Using Simulated Annealing for RNA 3D Structural Comparisons

Ryan R. Rahrig[1]
[1]Ohio Northern University, Ada, OH 45868

Abstract
In the study of RNA, an important task is finding the structural similarities and differences of two molecules. As the number of 3D structures available has been increasing dramatically in the past decade, it has become important and necessary for automated methods of RNA structural comparison to be developed. Some RNA molecules, such as ribosomal RNA, consist of hundreds or even thousands of nucleotides, making such a task computationally difficult. A method of RNA 3D alignment using a simulated annealing based approach has been developed and is presented. Resulting alignments are compared with those produced by other programs that are currently widely used.

Key Words: simulated annealing, RNA, alignment

1. Introduction

Comparing RNA molecules can yield important information regarding their structures, functions, and evolutionary histories. Thus, tools and methods capable of quickly and accurately comparing RNA molecules are in demand by those in the field of molecular biology.

1.1 RNA Structure

RNA is a single-stranded molecule consisting of basic units called nucleotides. Each nucleotide is made up of three components: 1) a phosphate, 2) a sugar, and 3) a base (there are 4 types of bases, which are represented simply by the letters A, C, G, and U.

1.1.1 Primary Sequence

Since it is single-stranded there is a distinct beginning and end. Therefore, the bases can be ordered and listed in a unique way. The sequence of bases for an RNA molecule is known as the *primary sequence*. For example, the primary sequence of the 5S rRNA molecule *Haloarcula Marismortui* (*H.m.*), which consists of 122 bases, is given by:

UUAGGCGGCCACAGCGGUGGGGUUGCCUCCCGUACCCAUCCCGAACACGGAAGAUAAGCCCACCAGCGUUCCGGGGAGUA
CUGGAGUGCGCGAGCCUCUGGGAAACCCGGUUCGCCGCCACC

1.1.2 Secondary Structure

RNA nucleotide sequences fold and pair with themselves to form what is known as the RNA *secondary structure*. The folding occurs in such a way that Watson-Crick complementary regions are paired together. Helical structural patterns are formed within these regions. Not all nucleotides are involved in Watson-Crick pairs. Other nucleotides

can be found in non-helical regions, which are represented in secondary structure diagrams as *loops*. There are three types of loops: 1) hairpin loops at the ends of helices, 2) internal loops between two helices, and 3) multi-helix junction loops linking three or more helices. Although not illustrated as such in the secondary structure diagram, these loops are often highly structured by the formation of several types of non-Watson-Crick basepairs, (Stombaugh, J., Zirbel, C.L., et al. 2009). The secondary structure of the *H.m.* molecule whose sequence was given above is illustrated in Figure 1.

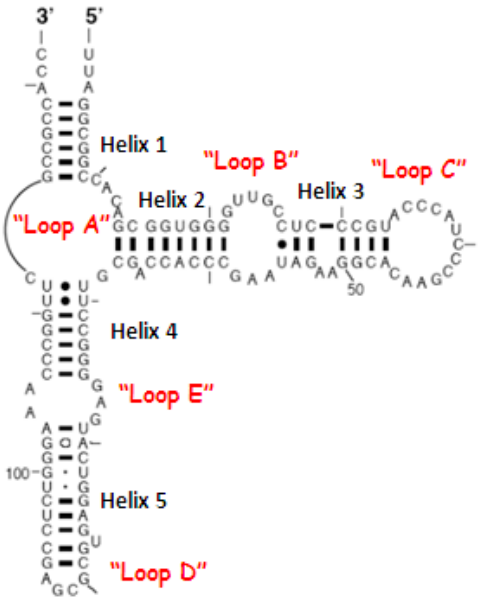

Figure 1: Secondary structure diagram of 5S rRNA *H.m.* G-C and A-U basepairs are common within helices. Loop A is an example of multi-helix junction loop, Loops B and E are examples of internal loops and Loops C and D are examples of hairpin loops.

