BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 5 2013, pages 588–596
doi:10.1093/bioinformatics/btt006

Structural bioinformatics

Advance Access publication January 12, 2013

Fast alignment and comparison of RNA structures

Tim Wiegels*, Stefan Bienert and Andrew E. Torda

Centre for Bioinformatics, University of Hamburg, Bundesstr. 43, D-20146 Hamburg, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: To recognize remote relationships between RNA molecules, one must be able to align structures without regard to sequence similarity. We have implemented a method, which is swift [$O(n^2)$], sensitive and tolerant of large gaps and insertions. Molecules are broken into overlapping fragments, which are characterized by their memberships in a probabilistic classification based on local geometry and H-bonding descriptors. This leads to a probabilistic similarity measure that is used in a conventional dynamic programming method.

Results: Examples are given of database searching, the detection of structural similarities, which would not be found using sequence based methods, and comparisons with a previously published approach.

Availability and implementation: Source code (C and perl) and binaries for linux are freely available at www.zbh.uni-hamburg.de/fries.

Contact: tim.wiegels@gmail.com

Received on January 5, 2012; revised on November 29, 2012; accepted on January 4, 2013

1 INTRODUCTION

RNA has diverse roles ranging from ligand binding/recognition to catalysis. These functions depend not just on sequence, but more on how a molecule positions groups in space (Cech and Bass, 1986; Coppins *et al.*, 2007; DeRose, 2002; Kim and Breaker, 2008; Lilly and Eckstein, 2008; Mandal and Breaker, 2004; Montange and Batey, 2008; Nudler and Mironov, 2004; Scott, 1999; Scott and Klug, 1996; Strobel and Cochrane, 2007; Waugh and Pace, 1986; Winkler, 2005). This means that if one wants to detect functional similarity, one must be able to align molecules structurally. We present a method for RNA-structure alignment, which seems to be sensitive, reliable and with a running time comparable with sequence alignment techniques. One could also claim that in evolutionary terms, structure is more conserved than sequence; therefore, structural searches and alignments will find relationships that would be missed at the sequence level (Holm and Sander, 1996).

If one allows for gaps and insertions, the problem of structurally aligning two biopolymers is NP-complete (Godzik, 1996). There is, however, an abundance of methods developed for proteins using various approximations that usually provide reasonable results (Alexandrov, 1996; Blankenbecler *et al.*, 2003; Carpentier *et al.*, 2005; Chen *et al.*, 2005; Gibrat *et al.*, 1996; Holm and Park, 2000; Holm and Sander, 1993; Ilyin *et al.*, 2004; Jung and Lee, 2000; Kawabata, 2003; Konagurthu *et al.*, 2006; Krissinel and Henrick, 2004; Lackner *et al.*, 2000; Lisewski and Lichtarge, 2006; Marchler-Bauer *et al.*, 2005; Ochagavia and Wodak, 2004; Oldfield, 2007; Orengo and Taylor, 1996; Ortiz *et al.*, 2002; Russell and Barton, 1992; Schenk *et al.*, 2008; Shapiro and Brutlag, 2004; Shatsky *et al.*, 2002; Shindyalov and Bourne, 1998; Subbiah *et al.*, 1993; Suyama *et al.*, 1997; Taubig *et al.*, 2006; Zhu and Weng, 2005; Zuker and Somorjai, 1989). Typically, these methods rely on some structural property specific to proteins. For example, most proteins can be reasonably characterized by their classic secondary structure elements or their sequence of backbone φ, ψ angles or a contact map based on C^{α} atoms. In the case of RNA, there is far less prior art, presumably because it is harder to describe, and there is little consensus as to what constitutes a characteristic structural motif.

In proteins, common structures are rather well described by neglecting the backbone ω dihedral angle and considering φ, ψ, perhaps with additional hydrogen bond information. In RNA, one could consider six backbone angles (α, β, γ, δ, ε, ζ) and perhaps the χ dihedral angle, which captures the orientation of the base. At the same time, not all these descriptors are necessary, as some of the angles are significantly correlated (Murthy *et al.*, 1999). Duarte and Pyle took advantage of this fact and pointed out that it may be more effective to define dihedral angles based on atoms such as C4′ and P atoms from successive bases even though they are not covalently connected (Duarte and Pyle, 1998). Using these angles as descriptors, they defined 10 structural motifs with which they were able to reconstruct known RNA structures with surprisingly high accuracy. These descriptors also seemed to be sufficient for finding motifs in collections of known structures (Duarte *et al.*, 2003; Wadley *et al.*, 2007).

Although there are methods for aligning RNA based on secondary structures (Havgaard *et al.*, 2005; Hochsmann *et al.*, 2004; Jossinet *et al.*, 2007; Will *et al.*, 2007), there is a distinct paucity of methods that use 3D properties. The approach implemented in ARTS uses the coordinates of phosphorus atoms and attempts to find and extend sets of atom pairs, which can be superimposed below some threshold (Dror *et al.*, 2005, 2006). The alignment calculation relies on the recognition of secondary structure elements. This means the method implicitly has some thresholds, but it has the advantage that the procedure is apparently fast enough to be used for searches of a structure against a database of known structures. Another approach is to label sites with some structural property or combined sequence/structure property. Sites from different molecules can be compared with each other giving a score, which can be used in a conventional dynamic programming algorithm. DIAL uses this approach and bases its score on sequence, nine dihedral angles and a measure

*To whom correspondence should be addressed.

of base pairing (Ferre *et al.*, 2007). SARSA, by Chang *et al.* (2008), used four dihedral angles to define a small structural alphabet. This approach was revisited in iPARTS (Wang *et al.*, 2010), with a similar structural alphabet limited to the pseudo-torsion angles η and θ as defined by Duarte and Pyle (1998). SARA (Capriotti and Marti-Renom, 2008) follows an approach inspired by the MAMMOTH programme for protein structure alignment (Ortiz *et al.*, 2002). The most recent approach, SETTER, divides RNA structures into non-overlapping secondary structure elements (Cech *et al.*, 2012). All of these methods are sensitive to the geometry of specific sites rather than larger pieces of structure.

