# iPARTS: An Improved Tool of Pairwise Alignment of RNA Tertiary Structures

Chih-Wei Wang[1†] Kun-Tze Chen[1,†] Chin Lung Lu[1,2*]

[1]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C.
[2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C.

## METHODS

As pointed out by Duarte and Pyle (1), the two-dimensional (2D) $\eta$-$\theta$ plot is a Ramachandran-like diagram that can provide us a graphic representation of quantitatively distinct structural features for analyzing and modeling RNA three-dimensional (3D) structures. Particularly, they showed that on this $\eta$-$\theta$ plot, clusters of nucleotides with similar $\eta$ and $\theta$ pseudo-torsional angles have similar conformational properties and *vice versa*. To depict this $\eta$-$\theta$ plot, we prepared a dataset that includes non-redundant crystal structures with minimum resolution of 3.0 Å from the PDB database (2). This dataset finally contains 117 crystal RNA structures, particularly including 74 structures used by Wadley *et al.* (3), with 9,527 nucleotides in total. We then used AMIGOS that was developed by Duarte and Pyle (1) to calculate the $\eta$ and $\theta$ pseudo-torsion angles for all non-terminal nucleotides (9,267 nt in total) from all RNA molecules in the above dataset and plotted these calculated pseudo-torsion angles on the axes of a 2D plot as illustrated in Figure 1.

Instead of using the vector quantization (VQ) approach as done in our previous work (4), we here applied the so-called *affinity propagation* (AP) clustering algorithm, introduced by Frey and Dueck recently (5), to classify all the non-terminal nucleotides in our prepared dataset according to their $\eta$ and $\theta$ pseudo-torsion angles. Like $k$-means clustering algorithms, the VQ approaches usually find locally optimum clusters and are sensitive to outliers and noise (6), although it can be used to classify high dimensional data points. Besides, the VQ methods need to keep track of a fixed set of candidate centers (or exemplars) while searching for good solutions. Basically, the AP algorithm is an *exemplar-based* clustering method for approximately solving the *exemplar learning problem* that aims to identify a set of data points as exemplars and assign every data point to an exemplar so as to maximize a fitness function, where notably the exemplar learning problem has been shown to be NP-hard (7). Denote the input data points by $x_1, x_2, \ldots, x_n$, the exemplar assigned to $x_i$ by $c_i$, and the similarity

†These two authors contributed equally to this work and should be considered co-first authors.
*To whom correspondence should be addressed. Tel: +886-3-5712121 ext. 56949; Fax: +886-3-5729288; Email: cllu@mail.nctu.edu.tw

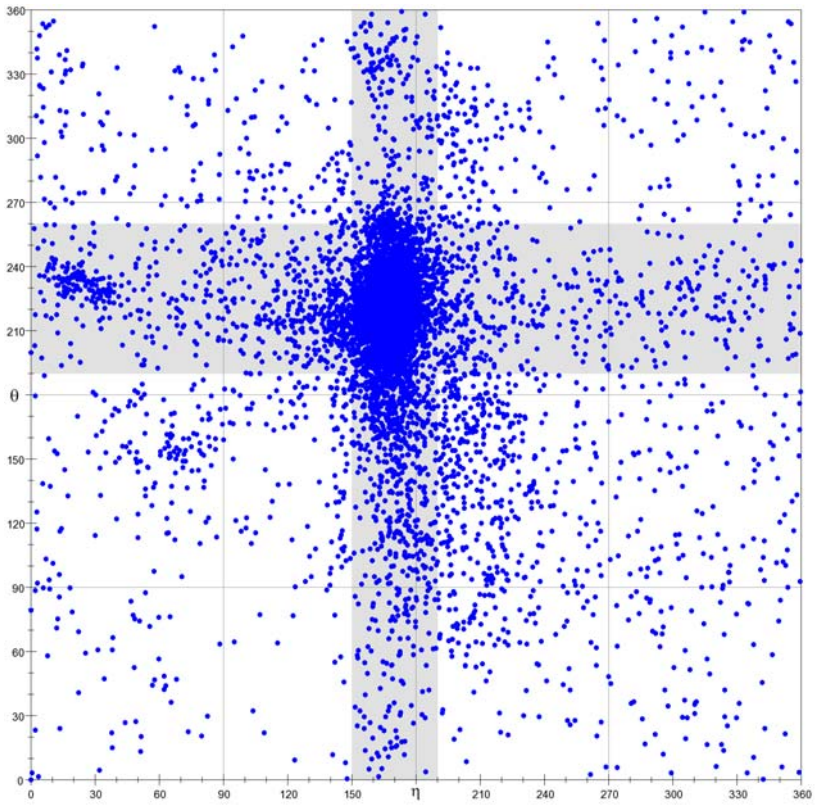


Figure 1: An $\eta$-$\theta$ plot of all non-terminal nucleotides from all RNA molecules in the dataset, where the intersection of the perpendicular gray regions ($150° \leq \eta \leq 190°$ and $190° \leq \theta \leq 260°$) is designated the helical region.

between $x_i$ and $c_i$ by $s(x_i, c_i)$. Then the *fitness function* mentioned above is defined to be $\sum_{i=1}^{n} s(x_i, c_i)$. Notably, if $x_i$ is an exemplar (i.e., $c_i = x_i$), then this fitness function includes the term $s(x_i, c_i)$.