1.1.3 Tertiary Structure

Finally, the elements of the secondary structure adopt a full three-dimensional structure in which distant elements of the secondary structure make "tertiary" contacts and fold up into the full 3D structure of the molecule, called the *tertiary structure*. These interactions create a tightly compacted and complex 3D structure. The tertiary structure (3D structure) for *H.m.* is given in Figure 2.

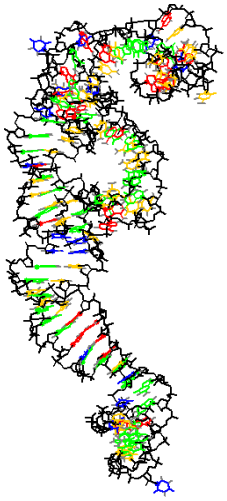

Figure 2: Three-dimensional structure diagram of 5S rRNA *H.m.*

1.1.3 Variability of RNA Structures and Sequences

To function properly, structured RNA molecules must fold into the correct 3D structure. As a result, the 3D structures are highly conserved over many, many generations. Secondary structures are also typically conserved, although less so than 3D structures. The primary nucleotide sequence is the most variable. Isosteric basepair substitutions can take place without affecting the overall 3D structure. While the case of covariation between CG, GC, AU, UA Watson-Crick basepairs is the most well known, all 12 basepair families have mutually isosteric subgroups, as illustrated and explained by Leontis (2002). Because basepairs can substitute without affecting the secondary or tertiary structure, RNA can accommodate high sequence variation with low structural variation. Consequently, methods that are able to identify structural similarities among RNA tertiary structures can produce a wealth of information regarding their functional properties that could not be found by analyzing sequence data alone.

1.2 RNA Alignment

An alignment of two RNA structures is a list of correspondences between individual nucleotides from the two RNA molecules. In order for an alignment to be valid, the *well-ordered* and *uniquely assigned* properties must both be satisfied. An alignment is considered to be *well-ordered* if nucleotides i and j aligned with nucleotides i' and j', respectively implies that $i < j$ if and only if $i' < j'$. If no nucleotide is aligned with more than one nucleotide from the other RNA, then the *uniquely assigned* property is satisfied.

It is common for alignments to be produced using only the information contained in the primary sequence (these are known as sequence alignments). For many years information regarding the sequences was the only information about an RNA structure that was known. Still today, because it is cheaper and faster to obtain sequential information than three-dimensional structural information, only the primary sequence is known for many RNA molecules.

Given two sequences, the objective of sequence alignment is to determine the alignment which produces an optimal score for a given scoring matrix, where identical or similar characters give a positive contribution and unaligned and dissimilar characters give a

negative contribution. The algorithm to find the optimal sequence alignment is known as the Needleman-Wunsch Algorithm (Needleman, 1970).

As described above, the structures (secondary and tertiary) of RNA molecules are more conserved over time than RNA sequences. Therefore, alignment methods that take structural information into account are typically more accurate than sequence alignment methods. More accurate alignments lead to more reliable and useful information regarding the functional properties of the molecule. However, RNA structures are quite complex, which makes structural alignment a more computationally challenging problem.

1.3 Review of 3D Structural Alignment Methods

The highest quality RNA 3D structural alignments are constructed by hand by experts. However, producing hand-crafted alignments, such as the alignment of 23S rRNAs by Stombaugh et. al (2009) is a labor-intensive and time-consuming process. Also, the number of RNA molecules whose 3D information has been made available has greatly increased in the past decade, making it essential to develop automatic tools that are capable of accurately discovering structural similarities among homologous RNA molecules.

Rahrig (2010) describes an automated method for RNA 3D structural alignment, implemented as the R3D Align program. Local alignments are merged to form a global alignment by employing a maximum clique algorithm on a specially defined graph. That paper also summarizes other alignment methods that are currently used. These include DIAL, described in Ferre et. al (2007); SARA, described in Capriotti and Marti-Renom (2008); SARSA, described in Chang et al. (2008); ARTS, described in Dror et al. (2005).