In this work, we have adapted a method from the protein literature, but one with little preconception as to what constitutes a structural motif (Margraf *et al.*, 2009; Schenk *et al.*, 2008). Instead, one says that a structure is built from overlapping fragments of some length ($k=5$ in this work). Each fragment is described by a set of structural descriptors (angles, distances and a measure of likelihood to be in a base pair). A large set of fragments from the protein data bank (PDB) (Berman *et al.*, 2000) was then subjected to a maximally parsimonious Bayesian classification yielding $\sim 10^2$ classes. To use this for alignments, one can take a fragment from a new molecule and calculate the probability that it is in class 1, 2,..., N_{class}. This vector of probabilities can be seen as a structural label for the site, but more importantly, the dot product of two such vectors is a direct measure of their similarity. These dot products can be used directly in a score matrix for a dynamic programming algorithm. The similarity measure is a continuous property ranging from 0 to 1; therefore, no thresholds are necessary. Calculating an optimal alignment has $O(n^2)$ running time albeit with a larger pre-factor than in sequence alignments. As we show below, the method seems to produce rather high quality alignments.

2 METHODS

2.1 Theory

The methodology is described only briefly, as it has been given in detail in the context of protein alignments (Schenk *et al.*, 2008). Imagine one has a probabilistic classification of RNA fragments, each of length k. Given a fragment, the aim is not to find the best class for the fragment, but rather to estimate the probability that the fragment is a member of class 1, 2,..., N_{class}. Each class is a set of probability distributions for each of the structural descriptors used. For example, one might consider an angle α. Over a large set of fragments, α may adopt many different values. Within one class, its distribution may be approximated by a single Gaussian distribution. The probability of seeing some particular value of α can be calculated if one knows the mean μ_α and standard deviation σ_α within each class. In this kind of work, one does not have a single α, but rather a set of descriptors containing several angles and distances and a measure of the probability of a base being in a base pair. The probability of being in a class is then given by the product over each descriptor and its corresponding distributions. Furthermore, angles may be correlated. This means the best statistical model is not a set of Gaussian distributions, but perhaps a set of multivariate Gaussian distributions. To formalize this briefly, note that one is interested in the probability $P(F_i \in c_j|F_i)$ that fragment F_i is in class c_j given by

$$P(F_i \in c_j|F_i) = \frac{P(F_i|F_i \in c_j)P(F_i \in c_j)}{P(F_i)} \quad (1)$$

where $P(F_i)$ and $P(F_i \in c_j)$ are prior probabilities of the descriptors of F_i and class c_j. Considering all classes, one has the normalization condition that probabilities associated with a fragment over all classes must sum to 1. These probabilities, of course, can only be calculated after the classification is known, which leads to the next step. A classification is really a statistical model for the data, and its likelihood can be calculated, given some dataset. The classification is then constructed from some training data, using expectation maximization, iteratively adjusting the parameters of the distributions, searching for an ever better model for the data (Dempster, 1977). This is a local minimization method; therefore, it is repeated many times from different starting points.

The number of classes is also an unknown parameter, but its value is optimized as with the rest of the classification. The probability of a classification includes products over all classes; thus, increasing the number of classes tends to make the statistical model less likely, unless it is strongly supported by the data. This means the method has a natural tendency to find the classification with the fewest classes.

Once a classification is calculated, it can be used to calculate a similarity score between fragments. For the fragment i, one calculates the vector $\mathbf{p}_i$ of probabilities that it is in each class using Equation (1). Given some fragment j and its corresponding vector, the similarity of the two fragments is just

$$s_{ij} = \mathbf{p}_i \cdot \mathbf{p}_j \quad (2)$$

after normalizing both vectors to length 1. This value will be near 1 if the two fragments have similar probability distributions and near zero if they are different. It is a graduated measure, and therefore it is sensitive to partial similarity. It is also tolerant of fragments, which have never been seen before (in the training set). Their characteristic $\mathbf{p}$ vectors will match best with $\mathbf{p}$ vectors from fragments with some similarity. Aside from 5′ and 3′ ends, each site participates in $2k-1$ fragments.

2.2 Datasets and calculations

The training dataset of RNA structures was taken from the PDB (Berman *et al.*, 2000; Rose *et al.*, 2011) and comprised all structures that only contained RNA without any DNA or protein and with a resolution of better than 5 Å. This resulted in 581 structures and just $>3.3 \times 10^5$ fragments of length $k=5$. For testing and searching, we used a larger set of coordinates, considering any structure, which contained RNA. This yielded 1400 coordinates files, corresponding to 2250 chains.

All alignments were calculated with Fast RNA structural Identity Estimator (FRIEs) based on wurst (Schenk *et al.*, 2008) using the Gotoh (1982) version of the Smith and Waterman (1981) algorithm. Structural differences are quoted after superposition with the method of Diamond (1988) and using the six backbone atoms C3′, C4′, C5′, O3′, O5′ and P as in Reijmers *et al.* (2001). As the root mean square differences (rmsd) of coordinates calculated for random polymers grows with the one-third power of structure length, rmsd values could be scaled by the cube root of the alignment length for ranking the significance of alignments (Maiorov and Crippen, 1994).

Given the model described later in the text, the classification was built using autoclass (Cheeseman and Stutz, 1995) and implemented in existing code. Comparisons were calculated with LaJolla version 2.1 (Bauer *et al.*, 2009).

