Published online 12 April 2013

Nucleic Acids Research, 2013, Vol. 41, No. 11 ***e114***
doi:10.1093/nar/gkt187

RNA global alignment in the joint sequence–structure space using elastic shape analysis

Jose Laborde[1], Daniel Robinson[2], Anuj Srivastava[1], Eric Klassen[2] and Jinfeng Zhang[1,*]

[1]Department of Statistics, Florida State University, FL, USA and [2]Department of Mathematics, Florida State University, FL, USA

Received August 9, 2012; Revised February 26, 2013; Accepted February 27, 2013

ABSTRACT

The functions of RNAs, like proteins, are determined by their structures, which, in turn, are determined by their sequences. Comparison/alignment of RNA molecules provides an effective means to predict their functions and understand their evolutionary relationships. For RNA sequence alignment, most methods developed for protein and DNA sequence alignment can be directly applied. RNA 3-dimensional structure alignment, on the other hand, tends to be more difficult than protein structure alignment due to the lack of regular secondary structures as observed in proteins. Most of the existing RNA 3D structure alignment methods use only the backbone geometry and ignore the sequence information. Using both the sequence and backbone geometry information in RNA alignment may not only produce more accurate classification, but also deepen our understanding of the sequence–structure–function relationship of RNA molecules. In this study, we developed a new RNA alignment method based on elastic shape analysis (ESA). ESA treats RNA structures as three dimensional curves with sequence information encoded on additional dimensions so that the alignment can be performed in the joint sequence–structure space. The similarity between two RNA molecules is quantified by a formal distance, geodesic distance. Based on ESA, a rigorous mathematical framework can be built for RNA structure comparison. Means and covariances of full structures can be defined and computed, and probability distributions on spaces of such structures can be constructed for a group of RNAs. Our method was further applied to predict functions of RNA molecules and showed superior performance compared with previous methods when tested on benchmark datasets. The programs are available at http://stat.fsu.edu/~jinfeng/ESA.html.

INTRODUCTION

Recent discoveries have shown that RNA molecules play important roles in many biological processes such as enzymatic activity, protein synthesis and transport, gene transcriptional regulation, RNA processing and splicing and chromosome replication (1–5). This changed the traditional view of RNA being solely a carrier of genetic information (1,6,7). Comparison of RNAs, including both sequence alignment and structure alignment, can reveal the conserved motifs important for RNA functions, the evolutionary relationships of RNAs and the sequence–structure–function relationship of RNAs in general (6).

Compared with protein alignments, the comparison/alignment of RNAs is much less studied (8–18). Although the alignment of RNA sequences can borrow directly those methods developed for protein or DNA sequence alignment, many methods designed for protein structure alignment cannot be readily used for alignment of RNA structures. This is partially due to the difference between the secondary structures of proteins and RNAs.

In the rest of the Introduction, current methods for RNA structure alignment will be briefly reviewed, followed by a conceptual description of our own method. Since local structural motifs of RNAs often have specific functions, instead of comparing the overall structures, some studies focus on detection of local structural motifs [such as NASSAM (17), COMPADRES (16), RNAMotifScan (19) and FR3D (15)]. This is analogous to the identification of functional domains in proteins. Methods that compare 3D RNA structures at scales larger than motifs can be divided largely into two types. The first type of methods represents nucleotide residues by some local structure features, which allow them to reduce the 3D structures to 1D sequences. The resulting 1D sequences can then be aligned using traditional sequence alignment methods. The second type of methods start

*To whom correspondence should be addressed. Tel: +1 850 644 3218; Fax: +1 850 644 5271; Email: jinfeng@stat.fsu.edu
Correspondence may also be addressed to Anuj Srivastava. Tel: +1 850 644 3218; Fax: +1 850 644 5271; Email: anuj@stat.fsu.edu

from alignment of similar local structures and then aims to obtain larger scale alignment by extending the initial local alignment. Among the first type of methods, SARA uses a set of unit vectors derived from consecutive nucleotides to represent each nucleotide. The difference of nucleotides can be compared via the unit vectors using unit-vector root mean square (URMS) as distance (8,9); iPARTS uses structural alphabet (SA) derived from backbone torsion angles, which are discretized into 23 states. A substitution matrix is then derived for the alphabet of 23 letters and used in the sequence alignment (10). LaJolla uses a n-gram model for analysing the sequences derived from the torsion angles of nucleotides (12). Similarly, PRIMOS/AMIGOS (13) and DIAL (14) also use torsion angles to represent nucleotides and align RNA on a sequence space encoded by the torsion angle representation. The above approaches do not necessarily produce globally similar alignment between two RNA structures [i.e. with small RMSDs (root mean square deviations)]. To achieve smaller RMSD for the aligned parts, extra steps need to be taken after the sequence alignment.

Among the second type of RNA structure alignment methods, ARTS used P (phosphor) atoms of two consecutive base pairs as seeds and aligned the overall structures based on the alignment of structurally similar seed quadrants between the two RNA molecules (11). In R3D Align, local alignments are merged to form a global alignment by using a maximum clique algorithm on a specially defined graph called a local alignment graph (18).

Sequence alignment uses information of side chains, which are reduced to single letters, and structure alignment uses information of the backbone geometry. Both methods have their advantages and disadvantages. Sequence alignment can be used to compare any two or multiple RNA molecules with sequence information, which is readily available. But such methods often perform poorly for remotely related sequences, and will not work for related structures without detectable sequence similarities. Structure alignment, on the other hand, fills in that gap by using structure information directly. Structure alignment also allows for detection of common functional motifs/domains in a set of structures. However, structure information is far more difficult to obtain than sequences. Even though the nucleotide sequences are available, most structure alignment methods do not use this information. A valid question is, can sequences provide additional useful information for structure alignment? Combining both structure and sequence information in RNA comparison may also provide more insight on the sequence–structure–function relationships of these molecules.

In this study, we address this problem using a novel approach, called elastic shape analysis (ESA). Recently, ESA has been successfully applied to protein structure comparison (20,21). ESA treats protein/RNA structures as three dimensional curves and uses a geometric framework that has been developed originally in image analysis and computer vision for shape analysis of parameterized curves and surfaces (22–25). The basic idea in this framework is to design an infinite-dimensional manifold of curves, endow it with a metric structure and compare any two objects by computing the distances between them on this manifold. Under this framework, we will be able to (i) quantify the similarity between any two protein/RNA structures by a formal distance, geodesic distance, computed on their respective shape manifolds. These geodesics can be seen as optimal deformations of objects into each other; (ii) compute intrinsic statistics of the shapes of a collection of protein/RNA conformations. For instance, one can compute their means and variances to study the statistical variability of shapes within the same group; and (iii) use the moments computed above to impose a full probability model on a shape group using a wrapped Gaussian density. This type of probability density can be used for statistical analysis and hypothesis testing of future structures. A nice property of ESA is that additional information can be readily added and the resulting distance can still be a formal one. In this study, we apply ESA to RNA structure alignment and extend the original framework to compare RNA molecules in the joint sequence-structure space.

