucleotide pair, contact edges are labeled by discretized distances in the following way: Each nucleotide is abstracted as two points, its phosphorus atom and the geometric center of its sugar ring (since most tertiary interactions involve the phosphate and sugar groups. We define the distance between two nucleotides as the shortest distance between their abstracted points, and discretize this into distance bins, as described in section 4.2.

We create one graph for each RNA structure, collect all the graphs into a graph database, and use the FFSM algorithm [8] to mine frequent subgraphs.

### 3.3. Constructing consensus motifs with Computational Geometry

The graph representation in Section 3.2 abstracts away some of the precise geometry of motifs. After obtaining frequent subgraphs, we construct the corresponding tertiary motifs by the atom coordinates in 3D structure, and develop a novel multiple structure alignment algorithm that classifies mirror symmetric motifs as right or left handed and finds the optimal alignment by minimizing the sum of root mean squared distance (RMSD), which is widely used in measuring structure similarities in bioinformatics.

Given $n$ motifs, $G_1$, $G_2$, …, $G_n$, each with $m$ points in correspondence, e.g. $p_{i1}$, $p_{i2}$, …, $p_{im}$ for motif $G_i$, we define the average motif $\overline{G}$ with points $\overline{p}_k = \frac{1}{n}\sum_{i=1}^{n} p_{ik}$ for $1 \le k \le m$, and define RMSD as the square root of the average of all squared pairwise distances between motifs, $\sqrt{\frac{2}{mn(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\left\|p_{ik}-p_{jk}\right\|^2}$, where $n(n–1)/2$ is the total number of motif pairs and $m$ is the number of points in each motif. Since $n$ and $m$ are fixed, we can look for rigid transformations that minimize the summation. Wang and Snoeyink [20] observe that the sum of all squared pairwise distances between $n$ motifs equals $n$ times the sum of squared distances to the average motif $\overline{G}$, i.e. $\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\left\|p_{ik}-p_{jk}\right\|^2 = n\sum_{i=1}^{n}\sum_{k=1}^{m}\left\|p_{ik}-\overline{p}\right\|^2$.

To minimize RMSD, we translate and rotate/reflect motifs in 3D space to minimize the target function $\arg\min_{R,T}\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\left\|R_i p_{ik} - T_i - R_j p_{jk} + T_j\right\|^2\right)$, where $R_i$ is a 3×3 rotation/reflection matrix and $T_i$ as a 3×1 translation vector for motif $G_i$. Matrix $R_i$ can be either a rotation (determinant = 1) or reflection (determinant = –1). After minimization, we classify all motifs into two handedness groups depending on whether reflection matrix gives better RMSD. The following algorithm iteratively aligns all motifs $G_i$ for $(1 \le i \le n)$ to $\overline{G}$, classifies mirror symmetric motifs, and updates the coordinates of $\overline{G}$ to minimize RMSD.

*Algorithm:* to classify and align motifs, perform the following steps:

1. Move the centroids of all $G_i$ for $(1 \le i \le n)$ to the origin.
2. Calculate the average motif $\overline{G}$ and $SD = \sum_{i=1}^{n}\sum_{k=1}^{m}\left\|p_{ik}-\overline{p}_k\right\|^2$.
3. Align $G_i$ for $(1 \le i \le n)$ to $\overline{G}$ by the optimal rotation or reflection matrix $R_i$, calculated by using the singular value decomposition (SVD) to determine the maximum eigenvalue of the covariance matrix $N$.
4. Calculate $SD^{new} = \sum_{i=1}^{n}\sum_{k=1}^{m}\left\|R_i p_{ik}-\overline{p}_k\right\|^2$.
5. If $SD - SD^{\text{new}} > \varepsilon$ ($1.0\times10^{-5}$ in tests), update $p_{ik} = R_i p_{ik}$ and $SD = SD^{\text{new}}$, calculate the average motif $\overline{G}$, and go to step 3; otherwise, go to step 6.
6. Set $R_i$ = the product of all the rotation or reflection matrices for G, and classify $G_i$ as right or left handed by the determinants of $R_i$ (either 1 or –1).

This algorithm extends the algorithm presented by Wang and Snoeyink [20], which finds optimal alignment in nearly linear time but does not classify the motifs into right and left handed.