2.3 Site descriptors

The final statistical model was based on fragments of length $k=5$ with five descriptors per base. This could be seen as 25 descriptors per fragment or four independent distributions as listed in Table 1. The H-bond score was calculated using a simple model for electrostatics. A Hydrogen bond can be seen as DH-AX where D and A are donor and acceptor

Table 1. Descriptors for model of fragments

Name	N_{dist}	Type of Gaussian
H-bond score	5	Univariate
distances $C4'_{n-1}$,P_{n+1} and P_n,$C4'_{n+1}$	1	10-dimensional multivariate
Angle δ	1	5-dimensional multivariate
Angle χ	1	5-dimensional multivariate

N_{dist} is the number of distributions used in the calculations. The H-bond score is E_H as in Equation (3). Standard nomenclature is used for the angles (IUPAC-IUB Commission on Biochemical Nomenclature, 1970).

atoms, and one can imagine some partial charge q, which has the same magnitude on all four sites. The H-bond score E_H was given by

$$E_H = \frac{q^2}{d_{DA}} + \frac{q^2}{d_{HX}} - \frac{q^2}{d_{HA}} - \frac{q^2}{d_{DX}} \quad (3)$$

where d_{AB} is the distance between atoms A and B. Hydrogen atoms were placed using standard geometry and a mean coordinate used in XH_2 groups. This measure does not require any threshold for H-bonds and values for E_H fall into two Gaussian distributions. As the scaling of the score is not important, q was set to 1. The two distances given in Table 1 were chosen, as they are rather sensitive to the compactness of a structure. Labels X_{n-1} and Y_n refer to atom X of the preceding residue and atom Y of the current residue. Rather than assume independence of the distances, they were treated as a 10D distribution with the full covariance matrix calculated. Similarly, we allowed for correlations within the sets of side-chain χ angles and δ angles. One could allow for even more correlations, but this would lead to more adjustable parameters. Angular data were treated by considering each angle twice (in two periods) and mapping all results back to the first period. The model can be seen in Figure 1.

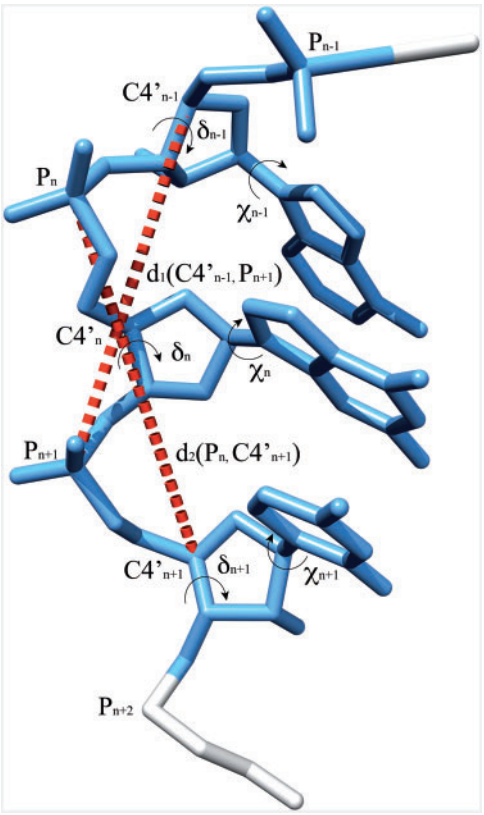

Fig. 1. Model for describing RNA structures. Black, dotted pseudo-bonds show the distance descriptors $C4'_{n-1}$, P_{n+1}.(and P_n, $C4'_{n+1}$). χ and δ, the angles used as descriptors, are also labelled

3 RESULTS

Before considering structural alignments, one can look at a few properties of the underlying fragment classification. This had 79 classes, although the largest 10 classes covered just >25% of the data. The relative importance of the classes varied from 3% down to <0.1% of the original data. As one would expect, the most populated classes reflect common RNA motifs. For example, the first two classes have similar helical structures (Fig. 2a and b), and, within statistical uncertainty, they are geometrically indistinguishable. They do however differ in the property, which cannot be seen in the diagram. The second class is distinguished by a much higher probability of being involved in H-bonds. The next most populated class also shows a common motif: a Hoogsteen pair (Hoogsteen, 1963) in which a base pairs with two other bases (Fig. 2c).

Given the machinery for comparison, we began with an all against all comparison within the set of 2250 chains from the PDB. For each structure, the alignments were sorted according to rmsd. The runtime for a pairwise comparison varied from less than a second to ~40 minutes. The runtime for a typical alignment [two transfer RNAs (tRNA) of ~70 bases]) was ~5 seconds. The most time-consuming examples were naturally those between two ribosomal structures of several thousand nucleotides. The biggest part of the computation time went into computing the H-bond terms. The times reflect a typical PC (Intel Dual Core 3.06 GHz processor).

This number of comparisons produced a mass of data, but we can show some interesting examples. The most important application is finding homology, which would not be detected using sequence-based methods. In the case of proteins, there are empirical studies showing the significance of the degree of sequence identity. For proteins of 100 residues, one would require 25–30% sequence identity before regarding a pair of sequences as related (Rost, 1999). For nucleotides, the value must be higher owing to the smaller alphabet. As an example of a structure pair, Figure 3 shows the structural alignment of 2det chain C and 1i9v (Mikkelsen *et al.*, 2001; Numata *et al.*, 2006). There is only 32% sequence identity, but the rmsd difference is only 4.9 Å calculated over the aligned 70 bases. Figure 4a shows the best global sequence alignment of the corresponding sequences. A superposition based on this alignment has an rmsd of 15 Å. Essentially, the correct alignment, based on FRIEs structural alignment Figure 4b, is completely missed with a sequence-based method. It was also impossible to find this alignment with BLAST, as the word size cannot be set less than four (Altschul *et al.*, 1997).