In this paper, a new methodology for the alignment of two RNA 3D structures is introduced. The method will be referred to as SA Align, which stands for 'Simulated Annealing Alignment'. Like R3D Align, a notable feature of SA Align is that it is able to accommodate the flexibility that exists among RNA 3D structures. Details on how this is done can be found in the Method section.

SA Alignment employs the simulated annealing method, which is an adaptation of the Metropolis-Hastings algorithm, described by N. Metropolis et al. (1953). The simulated annealing method is useful for locating a suitable *approximation* to the global optimum of a specified cost function when the search space is large (e.g., the space of all possible alignments).

As yet there are no generally agreed upon methods for evaluating RNA or protein 3D structural similarity, as there are for sequence similarity. This means that there typically is not one specific alignment that is considered be the "best" alignment. Any structural alignment produced by an automated program is typically subjected to a bit of post-hoc analysis to determine any manual adjustments that need to be made. For these reasons, although simulated annealing methods may only find a close approximation to the global optimum, this will be sufficient for the application described here. Also, the alignments produced by SA Align could be used as seed alignments for the R3D Align program in certain cases.

2. Method

2.1 Decompose RNA Molecules into Local Neighborhoods

The first step is to identify the local neighborhoods in each structure. Each local neighborhood will consist of four nucleotides. These smaller neighborhoods can be compared more easily than comparing the entire structures. So the final alignment will eventually be constructed using the information gathered when comparing local neighborhoods. The method used to construct the set of local neighborhoods is similar to that described by Rahrig (2010). The major ideas are described next.

Suppose that structure A is made up of nucleotides 1^A, 2^A,..., n^A, in 5' to 3' order. A set N^A is constructed in such a way that each nucleotide is a member of at least p neighborhoods. The value of p may be set by user, although 10 has been found to work well and is the default value. For each nucleotide i in A, the p neighborhoods of smallest diameter that include nucleotide i are found and added to the set N^A. The diameter of a neighborhood is the maximum distance between the geometric centers of the bases as described by Sarver, et. al. (2008). Four-nucleotide neighborhoods of small diameter will contain most pairwise interactions, including base-pairing, base-stacking, and base-backbone interactions. With respect to the sequence or secondary structure, both local and long-range interactions are included. The nucleotides of each neighborhood in N^A are listed in ascending order. The nucleotides of structure B are also decomposed into neighborhoods and N^B is then formed in a similar way as N^A.

2.2 Compare Local Neighborhoods

The neighborhoods of N^A and N^B are compared to determine which neighborhoods are structurally similar. The geometric discrepancy developed and described in Sarver, et. al (2008) serves as a useful quantitative measure of structural similarity. The geometric discrepancy is measured in Angstroms and is used to compare neighborhoods of the same size. Lower discrepancy values correspond to more structural similarity.

A cut-off discrepancy value d is used to determine the maximum discrepancy between two neighborhoods that are considered to be *structurally similar*. The parameter d may be set by the user, but .5 is usually considered sufficient and is the default value. If two neighborhoods have a discrepancy below d, then the neighborhood-pair is added to set L. Notice that each element of L is essentially an alignment of 4 pairs of nucleotides. The next step of the algorithm is to combine these small ("local") alignments in such a way to maximize a specified objective function. A simulated annealing method is employed for this purpose.

2.3 Combining Local Alignments with Simulated Annealing

The set S of possible solutions (i.e., possible alignments) is the set of all alignments that can be formed by merging various elements of set L. The simulated annealing algorithm consists of a discrete-time Markov chain, $x(t)$. The initial state, $x(0)$, is the alignment consisting of no aligned nucleotides.