In the following sections, we will first present the ESA method for RNA structure and sequence comparison. We then apply the method on benchmark datasets for RNA function prediction and compare its performance with that of previous methods. Finally, we provide a discussion and our conclusions.

METHODOLOGY

Elastic shape analysis in the joint sequence–structure space

A mathematical and statistical framework based on ESA has been recently developed for protein and RNA structure alignment where only the backbone geometric information was used in distance calculations (20, 21, 26). In this study, we extend the framework by incorporating additional dimensions containing sequence information so that the comparison can be done in the joint sequence–structure space. For the sake of completeness, we describe the key points of the original framework while focusing more on the new development.

For RNA structure comparison, ESA treats the backbone structures of RNA molecules as parameterized curves in R^3. Since the comparisons involve shapes, the resulting quantifications should not depend on the rigid motions and parameterizations of these curves. When incorporating additional information, such as sequences, we add extra coordinates resulting in curves in higher dimensions. Since the structure dimensions and sequence dimensions represent different types of information, they are treated differently at certain steps of the ESA procedure. We represent each parameterized curve with a special function called the square root velocity function (SRVF) and restrict to the manifold of such functions under the desired constraints. To compare shapes of curves, we remove all shape-preserving transformations from this representation. This is done using an algebraic technique—we form a quotient space of the original manifold with respect to these shape-preserving transformation groups. In the resulting quotient space, called

the shape space of elastic curves, one can perform statistical analysis of curves as if they are random variables. One can compare, match and deform one curve into another, or compute averages and covariances of curve populations, and perform hypothesis testing and classification of curves according to their shapes.

When incorporating sequence information as additional dimensions, nucleotides need to be converted to numerical values while the distances among them should still be sensible. A distance matrix or substitution matrix need to be used to derive the conversion rule. Here we use the Jukes and Cantor substitution matrix (27), a commonly used and also the simplest substitution model (JC model). JC model assumes equal base frequency and equal mutation rates, suggesting that the four nucleotides must have equal distance from one another. Consequently, we use a regular tetrahedron to represent the four nucleotides, where the numerical values of the four nucleotides are set to be the coordinates of the four vertices of the tetrahedron. Specifically, we have $A = (1, 1, 1)$, $C = (1, -1, -1)$, $U = (-1, 1, -1)$ and $G = (-1, -1, 1)$. The distance between any pair of these points is $2\sqrt{2}$. Note that it is impossible to embed four equidistant points in two dimensions and doing this in four dimensions is unnecessary. It is also possible to adopt other substitution matrices in a similar fashion.

Now suppose we have a sequence of n P (phosphor) atoms with coordinates $(x_1, y_1, z_1), \ldots, (x_i, y_i, z_i), \ldots, (x_n, y_n, z_n)$, taken from the PDB (28, 29) file of an RNA molecule. The corresponding sequence coordinates according to the representation above are (u_i, v_i, w_i). The RNA molecule is represented as a curve in R^6 going through the points

$$P_i := \begin{bmatrix} x_i & y_i & z_i & \lambda u_i & \lambda v_i & \lambda w_i \end{bmatrix}^T \in R^6, i = 1, 2, \ldots, n.$$

where λ is a weight that controls the contribution of the sequence information in the alignment. The greater the value of λ, the more influential the sequence information will be, and for a very large λ, only sequence information will be relevant. Conversely, if $\lambda = 0$, the alignment is done only in the structure space.

Using the representation above, given an RNA structure, we construct a continuous parameterized curve that interpolates the points $\{P_i\}_{i=1}^n$, which will map the interval [0,1] to R^6. We will denote this parameterizing variable as t. We assign the $\{t_i\}_{i=1}^n$ values corresponding to $\{P_i\}_{i=1}^n$ as follows:

$$t_1 = 0$$

$$t_{i+1} = t_i + \frac{1}{L}||(x_{i+1}, y_{i+1}, z_{i+1}) - (x_i, y_i, z_i)||,$$

$$\text{for } i = 1, 2, \ldots, n-1,$$

where L is the total length of the backbone. Note that the sequence information is ignored when selecting parameter values, and also that these are not necessarily uniformly spaced. Once parameter values have been assigned, we define our curve $\beta : [0, 1] \to R^6$ to be the piecewise-linear interpolating function mapping $t_i \mapsto P_i$ for $i = 1, 2, \ldots, n$. Since the β functions are absolutely continuous curves in R^6, all of the ESA techniques for R^n, described in (22, 23), can be readily applied to them. Since β is linear on each subinterval (t_{i-1}, t_i), the corresponding SRVF (which we discuss next) will be constant on each of these subintervals.

To analyse the (sequence-annotated) shape of a curve β, we represent β by its square-root velocity function: $q(t) = \dot{\beta}(t)/\sqrt{||\dot{\beta}(t)||}$ in R^6, where $||\cdot||$ is the standard Euclidean norm in R^6, and $\dot{\beta}(t) = \frac{d\beta(t)}{dt}$. In order for the shape analysis to be invariant to scales, we rescale each curve to length 1. This treatment is optional, but gave (surprisingly) better performance for protein structure alignment (20, 21). Restricting to the curves of interest, represented by their SRVFs, we obtain the set

$$C \equiv \{q : [0, 1] \to R^6 | \int_0^1 ||q(t)||^2 dt = 1\}. \quad (1)$$

C is called the pre-shape space and is the set of all SRVFs representing parameterized curves in R^6 of length one. It is actually a unit sphere in the Hilbert space $\mathbb{L}^2$.

Among four shape-preserving transformations, we have removed translation and scale (optional); the rotation and reparameterization are removed algebraically as follows. When analysing these curves in R^6, the sequence information is relevant only for the parameterization step, but not for the rotation step: when optimizing over rotations, we only use information of backbone structures. Since the nucleotide sequence is not considered geometric information, when dealing with rotation, we modify the original algorithm by letting $SO(3)$ be the group of 3×3 rotation matrices and Γ be the group of all re-parameterizations (they are actually positive diffeomorphisms of the interval [0,1]). Let $\bar{\Omega}$ be the embedding of $SO(3)$ in GL_6 (general linear group of 6×6 invertible matrices) by letting all elements $O \in SO(3)$ in GL_6 through $\Omega = \begin{pmatrix} O & 0 \\ 0 & I_3 \end{pmatrix}$. It is easy to show that $\bar{\Omega}$ is a subgroup of GL_6 or $SO(6)$.