One can also find examples to show that the method can find small structures, which are similar to parts of larger ones. Figure 5 shows the example of chain 1p5m and 1p5p (Lukavsky *et al.*, 2003) with an rmsd of 4.7 Å >55 aligned residues. In this

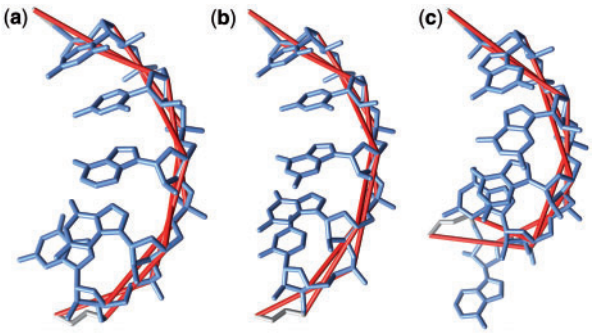

Fig. 2. Example structural classes. Dark pseudo-bonds denote distances used in classification

Fig. 3. Structural alignment of PDB files 2det and 1i9v, chain C. rmsd between tRNAs is 4.9 A with a sequence identity of only 32%

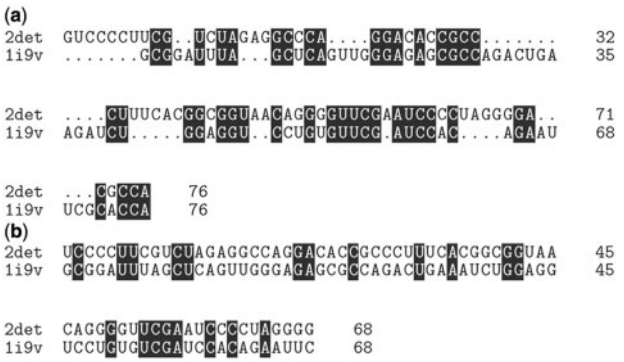

Fig. 4. Comparison of sequence versus structure alignments of 2det (chain C) and 1i9v. (**a**) The best sequence-based alignment; (**b**) the structure-based alignment

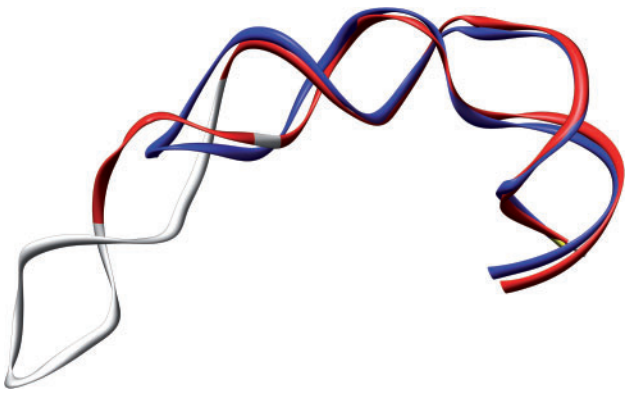

Fig. 5. Alignment of 1p5m and 1p5p. rmsd between the structures is 4.7 Å with a sequence identity of 87%

Fig. 6. Sequence alignment of PDB files 1p5m and 1p5p

case, the sequence identity is 87%, but a sequence-based method will only find the alignment if gap penalties are set low enough to allow a gap of 21 residues as shown in the corresponding sequence alignment (Fig. 6). In this work, one set of gap penalties was used for all calculations.

A more targeted application is a classic structure search. Given some query RNA structure, can one find related structures, regardless of sequence homology? Perhaps the clearest example comes from using a tRNA because this has been the most frequently solved structural family. The results are summarized in Figure 7. The tRNA used as the query structure was 1tra, and for each alignment, one has the alignment length, sequence identity (from the structural alignment) and rmsd. Alignments fall into two separate sets (or clouds): In the lower part of the plot are unrelated RNA structures, and the increasing deviation of the cloud reflects the way rmsd values grow with optimal alignments of randomly chosen molecules. In the upper part of the plot, there is a clear cluster of alignments with a minimum length of ~68 bases and rmsd values ranging from 0 to 10 Å (10 Å was the cut-off value used for finding matches). This cluster comprises related tRNA molecules and even includes tRNA chains buried within larger ribosome structures. It also includes examples such as the alignment of 1tra to 1wz2 chain D, which is an alignment of a phenylalanine-tRNA and a leucyl-tRNA (see Fig. 8). This alignment would also not be found by classical sequence alignment methods and had a sequence identity of only 48%.

There is often no such thing as a correct structural alignment (Godzik, 1996; Lackner *et al.*, 2000), but one can make some comparisons between different methods. Amongst the published methods, only LaJolla (Bauer *et al.*, 2009) could be locally installed for comparison. An all against all comparison was attempted. For each alignment, the rmsd and alignment length

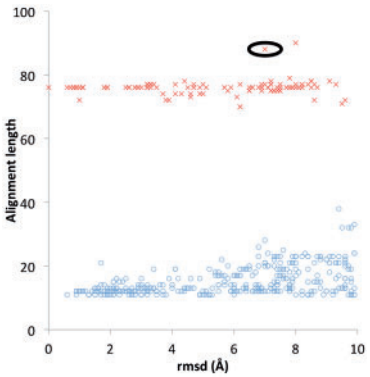

Fig. 7. Database search for PDB ID 1tra. The upper group is a set of alignments, which includes all tRNAs in the PDB. The structural superposition corresponding to the circled alignment is shown in Figure 8

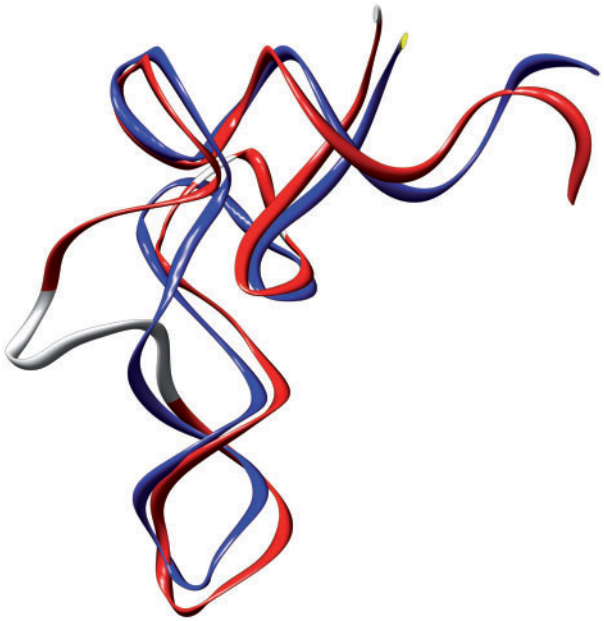

Fig. 8. Alignment of PDB files 1tra, chain A and 1wz2, chain D corresponding to the outlier in Figure 7

