RNA as a Permutation

Nilay Chheda
Qualcomm India
Hyderabad, India
Email:nchheda@qti.qualcomm.com

Manish K. Gupta
Laboratory of Natural Information Processing
Dhirubhai Ambani Institute of Information
and Communication Technology
Gandhinagar, Gujarat, 382007 India
Email: mankg@computer.org

***Abstract*—RNA secondary structure prediction and classification are two important problems in the field of RNA biology. Here, we propose a new permutation based approach to create logical non-disjoint clusters of different secondary structures of a single class or type. Many different types of techniques exist to classify RNA secondary structure data but none of them have ever used permutation based approach which is very simple and yet powerful. We have written a small JAVA program to generate permutation, apply our algorithm on those permutations and analyze the data and create different logical clusters. We believe that these clusters can be utilized to untangle the mystery of RNA secondary structure and analyze the development patterns of unknown RNA.**

***Keywords*—RNA Secondary Structure, RNA Classification, Permutation Based Approach, Similarity, RNA Clusters**

I. Introduction

RNA molecule is one of the three major macro molecules which is essential for all existing models of human life. RNA is a short form of Ribo nucleic acid. RNA molecules plays pivotal role in many biological functions. RNA can rebuild and transport genetic data [1], drive chemical reactions [2] and administer gene expressions [3]. RNA molecule's capability to perform bio-molecular computation through nanotechnology has escalated its importance among researchers from various fields ([4], [5] and [6]). RNAs form primary, secondary and tertiary structures very much like DNA and protein. In essence, primary structure is a simple one dimensional sequence of nucleotides whereas secondary and tertiary structures are nothing but two dimensional and three dimensional representation of that sequence respectively. There has been a debate going on for many years whether RNA is the only molecule that is responsible for evolution of life or RNA along with DNA and protein have facilitated the evolution.

A. Concept of RNA World

"RNA world" is theoretical time of the early ecosphere. During this time period, knowledge required for life and alchemical movement of lively organisms were accommodated by RNA molecules [7]. Another set of arguments indicated that compared to RNA, availability of DNA was rich in that time period. DNA is more stable in mildly primitive atmosphere. Existence of ammonia-rich sea and the seasoning of primitive crusts of earth also support former statement [8]. Another variant postulated that a different kind of nucleic acid came first which was called pre-RNA. It had a property to emulate itself and RNA, as we know it today, substituted it eventually. On contrary, new results prompted that pyrimidine ribonucleotides can be prepared under certain prebiotic setup [9].

B. Types of RNA

Majority variants of RNA fall under RNAs that are either involved in protein synthesis or DNA replication or Regulatory RNA. Some RNA has the double stranded structure like DNA they are named as dsRNA. Non-coding RNA (ncRNA) is one which does not translate itself into protein. Often it is supposed that large amount of genetic information is carried out by proteins. Modern studies have implied that unlike proteins, ncRNAs are generated by the translation of genome of many mammals and certain organisms [10]. Many regulatory RNA, tRNA, rRNA also fall under ncRNA category. Three major RNA involved in protein synthesis are following:

1) **mRNA** – mRNA stands for Messenger RNA. mRNA plays role to forward genetic information from DNA to the ribosome. In messages transferred, mRNA encodes details about amino acid strand for gene expression of protein derivatives. It is a one of the largest subset of RNA.
2) **rRNA** – rRNA stands for Ribosomal ribo nucleic acid. It is one of the important constituent of ribosome and fundamental for protein construction in all forms of life.
3) **tRNA** – tRNA stands for Transfer RNA. As the name suggests, it acts as an interconnection between the different types of chain of nucleotides (DNA & RNA) and amino acid string of proteins.

We do not know the correct answer of RNA world debate till date but what we know is that "RNA", irrespective of the fact whether it is the only molecule responsible for evolution of life, is certainly responsible for some of the biological functions. This very fact is motivating enough to analyze its behavior. Its behavior can be understood by finding patterns in its secondary structure in which it folds. This folding can be cumulative result of many different known and unknown biological, chemical, thermodynamic and mathematical parameters. Here we have tried to explore mathematical aspect of RNA secondary structures in the form of permutation.

After giving basic introduction in section I, we explain basics about RNA structure and different techniques to represent those structures in section II and III respectively. We also introduce our own representation of RNA secondary structure which is derived from one of the existing representation in section III. In section IV, we explain our algorithm to calculate two different types of similarity score based on the new representation that we proposed in the section III. We also show results of our analysis based on the algorithm. In

section V, we discuss what more can be build up based on the results that we got. Finally, we conclude our work in the section VI.

II. RNA Structure

As we mentioned earlier, RNA has three different types of structure. One dimensional structure is called primary whereas two and three dimensional structures are called secondary and tertiary respectively. Primary one is a random linear sequence of four nucleotide namely – Adenine, Cytosine, Guanine and Uracil (represented as alphabets A, C, G and U respectively). RNA is usually single stranded unlike DNA.

A. Chemical Structure

Chemically, RNA is a linear chain of polymers of ribose sugar having ringed structure. This ring has five carbons. Third ($3'$) and Fifth ($5'$) carbon are connected with phosphate group which act as a linkage forming chain of ribose sugar. First carbon is connected with one of the four base group (A,C,G & U) which are derivatives of purines and pyrimidines chemical structure. Now, this base group can pair with another base group with hydrogen bonds. A & T can be paired with two hydrogen bonds and C & U can be paired with three hydrogen bonds. Double stranded RNA (dsRNA) and DNA have complementary linear sequence means sequence which can be paired with original sequence with Watson-Crick pairing. Usually original sequence is observed from ($3'$) end to ($5'$) end and vice verse for complementary sequence. These relationships are called the rules of Watson-Crick base pairing [11]. There is a thermodynamic energy involved with each hydrogen bond according to which relationships between A & T and C & G are the most stable. Sometimes G is also paired with U with two hydrogen bond. This relationship is called as wobble pairing. This type of pairing is comparatively much less in numbers than Watson-Crick pairing.

