Inexact Graph Matching by "Geodesic Hashing" for the Alignment of Pseudoknoted RNA Secondary Structures

Mira Abraham and Haim J. Wolfson

Blavatnik School of Computer Science
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, Tel Aviv 69978, Israel
mira@post.tau.ac.il
wolfson@tau.ac.il (corresponding author)

Abstract. Non-coding RNAs are transcripts that do not encode proteins play key roles in many biological processes. The alignment of their secondary structures has become a major tool in RNA functional annotation. Many of the non-coding RNAs contain pseudoknots as a structural motif, which proved to be functionally important. We present HARP, a heuristic algorithm for the pairwise alignment of non-restricted (arbitrary) classes of pseudoknotted RNA secondary structures. HARP applies "geodesic hashing" to perform inexact matching of specially defined reduced RNA secondary structure graphs. The method proved to be efficient both in time and memory and was successfully tested on a benchmark of available experimental structures with known function. A web server is available at http://bioinfo3d.cs.tau.ac.il/HARP/.

Keywords: non-coding RNA, RNA structure alignment, secondary structure with pseudoknots, geometric hashing, geodesic distances, inexact graph matching

1 Introduction

Ribonucleic acid (RNA) molecules were once considered as mere carriers of the genetic information. Today it is known that many RNA transcripts or RNA segments within a transcript do not undergo translation. These sequences are often key players in numerous cellular processes such as chromosome replication, telomere maintenance, translation regulation and RNA modification. The increased interest in RNA function, and the assumption that the structure of a molecule reflects its function, motivates the development of efficient RNA structural alignment tools.

As in proteins, RNA structure is described at three levels. The **primary** structure is the RNA sequence drawn from an alphabet of 4 letters A,C,G,U. The **secondary** structure is a planar graph that consists of base-paired regions among single stranded regions. The **tertiary** structure consists of the 3D coordinates of the RNA molecule atom centers. Although there is statistical indication that tertiary structure similarity is more indicative of function similarity than secondary structure similarity [1], RNA structure usually means RNA secondary structure. This is due to the abundance of available modelled RNA secondary structures and the scarceness of RNA tertiary structures. The secondary structure is fully defined by the set of interacting base pairs $\{(i,j) : 1 \leq i < j \leq n\}$, where n is the length of the RNA molecule. Non-interacting consecutive nucleotides are called *loops*.

An RNA *helix (stem region)* is defined as a series of consecutive (in opposite directions) interacting base pairs $\{(i,j),(i+1,j-1)\ldots\}$. RNA helices, H_1 and H_2, are considered *crossing*, *non-nested* or *pseudoknotted* if $\exists(i_1,j_1) \in H_1, \exists(i_2,j_2) \in H_2$,

Mira Abraham, Haim J. Wolfson: Inexact Graph Matching by "Geodesic Hashing" for the Alignment of Pseudoknoted RNA Secondary Structures, pp. 45–58.
Proceedings of PSC 2011, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04870-2

so that $i_1 < i_2 < j_1 < j_2$ otherwise the helices are considered as *non-crossing* or *nested* (Fig. 1). The term *pseudoknot* refers to a set of crossing helices.

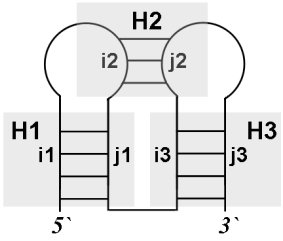


Figure 1. Pseudoknot example: H1 and H2 are crossing helices. H1 and H3 are non-crossing helices.

Many of the RNAs are known to have pseudoknots. Pseudoknots have been proved to be essential for the enzymatic activity of their RNAs, such as HDV ribozyme or self-splicing group I introns [31]. In other RNAs, such as 16S ribosomal RNA [27] and telomerase RNA [33], pseudoknots proved to be important for structural stability. Pseudoknots may also serve as drug targets as they are essential for induced frame-shifting in many viral RNAs [31].

Since an RNA secondary structure is represented by a graph, where vertices correspond to nucleotides and edges connect chemically interacting nucleotides/vertices, the pairwise RNA secondary structure alignment task is equivalent to detection of subgraph isomorphism. Moreover, due to possible insertions/deletions, which do not alter functionality, the algorithm should be tolerant enough to detect also "almost" isomorphic substructures. In the easier case, when pseudoknots are disregarded, the RNA graphs have been represented as rooted ordered trees, which led to polynomial algorithms for the detection of minimal edit distance between these trees [37,3,15,16,25,29]. Another approach for RNA secondary structure alignment that also disregards pseudoknots models the RNAs as "Multiple Graph Layers" (MiGaL) [4]. The case in which only one of the aligned structures is allowed to have pseudoknots [23] has been shown to be polynomial [17] as well. RNA structure alignment in the general case, where both structures can include pseudoknots, was shown to be NP-hard [38].

In order to circumvent the complexity challenge of the general alignment of RNA secondary structures, which include arbitrary pseudoknots, one can resort to the following strategies: (i) design algorithms that guarantee optimal solution for restricted classes of pseudoknots (limited problems). (ii) design heuristic algorithms that deal with the general problem (non-restricted classes of pseudoknots), however do not guarantee an optimal solution. Providing an optimal solution even for a limited problem is still complex, therefore, the currently available approaches suffer from both high time and high space complexity [26,12] which naturally creates limitations on both the complexity of the problem and the size of the input in addition to problem restrictions. In contrast to restricted problems algorithms, LARA [5], is a state-of-the-art heuristic method based on integer linear programming (ILP). LARA does not guarantee finding the optimal solution, however, it solves the general problem in relatively low time and space requirements. Heuristic algorithms, which tackle this, so called, "inexact graph matching task", have been also intensively studied in the structural pattern recognition community [7,19]. Another major application field is protein interaction network alignment in systems biology [11].