2.3.1 Forming Proposal Solutions

Each step of the algorithm attempts to replace the current solution by a randomly proposed solution. The proposed solution is constructed from solutions near the current solution. Let current state of Markov Chain, $x(t)$, be denoted by i. A proposal solution j is formed by randomly selecting a local alignment and incorporating the nucleotides into

the current solution. It is important recognize that incorporating a local alignment does not necessarily just add 4 more aligned nucleotides to the current alignment. This is because some of the correspondences to add may conflict with current correspondences. So once the 4 new correspondences are added, previously aligned nucleotides may have to be removed in order to create a valid (well-ordered and uniquely-assigned) alignment. This may result in a proposed alignment (solution) that actually aligns *fewer* nucleotides than the current alignment. This concept is illustrated in Figure 3.

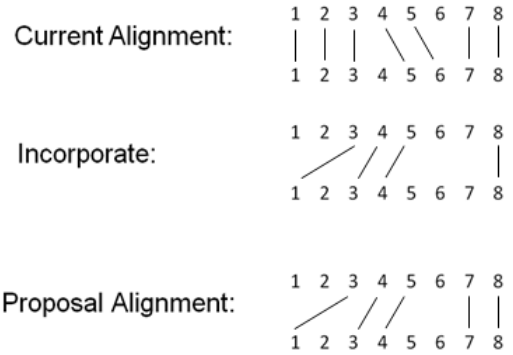

Figure 3: The current solution is the top alignment, which has 7 aligned nucleotides. The proposal solution is formed by incorporating the randomly selected local alignment (displayed in the middle) into the current solution. The resulting proposal alignment actually contains fewer aligned nucleotides (5).

2.3.2 Objective and Temperature Functions

Once the proposal solution *j* is constructed, it is accepted with a probability that depends on two things: 1) the change in the value of objective function from the current solution to the proposal solution; 2) the value of the global parameter *T* (called the *temperature*).

The objective function to be used is the following:

$$C = (\#\text{ nucleotides aligned}) + (\#\text{ basepairs aligned})$$

The rationale for this objective function is now explained. Since the proposal solutions are being constructed from the set of local alignments of structurally similar neighborhoods, we are ensured that combining these local alignments will still result in an alignment of structurally similar sets of nucleotides. Recall the goal of structural alignment which is to align as many nucleotides as possible for which the corresponding nucleotides are structurally similar. Also, if two nucleotides are forming a basepair in structure A and are aligned to two nucleotides in structure B, it should be the case that the two nucleotides in structure B are also forming a basepair if there really is structural similarity.

The function $T:N\rightarrow[0,\infty)$ is a nonincreasing function. The algorithm starts with a high value of *T* and is decreased according to an *annealing schedule*. As the value of T decreases, the proposal solution is less likely to be selected when it decreases the value of the objective function. This allows a broad region of solutions to be explored in the beginning, while moving toward regions that become more optimal as time increases (as the temperature decreases).

2.3.3 Acceptance Probabilities

If $C(j) \geq C(i)$, then $x(t+1) = j$ with probability 1. Thus, all "uphill" proposal solutions are accepted. However, some downhill proposal solutions will also be selected. This feature of simulated annealing helps avoid getting stuck in local optima. If $C(j) < C(i)$, then $x(t+1) = j$ with probability $e^{\frac{-(C(i)-C(j))}{T}}$.

Notice the probability of making such a downhill move decreases with time and as the difference between $C(i)$ and $C(j)$ increases.

2.3.4 Concluding the Algorithm

The simulated annealing algorithm may continue until any of the following occur:

1) A suitable solution is found.
2) $T=0$
3) A specified number of iterations without significant improvement pass.

What is considered to be a suitable solution often depends on the two structures being aligned, so typically criteria 2 and 3 will be used primarily for the conclusion of the algorithm.