B. Secondary & Tertiary Structure

As discussed in the previous subsection, RNA strand consist of bases which can pair with each other. This pairing take place in a way that the resulting structure becomes the most stable in thermodynamic aspects. There are only few possible substructure shapes in which RNA can fold. There can be multiple substructures of same type in single secondary structure and if not same they can be different in size (in terms of no. of paired and unpaired base). As shown in Figure 1, major substructure types are Stem, Bulge, Hairpin, Multiloop (Junction), Interior Loop and Pseudoknot. Any RNA secondary structure will be combination of these substructures only.

Secondary structure of an RNA is nothing but a planar view of actual real three dimensional (tertiary) structure. This tertiary structure is the actual mystery that many researchers want to solve but only possible way to understand it is by understanding two dimensional structure of it first. As shown in Figure 1, psuedoknots are the most complex and hardest to predict substructures because of its bonding nature. Psuedoknots are formed when two substructure other than pseudoknot come closer in three dimensional geometry and certain portion of strand in one substructure is complementary to some portion of strand in the other structure. Usually, unpaired portion of both the substructures are paired when psuedoknots are formed.

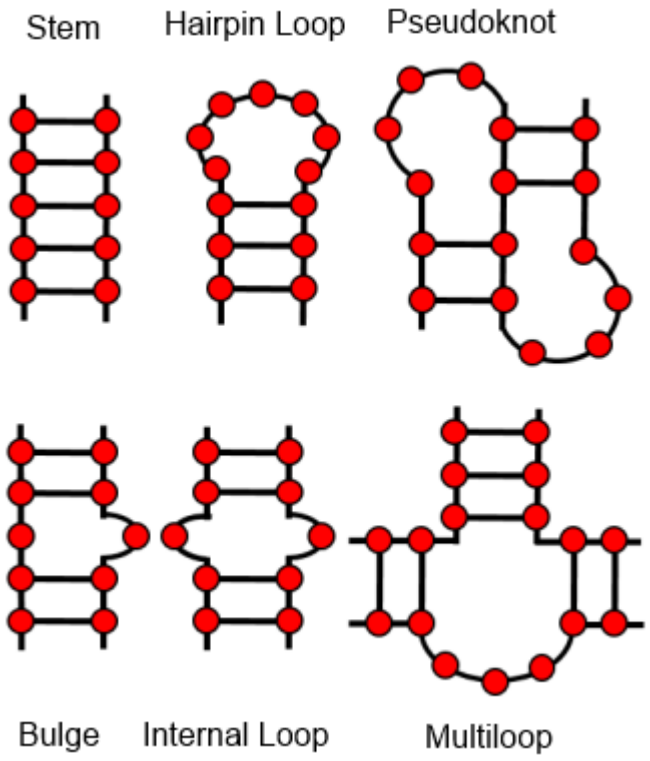

Fig. 1. Different types of RNA Secondary Structures. [1] First structure represents a stem with five paired bases. [2] Second structure represents Hairpin loop having three paired and six unpaired bases. [3] Third structure is of a HH type of pseudoknot which is one of the most common type in which unpaired base of two hairpins structure get paired with the other one. [4] In the first structure in the bottom row, we have shown a bulge of size one. There can be multiple unpaired bases in structures like this. [5] In fifth structrue we have shown internal loop which usually have double bulge like a structure. [6] In last image in the bottom right we have shown multiloop which joins multiple substructure.

C. Secondary Structure Prediction

The Human Genome Project which started in 1993 by the global research community have finished sequencing approximately 200 million base pairs by end of its first five year plan which ended in 1998 [12]. This kind of large scale project along with many similar small scale projects have generated immense amount of biological data in past few years. Now, newer challenge standing ahead of research community is to understand this information. In recent years importance of understanding secondary structure of RNA has become very rewarding and essential work. Efforts to accomplish this task has resulted in various techniques and approaches. RNA secondary structures are critical in many biological mechanisms and accurate techniques for prediction can provide crucial pathway to conduct effective research in the area [13].

First breakthrough algorithm was introduced in 1978 which was based on dynamic programming paradigm and designed to do sequence matching [14]. Based on this algorithm, efficient and precise method to predict secondary structure

was designed in 1980 [15].This algorithm is famously known as "Nussinov's algorithm". Now problem with this method was that it did not take into account thermodynamic and energy related factors affecting structure. Micheal Zucker and his colleagues came up with newer algorithm which was able to predict the energetically most stable structures of an RNA [16]. They also implemented this algorithm as a computer software [17]. They have been constantly improving their algorithm and software based on better free energy calculations [18]. As new standards and formats for representing RNA secondary structure were developed e.g. RNAML [19], they also updated their software to provide output in this standard format. As the usage of internet grew in early 20th century, they set up web server which provided their software as an online service. They called this server as mFold server [20]. This server, till date, is the most popular server for structure prediction due to its wide range of features and accuracy. Meanwhile another similar effort was carried out by group of researchers who also tried to develop a computationally faster techniques for prediction and comparison of different RNA structures [21]. They also set up a web server at almost the same time as mFold server [22]. Their server is called a Vienna Web Server.