Here we present HARP – Geodesic **H**ashing **A**lignment of **R**NA secondary structures including **P**seudoknots. It is a novel, efficient heuristic method for solving the general (none-restricted) pairwise pseudoknotted RNA secondary structure alignment. HARP's technique was motivated by the Geometric Hashing [22,36] ideas, which were introduced for object recognition in Computer Vision. The presented algorithm performs inexact matching of directed graphs. HARP was applied to a benchmark of 31 structures from 16 functional groups showing the algorithm's ability to efficiently and accurately align RNA secondary structures, which include pseudoknots, successfully distinguishing homologous structures from non-homologous ones.

2 Method

The input to the HARP algorithm are two RNA secondary structures represented by graphs, were the vertices represent the nucleic acid bases and edges connect bases, which are paired by chemical bonds. The output of the algorithm is a pair of large subgraphs (preferably of maximal size), one from each molecule, which are "almost" isomorphic.

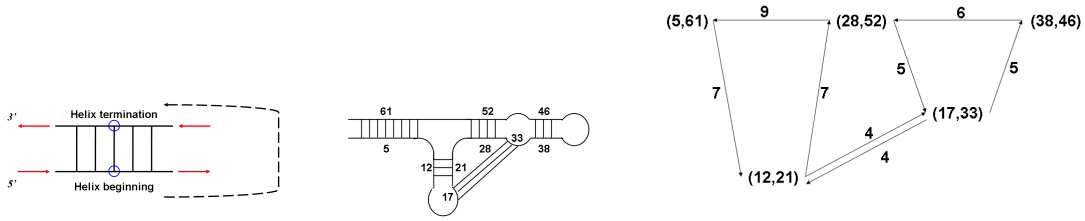
Informally, the algorithm performs as follows. First, the secondary structure graph is reduced to a directed graph, where the vertices represent significant enough helices and the edges connect chemically interacting helices. The edges are directed from 5' to 3'. Next, for each pair of vertices (nicknamed **basis**) a "local view" of the other accessible vertices is created. This is done by recording for each "viewed" vertex its directed "geodesic" distances from both basis vertices. The distance information is stored in a look-up table. Since similar substructures should produce identical "views" for each pair of corresponding bases, we are interested to align bases with similar "views". Such local alignments are efficiently retrieved from the look-up table. In the final stage the local alignments are clustered and merged to produce a maximal size alignment. Since this is done by a greedy procedure, maximality is not guaranteed. In the experimental benchmark the resulting alignments are large and biologically meaningful. The alignment technique is motivated by the Geometric Hashing method [22,36].

Our **Reduced Graph Representation** exploits the biological fact that RNA molecules usually form long double helices (stems), which are connected by stretches of non paired nucleic acids (loops). Extremely short double helices are usually unstable and therefore can be disregarded.

In the reduced graph representation each **stable helix** is represented by a **vertex**. It is described by the indices of the base pair at its middle by nicknaming the lower index of that base pair as *helix beginning* and the higher index as *helix termination* (see Fig. 2(a)). Each vertex stores the following triplet: (i) helix beginning; (ii) helix termination; (iii) helix length.

A **directed edge** connects two vertices if their corresponding helices are in contact. The edge direction is determined by the polymerization direction from 5' to 3'(see Fig. 2 (a)). The **weight of an edge** is set to the minimal number of nucleotides needed in order to connect the two vertices. This weight is correlated with the number of interactions that connect the helices and thereby provides an upper bound on their 3D distance. For two vertices v_i and v_j whose helix beginning points are i_b and j_b and termination points are i_t and j_t the distance is set to: $weight(v_i, v_j) = \min\{|i_b - j_b|, |i_b - j_t|, |i_t - j_b|, |i_t - j_t|\}$. An example of converting a secondary structure to the reduced graph representation is given in Fig. 2 (b) and

(c). Finally, for a given graph the directed distances between each pair of vertices are calculated by finding the all pairs shortest directed paths [9].



(a) (b) (c)

Figure 2. Graph representation. (a) Vertex representation: The helix is represented by the indices of the base-pair in its middle. The indices are determined by the position of the nucleotides on the RNA chain according to the polymerization direction (5‘ to 3‘) described by a dashed arrow. Edge direction also corresponds to the polymerization direction. **(b)** and **(c) Converting an RNA molecule to a graph:** The RNA secondary structure and its corresponding directed weighted graph. For each helix and vertex the *helix beginning* and *helix termination* are indicated (helix length is absent). The edge direction and weight are calculated as explained in the text.

Detection of Similar Local Environments: Let $G_1 = \langle V_1, E_1 \rangle$ and $G_2 = \langle V_2, E_2 \rangle$ be the reduced graph representations of the input RNA molecules. In order to detect locally similar substructures we follow the ideas of Geometric Hashing ([22,36]) adapting them for directed geodesic distances in graphs. First, each graph is pre-processed to encode the directed distances of each vertex from any pair of other vertices, which will be nicknamed – basis pairs. This redundant encoding is stored in a look-up table for efficient retrieval.

Specifically, let $v_i, v_j \in V_1$ where $i < j$, be a basis. For each $v_k \in V_1, k \neq i, j$, we consider two directed triangles – the **forward triangle:** $v_i v_k v_j$ and the **backward triangle:** $v_j v_k v_i$. The “length” of a directed triangle edge vu is the sum of weights in the shortest path connecting v to u. The triangle edges that touch vertex v_k are called *indexing edges*, while the edge connecting the basis vertices is called the *basis edge*. The lengths of the indexing edges are used to access a two dimensional look-up table, where the following information is stored: the vertex v_k, the “forward” or “backward” triangle type, and the basis pair $v_i v_j$. Each such vertex v_k will be encoded as $(x^f_{v_k}, y^f_{v_k})$ for the forward triangle and $(x^b_{v_k}, y^b_{v_k})$ for the backward triangle. Since this representation is done for each basis pair it is highly redundant. This redundancy ensures, that significant local alignments will not be missed.