For a curve β, an (embedded) rotation $\Omega \in \bar{\Omega}$ and a re-parameterization $\gamma \in \Gamma$, the transformed curve is given by $\Omega(\beta^\circ\gamma)$. The SRVF of the transformed curve is given by $\sqrt{\dot{\gamma}}\Omega(q^\circ\gamma)$. To unify all elements that denote the same shape in C, we define equivalence classes of the type: $[q] = \{\Omega(q^\circ\gamma)\sqrt{\dot{\gamma}} | \Omega \in \bar{\Omega}, \gamma \in \Gamma\}$. Each such class $[q]$ is uniquely associated with a shape and vice versa. The set of all these equivalence classes is called the shape space S. Mathematically, it is a quotient space of the pre-shape space: $S \equiv C/(\bar{\Omega} \times \Gamma) = \{[q] | q \in C\}$.

With the above modifications of the original framework, we can obtain geodesic paths and distances between two RNA structures, represented by 3D curves β_1 and β_2, as done in (20, 21, 26). Details of the mathematical framework can be found in the appendix.

Obtaining a sequence alignment from the elastic matching

Now suppose that $\beta_1, \beta_2 : [0, 1] \to R^6$ are two curves obtained as above. Let $t_1^1, t_2^1, \ldots, t_{n1}^1$ and $t_1^2, t_2^2, \ldots, t_{n2}^2$ be the parameter values assigned to the original sample

points for β_1 and β_2, respectively, and let q_1 and q_2 be the corresponding SRVFs. Our approach to finding an optimal re-parameterization γ^* as done in (20,21,26) is basically the same as the dynamic programming algorithm used in (25), but in our case the nucleotide sequence information gives rise to a few considerations which we briefly discuss here.

In the original ESA framework for RNA/protein alignment (21,26), we added a pre-processing step: any two RNAs/proteins to be compared had to have their 3D coordinates re-sampled using interpolations, and we obtained smooth curves with equal number of points on each structure. That treatment was convenient since it facilitated the use of a uniform grid that evenly partitions $[0,1] \times [0,1]$ to search for the optimal γ^*. In the new framework, each point is mapped to a nucleotide and re-sampling of points is not necessary. With this modification, sequence alignments are obtained together with structure alignments, like most of the other structure alignment methods do. As in the original algorithm, we create a grid on the unit square $[0,1] \times [0,1]$ and search for an optimal path through the grid from (0,0) to (1,1). However, our grid is non-uniform: the vertical lines are placed at x-coordinates $t_1^1, t_2^1, \ldots, t_{n1}^1$ and the horizontal lines are placed at y-coordinates $t_1^2, t_2^2, \ldots, t_{n2}^2$ (see Figure 1). In our case, the gridpoints have special significance: each gridpoint (t_i^1, t_j^2) represents a match between one of the P atoms on the backbone of the molecule represented by β_1 and a P atom on the backbone of the molecule represented by β_2. As in the original algorithm, a list of gridpoints $(t_{i1}^1, t_{j1}^2), (t_{i2}^1, t_{j2}^2), \ldots, (t_{im}^1, t_{jm}^2)$ which starts at (0,0), ends at (1,1), and is strictly increasing in both the x and y components can be thought of as the graph of an increasing piecewise-linear re-parameterization $\gamma : [0,1] \to [0,1]$. In our case, such a list also gives us an alignment between the two nucleotide sequences. Here, parameter values t_j^i which do not appear in the list of gridpoints correspond to alignment gaps (elements in the nucleotide sequence which are unmatched). Dynamic programming is used as in the original algorithm to search over the space of all such lists for one which minimizes the energy in equation A.1, and the result gives us both a full elastic matching between β_1 and β_2, as well as a nucleotide sequence alignment. Figure 2 shows an example of a sequence alignment obtained in this way.

Since paths in the grid are required to start at (0,0) and end at (1,1), the alignments produced by our method always match the first nucleotide in β_1 to the first nucleotide in β_2; similarly, the last nucleotides are always matched together. To work around this restriction, we add a preprocessing step in which a dummy point is added to the beginning and ending of each curve in R^6. The 3D coordinates (x_i, y_i, z_i) of these points are obtained by linear extrapolation, and the sequence coordinates, (u_i, v_i, w_i), are duplicates of the sequence coordinates of the first (or last) real point. For the example in Figure 1, we used one dummy point at the beginning and ending of each curve. The first row and column and the last row and column in the grid correspond to these extra points. In Figure 2, the extra points have been removed for display, along with any matches involving them. The number of added dummy points is arbitrary and quite easy to change, but one or two should be enough to get around the 'endpoints' matching restriction. In this example, we scale the structures to unit length, let the sequence weight in the matching be $\lambda = 5$ and the resulting geodesic distance between both structures is 0.8966.

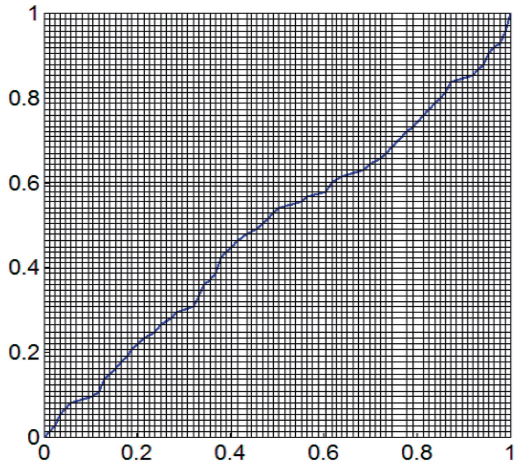

Figure 1. Piecewise-linear re-parameterization γ^* yielding an optimal matching between 1h3e chain B and 1gtr chain B. The matching grid has 82 columns, corresponding to the 80 points on the backbone of 1h3eB plus the dummy points at each end. Similarly, the grid has 76 rows, corresponding to the 74 points on the backbone of 1gtrB plus the dummy points at each end.

The range of ESA distances for any two unscaled-length RNAs is $[0, \infty)$. The geodesic distance of a structure to itself is zero. In many applications of shape analysis, the lengths of objects are often scaled to unit length. In such cases, distances have an upper bound of $\pi/2$.

Statistics of tertiary structures using 3D shape and sequence

With a formal way to measure the distances between RNA structures, we can compute some important statistics for these shapes. Note that we can only do statistics of shapes if there is a true notion of shape distance. In particular, we would like to calculate mean and covariance and even impose probability distributions for a given set of RNA structures and, if desired, their nucleotide sequences, as well.