In each iteration, steps 1-5 need O($nm$) each and step 6 needs O($n$). The proof of convergence in Wang and Snoeyink [20] also applies to this algorithm. In our experiments reported below, the number of iterations is small and the values reached are stable.

## 4. Experiments

### 4.1. Data sets

A list of selected tRNAs and rRNAs used in this paper is shown in Table 1. In total we have 20 tRNAs, 3 5s rRNAs, 2 16s rRNAs, and 4 23s rRNAs. There are many examples of same RNA from same species binding to different proteins in NDB [19]. We manually cleansed the data set with the following criteria to remove redundant ones:

A. From NDB with cutoff date December 22nd, 2005, we choose RNA with more than 90% nucleotides present.
B. For duplicated structures (from same species with same function), we keep the most recent one. If the time is the same, we keep the one with highest resolution.
C. For two structures with more than 70% of sequence similarity, we keep the more recent one. If the time is the same, we keep the one with higher resolution.

We keep wild-type RNA and remove mutated RNA and synthesized RNA.

The tRNAs and rRNAs in Table 1 are the only available RNA molecules in NDB. Each RNA

molecule is represented by a four-letter string, known as the protein databank identifiers (PDB ID). The fact that we have relatively few structures of rRNAs, some of which are large (especially the 23s rRNA), is a potential problem. FFSM determines frequent subgraphs by the number of graphs (structures) that have a subgraph, rather than the number of times a subgraph is found. This makes sense for identifying common structures in large families of related molecules, but we plan in future work to try to modify FFSM to count frequency by number of subgraphs for RNA.

**Table 1. List of selected tRNAs and rRNAs**

| Type | | Pdb Name |
|---|---|---|
| tRNA | | 1ehz, 1yfg, 1fir, 1qf6, 1qu2, 1eiy, 1f7u, 1il2, 1h4s, 2fmt, 1ivs, 1n78, 1j1u, 1j2b, 1u0b, 1wz2, 1zjw, 1h3e, 2csx, 1ser |
| rRNA | 5s: | 1nkw (chain 9), 1s72 (chain 9), 1yl3 (chain B) |
| | 16s: | 1fjg, 1pns |
| | 23s: | 1nkw (chain 0), 1pnu (chain 0), 1s72 (chain 0), 1yl3 (chain A) |

## 4.2. Identifying tertiary motifs

We identify motifs for tRNAs and rRNAs (5s, 16s and 23s) in two separate groups. For each group, we generate three different graphs using different bin sizes for contact edges (3, 4, or 5Å), with cutoff distance 8Å. This cutoff distance is large enough to capture the edges of most known tertiary motifs; we have tried larger cutoff distances but found too many contact edges, causing "noisy" occurrences of motifs. Lists of all the mined motifs can be found at http://www.cs.unc.edu/~xwang/RNAGraph/.

Most of the mined motifs contain 4 nucleotides. RNA molecules are quite flexible and large frequent motifs are less likely to be found in the same topology. In trying larger cutoff distance (e.g. 18Å) for rRNAs, we find the largest mined motifs contain 8 nucleotides.

We compare our results to SCOR [9], which is a comprehensive database of RNA motifs identified by manual. As mentioned in section 3.2, our focus is to identify motifs that are involved in backbone interactions within a single chain, which fall into the tertiary motifs category in SCOR. Because we use the phosphorus, which is between two nucleotides, as one of our two points representing a nucleotide, we allow a shift of one nucleotide when comparing mined motifs to those of SCOR.

Note that all the motifs discussed in this paper involve backbone interactions only. We do not consider the backbone-base interactions. The contact distances are longer in the backbone-base interactions than the backbone-backbone interactions, and the number of contact edges and the noise in the data (motifs without biological meaning) significantly increase.

For rRNA, we choose a support threshold $\sigma$ of 70% – that is, motifs must occur in 7 of the 9 graphs in the family to be considered frequent. The threshold is high because the 16s and 23s rRNAs are large and have many motifs. For example, for bin sizes of 3, 4, and 5Å we find 75, 260 and 152 distinct subgraphs in the 23s rRNA 1s72, respectively.

The ribose zipper is a tertiary motif formed by hydrogen bonds among the 2'–OH groups of sugars at two anti-parallel backbone strands. We identify 37 of the 43 ribose zippers recorded in SCOR (86%) for 23s rRNA 1s72. The number of found ribose zippers using different bin sizes is shown in Table 2. Note that all 37 identified ribose zippers are found with bin size 4Å, which occupies 14% of 260 total distinct subgraphs.