3. Results

SA Align was implemented in the Matlab programming language and was tested by aligning the same 3D structures used to test R3D Align so that a comparative analysis could be performed. 3D structures of 5S, 16S, and 23S rRNAs of *E. Coli* (*E.c.*) and *Thermus thermophilus* (*T.th.*) were aligned. Also, 3D structures of 5S and 23S rRNAs of E.*c.* and *Haloarcula marismortui* (*H.m.*) were aligned. PDB file 2aw4, described in Schuwirth, et. al (2005), was used for 5S and 23S *E.c.* PDB file 2j01, described in Selmer, et. al (2006), was used for 5S and 23S *T.th.* PDB file 1s72, described in Klein, et. al (2004), was used for 5S and 23S *H.m.*

In this section we focus on the alignment of the 16S rRNA molecules of *E.c.* and *T.th.*, since this was the alignment most fully described in the R3D Align paper. PDB file 2avy, described in Schuwirth, et. al (2005), was used for 16S *E.c.* and PDB file 1j5e, described in Wimberly, et. al (2000), was used for 16S *T.th.* Bar diagrams as introduced by Rahrig (2010) will be used to the display the alignments.

The bar diagrams in Figure 4 display the alignments produced by SA Align and several other methods. As in Figure 3, straight line segments connect corresponding nucleotides in the two structures. For each aligned nucleotide in *T.th.*, the nearest four nucleotides in *T.th.* which were aligned to nucleotides in *E.c.*, are found and superimposed onto the corresponding five nucleotides in *E.c.* The geometric discrepancy between the two five-nucleotide neighborhoods is then found. The color of the line indicates the value of the geometric discrepancy. The color bar indicates that the colors change from blue to red as the discrepancy increases.

The bar diagram shows that SA Align produces a very accurate alignment since there are many correspondences and most of the lines are colored dark blue. The comparison of SA Align and R3D Align will be focused on since the comparison of the other methods was previously discussed in the R3D Align paper.

SA Align performs nearly as well as R3D Align in many regions but fails to align nucleotides in the 1400-1500 range. However, as can be seen from the correspondences for this region indicated by other alignment methods, there is a great deal of structural dissimilarity in this region. Indeed, none of the methods produce an alignment of nucleotides in this region represented by many dark blue lines. However, if one argued that the amount of dissimilarity was acceptable, a larger value than 0.5 for the geometric discrepancy threshold parameter can be used for SA Align. This would result in more nucleotides being aligned.