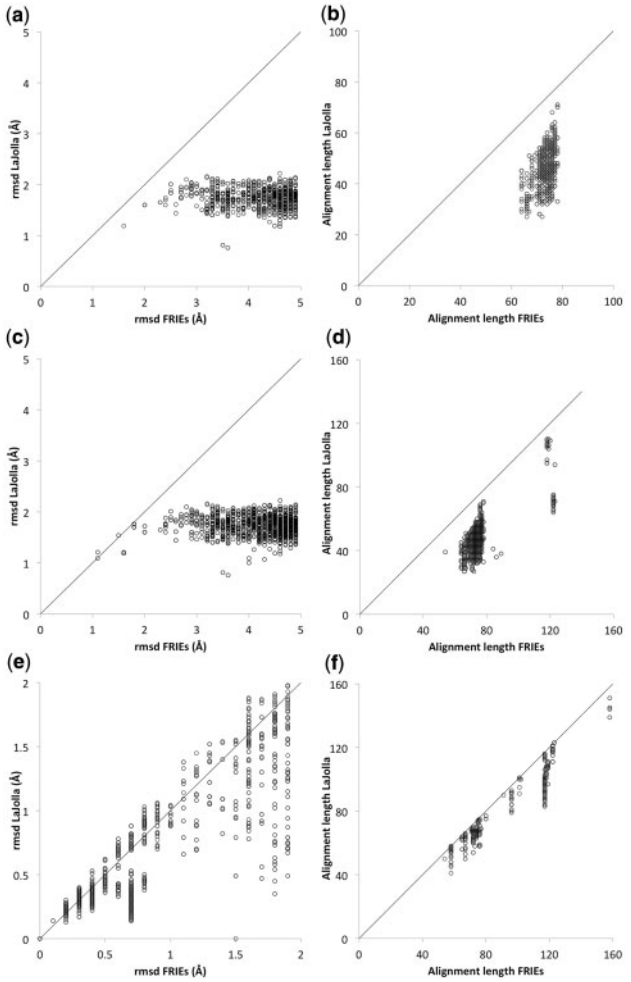

Fig. 9. Comparison of FRIEs and LaJolla regarding alignment length and rmsd is shown for all possible alignments of three sets with different thresholds for the FRIEs results. Alignments with a size between 60 and 80 nucleotides (i.e. tRNAs) in (**a**) and (**b**), alignments with an rmsd below 5 Å and a sequence identity of <50% in (**c**) and (**d**), and alignments with an rmsd below 2 Å in both FRIEs and LaJolla in (**e**) and (**f**)

generated by FRIEs and the corresponding values generated by LaJolla were compared and are shown in Figure 9a–f. We tested LaJolla on three different sets, the first one comprising all possible alignments of structures between 60 and 80 nucleotides (i.e. tRNAs, see Fig. 9a and b). The second and third sets were not limited to a certain size, but to all possible alignments for a sequence identity of ≤50% and rmsd <5 Å in FRIEs (Fig. 9c and d) and with an rmsd of ≤2 Å with FRIEs (Fig. 9e and f).

Several features are clear. First, LaJolla caps rmsd values at 2 Å (calculated using its selection of atoms); thus, many alignments found by FRIEs would not be seen with LaJolla. Second, LaJolla did not return any results for structures larger than 1000 nucleotides. The plots further show that, given the capping, LaJolla generally produces alignments with a smaller rmsd (9 a, c). However, for all alignments, the number of aligned nucleotides was significantly higher for alignments from FRIEs (9 b, d, f). We also ran a test taking into account only alignments with rmsd values of ≤2 Å in FRIEs. This showed that FRIEs could produce alignments with better rmsd values than LaJolla, but with a significantly larger number of aligned nucleotides. Higher rmsd values are obviously acceptable when they reflect longer and more meaningful alignments.

Finally, some programs offer web interfaces; thus, it is possible to show some comparisons with ARTS (Dror *et al.*, 2005, 2006), SARA (Capriotti and Marti-Renom, 2008) and SARSA (Chang *et al.*, 2008). For the alignment of 2det chain C and 1i9v, ARTS provided an alignment that comprised only 52 residues (in comparison with 68 residues for FRIEs) but had a smaller rmsd of 1.9 Å. The distinctive features of ARTS can explain this difference; its alignment method is exclusively based on the comparison of coordinates of phosphate atoms, and no further backbone information is taken into account. Combined with the use of a distance matching error, this can lead to entirely different

Table 2. Comparison of the structural alignment of 16S rRNA of *Escherichia coli* (2avy, chain A, total number residues 1542) and *T. thermophiles* (1j5e, chain A, total number of residues 1522)

	FRIES	SETTER	R3D Align	ARTS	SARA	DIAL
Number of nucleotides aligned	1473	1183	1400	1115	1460	1506
Number of exact base matches	1107	947	1089	867	1075	1017

Results for SETTER are from Cech *et al.* (2012). Results for R3D Align, ARTS, SARA and DIAL are taken from Rahrig *et al.* (2010).

structural alignments. Furthermore, owing to the lack of an option to choose between local and global alignments in ARTS, it was not possible to reproduce our alignment. ARTS will always find a local alignment with a much better rmsd value instead of our global one. The results for SARSA are similar with an alignment showing many gaps. The alignment obtained by SARA also contains gaps; however, these are located mainly in the anticodon region of both tRNAs. Given the fact that 1i9v describes a Phenylalanine-tRNA and 2det contains a Glutamine-tRNA, they are most likely to differ in this area. Using 2det chain C as query for the database search of ARTS, 1i9v will not be found at all. This is caused by the nature of the dataset used for database searches in ARTS, which consists of only 250 representative RNA structures. This obviously improves the computation time but misses many results. SARSA offers no database search. The alignment of 1tra and 1wz2 chain D gave similar results. Here, the alignments with SARSA and ARTS are slightly shorter with a correspondingly smaller rmsd value. The alignment from SARA is similar to the one obtained from FRIEs with two more gaps, which also results in a better rmsd value.