Local Alignment Seeds: In this stage we efficiently detect bases in G_1 and in G_2 that have “similar views”, namely “almost” coincide on the distances of other vertices from them.

For a specific basis $v_i, v_j \in V_2$ where $i < j$, and for each node $v_k \in V_2$, the forward and backward triangles are calculated. We examine the look-up table entries within ε-vicinity[1] from the entry that corresponds to the lengths of the indexing edges of the forward and backward triangle respectively. We now list all triangles that satisfy the following: (i) **Type:** the triangle must be of the same type (forward or backward) as

[1] The ε defines our search radius (see Table 3 in the appendix for RADIUS SEARCH default parameter). The calculated distance between the entries is the l_2 distance.

the query triangle; (ii) **Length:** its basis edge is of similar length to the basis edge of the query triangle (up to an ε).

The set of triangles of some basis in V_1 that satisfy these conditions defines a "similar view" or a correspondence list between G_1 and G_2.

Refinement of Local Alignments: The correspondence lists computed at the previous stage are not necessarily one-to-one mappings, since a triangle in one graph may have more than one distance-congruent triangle in the other graph under the same "view". To resolve the conflicts we apply bipartite graph matching to refine the correspondence lists for pairs of bases, which scored more than 3 hits in the local alignment procedure.

For such a pair of bases, we define a bipartite graph $G_t = \langle V_{G_1} \cup V_{G_2}, E_t \rangle$ where V_{G_1} and V_{G_2} are the vertices of G_1 and G_2 respectively. We connect $v_i \in V_{G_1}$ with $v_j \in V_{G_2}$ if (for the given bases) the l_2 distance between the forward points $(x^f_{v_i}, y^f_{v_i})$ and $(x^f_{v_j}, y^f_{v_j})$ as well as the the backward points $(x^b_{v_i}, y^b_{v_i})$ and $(x^b_{v_j}, y^b_{v_j})$ is less than ε. We set the weight of edge, e_{v_i,v_j}, connecting v_i and v_j to: $w(e_{v_i,v_j}) = \frac{1}{C_f(1+d(v_i,v_j))} \frac{l(v_i)l(v_j)}{1+|l(v_i)-l(v_j)|^2}$, where C_f is a constant factor, $l(a)$ is the length of the helix represented by vertex a and d is the average l_2 distance between the forward and backward representation of the vertices. The weight $w(e_{v_i,v_j})$ is inversely proportional to the distance between the matched vertices and directly proportional to the length of the helices, which they represent. Based on experiments, the default value of C_f was set to 10.

The correspondence list (for a given pair of bases), which is calculated by the bipartite matching algorithm, is further pruned by removing pairs of vertices with non matching neighbors, namely, with the majority of their surrounding vertices not matched, as well as removing matched small connected components.

Alignment Extension: The local alignments of the previous stage should be further combined and extended to the largest biologically significant alignment. In the current development stage of the algorithm we have adopted a straightforward greedy approach, which starts with the largest matched set of vertices and at each step adds the local alignment that: (i) overlaps the currently matched set; (ii) adds the biggest number of vertex matches (disregarding the overlap). The procedure terminates when there are no new matches that can be added to the set. The resulting final alignment is the output of our algorithm.

Scoring: We evaluate the quality of the one-to-one mapping by the sum of matched nucleotide base-pairs ($S_{bp}(R_1, R_2)$) as a fraction of the total number of base pairs in the stable helices, which is $NS_{bp}(R_1, R_2) = \frac{S_{bp}(R_1,R_2)}{min(bp_1,bp_2)}$, where bp_1 and bp_2 are the total numbers of base-pairs in the stable helices in R_1 and R_2 respectively.

3 Results

In this section we have conducted an all-against-all alignment experiment on a benchmark of RNAs. First we present the performance of HARP on a benchmark of 31 RNA structures and present in parallel the performance of LARA, a current state-of-the-art arbitrary class pseudoknot alignment method. Then, we provide a more detailed analysis of HARP's all-against-all alignment scores. Finally, we provide a thorough examination of the HARP's alignments for some biologically interesting functional groups.

The all-against-all alignment experiment was conducted on a benchmark of 31 RNA structures belonging to 16 functional groups. All of these have high resolution 3D structures in the PDB [6]. The secondary structures were extracted by the X3DNA program from the 3DNA package [24]. To avoid redundancy the structures have less than 75 % sequence identity [34][2]. In order to avoid bias due to the size, functional group size was limited to 4. The complete list of PDBs of each functional group is listed in Table 2 of the Appendix.

3.1 Performance of HARP and comparison to LARA

We have calculated for both methods and for each functional group an average pairwise alignment score and a p-value. Since our run of LARA failed on the functional group of largest size (the 23S ribosomal RNA (rRNA) subunit whose average size is ∼2800 nucleotides) due to high memory requirements[3], this functional group was omitted from the comparison.

The **average score** of a functional group is calculated as the average of the normalized scores for each pair of structures belonging to that functional group. The normalized score of HARP is the calculated NS_{bp} (see "Scoring" paragraph of the "Methods" section). The corresponding normalized score for LARA was calculated as the alignment's identity score for a pair of structures divided by the size of the shorter structure of that pair.