Major shortcoming of Zucker's algorithm was that it was not able to predict pseudoknot within a secondary structure. This limitation motivated researchers to design algorithm which can also predict structures with pseudoknot as RNA pseudoknots are quiet influential for various existing RNAs [23]. The very first such algorithms was developed in 1999 and named as "PKNOTS" [24]. This algorithm was capable of computing secondary structure containing widest class of pseudoknots but it was computationally very inefficient ($O(n^6)$) although it was developed based on the Zucker's original algorithm which runs in $O(n^3)$. There are different types of pseudoknots that occur in secondary structures ([24], [18]). There were many variants of [24] algorithm ([25], [26], [27], [28]) developed during the time frame of five years . Most of them discarded one or more types of pseudoknots from their methods to bring down the computational complexity. But after all these efforts computational complexity was ranging from $O(n^4)$ to $O(n^6)$ which was still better than Zucker's $O(n^3)$. Another problem with these methods was is that they were producing only the optimal solution while ignoring sub optimal solution which may reveal the true structure [29]. On the contrary, computationally efficient (within $O(n^3)$) methods ([30], [31]) have also been developed which takes heuristic approach without restricting to any particular class of pseudoknots. These methods do not guarantee optimality and quality for the results of prediction though their computational complexity is much better than previous methods [29].

III. Representation of RNA secondary structures

After development of so many RNA secondary structure prediction techniques, there were three main areas which emerged as a logical extension to mankind's aspiration to solve the mystery of life which are following:

1) **Classification of predicted structures** – Classification brings RNA secondary structure with similar characteristic in one or more ways together. This field encouraged creation of database having basic database functions like search, sort, add, delete. This database operations can be helpful to carrying out further scientific research on the data. Database containing one dimensional sequence data of various kind of proteins, viruses, RNAs, DNAs etc. already exist. Very recent work [32] has brought together almost all the one dimensional sequential data that is known to the world which is just 10% of total possible data.
2) **File Format** – Today, we are living in the computer age. Practically every research area in the world need computers for carrying out various tasks of computation, simulation etc. Biological data size is one of the largest in the world and large data size are nearly impossible for humans to process. RNA secondary structure data is also quiet big which has created requirement of proper file formats that computers can interpret.
3) **Display Tools** – Computers can do computation with the data but human touch (in technical terms input) is always required to do meaningful computation. Computer needs to display data in a way that a human can understand e.g. showing secondary structure drawing along with a text file having thousands of alphanumeric characters is much more convenient & informative for human than simply showing the same text file without visual output. Display tools developers are constantly trying to make visuals clearer and more informative.

Some of the RNA secondary database existing today are NDB [33], Rfam [34], RNasePDB [35], sprinzIDB [36], tmRDB & SRPDB [37], CRW [38] and PDB [39]. Pseudobase is a database which only contain structures having pseudoknot ([40], [41]). Based on various minimum free energy (mfe) based prediction algorithm, [42] has also tried to classify pseudoknotted structures. All of these databases are constantly being updated with more data, better accessibility and useful features. Very recent example can be Rfam [34] whose 11th version has just been released [43]. We have used RNA STRAND [44] database for our analysis purpose, as it has collaborated almost all of above mentioned databases and combined all of them together.

Different notation used to represent secondary structures are String Notation, Bracket Dot Notation, Linked Graph Notation, Circular Notation, Dot Plot notation, Mountain Plot Notation, Mountain Metric notation, Tree Notation. Bracket notation has become quite popular and an extended dot bracket notation [45] has been developed which can also represent pseudoknots. Dot bracket notation is just a sequence of dots and brackets in which dot represents unpaired base and brackets (parentheses, square brackets, and curly braces - depending upon the base pairing and structures like pseudoknot) represent paired base. Originally dot bracket notation was used by Vienna Web Server [22]. RNADraw [46] and RNAviz [47] are software that can visualize secondary structures. Most of these kind of software use XML based RNAML format [19]. Currently, Psedoviewer [48] is the most popular visualization software as it can draw pseudoknotted structures. It has constantly updated its software to support more and more pseudoknots and better and efficient 2D and 3D visualization [49]. One of the popular file format is ".ct". This format is nothing but the table containing six columns and along with one common

header text. Common columns of the tables are index of base in the sequence, index of paired base in the sequence for corresponding base, alphabet representing base etc. There are many variants of .ct format like "Vienna .ct", "RnaViz ct", "Mac ct" and ".bpseq". All of them have table with minimum three columns which we just mentioned. Some of them have header containing energy related information and some of them have columns that contain energy related information. ".bpseq" is the file format which has just three column table and a simple header containing information like name of the file, source of the file and accession number. All of the formats are supported by mFold [20] and Vienna Server [22]. We have used ".bpseq" file format to create our hybrid notation which we use for our analysis.

A. New Hybrid Representation

Our new hybrid notation is developed based on the permutation concept of mathematics. In the subsection below, we will explain what permutation is, how we are using permutation based representation for creating clusters from the given data set and finally how permutation notation can be helpful to propose a new kind of classification.

B. Permutation

Before, we explain about our hybrid notation, we would like to first explain what permutation is. Dictionary meaning of permutation is rearrangement. For n number of different objects, there are $n!$ different arrangement or permutation possible. In combinations theory, a permutation is a sequence having all the elements of a finite set just once.