Some literature results are based on very few structures, but one can also compare against them. Both SETTER and R3D Align (Cech *et al.*, 2012; Rahrig *et al.*, 2010) were compared on the basis of the alignment of two 16S rRNA structures against ARTS, SARA and DIAL. Table 2 shows little difference between the results from FRIEs and the other approaches.

In addition to comparing FRIEs against programmes using tertiary structure, we also compared it with two methods that compute alignments using only secondary structure. RNAforester hierarchically ordered multiple sequence alignments to generate alignments of RNA secondary structure strings (Hochsmann *et al.*, 2004), and locaRNA (Will *et al.*, 2007) uses a Sankoff (1985) style alignment and folding. We compared the resulting alignments in terms of two properties: (i) the number of indels (insertions and deletions) in the resulting sequence alignment and (ii) the structural alignment implied by the secondary structure alignment. Property (i) shows, that for the example of 1p5m and 1p5p, there is a big difference in the number of indels for RNAforester and locaRNA/FRIEs (Fig. 10). The alignment obtained from RNAforester contains 30 indels in 10 gaps (Fig. 10a), whereas the alignments of locaRNA and FRIEs both show only 22 indels in two gaps (Fig. 10b and c). Considering the implied structural alignments, the large differences between the alignment obtained from RNAforester (Fig. 11a) and the ones from locaRNA (Fig. 11b) and FRIEs (Fig. 5) are even more apparent. The alignments from FRIEs and locaRNA are of similar quality, with a slightly lower rmsd for locaRNA (3.1 Å as compared with 4.7 Å for FRIEs). The rmsd for the alignment obtained from RNAforester is 13.4 Å.

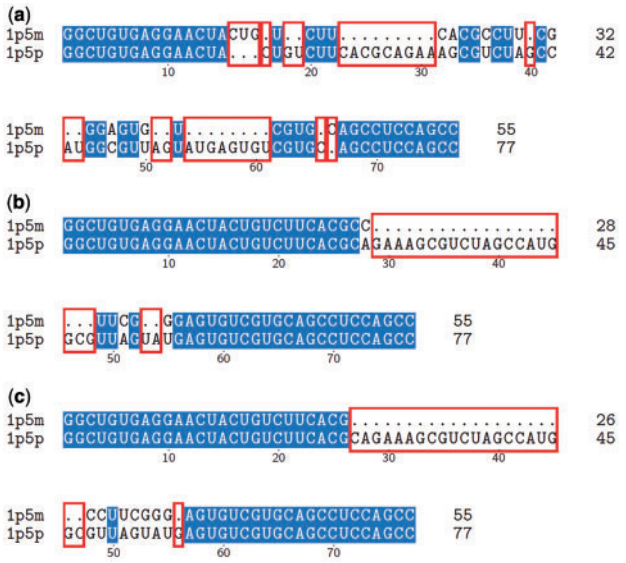

Fig. 10. Sequence alignments of 1p5m and 1p5p based on alignments generated by (**a**) RNAforester and (**b**) locaRNA compared with (**c**) the sequence alignment based on the tertiary structure alignment obtained from FRIEs

4 DISCUSSION

Because the structural alignment of biopolymers is fundamentally NP complete (Godzik, 1996), all methods use heuristics and approximations. The method in this work might be closest to those which rely on some kind of structural alphabet to label fragments. However, the method here is different in the type of labelling used. Firstly, each site in the chain is influenced by all overlapping fragments. This means that each pair of aligned residues reflects a region of nine bases. Next, the labels come from an unsupervised learning approach (Cheeseman and Stutz, 1995); thus, there are few preconceptions as to what constitutes a structural element. This is a major difference compared with RNA secondary-structure aligners (locaRNA, RNAforrester), which cannot take advantage of other common structural motifs. Most importantly, the labels are not discrete as in a typical

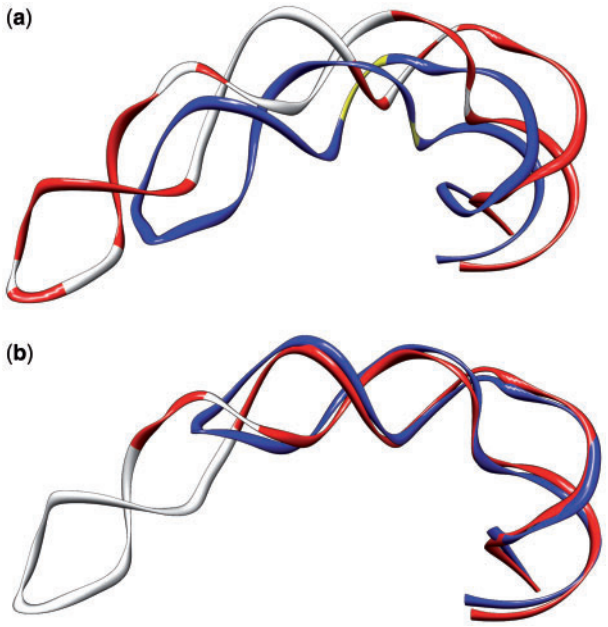

Fig. 11. Superposition of 1p5m and 1p5p based on alignment generated by RNAforester, rmsd of 13.4 Å (**a**) and locaRNA, rmsd of 3.1 Å (**b**). The structure alignment obtained from FRIEs is shown in Figure 5

structural alphabet (Chang *et al.*, 2008) or the much smaller alphabet (paired/unpaired) used by RNA secondary-structure aligners. In this work, the product of the probability vectors [Equation (2)] leads to a smooth measure of similarity; therefore, the method works when confronted with a completely new structural element.