**Table 2. Ribose zippers found in 23s rRNA 1s72**

| Bin size | 3Å | 4Å | 5Å |
|---|---|---|---|
| Number of identified ribose zippers | 12 | 37 | 8 |
| Total found distinct subgraphs | 75 | 260 | 152 |

There are five subcategories of ribose zippers (canonical, single, reverse single, naked and Cis) in 1s72 and we identify instances of each of them. Figure 3 shows a canonical ribose zipper (nucleotides 1078-1079 and 2077-2078, 23s rRNA 1s72).

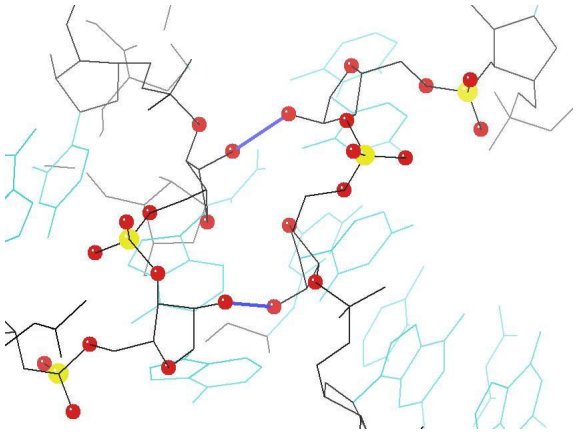


**Figure 3. Canonical ribose zipper (nucleotides 1078-1079 and 2077-2078, 23s rRNA 1s72). Yellow ball is phosphorus, red ball is oxygen, and blue line is hydrogen bond.**

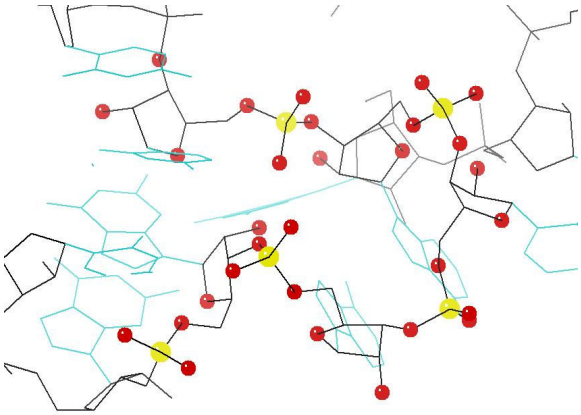


**Figure 4. U turn motifs form by 5 continuous nucleotides (nucleotides 394-398, 23s rRNA 1s72), found by bin size = 4Å. Yellow ball is phosphorus and red ball is oxygen**

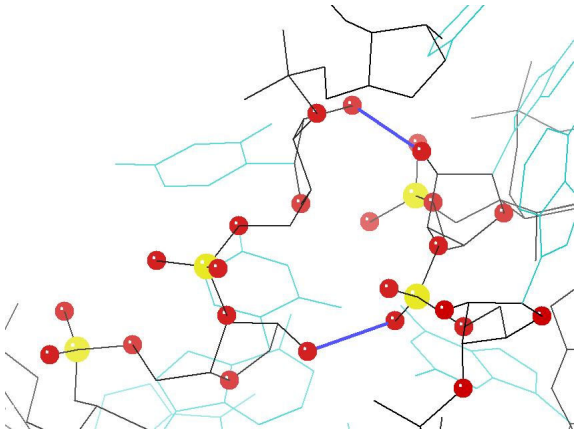


**Figure 5. Tertiary interaction formed by a hydrogen bond (blue line) between two sugars and a hydrogen bond (blue line) between sugar and phosphorus (nucleotides 66-67 and 107-108, 23s rRNA 1s72), found by bin size = 3Å.**

We have also identified motifs classified as secondary motifs in SCOR. For example, Figure 4 shows a U-turn motif formed by five contiguous nucleotides (394-398); our method identifies four of them (nucleotides 394-395 and 397-398, 23s rRNA 1s72).

By carefully checking the mined motifs that do not match any existing motifs in SCOR, we find some interesting structures that could be good candidates for tertiary motifs. For example, Figure 5 shows a tertiary motif with one hydrogen bond between two sugars and another hydrogen bond between sugar and phosphorus (nucleotides 66-67 and 107-108, 23s rRNA 1s72).