A **p-value** is assigned to the score a of a pair of RNA structures within the same functional group f_i. $p(a) = \frac{N_a(R)}{\|R\|}$ where the random group R is the set of all scores between RNA structures R_i and R_j where $R_i \in f_i$ and $R_j \notin f_i$. $N_a(R)$ is the number of scores in R that have a score value above a. The size of the random group used for the p-value calculations includes functional groups of size 1. Specifically, HARP's random groups size are $\|R\| = 58$ or $\|R\| = 108$ for functional group size of 2 and 4 respectively. LARA's random groups size (due to exclusion of the 23S rRNA) are $\|R\| = 50$ or $\|R\| = 92$ for functional group size of 2 and 4 respectively.

The average scores and p-values for the different functional groups are presented in Figure 3(a) and (b). Generally, on the given benchmark for most functional groups HARP's average identity score is higher than LARA's score. Figure 3(b) also indicates that HARP's p-values are generally better than LARA's.

Ideally, an alignment method should successfully differentiate between pairs of structures from the same functional group and pairs of structures from different functional groups. In other words, the method's alignment scores within functional groups should be distinguishable from alignment scores between functional groups. This quality is well reflected in the receiver operating characteristic (ROC) curve. The ROC curve, which is widely used to evaluate a method's performance, plots the TPR (True Positive Rate) vs. the FPR (False Positive Rate) for different thresholds. The TPR corresponds to the sensitivity of the method and calculated TPR = TP + FN where TP is the number of correctly predicted pairs of the same functional group and FN is the number of incorrectly predicted pairs of different functional groups. The FPR corresponds to (1- specificity) of the method and calculated FPR = FP + TN where

[2] The sequence identity score was calculated as the number of matched nucleotides divided by the size of the smaller structure.

[3] All runtests for both methods were performed on the same PC workstation (Pentium© 4 1800 MHz processor with 1 GB internal memory) under the Linux operating system.

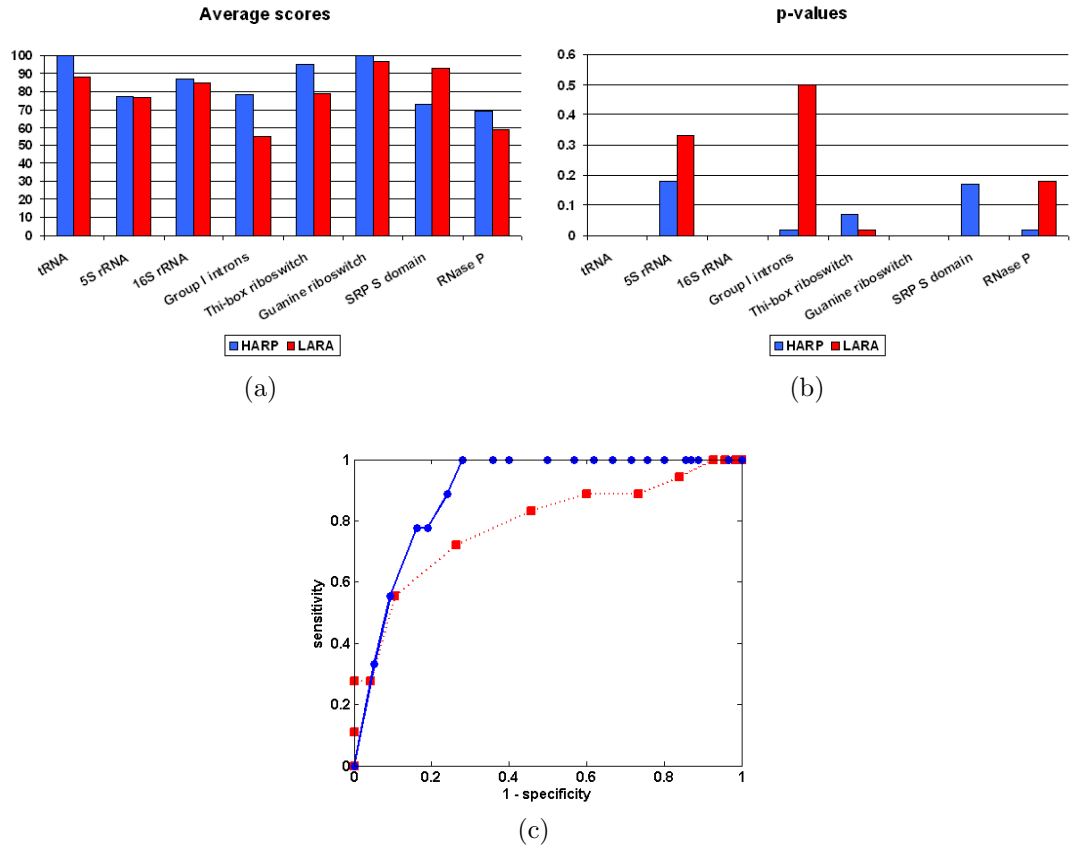


Figure 3. Comparison between HARP and LARA: (a)–(b) Average scores and p-values by functional groups: The average score and p-values presented here are only of functional of at least two structures. The group I introns is an abbreviation for group I self splicing introns pre-cleavage. RNase P is an abbreviation for RNase P catalytic domain. Average scores are given in percents. **(c) ROC curves for similar function predictors:** The curves are represented as blue circles and red rectangles for HARP and LARA respectively. The area underneath the curves is 0.89 and 0.79 for HARP and LARA respectively.

FP is the number of incorrectly predicted pairs of the same functional group and TN is the number of correctly predicted pairs of different functional groups.