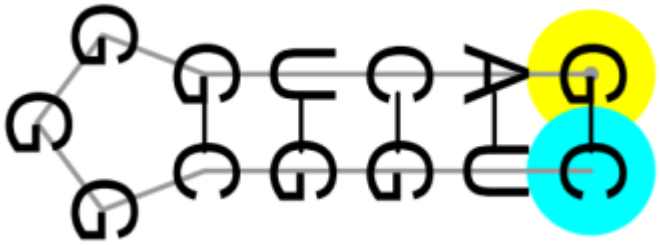

Fig. 2. Secondary Structure for PDB_00226

TABLE I. PERMUTATION BASED HYBRID NOTATION OF RNA SECONDARY STRUCTURE

1	2	3	4	5	6	7	8	9	10	11	12	13
G	A	C	U	G	G	G	G	C	G	G	U	C
13	12	11	10	9	0	0	0	5	4	3	2	1
13	12	11	10	9	6	7	8	5	4	3	2	1
0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	1	1	0	1	1	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0

In Table I, first three rows are simply horizontal representation of ".bpseq" file format of secondary structure of tRNA with accession number PDB_00226 (shown in Figure 2) taken from [44]. Fourth row is nothing but the permutation that we have derived from the third row by simply replacing 0 with its corresponding index. Here indexes 6, 7 and 8 are unpaired which is why their value in third row is 0 unlike in row 4. This way, we can get unique permutation defined on a set S. Here length of RNA sequence is 13. In fourth row integers from 1 to 13 appear exactly once which is why it is a permutation. Next four rows are indicator sequence [50] of A, C, G and U respectively. Let us call this eight sequence notation/format/representation as "permutSeq".

Permutation defined on the RNA sequence has a very special property that permutation always form two cycles or one cycle. This is because of the nature of RNA, where there can be either paired base or unpaired base. For unpaired base we always get one cycle and for paired base we always get two cycles.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\ 13 & 12 & 11 & 10 & 9 & 6 & 7 & 8 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$$

$$= (1\ 13)\,(2\ 12)\,(3\ 11)\,(4\ 10)\,(5\ 9)\,(6)\,(7)\,(8)$$

Fig. 3. A permutation representation of PDB_00226 of Figure 2

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 3 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

Fig. 4. All 9 possible permutations of length 4 used in counting

One can count such permutations. Suppose, there is an RNA sequence of length 3. If we hypothetically enumerate all the possible permutation of length 3, there will be some paired bases and some unpaired bases. Paired bases will always be in even number as they need a partner to pair with. So, essentially it is a partition problem. 3 can be partitioned into unpaired and paired (u, p) as $(0, 3)$ or $(1, 2)$. For $(0, 3)$ there will be only one possible structure. For $(1, 2)$, there will be three possible structures calculated as $\binom{3}{1} \times 1$. Let us take example of length 4. 4 can be partitioned into $(0, 4)$ or $(2, 2)$ or $(4, 0)$. There will be only one structure possible for $(0, 4)$ as all of them are unpaired. For $(2, 2)$, there will be 6 different structure possible which can also be obtained by the calculation $\binom{4}{2} \times 1$. Similarly, for $(0, 4)$ case, there are 3 different possible arrangements which can also be computed by $\binom{4}{0} \times 3 \times 1$. We have shown all 9 possible permutation in Figure 4. We are discarding the trivial case in which all the base are unpaired. For any n, this case will result in only one structure. That can be included by simply adding 1 to the above formula. Now, generalizing above results for any integer n give us the following formula,

$$No.\ of\ structures\ (\ \theta \in \mathbb{N})\ and\ n \geq 2 =$$

$$\sum_{k=0}^{\theta-1} \binom{n}{2k} \times \Big\{ [n-(2k+1)][n-(2k+1)-2] \dots 1 \Big\}$$
$$n \in \mathbb{N},\ n \textit{ is even and } n = 2\theta$$
$$\sum_{k=0}^{\theta-1} \binom{n}{2k+1} \times \Big\{ [n-(2k+2)][n-(2k+2)-2] \dots 1 \Big\}$$
$$n \in \mathbb{N},\ n \textit{ is odd and } n = 2\theta+1$$

Above formula can be simplified as follows:

Theorem 1. *Let $S(n)$ be the number of secondary structures represented as a permutation of n length. Then for $n \geq 2$, $S(n)$ is given by,*

$$\sum_{k=0}^{\theta-1} \binom{n}{m-1} \times \left\{ \prod_{p=0}^{(\frac{n-m-1}{2})} (n-m-2p) \right\}$$

Where if n is odd then $n = 2\theta+1, m = 2k+2$. If n is even then $n = 2\theta, m = 2k+1$

In the Table II, we have listed the results of above formula from $n = 2$ to $n = 9$. This series is often known as number of degree - n permutation of order exactly 2. There are alternative ways describe this formula. You can look at the values of this series upto $n = 300$ from web url http://oeis.org/A001189/b001189.txt.

TABLE II. NO. OF DIFFERENT SECONDARY STRUCTURE PERMUTATION POSSIBLE DIFFERENT VALUES OF N

n	2	3	4	5	6	7	8	9
structures	1	3	9	25	75	231	763	2619

IV. CLASSIFICATION OF RNA

Classification of RNA is a very subjective topic and it is very closely related with the database development of data. Many approaches have been taken to carry out this work. Some of the approaches ([44], [41]) are based on the biological aspect of the secondary structure whereas some ([51], [52]) are based on mathematics. Our permutation based approach also fall under the later category. In previous section, we talked about new hybrid notation, "permutSeq" which will contain eight different sequences as shown in Table I. Now, we will explain two simple algorithm we have developed to compute two types of similarity score and create clusters based on these two similarity score.

A. Algorithm

Algorithm 1 computes two similarity score, first based on character sequence (containing A,C,G&U) and the other based on permutSeq. "getSequence(i)" procedure call on permutSeq returns i^{th} character of an original sequence (row 1), "getIndicatorX(i)" procedure call on permutSeq returns i^{th} character of indicator sequence (row 5-8) of X (X = A, C, G & U) and "getPermutation(i)" procedure call on permutSeq returns i^{th} character of permutation sequence (row 4).