One should also note that the classification of fragments is certainly not optimal and need not be. It must only serve to provide a fingerprint for recognizing similarity. This can be seen by a simple numerical argument. Imagine a class in the classification was incorrectly split into two classes, and a calculation is then faced with a fragment, which fits perfectly into this class. The calculated probability vector would have two entries with values of 0.5. When compared with a similar vector, the dot product would still be near 1.0.

These might be technical aspects only relevant to this method, but there is a general technical problem with this kind of approach. Ultimately, most methods are based on some choice of atoms and structural properties. The classic statistical mixture model used in this work does impose one important restriction. It must be possible to approximate the distributions of the descriptors by the sums of simple distributions. For example, the quasi-electrostatic score of Equation (3) would not be a good model for physics, but it yields relatively well-separated sets for H-bonded and not H-bonded sites. Similarly, the distances and angles used here fit better to Gaussian distributions than examples from the literature (Duarte and Pyle, 1998; Duarte *et al.*, 2003; Wadley *et al.*, 2007). It is also worth noting that the choice of descriptors can have a large influence on the kind of alignments generated. Most of the methods working with RNA structures are dominated or begin with local structural properties. In contrast, our method included a contribution from H-bonding, which is a distinctly non-local property in terms of sequence. This means that Equation (3) will yield a good score for local similarity, but there is a preference to align H-bonded sites with H-bonded sites and vice versa. One could also say that in an alignment calculation, one never requires a perfect match, and an optimal alignment may involve many less than optimal local matches.

The underlying classification is entirely automatic and, although not perfect, a model for the most important features of RNA structure (given a set of descriptors). One could, for example, interpret the most populated classes in structural terms. Looking at Figure 2, one would say that the single most common motif is a helix with well-described angles, but one can statistically justify splitting it into two classes based on H-bonding probability. The third class is also similar but is characterized by one particular distance and a particular χ value at one site.

The results in this work highlight the ability to find remote structural similarities, but the comparisons suffer from a typical problem. It is often difficult to claim one alignment is better than another. Method 1 may truncate alignments and return results with a small rmsd value (structural similarity), whereas method 2 returns longer alignments, which will appear to have a larger rmsd value. It is completely arbitrary which method one prefers. The results do, however, suggest that our proposal often finds rather long alignments without a great cost in rmsd values. This is also related to a second problem. When searching a database for related structures, there are several reasonable ways to rank the hits. A naïve structural measure such as rmsd will certainly retrieve similar structures, but these will penalize longer and more interesting alignments. One can invent similarity measures based on multiple factors, although these will entail decisions as to the importance of different properties.

When comparing methods, one can also see some differences in principle. If a method tries to optimize rmsd values, it will be sensitive to hinge movements within structures. In our approach, rmsd values are only calculated for comparisons. This means that the results are tolerant of hinge movements and, in particular, one has no problem with alignments in large structures such as ribosomes.

To make comparison with literature methods easiest, the results here were based on Smith and Waterman (1981) style alignments, which emphasize core regions of similarity. Should one want to force globally optimal alignments, it is only a switch to choose Needleman and Wunsch (1970) style alignments. If one is interested in extending the method, there are a number of possibilities. The classification used here was based on fragments of length $k = 5$. It is possible that one would capture more structural information with longer fragments, but the classification time does become prohibitively expensive. One should, in principle update the classification, as more structures become available. This can do no harm but will only make a difference with exotic and uncommon pieces of structure. At the moment, the largest 25 classes account for most of the probability density. If one doubled the amount of data, this would lead to a slightly better statistical model for the less commonly seen fragments. One could also say that the dataset used for the classification is biased. It includes too many tRNA structures because there are

so many structures in the PDB. It includes too many bases from ribosomes because ribosome structures are so large. If one removed these biases, it would change the prior probabilities associated with each class and yield a classification that was in some chemical sense better. It would not make a great deal of difference in this application where the aim is simply to apply unique labels to structural fragments.

A more interesting extension would be to combine sequence and structure information. Statistically, this means one uses a multi-way Bernoulli description for the probability of base types at each site. This is technically possible and has been tested on proteins (Schenk *et al.*, 2008), but it does raise a question that is difficult to answer. One is usually dealing with cases where sequence similarity is so small, that one has to use structural alignments. In this regime, it is not clear as to whether adding sequence information adds more noise or more information. Finally, one must make the obvious concession. Structure-based alignments can only complement sequence-based methods, as they can only be applied to the fraction of RNAs for which a structure has been determined.

The method presented here seems to be swift, sensitive to remote similarities and capable of producing accurate structural alignments of RNA structures. The code is available for download and should prove useful as more RNA structures enter the databases and one is interested in remote relationships.

Conflict of Interest: none declared.

REFERENCES

Alexandrov,N.N. (1996) SARFing the PDB. *Protein Eng.*, **9**, 727–732.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bauer,R.A. *et al.* (2009) Fast structural alignment of biomolecules using a hash table, N-grams and string descriptors. *Algorithms*, **2**, 692–709.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blankenbecler,R. *et al.* (2003) Matching protein structures with fuzzy alignments. *Proc. Natl Acad. Sci. USA*, **100**, 11936–11940.

Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, I112–I118.

Carpentier,M. *et al.* (2005) YAKUSA: a fast structural database scanning method. *Proteins*, **61**, 137–151.

Cech,P. *et al.* (2012) SETTER: web server for RNA structure comparison. *Nucleic Acids Res.*, **40**, W42–48.

Cech,T.R. and Bass,B.L. (1986) Biological catalysis by RNA. *Annu. Rev. Biochem.*, **55**, 599–629.

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

Cheeseman,P. and Stutz,J. (1995) Bayesian classification (autoclass): theory and results. In: Fayyad,U. *et al.* (ed.) *Advances in Knowledge Discovery and Data Mining*. The AAAI Press, Menlo Park, CA, pp. 61–83.