The area under a ROC curve is used to evaluate the method's performance. The bigger the area underneath the ROC curve the more successful is the method in differentiating pairs of the same functional group from pairs of different functional groups. The ROC curves of HARP and LARA on the discussed benchmark are illustrated in Figure 3(c). The areas underneath the curves of HARP and LARA are 0.89 and 0.79 respectively[4].

In conclusion, for the presented benchmark, HARP's performance is better than LARA's as expressed by the following measures. First, HARP is able to align structures of very large size (e.g, the 23S rRNA). Second, HARP generally maintains higher average scores with lower p-values. Third, compared with LARA, HARP has improved ability to differentiate between pairs of structures of the same functional group from pairs of structures of different functional groups.

[4] The ROC curves were calculated for an identical data set for both HARP and LARA, omitting the 23S rRNA. Adding the 23S rRNA to HARP's data set improves its performance slightly as the area underneath the curve increases to 0.9 instead of 0.89.

3.2 Detailed analysis of HARP's scores

Table 1 details HARP's average scores and p-values for all functional groups. The highest average score (100 %) was achieved in two functional groups: tRNA and the Guanine riboswitch. Both functional groups include small structures consisting of 4 helices that differ in their topology: the guanine riboswitch contains a pseudoknot while tRNA does not. The average normalized sequence identity scores [5] for both functional groups as determined by ClustalW [34] are much lower: 50.7 % and 59.0 % for the tRNA and the Guanine riboswitch functional groups respectively. This could be accounted for compensatory mutations that do not alter the overall structure. In both groups the average pairwise tertiary identity scores as determined by ARTS [10] are also lower than those of HARP: 75.1 % and 81.7 % for the tRNA and the Guanine riboswitch functional groups respectively. This can be explained by hinges that alter the tertiary structure but have no effect on the secondary structure nor on the general functionality of the structure.

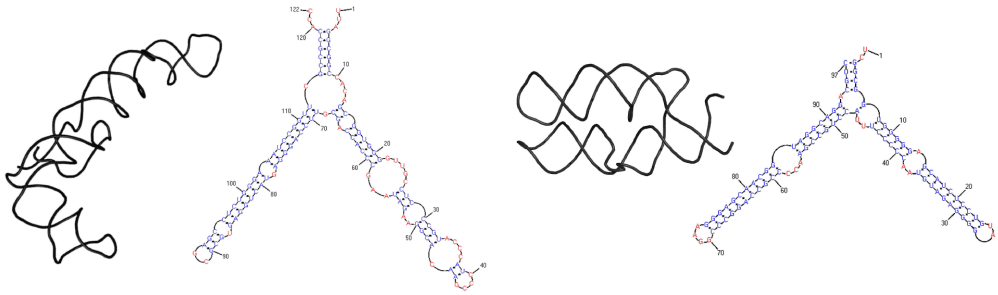
The lowest average score was achieved in the RNase P catalytic domain functional group (68.9 %). This can be explained by the fact that the molecules of RNase P, though sharing overall similar secondary structure, have some insertions and deletions (as the two differ in the number of stable helices, 16 and 19). These insertion/deletions are also expressed in LARA's low score (59.0 %) for the same structures.

Functional group	Group size	Average size (nucleotides)	Average score	p-value
tRNA	4	78	100 %	0
23S rRNA	4	2852	71.9 %	0
5S rRNA	4	120	77.2 %	0.18
16S rRNA	2	1530	86.7 %	0
Self splicing group I introns pre-cleaving	2	224	78.0 %	0.02
Thi-box riboswitch	2	80	95.0 %	0.07
Guanine riboswitch	2	69	100 %	0
SRP S domain	2	114	73.2 %	0.17
RNase P catalytic domain	2	298	68.9 %	0.02

Table 1. HARP's Detailed Statistics

Only three functional groups received a p-value bigger than 0.05: 5S rRNA, SRP S domain and the Thi-box riboswitch with p-values 0.18, 0.17 and 0.07 respectively. These p-values are attributed to high scores between the three functional groups. The biggest overall similarity between functional groups was observed between the 5S rRNA and the SRP S domain. These molecules, though having an overall similar secondary structure, do not have a common function nor a common tertiary structure (see Figure 4). Since these molecules do not contain pseudoknots and therefore are applicable to tree editing distance based methods, we have aligned them with RNAforester receiving quite similar results to those of HARP's alignment: average normalized RNAforester identity score 92 % compared with 75 % normalized HARP identity score (NS_{bp}). The high p-value of Thi-box riboswitch is explained in the same manner. This functional group has an overall similar secondary structure to both 5S rRNA and SRP S domain.

[5] The average normalized sequence identity score is calculated as the average of the normalized scores for all the structures pairs within the functional group. The normalized scores are calculated as the identity score divided by the size of the smallest structure of the pair.



(a) (b)

Figure 4. Different 3D Structures with Similar 2D Structures. (a) Secondary structure and tertiary structure of a 5S rRNA molecule (PDB:1yjw, chain 9). **(b)** Secondary structure and tertiary structure of a Signal Recognition Particle (SRP) molecule (PDB:1lng, chain B). The two RNA molecules share very little spatial similarity and have no known function in common. Nevertheless, the two molecules have very similar 2D structures with over 90 % RNAforester identity score.