Algorithm 2 describes procedure for creating clusters based on the similarity score data given as an input. Cluster creation can be applied only on a set of Similarity Score S where,

$$S = \{(m,n) \mid m,n \in \mathbb{N}, (m,n) = similarityScore(P1,P2)\}$$

Algorithm 1 Similarity Score Computation

```
1:  Procedure (permutSeq P1, permutSeq P2)
2:  Score1 ← 0 \\StrandSimilarityScore
3:  Score2 ← 0 \\PermutationSimilarityScore
4:  while Character i in P1 and Character j in P2 do
5:    if P1.getSequence(i) = P2.getSequence(j) then
6:      Score1 ← Score1 + 1
7:    end if
8:    if P1.getPermutation(i) = P2.getPermutation(j) then
9:      if P1.getIndicatorX(i) = P2.getIndicatoX(j) then
10:       \\X = A, C, G and U
11:       Score2 ← Score2 + 2
12:     end if
13:     Score2 ← Score2 + 1
14:   end if
15: end while
16: return Score1, Score2\\ Score1, Score2 ∈ ℕ
```

Here, $SimilarityScore(P1,P2)$ procedure refers to output of Algorithm 1 for inputs P1 and P2 whose type is permutSeq. We have given certain type of RNA to Algorithm 1 and generated one single output file listing all permutSeq. If there are n files in an input folder for Algorithm 1, it generates ${}^{n}C_{2}$ number of permutSeq entries in output file which is used as an input for Algorithm 2.

Algorithm 2 Cluster Creation

```
1:  Procedure (Score1[ ], Score2[ ])
2:  clusters[ ] \\Array of Clusters
3:  for all Integer i in Score1[ ] and Score2[ ] do
4:    Integer j ← i
5:    for j in Score1[ ] and Score2[ ] do
6:      if Score1[j] = Score1[i] and Score2[j] = Score2[i]
        then
7:        Create Cluster Rule_Score1[j]_Score2[j]
8:      end if
9:      j ← j + 1
10:   end for
11: end for
12: return clusters
```

Main features of the above two algorithms are:

- It can rank all types of sub structures including pseudonots.
- Score computation runs in $O(n)$ which is much faster than any existing algorithm for doing similar kind of work.
- It scores binding and binding with same base pairs in $1:2$ ratio.
- It captures structural similarity between any two RNA secondary structure.
- It can be used to classify entire secondary structure data available or sub-classify existing classes of secondary structure.
- It can be used to study the evolution of a particular species like HIV. For example, one can study which particular structural portions are not evolving among

all different variants of HIV virus and infer which structural portions are actually tweaked by nature.

- Right now we have done exhaustive analysis of a given data set but it can also be used to analyze entire data set with respect one particular strand.
- Problem with the above point is that it is hard to analyze evolution if bases are being added or removed. It can be done only if modification of base is taking place.
- One of the limitation is that it can be applied only on secondary structure data which is very less compared to actual sequence data.
- Similarity Score obtained are overall score, as of now there is no way to determine whether that score is result of contiguous portion of secondary structure or not.

B. Results

We, have applied both of the above algorithm on various data available on [44]. We have got some really interesting results which we have tabulated in Table III. One very important point here is that set of all clusters created for any particular set of data is not a disjoint set because of the nature of the computation which is happening between every two entries from given data set. In Figure 5, we have shown two RNA secondary structure (accession number CRW_00609 and CRW_00699 [44]) side by side. From the given drawing of the structure it is very hard to find any similarity between these two structures. But both of them belong to the same cluster having strand similarity score of 98. Their total length are 373 and 429 and both of them contain a pseudoknot. Pseudoknot bindings are shown in green color in both the structure. Cluster in which they belong has a permutation similarity score of 352 which is very high score compared to total length of these two structres.

Here let us clarify one thing, scores associated with any cluster are fixed and there can be many structures belonging to the same cluster but that does not mean all the members of a single cluster has the same similarity score with every other member of that cluster. But there are at least two members in the same cluster which can give exactly same similarity score as defined for that particular cluster. Another very useful observation regarding scores is that if permutation similarity score between any two structure is more than twice the strand similarity score than it is more probable that both the structures have similar sub structures.

C. Comparison

Now, we would like to compare our classification methodology which uses clusters and permutation with one of the very crude but yet useful type of classification [51]. This is very famously known as RAG or RNA-As-Graphs. We are calling these classification very crude because it is loosing so much information about actual sequence like base count, type of base & base pairing. It is useful because it is capturing all kind of structural information including pseudoknots along with their orientation in two dimensional geometry. Another reason for selecting [51] is that this approach is using graph theory which is not a biological concept but a mathematical concept like ours.

TABLE III. CLUSTERS GENERATED FOR DIFFERENT TYPES OF RNA

RNA Type	# RNA	RNA Length	# Clusters	Cluster Range
Transfer Messenger RNA	726	102 to 1331	13596	842_421 to 38_22
Synthetic RNA	450	4 to 302	590	172_86 to 2_1
Signal Recognition Particle RNA	394	12 to 533	4184	600_298 to 2_2
23S Ribosomal RNA	205	18 to 4381	3536	5872_2958 to 2_2
5S Ribosomal RNA	161	24 to 135	923	244_121 to 2_5
Group I Intron	152	14 to 2630	2236	1528_263 to 2_1
Hammerhead Ribozyme	146	40 to 119	451	236_116 to 2_5
Other Ribosomal RNA	64	12 to 500	196	232_108 to 2_2
Other Ribozyme	53	17 to 968	147	284_142 to 2_2
Group II Intron	42	27 to 2729	99	5040_2520 to 4_8
Cis-regulatory element	41	65 to 102	124	202_100 to 24_16

Their database contain majorly two types of graph called Tree Graph and Dual Graph. RNA structures having pseudoknots can not be represented by the simple tree graph which is why they have two major classification as tree graph and dual graph. In tree graph, they have classified motifs into nine sub category which is based on the number of vertex in the tree graph. Among these nine, smallest class is the one which has two vertex and there is only one possible topology for that configuration. On the other end, they have listed sub class having tree graphs with 10 vertex. There are 106 different possible graphs having 10 vertex out of which only 10 have been found so far. Whereas for dual graph they have sub category ranging from 2 vertex to 9 vertex. Here, there are 38595 different possible motifs for 9 vertex, out of which only 4 have been discovered till today. Now, if we compare this with our permutation based motifs, there are 2619 different structures having length 9. Currently, we do not have data or analysis to show how many of them have already been discovered. In cluster based classification, we have found 13596 different clusters for the "transfer messenger RNA" type of RNA STRAND [44] which is like a sub classification of biologically classified type of RNA which is not possible to do in RAG.