Computing mean and covariance in non-linear (in this case, spherical) manifolds is not straightforward, as the shapes are not in a vector space. To get around this limitation, we use the linear properties of the tangent space at each point of S. Specifically, let $\beta_1, \beta_2, \ldots, \beta_n$ be a given set of RNA structures, represented by their SRVFs $q_1, q_2, \ldots, q_n$. The sample mean is defined as the Karcher mean, and the covariance structure is obtained using the differential geometry of the q-function space. For mathematical details, see the appendix.

Mean shapes and probability distributions of RNA structure families/classes can be very useful in automatic classifications of new structures. For example, mean shapes can serve as filters to quickly narrow down the

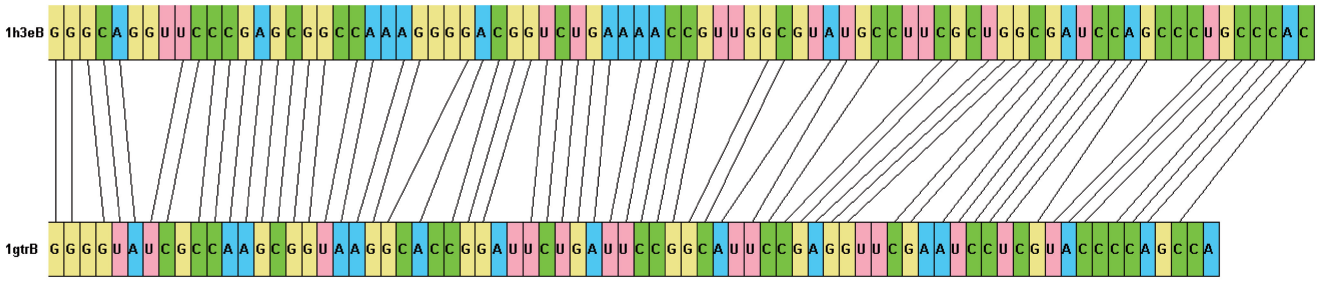

Figure 2. The nucleotide sequence alignment between 1h3eB and 1gtrB obtained from the matching shown above. The black lines represent the matching between nucleotides of each structure which are labelled and colour coded.

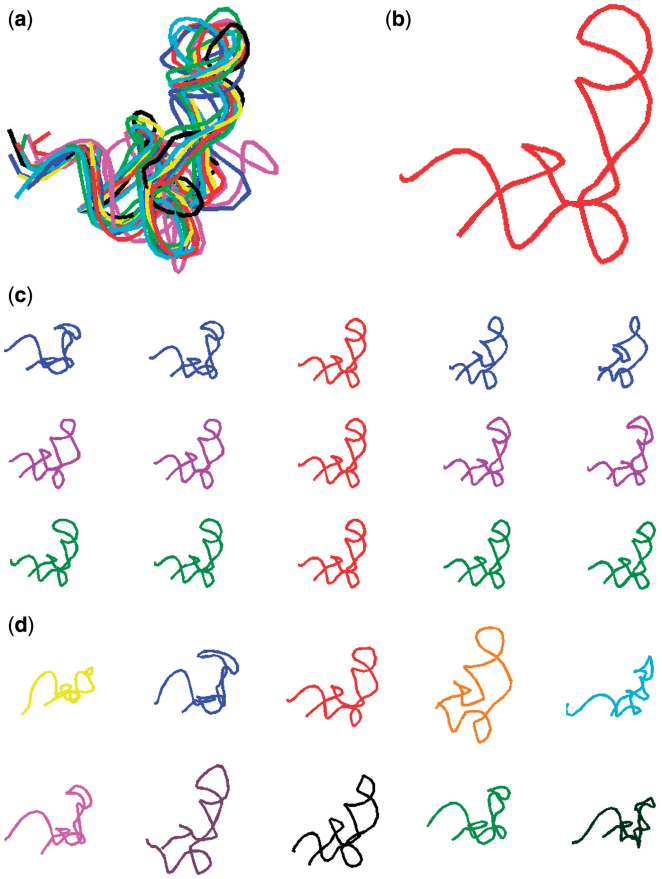

Figure 3. Structures, mean, variability and random Gaussian samples of a group of tRNA synthetase complexes. (**a**) RNA structures in *tRNA synthetase complex* class. (**b**) The mean structure of the structures in (a), which is the sample Karcher mean. (**c**) Samples from the directions of the three main variance components U_1, U_2 and U_3. They represent the amount of variation in the set of structures. (**d**) Randomly sampled structures from Gaussian distributions with mean and variance estimated from the set of structures in (a).

list of more likely RNA families, which can then be studied in more detail. We can also obtain confidence regions in the directions of main variability and apply likelihood ratio tests in structure classifications. Moreover, random RNA structures can be generated for a given class of structures. As an example of the latter application, Figure 3a shows the set of RNA structures in tRNA synthetase complexes family; Figure 3b is the mean shape calculated from the structures in 3a; Figure 3c shows five sampled structures on each of the top three variance components; and Figure 3d shows some randomly sampled structures from the distribution derived from the set of structures shown in Figure 3a.

In this study, the calculated statistics are on the shapes only, and the sequences are used to refine the overall registration of the structures to their sample mean. In principle, these statistics can also be calculated using a combination of nucleotide sequences and structures, if desired.

RESULTS

Benchmark dataset

To test the performance of ESA and compare it with previous methods, we use a benchmark dataset, FSCOR (8, 9), compiled from the SCOR database (30). FSCOR contains 419 RNA structures in 168 functional classes. The histograms of chain lengths for the 419 RNA molecules are plotted in Figure 4a, and the histogram of number of members in each class is plotted in Figure 4b. We can see that most of the RNA molecules have fairly small sizes, with <200 residues, and the class frequencies are very unbalanced, with many classes containing only one member.