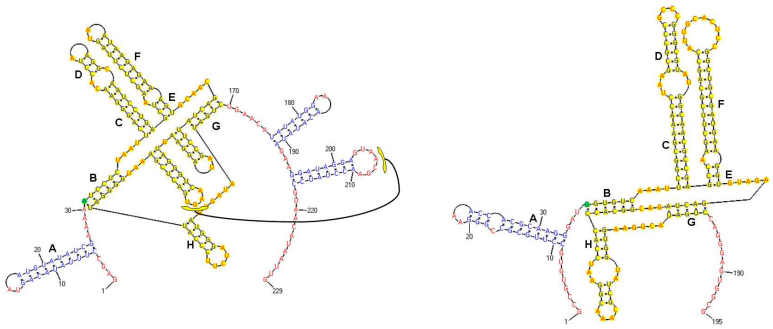
3.3 Detailed examination of HARP's alignments

Below, we provide a thorough examination of the HARP's alignments for some biologically interesting functional groups. We focus on the more challenging alignments, alignments of functional groups of size greater than 200.

Self splicing group I intron Self splicing group I introns catalyze their own excision from precursor RNA transcripts. The pseudoknotted region of the molecule is conserved throughout the different catalytic stages [30,13]. The pseudoknot actually establishes the ribozyme's catalytic core [2]. The secondary structure alignment of two self splicing group I introns done by HARP captures the entire pseudoknotted area (Figure 5). The self splicing group I introns structure variability is mainly in the 3D peripheral helices that are relatively remote from the active site. The main structure, consisting of the helices adjacent to the catalytic site, is conserved over all organisms [8]. The HARP alignment illustrated in Figure 5 includes all the helices that are 3D adjacent to the active site. Specifically, the matched helix labeled A corresponds to P2 in the P1-P2 domain. The matched helices labeled B, G and H correspond to helices P3, P7 and P8 in the P3-P9 domain. The matched helices labeled C, D, E, F correspond to helices P4, P5, P6 and P6a in the P4-P6 domain. The alignment does not include helices P7.1, P7.2, P9 and P9.1. These helices are present in only one of the structures, reflecting the variance between remote species: Homo Sapiens (PDB id 1zzn chain B) and Bacteriophage twort (PDB id 1y0q chain A).

Ribonuclease P Ribonuclease P is the enzyme that cleaves the tRNA at its 5′ and is therefore essential for the tRNA maturation. It is composed of two domains: the specificity domain and the catalytic domain. The currently solved ribonuclease P structures belong to two types according to its organism: Ancestral bacteria (A-type) and Bacillus (B-type). The two types have relatively similar secondary structures of the catalytic domain, while having considerable differences in the secondary and tertiary structures of the specificity domains [21,10].

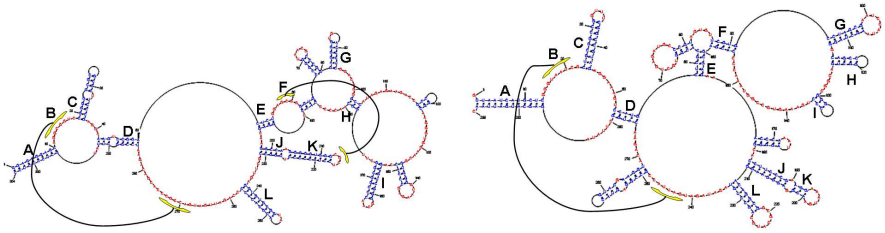
The solved tertiary structures are: PDB ids 2a2e chain A and 2a64 chain A, belonging to types A and B in correspondence. The alignment of the catalytic domains is illustrated in Fig. 6. The pseudoknot region that is conserved in both molecules was previously postulated to be important for the dynamics of the molecule [18].



(a) (b)

Figure 5. Alignment of the secondary structures of self splicing group I introns: (a) PDB id 1zzn chain B, 10 stable helices.**(b)** PDB id 1y0q chain A, 13 stable helices. The yellow arcs in the 1zzn chain B structure correspond to a helix. Matched helices are labelled by identical letters.

Even though there are two mismatches in this alignment (helices H, F) the rest of the alignment (matched 10 helices) including the conserved pseudoknot is consistent with the literature [18,35].



(a) (b)

Figure 6. Alignment of the secondary structures of the catalytic domains of ribonucleaseP: (a) PDB id 2a2e chain A, 19 stable helices **(b)** PDB id 2a64 chain A, 16 stable helices. The yellow arcs in each structure connects a helix. Matched helices are labelled by identical letters.

Ribosomal RNA Ribosomal RNA is the central component of the protein synthesis process. The prokaryotic ribosome is composed of a small unit containing the small subunit (30S) an the large subunit (50S). The small subunit includes the 16S ribosomal RNA that is ∼1500 nucleotides long that constitutes ∼85 stable helices. The large subunit includes the 5S and the 23S ribosomal RNAs that are ∼120 and ∼2800 nucleotides long and contain ∼7 and ∼140 stable helices respectively. Below are presented the results for the 16S and the 23S subunits. We focus on these units as they are of greater challenge, being very large and containing pseudoknots. Due to the large size and high complexity of the structures their alignment could not be illustrated but only literally described.

The alignment of the 16S subunit was performed on the two 16S structures, PDB ids 1yl4 chain A and 2i2u chain, containing 82 and 85 stable helices and 436 and 501 base-pairs in correspondence. The resulting alignment contains a match of 75 helices that overall include 378 base-pairs. The alignment contains 4 mismatched helices resulting in a correct match of 71 helices of 358 base-pairs (error rate of 5.3 %).

The correctly aligned helices include the central pseudoknot in 16S ribosomal RNA (indices (17, 918) in the 1yl4 chain A structure). This pseudoknot is known to be essential for ribosome stability [27]. The helices that were not matched by HARP are due to insertions/deletions.

The 23S functional group contains four structures. The pairwise alignments within the functional group had an average score of 71.9 % and 67.3 % true matching base-pairs (disregarding mismatches). In all alignments in this group the two pseudoknots are conserved. One of these pseudoknots is described by Steinberg et al. [32] as a G-motif. This motif is a highly conserved structural motif in ribosomal RNA. The other pseudoknot is also highly conserved connecting the molecule's beginning and ending.