V. FUTURE WORK

There are lot of possibilities one can think of related to the work that we have done. We will list down some of the idea that can be worked upon to extend our current work. Since, we are able to get unique representation in the form of permutation, one can explore many prospects of permutation. For example, one can define an operation on a permutation which will transform permutation into newer permutation and that newer permutation will uniquely represent a secondary structure if indicator sequences are known. One can experiment and obtain a heuristic results for selection of indicator sequence and its value to feed in after doing transformation on permutation. After all, two RNA secondary structure of related species are nothing but the transformation of each other which

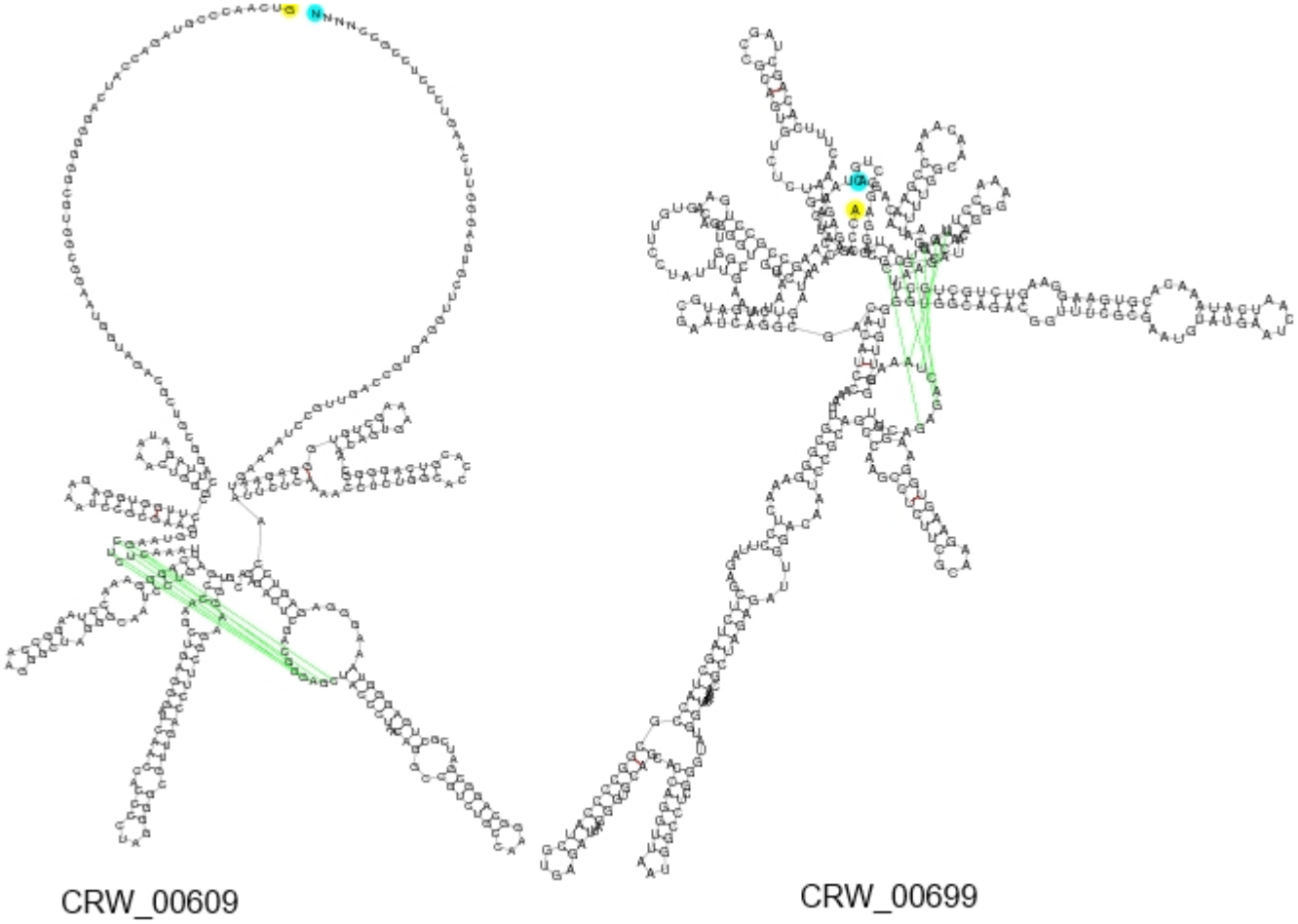

Fig. 5. Secondary Structures of CRW_00609 & CRW_00699. They belong to a cluster having permutation and strand similarity score of 352 and 98 respectively.

is driven by some unknown nature's law. This was mathematics based idea but one can also work on the software front of the current work.