Parameter optimization

When we combine sequence and structure information, we introduce a weight parameter, λ, controlling the relative influence of sequence information, which is subject to tuning. To search for the range of plausible values of λ, we randomly sample 80% of RNA structures with <200 residues in FSCOR dataset and use leave-one-out cross validation (LOOCV) to find the optimal λ among some selected values between zero and 70. Specifically, there are 369 out of 419 structures with <200 residues. This subset was selected mainly to save computational time, since some of the remaining structures are very large. In LOOCV, we compute the area under ROC curve (AUC) and use this criterion for performance evaluation throughout this article. This randomization was performed 10 times to assess the robustness of estimated optimal λ values for different sets of RNA structures. AUC values obtained with different λ from this experiment were plotted in Figure 5. In addition to this set of λ values, we also performed RNA comparison using sequence

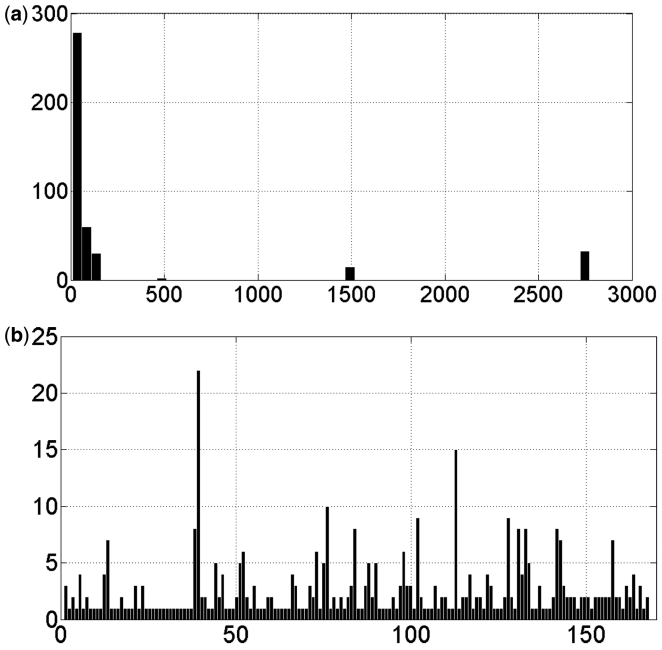

Figure 4. (**a**) Distribution of the lengths of RNAs in FSCOR dataset; and (**b**) Number of members in each of the classes in the FSCOR dataset.

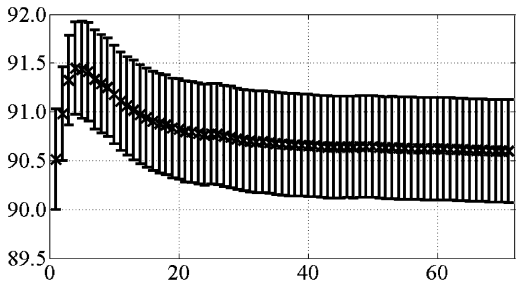

Figure 5. AUC versus lambda values obtained from the randomization experiment. Error bars are also shown together for each lambda values. λ values are from 0 to 70 with a step size of 1. The last value on x-axis corresponds to the λ_∞ case, which is the case of alignment using pure nucleotide sequence.

information alone, denoted as λ_∞. When using only sequence information in alignment, equation A.1 is replaced by:

$$\gamma^* = \underset{\gamma\in\Gamma}{\operatorname{argmin}} ||q_1 - \sqrt{\dot{\gamma}}(q_2^{\circ}\gamma)||^2. \quad (2)$$

The AUCs for different λ for the full FSCOR dataset are shown in Figure 6. Both results show that λs around 5 tend to give the best AUCs, which is the value we use for subsequent performance evaluation on the whole FSCOR dataset. It is worth noting that both $\lambda = 0$ and λ_∞ give worse classification performances, showing that alignment in the joint sequence–structure space indeed performs better than using either sequence or structure information alone.

To study the effect of scaling of SRVFs, we also computed the geodesic distances without scaling by

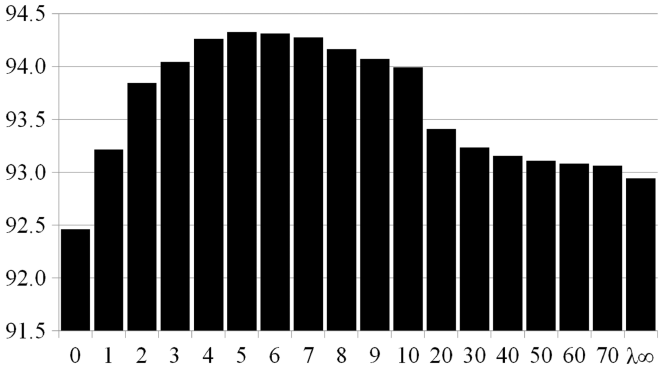

Figure 6. AUC for selected values of λ using the full FSCOR dataset. The selected λ values are 0, 1, 2…, 10, 20, 30…, 70, λ_∞ (0 to 10 with a step-size of 1, 10 to 70 with a step-size of 10, and sequence only alignment).

replacing equations of the great circle 4 and distance 5 (see appendix) by the equations of the straight line and L^2 distance, respectively. The performance without scaling turns out to be worse, with AUCs ranging from 84% to 87%, for different λ values. This observation is consistent with what we have found in protein structure comparison using ESA (21).

RNA function classification and comparison with existing methods

We performed RNA function classification on the whole FSCOR dataset with optimal parameter setting and compared its AUC [following the same procedure as in (10)] performance with several existing methods, including SARA (8, 9), iPARTS (10), LaJolla (12) and ARTS (11). The functional classes were obtained from SCOR database (30). We ran the executable programs locally when possible. Otherwise we obtained the results directly from the authors. SARA's original score is the mean log *P*-value (MLP), which combines *P*-values for percentage identity of primary, secondary and tertiary structure in their alignments. iPARTS' original score is the structural alignment score (SAS), which is based on their native alignment score and geometric match measure. The score used for ARTS is their top alignment score chosen from their *k* non-redundant top-ranking matches. LaJolla's score is their TM-Score, which is based on lengths of the target and aligned structures along with their euclidean residue distances. In the case of ARTS and LaJolla, if the scores were non-existing due to their limitations discussed in (9) and (10), we assigned a low similarity score.

The all-against-all distance matrices calculated from the above methods are shown in Figure 7, where RNAs in the same classes are placed adjacent to one another.