4 Conclusions

We have presented a new heuristic algorithm, HARP, that aligns RNA secondary structures of non-restricted (arbitrary) classes of pseudoknots. HARP introduces a reduced graph representation of the secondary structures and aligns these reduced graphs by "geodesic hashing".

Evaluation of the experiments carried out on a relatively large benchmark demonstrates the biological significance of the obtained alignments. Those high quality alignments, are competitive with the results of a current state-of-the-art available RNA alignment method dealing with non-restricted classes of pseudoknots.

HARP is highly efficient: the average running time for a pair of pseudoknotted structures of the 23S ribosomal subunit ($\sim$ 2800 nucleotides) is less than a minute on a relatively weak single processor PC (Pentium© 4 1800 MHz processor with 1 GB internal memory under the Linux operating system). Currently, HARP is the only arbitrary class pseudoknots alignment method capable of aligning such big structures.

The presented algorithm is a general method for inexact matching of directed (and non directed) graphs. It detects large local "almost isomorphic" sub-structures. The algorithm's performance in other application domains will be examined.

5 Acknowledgements

This work is part of MA M.Sc. thesis. The research of HJW has been supported in part by the Israel Science Foundation (grant no. 1403/09) and the TAU Minerva-Minkowski Geometry center.

References

1. M. Abraham, O. Dror, R. Nussinov, and H. J. Wolfson: *Analysis and classification of RNA tertiary structures.* RNA, 14 2008, pp. 2274–2289.
2. P. L. Adams, M. R. Stahley, A. B. Kosek, J. Wang, and S. A. Strobel: *Crystal structure of a self-splicing group I intron with both exons.* Nature, 430 2004, pp. 45–50.
3. J. Allali and M.-F. Sagot: *Novel tree edit operations for RNA secondary structure comparison*, vol. 3240, Springer, Berlin/Heidelberg, 2004, pp. 412–425.
4. J. Allali and M.-F. Sagot: *String Processing and Information Retrieval*, Springer, Berlin/Heidelberg, 2005, ch. A Multiple Graph Layers Model with Application to RNA Secondary Structures Comparison, pp. 348–359.
5. M. Bauer, G. W. Klau, and K. Reinert: *Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization.* BMC Bioinformatics, 8 2007, p. 271.

6. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne: *The protein data bank.* Nucleic Acids Res., 28 2000, pp. 235–242.
7. H. Bunke and G. Allermann: *Inexact graph matching for structural pattern recognition.* Pattern Recognition Letters, 1 1983, pp. 245–253.
8. J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna: *Crystal structure of a group I ribozyme domain: Principles of RNA packing.* Science, 273 1996, pp. 1678–1685.
9. T. Cormen, C. Leiserson, and R. Rivest: *Introduction to Algorithms*, MIT Press, U.S.A, 1990.
10. O. Dror, R. Nussinov, and H. J. Wolfson: *ARTS: Alignment of RNA tertiary structures.* Bioinformatics, 21 Suppl. 2 2005, pp. ii1–ii7, `http://bioinfo3d.cs.tau.ac.il/ARTS`.
11. B. K. et al.: *Pathblast: a tool for alignment of protein interaction networks.* Nucleic Acids Res., 32 2004, pp. W83–W88.
12. P. A. Evans: *Finding Common RNA Pseudoknot Structures in Polynomial Time*, vol. 4009, Springer, 2006, pp. 223–232.
13. B. L. Golden, H. Kim, and E. Chase: *Crystal structure of a phage Twort group I ribozyme-product complex.* Nat. Struct. Mol. Biol., 12 2005, pp. 82–89.
14. F. Guo, A. R. Gooding, and T. R. Cech: *Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site.* Mol. Cell, 16 2004, pp. 351–362.
15. M. Höchsmann, T. Töller, R. Giegerich, and S. Kurtz: *Local similarity in RNA secondary structures*, in Proceedings of Computational Systems Bioinformatics (CSB 2003), C. Stanford, ed., 2003, pp. 159–168.
16. I. L. Hofacker: *Vienna RNA secondary structure server.* Nucleic Acids Res., 31 2003, pp. 3429–3431.
17. T. Jiang, G. Lin, B. Ma, and K. Zhang: *A general edit distance between RNA structures.* Journal of Computational Biology, 9(2) 2002, pp. 371–388.
18. A. V. Kazantsev, A. A. Krivenko, D. J. Harrington, S. R. Holbrook, P. D. Adams, and N. R. Pace: *Crystal structure of a bacterial ribonuclease P RNA.* Proc. Natl. Acad. Sci. USA, 102 2005, pp. 13392–13397.
19. S. Kosinov and T. Caelli: *Inexact multisubgraph matching using graph eigenspace and clustering models*, in SSPR & SPR, vol. 2396 of Lecture Notes in Computer Science, Springer Verlag, 2002, pp. 133–142.
20. A. S. Krasilnikov, Y. Xiao, T. Pan, and A. Mondragon: *Basis for structural diversity in homologous RNAs.* Science, 1 2004, pp. 104–107.
21. A. S. Krasilnikov, X. Yang, T. Pan, and A. Mondragon: *Crystal structure of the specificity domain of ribonuclease P.* Nature, 13 2003, pp. 760–764.
22. Y. Lamdan and H. Wolfson: *Geometric Hashing: A General and Efficient Model-Based Recognition Scheme*, IEEE Computer Society Press, Tampa, Florida, USA, December 1988, pp. 238–249.
23. G. H. Lin, B. Ma, and K. Zhang: *Edit distance between two RNA structures.* Proceedings of the fifth annual international conference on Computational biology, ACM Press, 2001, pp. 211–220.
24. X.-J. Lu and W. K. Olson: *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic Acids Res., 31 2003, pp. 5108–5121.
25. R. Mikhaiel, G. Lin, and E. Stroulia: *Simplicity in RNA Secondary Structure Alignment: Towards biologically plausible alignments*, IEEE Computer Society Washington, DC, USA, 2006, pp. 149–158.
26. M. Mhl, S. Will, and R. Backofen: *Lifting Prediction to Alignment of RNA Pseudoknots*, Springer Berlin, Heidelberg, 2009, ch. 5541, pp. 285–301.
27. R. A. Poot, C. W. A. Pleij, and J. van Duin: *The central pseudoknot in 16S ribosomal RNA is needed for ribosome stability but is not essential for 30S initiation complex formation.* Nucleic Acids Res., 24 1996, pp. 3670–3676.
28. P. Schimmel and K. Tamura: *tRNA structure goes from L to λ.* Cell, 113 2003, pp. 276–278.
29. S. Siebert and R. Backofen: *MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons.* Bioinformatics, 21 2005, pp. 3352–3359.