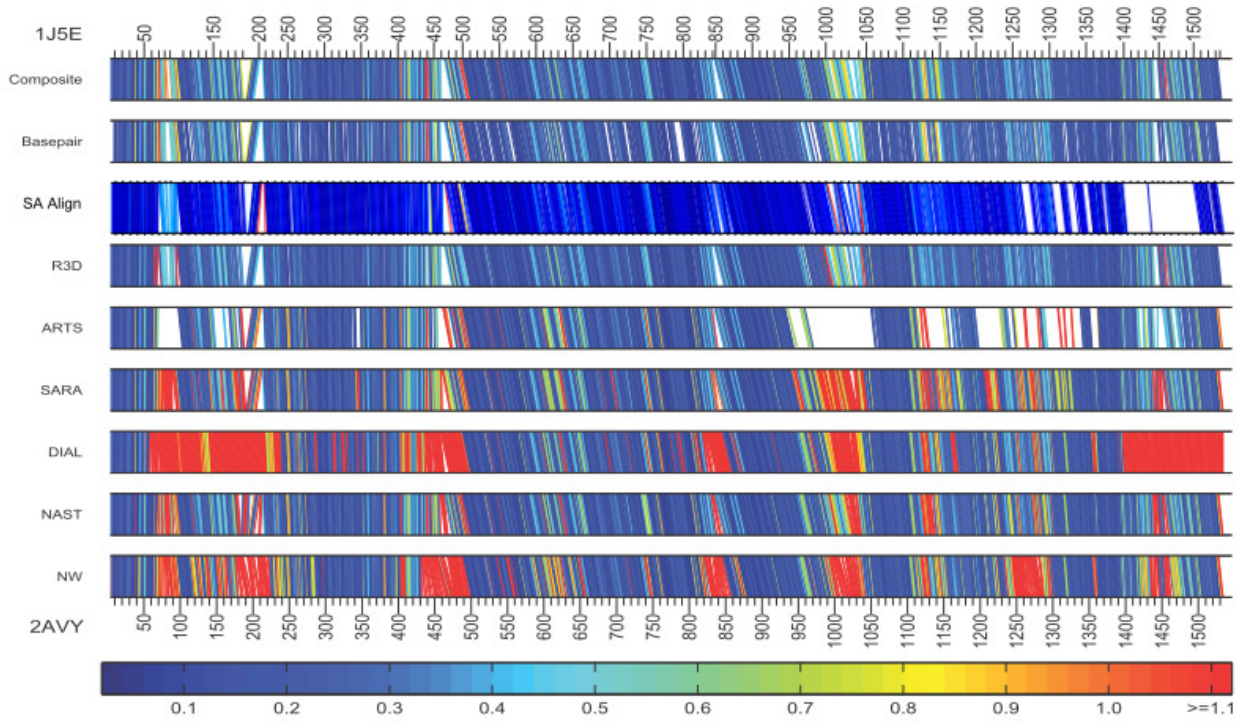

Figure 4. Bar diagrams illustrating the alignments produced by several different methods of *T.th.* (1J5E) and *E.c.* (2AVY). The nucleotides of 1J5E are listed along the top of each bar and nucleotides of 2AVY are listed along the bottom. The lines are colored according to the geometric discrepancy between the aligned nucleotide and the set of nearest four aligned nucleotides.

4. Conclusion

A new method for structural alignment of two RNA 3D molecules has been introduced and implemented as the "SA Align" program suite in Matlab. Like R3D Align, it can be used to accurately align large RNA structures. While not as accurate as R3D Align, SA Align has a shorter running time. For the 16S rRNA alignment discussed in Section 3, SA Align aligns the structures in 7 minutes while R3D Align takes 13 minutes. Because R3D Align is more accurate and is capable of inputting a seed alignment (which decrease the overall execution time), a useful application of SA Align may be to quickly produce an alignment that is then fed into R3D Align for refinement and improvement.

Acknowledgements

Thanks goes to Craig Zirbel and Neocles Leontis for their work and help in developing R3D Align. The research and programs developed for R3D Align provided a starting point for this project.

References

Capriotti,E. and Marti-Renom,M.A. (2008) SARA: RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24, i112-8.

Chang,Y.F., Huang,Y.L., et al. (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res.*, 36, W19-24.

Dror,O., Nussinov,R., et al. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21 Suppl 2, ii47-53.

Ferre,F., Ponty,Y., et al. (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res.*, 35, W659-68.

Klein D.J., Moore P.B., et al. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol.*, 340, 141-177.

Leontis,N.B., Stombaugh,J., et al. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, 30, 3497-3531.

Metropolis,N., Rosenbluth, A.W., Rosenbluth, M.N., et al. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys., 21, 1087-1092.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J.Mol.Biol.*, 48, 443-453.

Rahrig,R., Zirbel, C.L., Leontis,N.B., (2010) R3D Align: Global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics,* 26, 2689-2697.

Sarver,M., Zirbel,C.L., et al. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J.Math.Biol.*, 56, 215-252.Schuwirth,B.S., et al. (2005) Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science*, 310, 827-834.

Schuwirth,B.S., et al. (2005) Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science*, 310, 827-834.

Selmer,M., Dunham,C.M., et al. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, 313, 1935-1942.

Stombaugh,J., Zirbel,C.L., et al. (2009) Frequency and isostericity of RNA base pairs. *Nucl.Acids Res.*, 37, 2294-2312.

Wimberly,B.T., *et al.* (2000) Structure of the 30S ribosomal subunit. *Nature*, 407, 327-339.