In an ideal clustering, one should expect to see small distances (darker rectangles) only along the diagonal of each matrix. We can see from the matrices that some of the classes can be accurately clustered by all methods, and ESA appears to generate better clustering overall. To quantitatively evaluate clustering performance we used the *F-measure* criterion described in (26). Table 1

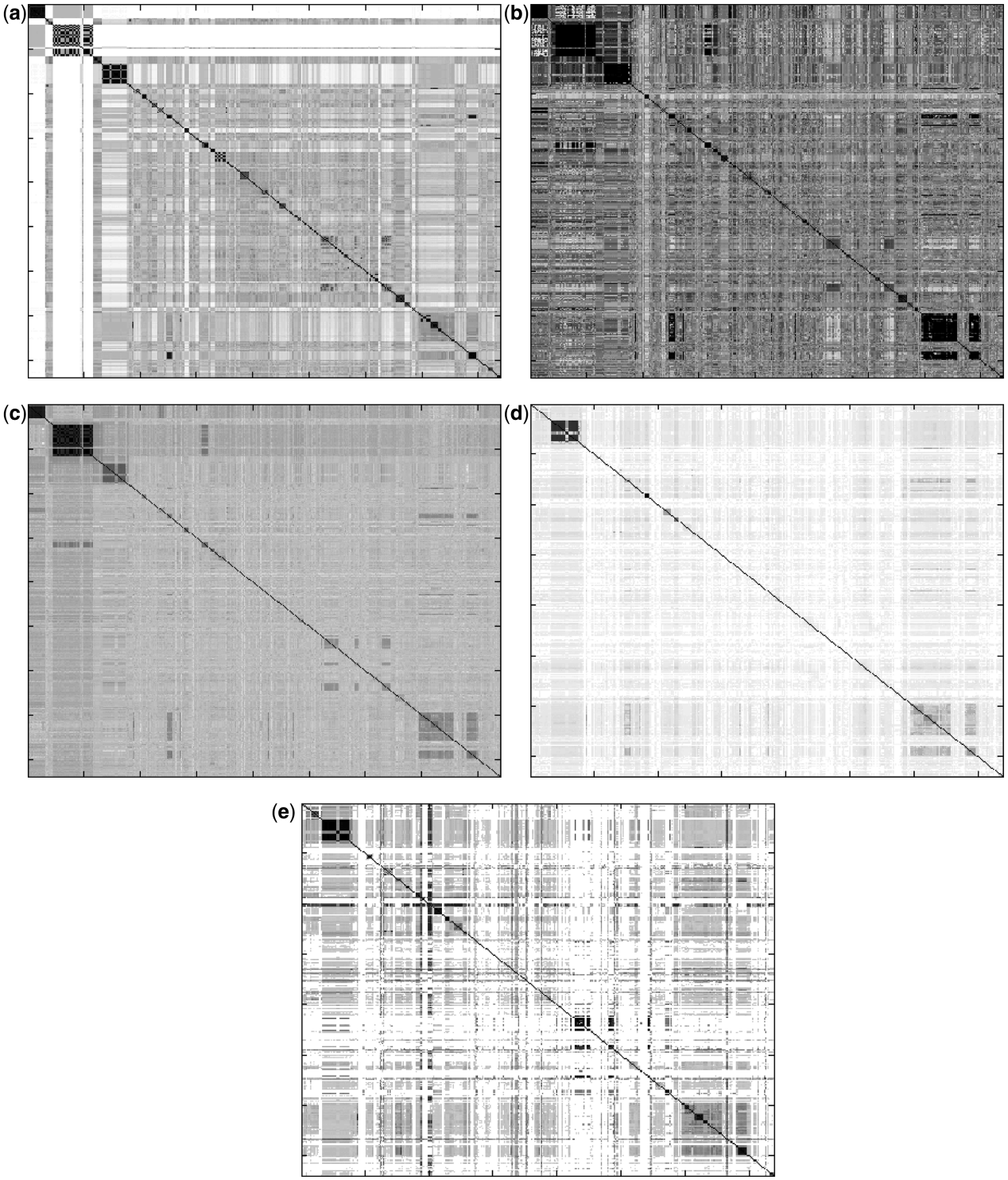

Figure 7. (**a–e**) Distance matrices for all methods. Darker colour means smaller distances. (a) ESA ($\lambda = 5$), (b) SARA, (c) iPARTS, (d) ARTS, and (e) LaJolla.

contains *F-measure* results for all methods, which shows that ESA has the best performance among all the methods.

The ROC curves for function prediction for all the methods are plotted in Figure 8 and the corresponding AUC values are shown in Table 2. Among all the methods, ESA has the highest AUC value of 94.33.

We also compared the average CPU running time for all methods using only the structures that are <200 nucleotides long. These are given in Table 3. SARA, ARTS and LaJolla were run on a standard computer with Linux OS and Intel® Pentium 4™ 3.0 GHz CPU. ESA was run using MATLAB™ on a standard computer with Windows 7™ OS and Intel® Core 2™ 2.3 GHz CPU.

Table 1. Comparison of F-Measure for all methods

ESA	SARA	iPARTS	ARTS	LaJolla
56.38	46.65	27.37	5.09	7.50

The values are given in (%).

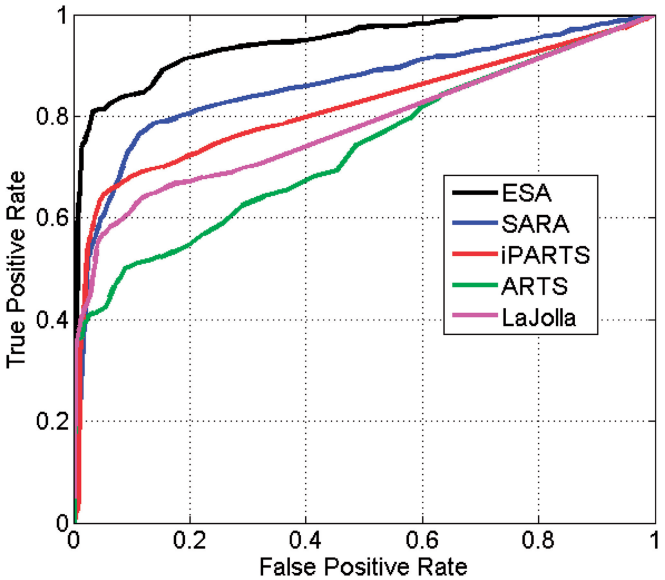

Figure 8. Comparison of ROC curves between different methods. ESA shown here uses $\lambda = 5$.

Table 2. Comparison of AUC for all methods

ESA	SARA	iPARTS	ARTS	LaJolla
94.33	85.78	81.43	73.21	78.04

The values are given in (%).

Table 3. Comparison of running time[a] for all methods

ESA	SARA	iPARTS	ARTS	LaJolla
0.0106	1.0805	0.3281	0.0745	1.4285

[a]Times are measured in average seconds per comparison.

In the case of iPARTS, we obtained the pairwise comparison results and the running time information from the authors (Intel® Xeon™ 2.5 GHz CPU). On average, it takes less time for ESA to perform an alignment than the other methods.

CONCLUSION AND DISCUSSION

In this study, we developed a more general ESA framework for RNA structure alignment, which allows us to add new information in addition to the backbone geometric information of RNA molecules. We applied the method to RNA structure comparison and function classification. Tested on a benchmark dataset, our method achieved better accuracy and higher computational efficiency than several previous methods.