A software based tool can be developed which can take any unknown (one dimensional sequence is known) RNA sequence (Let us call it S) and predict its secondary structure based on a cluster information. First, we can find similarity score of given sequence with entire database. Suppose there are n entries in database of RNA secondary structure. We will find n different similarity score based on this analysis. We choose the highest similarity score. Suppose sequence S' is the one which was used for comparison to obtain highest strand similarity score. Now, we will search for the clusters which contain the same strand similarity score. In each matching cluster, we will search for the sequence S', if found we qualify that cluster otherwise discard it. Finally, we will select the one or more structures from qualified clusters which are nearest in the size with given unknown sequence S. Currently, this is just a concept, we need to verify output of these procedure with the output of the prediction software like mFold [20] for the same unknown sequence input.

Another software can be developed which takes any two sequences with known secondary structure and find six open reading frame portions. Cut all the ORFs and create a data set which can be used to do the exact same analysis that we have done for data sets of [44]. Similar work can also be done for two unknown sequence. Currently, our JAVA program implementation of two proposed algorithm does not show portions or index values of matching portions. A proper web based database can also be implemented which will classify all the data based on our permutation based approach and clusters will be put online for free public access. It can even be integrated for any existing databases for sub classification or it can be simply added to [44] which has accumulated all different types of secondary structure databases.

VI. Conclusion

We have got some very interesting results in the form of clusters. For example, there are total 152 Group I Intron secondary structure known and they are so tightly coupled that they have generated 2236 clusters. Group II Intron class of RNA has total count of just 42 known motifs but their size range from 27 to 2700 and its permutation similarity score varies from 4 to 5000. We have also been able to derive a formula that can count all possible secondary structure in terms of permutation. That can also be used to do the detailed

analysis along with cluster data available for different sub classes of RNA.

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*, vol. 434, no. 7032, pp. 505–509, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1038/nature03380

[2] D. M. Lilley, "Structure, folding and mechanisms of ribozymes," *Current opinion in structural biology*, vol. 15, no. 3, pp. 313–323, 2005.

[3] Y. Tomari and P. D. Zamore, "Perspective: machines for RNAi," *Genes and Development*, vol. 19, no. 5, pp. 517–529, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1101/gad.1284105

[4] J. Couzin, "Small RNAs make big splash," *Science*, vol. 298, Dec. 2002.

[5] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," vol. 2, no. 12, pp. 919–29+, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11733745

[6] B. A. Sullenger *et al.*, "Series introduction: Emerging clinical applications of nucleic acids," *Journal of Clinical Investigation*, vol. 106, no. 8, pp. 921–922, 2000.

[7] W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, no. 6055, p. 618, Feb. 1986. [Online]. Available: http://dx.doi.org/10.1038/319618a0

[8] J. P. Dworkin, A. Lazcano, and S. L. Miller, "The roads to and from the RNA world," *JOURNAL OF THEORETICAL BIOLOGY*, vol. 222, no. 1, May 2003.

[9] M. W. Powner, B. Gerland, and J. D. Sutherland, "Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions," *Nature*, vol. 459, no. 7244, pp. 239–242, May 2009. [Online]. Available: http://dx.doi.org/10.1038/nature08013

[10] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human Molecular Genetics*, vol. 15, no. suppl 1, pp. R17–R29, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1093/hmg/ddl046

[11] J. D. Watson and F. H. C. Crick, "A structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[12] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, the members of the DOE, and N. planning groups, "New goals for the U.S. human genome project: 1998-2003," *Science*, vol. 282, no. 5389, pp. 682–689, Oct. 1998. [Online]. Available: http://dx.doi.org/10.1126/science.282.5389.682

[13] B. Knudsen and J. Hein, "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." *Bioinformatics*, vol. 15, no. 6, pp. 446–454, Jun. 1999. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/15.6.446

[14] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, "Algorithms for loop matchings," *SIAM Journal on Applied Mathematics*, vol. 35, no. 1, pp. 68–82, Jul. 1978.

[15] R. Nussinov and A. B. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 11, pp. 6309–6313, Nov. 1980. [Online]. Available: http://dx.doi.org/10.1073/pnas.77.11.6309

[16] M. Zuker and D. Sankoff, "RNA secondary structures and their prediction," *Bulletin of Mathematical Biology*, vol. 46, no. 4, pp. 591–621, Jul. 1984. [Online]. Available: http://dx.doi.org/10.1007/bf02459506

[17] M. Zuker, "On finding all suboptimal foldings of an RNA molecule." *Science*, vol. 244, no. 4900, pp. 48–52, Apr. 1989. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/2468181

[18] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." *Journal of molecular biology*, vol. 288, no. 5, pp. 911–940, May 1999. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1999.2700

[19] A. Waugh, P. Gendron, R. Altman, Jw, D. Case, D. Gautheret, Sc, N. Leontis, J. Westbrook, E. Westhof, M. Zuker, and F. Major, "RNAML: a standard syntax for exchanging RNA information," *RNA*, vol. 8, no. 6, pp. 707–717, 2002.

[20] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction." *Nucleic acids research*, vol. 31, no. 13, pp. 3406–3415, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg595

[21] I. L. Hofacker, W. Fontana, P. F. Stadler, S. L. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatsh. Chem.*, vol. 125, pp. 167–188, 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00818163

[22] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3429–3431, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg599

[23] E. Ten Dam, K. Pleij, and D. E. Draper, "Structural and functional aspects of RNA pseudoknots," *Biochemistry*, vol. 31, no. 47, pp. 11 665–11 676, Dec. 1992. [Online]. Available: http://dx.doi.org/10.1021/bi00162a001

[24] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots." *Journal of molecular biology*, vol. 285, no. 5, pp. 2053–2068, Feb. 1999. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1998.2436

[25] T. Akutsu, "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots," *Discrete Appl. Math.*, vol. 104, no. 1-3, pp. 45–62, 2000. [Online]. Available: http://portal.acm.org/citation.cfm?id=351046