For tRNA, we choose a support threshold $\sigma$ of 20%, that is, motifs must occur in 4 of the 20 graphs in the family to be considered frequent. The threshold is much lower because tRNA is quite flexible and is much smaller than the large rRNA. We find several good candidates for tertiary motifs, which are available at http://www.cs.unc.edu/~xwang/RNAGraph/.

For the 20 tRNAs we choose, SCOR records only 5 tertiary motifs in 3 tRNA: 1ehz, 1yfg and 1fir. All the tertiary motifs are large (the smallest having 7 nucleotides), and no two tertiary motifs share the same topology. So for tRNA, we cannot compare our mined tRNA motifs with SCOR, because the support threshold of tertiary motifs of tRNAs in SCOR is too low ($\sigma < 5\%$).

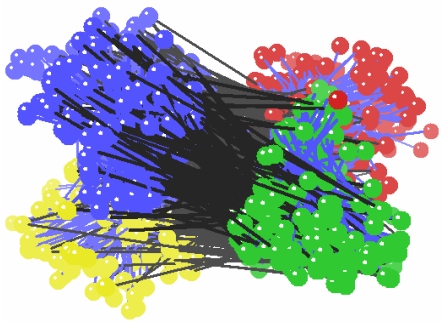
### 4.3. Consensus motifs

We apply the multiple structure alignment algorithm to classify the structures of found tertiary motifs and generate consensus motifs. The alignment is done on a laptop with Pentium M 1.8GHz CPU and 784M memory. Table 3 shows the performance of aligning 12 motif groups by bin size = 4Å. The running time is collected by 1,000 tests on each motif group. We can see that when we classify mirror symmetric motifs, the RMSD is significantly decreased, along with the number of iterations and running time.

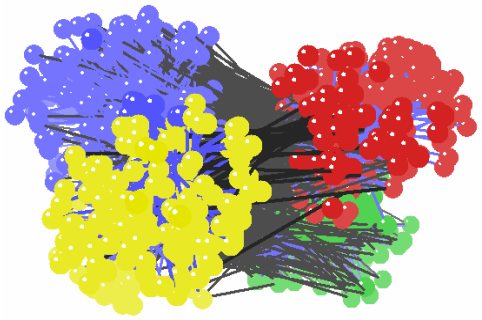
All the motif groups contain mirror symmetric motifs and we achieve better alignment when we use our algorithm to classify and separate motifs by handedness, as shown in Figure 6. As we have verified, the handedness has no relation with the functions of motifs and the type of motifs. For example, all five types of ribose zippers can occur in both right and left handedness. But it is an interesting problem whether all the tertiary motifs are independent of handedness.

**Table 3. Performance for 12 mined motifs by bin size = 4Å**

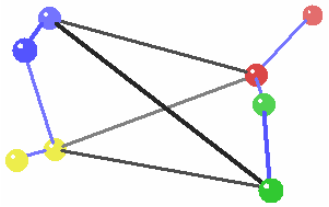
| Motif ID | # of subgraphs all, 1s72, zipper | Motif RMSD | Motifs with reflection RMSD, iterations, time(s) | Right handed # | Right handed RMSD | Left handed # | Left handed RMSD |
|---|---|---|---|---|---|---|---|
| 1 | 160, 19, 0 | 4.11 | 3.52, 6, 0.095 ± 0.005 | 81 | 3.44 | 79 | 3.47 |
| 2 | 202, 45, 7 | 3.93 | 3.53, 8, 0.158 ± 0.004 | 106 | 3.38 | 96 | 3.44 |
| 3 | 38, 8, 0 | 4.40 | 3.64, 4, 0.016 ± 0.005 | 20 | 3.56 | 18 | 3.72 |
| 4 | 10, 1, 0 | 3.54 | 3.35, 5, 0.005 ± 0.005 | 6 | 2.95 | 4 | 2.78 |
| 5 | 79, 21, 0 | 4.50 | 3.91, 5, 0.039 ± 0.003 | 41 | 3.74 | 38 | 3.93 |
| 6 | 53, 10, 0 | 3.81 | 3.60, 8, 0.041 ± 0.003 | 28 | 3.49 | 25 | 3.59 |
| 7 | 27, 7, 0 | 3.73 | 3.36, 7, 0.018 ± 0.004 | 15 | 3.29 | 12 | 3.04 |
| 8 | 396, 116, 5 | 4.32 | 3.80, 7, 0.288 ± 0.013 | 219 | 3.76 | 177 | 3.79 |
| 9 | 28, 7, 0 | 4.40 | 3.94, 4, 0.011 ± 0.004 | 15 | 3.83 | 13 | 3.92 |
| 10 | 16, 5, 0 | 3.94 | 3.85, 9, 0.014 ± 0.005 | 10 | 3.88 | 6 | 3.50 |
| 11 | 353, 76, 11 | 3.89 | 3.76, 16, 0.950 ± 0.576 | 192 | 3.54 | 161 | 3.72 |
| 12 | 361, 86, 24 | 4.05 | 3.56, 8, 0.382 ± 0.230 | 218 | 3.51 | 143 | 3.54 |