As mentioned earlier, ESA offers several unique advantages compared with other methods. Firstly, it calculates a formal distance, geodesic distance, between two RNA structures; secondly, the method is able to use both sequence and structure information in RNA comparison, which is, in principle, more desirable than using only sequence or structure information alone; thirdly, under the mathematical framework, one can calculate the mean and covariances for a group of structures, which allows for the construction of probability distributions for a group of RNA structures. This unique feature will be further investigated in our future studies. Finally, ESA performs global alignment and computes the global similarity for a given pair of RNA structures. To the best of our knowledge, there has been no global structure alignment methods for RNAs developed in the previous literature.

It is a common practice in most previous studies to evaluate alignment quality in terms of structural similarity of the aligned parts. We did not provide such evaluation in this study due to the following reasons. Firstly, the ESA method is a global alignment procedure that measures the global similarity between RNA or protein structures. Most of the current metrics for evaluating alignment quality are designed for local alignments, which are not appropriate for global alignment. Secondly, it is well acknowledged that there is no universal measure for structure similarity, which works for all purposes. Finally, the evaluation we currently use, although not based on structural similarity, is an objective criterion based on the similarity of the functions of RNAs, which are determined experimentally. One of the most important applications of structure/sequence alignment is function inference.

We have demonstrated in this study that one can achieve better classification performance in RNA function prediction by combining sequence and structure (backbone geometry) information than using either sequence or structure information alone. In principle, one can infer functions given complete structure information, which can be decomposed into three parts: the backbone geometry, the type of side chains (the same as sequence information) and side chain conformations. The last two types of information are strongly correlated, as one cannot talk about side chain conformations without knowing the identities of the side chains. All the three types of information are determining factors of RNA functions. Traditional structure alignments use only backbone geometry information, which limits the accuracy they can achieve. By incorporating sequences, we add the second type of information, which explains the superior performance our method has achieved. One can probably further improve classification performance by incorporating the information of side chain conformations. An alternative approach for function inference is through comparing molecular surfaces, which uses side chain conformations (and some information from backbone as well).

Local structure alignments can be conducted using the algorithms outlined in this article. However, under the current framework of ESA, the distances generated will not be formal distances. To maintain formal distances while performing local alignments is an interesting topic for future studies.

As shown in an earlier study (26), the distance distribution of RNA structure alignments by ESA shows a clear bi-modal pattern. A cut-off can be easily obtained using Expectation Maximization algorithm to distinguish similar RNA structures and non-similar RNA structures. This cut-off value can be used to discover new RNA structure classes if a new RNA structure with unknown class has distances larger than the cut-off to all the existing RNA structures. A similar approach can be used when aligning RNAs in the joint sequence–structure space.

ACKNOWLEDGEMENTS

The authors will like to thank Dr. Sebastian Kurtek (FSU-Statistics) for his valuable discussions.

FUNDING

National Institutes of Health (NIH) [1R21GM101552 to J.Z. and A.S.]. Funding for open access charge: NIH in part.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Reviews Gen*., **2**, 919–929.
3. Dorsett,Y. and Tuschl,T. (2004) siRNAs: applications in functional genomics and potential as therapeutics. *Nat. Rev. Drug Discov*., **3**, 318–329.
4. Doudna,J. and Cech,T. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
5. Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*., **3**, 956–959.
6. Doudna,J. (2000) Structural genomics of RNA. *Nat. Struc. Biol*., **7(Suppl. S)**, 954–956.
7. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
8. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, I112–I118, Joint Meeting of the 7th European Conference on Computational Biology/5th Meeting of the Bioinformatics-Italian-Society, Cagliari, ITALY, SEP 22–26, 2008.
9. Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res*., **37**, W260–W265.
10. Wang,C., Chen,K. and Lu,C. (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res*., **38**, W340–W347.
11. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21(Suppl. 2)**, 47–53.
12. Bauer,R.A., Rother,K., Moor,P., Reinert,K., Steinke,T., Bujnicki,J.M. and Preissner,R. (2009) Fast structural alignment of biomolecules using a hash table, n-Grams and string descriptors. *Algorithms*, **2**, 692–709.
13. Duarte,C.M., Wadley,L. and Pyle,A. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res*., **31**, 4755–4761.
14. Ferre,F., Ponty,Y., Lorenz,W.A. and Clote,P. (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–W668.
15. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol*., **56**, 215–252.
16. Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res*., **32**, 6650–6659.
17. Harrison,A.M., South,D.R., Willett,P. and Artymiuk,P.J. (2003) Representation, searching and discovery of patterns of bases in complex RNA structures. *J. Comput. Aided Mol. Design*, **17**, 537–549.
18. Rahrig,R.R., Leontis,N.B. and Zirbel,C.L. (2010) R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, **26**, 2689–2697.
19. Zhong,C.T.H. and Zhang,S. (2010) RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res*., **38**, e176.
20. Liu,W., Srivastava,A. and Zhang,J. (2010) Protein structure alignment using elastic shape analysis. In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, 2010*. ACM, Niagara Falls, New York, pp. 62–70.
21. Liu,W., Srivastava,A. and Zhang,J. (2011) A mathematical framework for protein structure comparison. *PLoS Comp. Biol*., **7**, e1001075.
22. Srivastava,A., Klassen,E., Joshi,S. and Jermyn,I. (2011) Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell*., **33**, 1415–1428.
23. Joshi,S.H., Klassen,E., Srivastava,A. and Jermyn,I.H. (2007) A novel representation for Riemannian analysis of elastic curves in R. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*., **2007**, 1–7.
24. Klassen,E., Srivastava,A., Mio,W. and Joshi,S.H. (2004) Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell*., **26**, 372–383.
25. Mio,W., Srivastava,A. and Joshi,S. (2007) On shape of plane elastic curves. *Int. J. Comput. Vis*., **73**, 307–324.
26. Laborde,J., Srivastava,A. and Zhang,J. (2011) Structure-based RNA function prediction using elastic shape analysis. *2011 IEEE International Conference on Bioinformatics and Biomedicine*. Atlanta, GA, pp. 16–21.
27. Jukes,T.H. and Cantor,C.R. (1969) *Mammalian Protein Metabolism*, 3rd edn. Academic Press, New York.
28. Bernstein,F., Koetzle,T., Williams,G., Meyer,E. Jr, Brice,M., Rodgers,J., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535.
29. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Iype,L., Jain,S., Fagan,P., Marvin,J. *et al.* (2002) The protein data bank. *Acta. Crystallographica*., **58(Part 6)**, 899–907.
30. Tamura,M., Hendrix,D.K., Klosterman,P.S., Schimmelman,N., Brenner,S.E. and Holbrook,S.R. (2004) SCOR: structural classification of RNA, version 2.0. *Nucleic Acids Res*., **32**, D182–D184.