[26] R. M. Dirks and N. A. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," *J. Comput. Chem.*, vol. 24, no. 13, pp. 1664–1677, Oct. 2003. [Online]. Available: http://dx.doi.org/10.1002/jcc.10296

[27] R. B. Lyngsø and C. N. S. Pedersen, "Pseudoknots in RNA secondary structures," in *Proceedings of the fourth annual international conference on Computational molecular biology*, ser. RECOMB '00. New York, NY, USA: ACM, 2000, pp. 201–209. [Online]. Available: http://dx.doi.org/10.1145/332306.332551

[28] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori, "Tree adjoining grammars for RNA structure prediction," *Theoretical Computer Science*, vol. 210, no. 2, pp. 277–303, Jan. 1999. [Online]. Available: http://dx.doi.org/10.1016/s0304-3975(98)00090-5

[29] J. Zhao, R. L. Malmberg, and L. Cai, *Rapid ab initio RNA Folding Including Pseudoknots Via Graph Tree Decomposition*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2006, vol. 4175, pp. 262–273.

[30] J. Ruan, G. D. Stormo, and W. Zhang, "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots," *Bioinformatics*, vol. 20, no. 1, pp. 58–66, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg373

[31] J. Ren, B. Rastegari, A. Condon, and H. H. Hoos, "HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots," *RNA*, vol. 11, no. 10, pp. 1494–1504, Oct. 2005. [Online]. Available: http://dx.doi.org/10.1261/rna.7284905

[32] S. Federhen, "The NCBI taxonomy database." *Nucleic acids research*, vol. 40, no. Database issue, pp. D136–D143, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1093/nar/gkr1178

[33] 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*, vol. 63, no. 3, pp. 751–759, Sep. 1992. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/1384741

[34] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes." *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D121–4, 2005. [Online]. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15608160

[35] J. W. Brown, "The ribonuclease p database," *Nucleic Acids Research*, vol. 26, no. 1, pp. 351–352, Jan. 1998. [Online]. Available: http://dx.doi.org/10.1093/nar/26.1.351

[36] S. Steinberg, A. Misch, and M. Sprinzl, "Compilation of tRNA sequences and sequences of tRNA genes." *Nucleic acids research*, vol. 21, no. 13, pp. 3011–3015, Jul. 1993. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/7687348

[37] E. S. Andersen, M. A. Rosenblad, N. Larsen, J. C. Westergaard, J. Burks, I. K. Wower, J. Wower, J. Gorodkin, T. Samuelsson,

and C. Zwieb, “The tmRDB and SRPDB resources,” *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D163–D168, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1093/nar/gkj142

[38] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. D’Souza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Muller, N. Pande, Z. Shang, N. Yu, and R. Gutell, “The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs,” *BMC Bioinformatics*, vol. 3, no. 1, pp. 2+, 2002. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-3-2

[39] J. Westbrook, Z. Feng, L. Chen, H. Yang, and H. M. Berman, “The protein data bank and structural genomics,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 489–491, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg068

[40] F. Batenburg, A. P. Gultyaev, C. W. A. Pleij, J. Ng, and J. Oliehoek, “PseudoBase: a database with RNA pseudoknots,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 201–204, 2000. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102383

[41] M. Taufer, A. Licon, R. Araiza, D. Mireles, F. H. D. van Batenburg, A. P. Gultyaev, and M.-Y. Leung, “PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots,” *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D127–D135, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn806

[42] A. Condon, B. Davy, B. Rastegari, S. Zhao, and F. Tarrant, “Classifying RNA pseudoknotted structures,” *Theoretical Computer Science*, vol. 320, no. 1, pp. 35–50, Jun. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.tcs.2004.03.042

[43] S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman, “Rfam 11.0: 10 years of RNA families,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D226–D232, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1093/nar/gks1005

[44] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon, “RNA STRAND: the RNA secondary structure and statistical analysis database.” *BMC bioinformatics*, vol. 9, no. 1, pp. 340+, Aug. 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-340

[45] E. I. Ramlan and K.-P. Zauner, “An extended dot-bracket-notation for functional nucleic acids,” 2008.

[46] O. Matzura and A. Wennborg, “Rnadraw: an integrated program for rna secondary structure calculation and analysis under 32-bit microsoft windows,” *Computer applications in the biosciences: CABIOS*, vol. 12, no. 3, pp. 247–249, 1996.

[47] P. De Rijk, J. Wuyts, and R. De Wachter, “Rnaviz 2: an improved representation of RNA secondary structure.” *Bioinformatics*, vol. 19, no. 2, pp. 299–300, 2003.

[48] K. Han, Y. Lee, and W. Kim, “PseudoViewer: automatic visualization of RNA pseudoknots.” *Bioinformatics*, vol. 18 Suppl 1, pp. s321–s328, 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12169562

[49] Y. Byun and K. Han, “PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots,” *Bioinformatics*, vol. 25, no. 11, pp. 1435–1437, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp252

[50] R. F. Voss, “Evolution of long-range fractal correlations and 1/f noise in DNA base sequences,” *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, Jun. 1992. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.68.3805

[51] H. H. Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick, “Rag: Rna-as-graphs databaseâ??concepts, analysis, and features,” *Bioinformatics*, vol. 20, no. 8, pp. 1285–91, 2004. [Online]. Available: http://bioinformatics.oupjournals.org/cgi/content/abstract/20/8/1285

[52] J. E. Andersen, R. C. Penner, C. M. Reidys, and M. S. Waterman, “Topological classification and enumeration of RNA structures by genus,” *Journal of Mathematical Biology*, pp. 1–18, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1007/s00285-012-0594-x