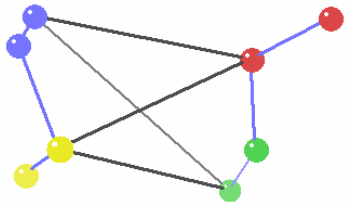
**6a. Aligning right hand occurrences of motif #12**



**6b. Aligning left hand occurrences of motif #12**



**6c. Consensus motif for right handed occurrences**



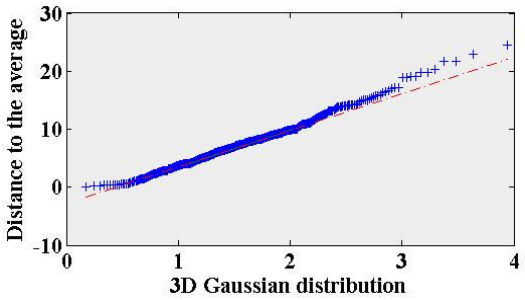
**6d. Consensus motif for left handed occurrences**

**Figure 6. Example of aligning instances of motif #12. Two points in each of the four nucleotides are colored as yellow, red, green and blue. Blue line is backbone edges and black line is contact edges.**

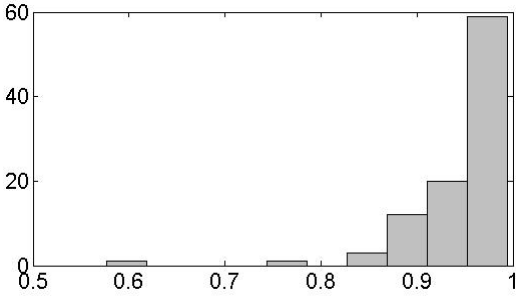
### 4.4. Statistical analysis of consensus motifs

Deriving the statistical description of the aligned motifs is an intriguing question that has significant theoretical and practical implications. We test the null hypothesis that the distances of $n$ atoms at a fixed position $k$ to the average $\overline{p}_k$ are consistent with the distances from a 3D Gaussian distribution. The Gaussian is most used distribution function due to the central limit theorem of statistics, and previous studies hint that Gaussian is the best model to describe the aligned structures [21]. We adopt the Quantile-Quantile Plot (QQ plot) procedure [22] to test the fitness of our data to the 3D Gaussian model. Figure 7a shows QQ plot for phosphorus of first node in motif #12. The y-axis is the distance from each motif to the average for a fixed position and the x-axis is the quantile data from 3D Gaussian. The correlation coefficient $R^2$ = 0.993, which suggests that the data fits a 3D Gaussian model reasonably well. We carried out the same experiments for all the positions and the collected histogram of the correlation coefficient $R^2$ is shown in figure 7b. We

identify that more than 88% of the positions we check have $R^2 > 0.9$.




**7a. QQ plot for phosphorus of first node in motif #12**




**7b. Histogram of $R^2$ for all aligned positions**

**Figure 7. 3D Gaussian distribution analysis of the distances from each point to average motif**

## 5. Conclusion and future work

We present an automated method of mining graph database to identify tertiary motifs in RNA structures. In our method, we defined a graph representation of RNA molecules and applied frequent subgraph mining algorithm for mining tertiary motifs. In post-processing of the tertiary motifs, we develop a multiple structure alignment algorithm for classifying mirror symmetric motifs and finding consensus motifs, and show that the aligned motifs follow 3D Gaussian distribution model. Our results show that the automated method can discover tertiary motifs in RNA molecules, despite limitations on the number of available RNA structures, and the fact that we included RNA only, but not the proteins that rRNA, in particular, interacts with.