30. M. R. Stahley and S. A. Strobel: *Structural evidence for a two-metal-ion mechanism of group I intron splicing.* Science, 309 2005, pp. 1587–1590.
31. D. W. Staple and S. E. Butcher: *Pseudoknots: RNA structures with diverse functions.* PLoS Biology, 3(6) 2005, p. 213.
32. S. V. Steinberg and Y. I. Boutorine: *G-ribo: a new structural motif in ribosomal RNA.* RNA, 13 2007, pp. 549–554.
33. C. A. Theimer, C. A. Blois, and J. Feigon: *Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function.* Molecular Cell, 17 2005, pp. 671–682.
34. J. D. Thompson, D. G. Higgins, and T. J. Gibson: *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res., 22 1994, pp. 4673–4680.
35. A. Torres-Larios, K. K. Swinger, A. S. Krasilnikov, T. Pan, and A. Mondragon: *Crystal structure of the RNA component of bacterial ribonuclease P.* Nature, 437 2005, pp. 584–587.
36. H. Wolfson and I. Rigoutsos: *Geometric hashing: An overview.* IEEE Computational Science and Eng, 4(4) 1997, pp. 10–21.
37. K. Zhang and D. Shasha: *Simple fast algorithms for the editing distance between trees and related problems.* SIAM J. Computing, 18(6) 1989, pp. 1245–1262.
38. K. Zhang, L. Wang, and B. Ma: *Computing similarity between RNA structures*, vol. 1654, Springer, 1999, pp. 281–293.

6 Appendix

Functional group	**PDB Codes**
23S rRNA	1vqm chain 0, 1vp0 chain B 2hgu chain A, 2i2v chain B
5S rRNA	1vp0 chain A, 1yjw chain 9 2hgu chain B, 2awb chain A
16S rRNA	1yl4 chain A, 2i2u chain A
tRNA	1efw chain D, 1ttt chain F 2dxi chain D, 2hgr chain D
lambda form tRNA	1j2b chain C
Self splicing group I introns pre-cleaving	1y0q chain A, 1zzn chain B
Self splicing group I introns post-cleaving	1x8w chain B
Thiamine pyrophosphate (thi-box riboswitch)	2cky chain B, 2hoo chain A
Guanine riboswitch	1y26 chain X, 1u8d chain A
Signal recognition partical (SRP) S domain	1mfq chain A, 1z43 chain A
Ribonuclease P (RNase P) catalytic domain type A and B	2a2e chain A, 2a64 chain A
Ribonuclease P (RNase P) specificity domain type A	1u9s chain A
Ribonuclease P (RNase P) specificity domain type B	1nbs chain B
MLV Psi site	1s9s chain A
muPsi	2ihx chain B
S-adenosylmethionine riboswitch	2gis chain A

Table 2. HARP Data Set: The first four letters of an RNA name are the PDB code, followed by the chain id. Ribonuclease P of type A (Archeal) or B (Bacterial) are known to have different secondary (and tertiary) structure. The differences between the two types are more expressed in the specificity domain [20]. Therefore the Ribonuclease P structures were divided to the above three functional groups. Self splicing group I introns also change their secondary (and tertiary) structure upon cleavage, loosing both their exons and therefore having different functional groups for different stages of the catalysis [14]. Lambda form tRNA differs from the canonical "L shaped" tRNA in both secondary and tertiary structure [28] and therefor was considered separately.

Parameter name	**Parameter description**	**Default value**
MIN HELIX LENGTH	The minimal number of consecutive base-pairs that define a stable helix.	3
MAX DIST BASIS	The maximal (geodesic) distance between two vertices. Above this threshold the two vertices will not be considered as basis. Note that the graph diameter of the largest single structure is 640.	50
RADIUS SEARCH (ε)	The maximal l_2 distance between the positions of two matched points under a certain transformation.	15
C_f	The ratio between the distance constraint and size constraint in the determination of the bipartite edges weights.	10
MIN SOLUTION SIZE	The minimal number of matched helices under a given transformation. Under this threshold the transformation is ignored.	3
MIN NEIGHBOR NUM	The minimal number of matched immediate neighbors in order to consider the pair of vertices as matched.	2
MIN SIZE CONNECTED COMPONENT	The minimal size of connected component of the matched vertices. Under this size vertices matches are ignored.	3

Table 3. HARP parameters description and default values: This table presents the default values used by the HARP program. The parameters enable flexibility of the program according to various sets of demands coming from end-users. However, for most experiments the default parameter values were sufficient. These parameters were determined by a trial and error process.