Chen,L. *et al.* (2005) Protein structure alignment by deterministic annealing. *Bioinformatics*, **21**, 51–62.

Coppins,R.L. *et al.* (2007) The intricate world of riboswitches. *Curr. Opin. Microbiol.*, **10**, 176–181.

Dempster,A. (1977) A maximum likelihood from incomplete data via the EM algorithm. *R. J. Stat. Soc.*, **39**, 1–38.

DeRose,V.J. (2002) Two decades of RNA catalysis. *Chem. Biol.*, **9**, 961–969.

Diamond,R. (1988) A note on the rotational superposition problem. *Acta Cryst.*, **A**, 211–216.

Dror,O. *et al.* (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21 (Suppl 2)**, ii47–ii53.

Dror,O. *et al.* (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.*, **34**, W412–W415.

Duarte,C.M. and Pyle,A.M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.

Duarte,C.M. *et al.* (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.

Ferre,F. *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–W668.

Gibrat,J.-F. *et al.* (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Havgaard,J.H. *et al.* (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Hochsmann,M. *et al.* (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **1**, 53–62.

Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.

Holm,L. and Sander,C. (1993) Protein-structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Hoogsteen,K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Cryst.*, **16**, 907–916.

Ilyin,V.A. *et al.* (2004) Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.*, **13**, 1865–1874.

IUPAC-IUB Commission on Biochemical Nomenclature (1970) Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules. *Biochemistry*, **9**, 3471–3479.

Jossinet,F. *et al.* (2007) RNA structure: bioinformatic analysis. *Curr. Opin. Microbiol.*, **10**, 279–285.

Jung,J. and Lee,B. (2000) Use of residue pairs in protein sequence-sequence and sequence-structure alignments. *Protein Sci.*, **9**, 1576–1588.

Kawabata,T. (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.

Kim,J.N. and Breaker,R.R. (2008) Purine sensing by riboswitches. *Biol. Cell*, **100**, 1–11.

Konagurthu,A.S. *et al.* (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst.*, **D60**, 2256–2268.

Lackner,P. *et al.* (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.

Lilly,D.M.J. and Eckstein,F. (2008) *Ribozymes and RNA Catalysis. RSC Biomolecular Sciences*. RSC Publishing, Cambridge, UK, p. 318.

Lisewski,A.M. and Lichtarge,O. (2006) Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Res.*, **34**, E152.

Lukavsky,P.J. *et al.* (2003) Structure of HCV IRES domain II determined by NMR. *Nat. Struct. Mol. Biol.*, **10**, 1033–1038.

Maiorov,V.N. and Crippen,G.M. (1994) Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.*, **235**, 625–634.

Mandal,G. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, **5**, 451–463.

Marchler-Bauer,A. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucl. Acids Res.*, **33**, D192–D196.

Margraf,T. *et al.* (2009) The SALAMI protein structure search server. *Nucleic Acids Res.*, **37**, W480–W484.

Mikkelsen,N.E. *et al.* (2001) Aminoglycoside binding displaces a divalent metal ion in a tRNA-neomycin B complex. *Nat. Struct. Mol. Biol.*, **8**, 510–514.

Montange,R.K. and Batey,R.T. (2008) Riboswitches: emerging themes in RNA structure and function. *Ann. Rev. Biophys.*, **37**, 117–133.

Murthy,V.L. *et al.* (1999) A complete conformational map for RNA. *J. Mol. Biol.*, **291**, 313–327.

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.
Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.
Numata,T. *et al.* (2006) Snapshots of tRNA sulphuration via an adenylated intermediate. *Nature*, **442**, 419–424.
Ochagavia,M.E. and Wodak,H. (2004) Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins*, **55**, 436–454.
Oldfield,T.J. (2007) CAALIGN: a program for pairwise and multiple protein-structure alignment. *Acta. Cryst.*, **D63**, 514–525.
Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.
Ortiz,A.R. *et al.* (2002) MAMMOTH (Matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
Rahrig,R.R. *et al.* (2010) R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, **26**, 2689–2697.
Reijmers,T.H. *et al.* (2001) The influence of different structure representations on the clustering of an RNA nucleotides data set. *J. Chem. Inf. Comput. Sci.*, **41**, 1388–1394.
Rose,P.W. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison. *Proteins*, **14**, 309–323.
Sankoff,D. (1985) Simultaneous solution of the Rna folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
Schenk,G. *et al.* (2008) Protein sequence and structure alignments within one framework. *Algorithms Mol. Biol.*, **3**, 4.
Scott,W.G. (1999) RNA structure, metal ions, and catalysis. *Curr. Opin. Chem. Biol.*, **3**, 705–709.
Scott,W.G. and Klug,A. (1996) Ribozymes: structure and mechanism in RNA catalysis. *Trends Biochem. Sci.*, **21**, 220–224.
Shapiro,J. and Brutlag,D. (2004) FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res.*, **32**, W536–W541.
Shatsky,M. *et al.* (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.
Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Strobel,S.A. and Cochrane,J.C. (2007) RNA catalysis: ribozymes, ribosomes, and riboswitches. *Curr. Opin. Chem. Biol.*, **11**, 636–643.
Subbiah,S. *et al.* (1993) Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, **3**, 141–148.
Suyama,M. *et al.* (1997) Comparison of protein structures using 3D profile alignment. *J. Mol. Evol.*, **44**, S163–S173.
Taubig,H. *et al.* (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.
Wadley,L.M. *et al.* (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol.*, **372**, 942–957.
Wang,C.W. *et al.* (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.*, **38**, W340–W347.
Waugh,D.S. and Pace,N.R. (1986) Catalysis by RNA. *BioEssays*, **4**, 56–61.
Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, 680–691.
Winkler,W.C. (2005) Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.*, **9**, 594–602.
Zhu,J.H. and Weng,Z.P. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
Zuker,M. and Somorjai,R.L. (1989) The alignment of protein structures in three dimensions. *Bull. Math. Biol.*, **51**, 55–78.