APPENDIX

ESA framework

For any two points in S, the distance between them is given by the length of the shortest path (called a geodesic) connecting them in that manifold. An interesting feature of this framework is that it not only provides a distance between two RNA structures, thus quantifying

differences between their shapes, but also a geodesic path between them in S. This path has the interpretation that it provides the optimal deformation of one shape into another. The geodesics are actually computed using the differential geometry of the underlying space S. Consider two curves β_1 and β_2, represented by their SRVFs q_1 and q_2. To compute geodesics between their equivalence classes $[q_1]$ and $[q_2]$, we fix q_1 and find the optimal rotation and re-parameterization of q_2 to solve:

$$(\Omega^*, \gamma^*) = \underset{\Omega\in\bar{\Omega},\,\gamma\in\Gamma}{\operatorname{argmin}} ||q_1 - \sqrt{\dot{\gamma}}\Omega(q_2^\circ\gamma)||^2. \quad \textbf{(A.1)}$$

The optimization over rotation is done using singular value decomposition (SVD) on the O component of Ω, but the optimization over the re-parameterization requires a dynamic programming algorithm. The optimal γ^* is the matching function between the two backbones. Define $q_2^* = \sqrt{\dot{\gamma}^*}\Omega^*(q_2^\circ\gamma*)$ and compute a geodesic path between q_1 and q_2^* in C. Since C is a sphere, the geodesic between any two points is given by a great circle whose equation is:

$$\alpha(\tau) = \frac{1}{sin(\theta)}(sin((1-\tau)\theta)q_1 + sin(\tau\theta)q_2^*), \quad \textbf{(A.2)}$$

where α is a geodesic path between the given two shapes such that it is in $[q_1]$ at $\tau = 0$ and in $[q_2]$ at $\tau = 1$. Here

$$\theta = cos^{-1} < q_1(t), q_2^*(t) > \quad \textbf{(A.3)}$$

is the distance between the two equivalence classes in S, i.e. $d([q_1],[q_2]) = \theta$. This θ is a **proper distance** in the shape space, as it satisfies all three properties of a distance function, including the triangle inequality.

Statistics under the SRVF representation

Given a set of SRVFs $q_1, q_2, \ldots, q_n$, their Sample Karcher Mean is given by:

$$\hat{\mu} = \underset{[q]\in S}{\operatorname{argmin}} \sum_{i=1}^{n} d([q],[q_i])^2 \quad \textbf{(A.4)}$$

where $d([q],[q_i])$ in equation A.4 is the geodesic distance between q, and q_i. This definition is equivalent to the sample mean $\bar{X}$ in a Euclidean vector space. The actual minimizer is found using an iterated gradient-approach (23). To get an average shape $\beta_{\hat{\mu}}$ we simply use integration:

$$\beta_{\hat{\mu}}(t) = \int_0^t \hat{\mu}(s)||\hat{\mu}(s)||ds \quad \textbf{(A.5)}$$

To define sample covariance, we approximate the shape space S in a neighbourhood of $\hat{\mu}$ by a flat space $T_{\hat{\mu}}(S)$. Each of the observed SRVFs are then projected to the flat space $T_{\hat{\mu}}(S)$ using the *inverse exponential mapping*:

$$q_i \mapsto v_i \equiv \frac{\theta_i}{sin(\theta_i)}(q_i - cos(\theta_i)\hat{\mu}). \quad \textbf{(A.6)}$$

where $\theta_i = d([q_i],[\hat{\mu}])$. Intuitively, v_i's are simply the directions of the geodesics from the mean $\hat{\mu}$ to q_i's.

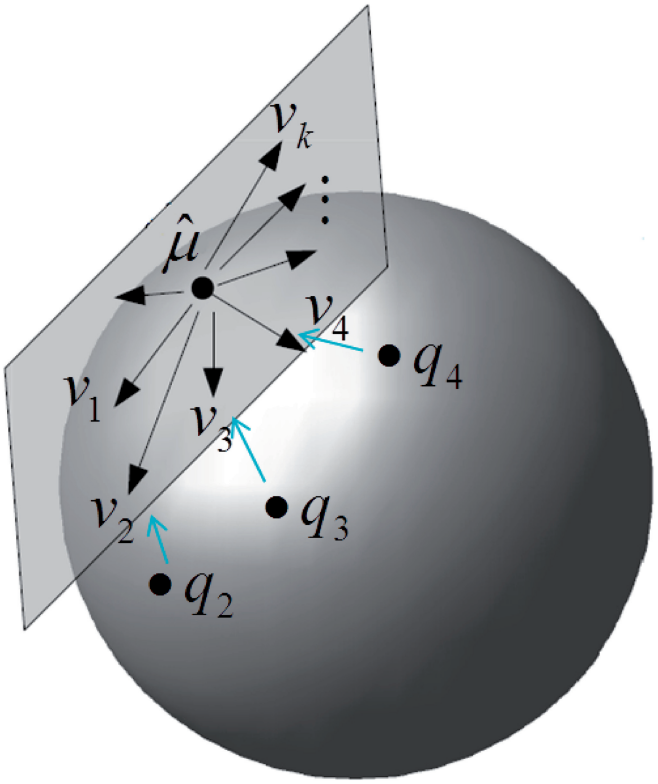

Figure A1. q functions, their Karcher mean $\hat{\mu}$ and their projection onto $\hat{\mu}$'s tangent space.

Figure A1 shows a depiction of this using a sphere. We can compute the standard sample covariance matrix $K = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \hat{\mu})^T(v_i - \hat{\mu})$ of v_is and take its SVD $K = U\Sigma U^t$. Here Σ is a diagonal matrix of singular values (which for convenience, we can denote and arrange as $\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq \ldots$), and U contains the corresponding singular vectors. Since the singular values are arranged in decreasing order, the first few, say k, columns of U represent the directions of major variation, or the principal components, of the underlying population. If we let $z_1, z_2, \ldots z_k$ be independent standard normal random variables, we can define a multivariate normal density on the direction v according to: $v = \sum_{i=1}^{k} z_i\sigma_i U_i$. Then, this random direction can be converted into the SRVF of a random shape using the *exponential mapping*:

$$v \mapsto q \equiv cos(||v||)\hat{\mu} + \frac{sin(||v||)}{||v||}v, \quad \textbf{(A.7)}$$

which can be further converted into a shape using integration: $\beta(t) = \int_0^t q(s)||q(s)||ds$. This defines a formal Gaussian probability model on the shape space S and one can sample random structures from it using the steps outlined above.