Our plans for future work include extending FFSM to count frequency by subgraphs, considering RNA + protein, and finding fingerprint (i.e. distinct motif) candidates for RNA families. We also plan to investigate evolutionary relations among the tRNAs. Statistical analysis of the aligned RNA subgraphs is intriguing and we plan to investigate how Gaussian distribution model may help cluster RNA tertiary motifs.

## Acknowledgements

This work is partially supported by NIH grant GM-074127.

## References

[1] S.J. Lolle, J.L. Victor, J.M. Young, and R.E. Pruitt, "Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis", *Nature*, 434(7032): 505-509, 2005.

[2] D.M. Lilley, "Structure, folding and mechanisms of ribozymes", *Curr Opin Struct Biol*. 15(3): 313-323, 2005.

[3] Y. Tomari and P.D. Zamore, "Perspective: machines for RNAi", *Genes Dev*. 19(5): 517-529, 2005.

[4] N.B. Leontis and E. Westhof, "Analysis of RNA motifs", *Curr Opin Struct Biol*. 13(3):300-308, 2003.

[5] M. Tamura and S.R. Holbrook, "Sequence and structural conservation in RNA ribose zippers", *J Mol Biol*. 320(3):455-474, 2002.

[6] R.T. Batey, R.P. Rambo, and J.A. Doudna, "Tertiary motifs in RNA structure and folding", *Angew Chem Int Ed*. 38(16):2326-2343, 1999.

[7] T. Hermann and D.J. Patel, "Stitching together RNA tertiary architectures", *J Mol Biol*. 294(4):829-849, 1999.

[8] J. Huan, W. Wang, and J. Prins, "Efficient Mining of Frequent Subgraph in the Presence of Isomorphism", in *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*, pp.549-552, 2003.

[9] M. Tamura, D.K. Hendrix, P.S. Klosterman, N.R.B. Schimmelman, S.E. Brenner, and S.R. Holbrook, "SCOR: Structural Classification of RNA, version 2.0", *Nucleic Acid Res*. 32:D182-184, 2004.

[10] P.S. Klosterman, D.K. Hendrix, M. Tamura, S.R. Holbrook, and S.E. Brenner, "Three-dimensional motifs from the SCOR: structural classification of RNA database: extruded strands, base triples, tetraloops and U-turn", *Nucleic Acids Res*. 32(8):2342-2352, 2004.

[11] M. Morgante, A. Policriti, N. Vitacolonna, and A. Zuccolo, "Structured motifs search", *J Comput Biol*. 12(8):1065-1082, 2005.

[12] S. Rajasekaran, S. Balla, and C.H. Huang, "Exact algorithms for planted motif problems", *J Comput Biol*. 12(8):1117-1128, 2005.

[13] X. Zhao, H. Huang, and T.P. Speed, "Finding short DNA motifs using permuted Markov models", *Journal of Computational Biology*, 12(6):894-906, 2005.

[14] L.M. Wadley and A.M. Pyle, "The identification of novel RNA structure motifs using COMPADRES: an automated approach to structural discovery", *Nucleic Acids Res*. 32(22):6650-6659, 2004.

[15] H.C. Huang, U. Nagaswamy, and G.E. Fox, "The application of cluster analysis in the intercomparison of loop structures in RNA", *RNA*. 11(4):412-423, 2005.

[16] O. Dror, R. Nussinov, and H. Wolfson, "ARTS: alignment of RNA tertiary structures", *Bioinformatics*, 21 Suppl 2:ii47-ii53, 2005.
[17] H.H. Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick, "RAG: RNA-As-Graphs database--concepts, analysis, and features", *Bioinformatics*, 20(8):1285-1291, 2004.
[18] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha, "Mining Family Specific Residue Packing Patterns from Protein Structure Graphs", *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp.308-315, 2004.
[19] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, and B. Schneider, "The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids", *Biophys. J.*, 63(3):751-759, 1992.
[20] X. Wang and J.S. Snoeyink, "Multiple structure alignment by optimal RMSD implies that the average structure is a consensus", *Proceedings of 2006 LSS Computational Systems Bioinformatics Conference*, pp.79-87, 2006.
[21] V. Alexandrov and M. Gerstein, "Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures", *BMC Bioinformatics*, 5:2, 2004.
[22] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, 3rd ed. New York, Wiley, 2000.