Basically, the AP algorithm operates by simultaneously considering all input data points $x_1, x_2, \ldots, x_n$ as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges. For simplicity, the similarity $s(x_i, x_j)$ between two points $x_i$ and $x_j$ is also denoted as $s(i, j)$. In each iteration, two kinds of messages, called responsibility and availability, respectively, are exchanged between data points. The *responsibility* $r(i, k)$, which is sent from point $x_i$ to point $x_k$, indicates the accumulated evidence for how proper it would be for $x_k$ to serve as the exemplar of $x_i$ with taking into account other potential exemplars for $x_i$. Before being sent, the value of $r(i, k)$ is updated according to the following rule: $r(i, k) = s(i, k) - \max_{k':k' \neq k} \{a(i, k') + s(i, k')\}$. The *availability* $a(i, k)$, which is sent from point $x_k$ to point $x_i$, indicates the accumulated evidence for how proper it would be for $x_i$ to choose $x_k$ as its exemplar with taking into account the support from other points that $x_k$ should be an exemplar. The value of $a(i, k)$ is updated as follows: if $i \neq k$, then $a(i, k) = \min\{0, r(k, k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i', k)\}\}$; otherwise, $a(k, k) = \sum_{i':i' \neq k} \max\{0, r(i', k)\}$. It should be noted that numerical oscillations may arise in some circumstances when updating the above two messages. To avoid such oscillations, therefore, each message is set to $\lambda$ times its value from the previous iteration plus $1 - \lambda$ times its currently prescribed updated value,

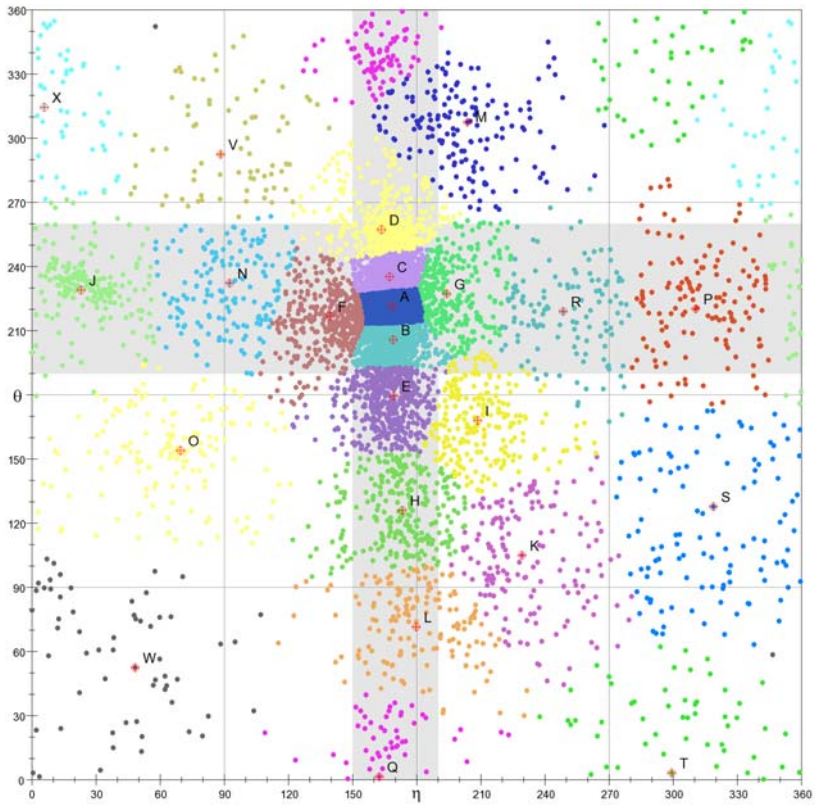


Figure 2: Twenty-three clusters classified by the AP algorithm.

where $\lambda$ is a *damping factor* whose value is between 0 and 1. In this study, we used a default damping factor of $\lambda = 0.5$. The above message-passing scheme is therefore referred to as *affinity propagation*. At any point during the affinity propagation, responsibilities and availabilities are combined to identify exemplars. That is, for data point $x_i$, the $k$ that maximizes $r(i,k) + a(i,k)$ indicates that $x_k$ is the exemplar of $x_i$. Finally, the message-passing procedure may be terminated after a fixed number of iterations (or after the changes in the messages fall below a threshold or the local decisions stay constant for some number of iterations).

Note that each data point in this study corresponds to a non-terminal nucleotide of an RNA 3D structure on the 2-dimensional $\eta$-$\theta$ plot and, therefore, the similarity between data point $x_i$ and its exemplar $c_i$ defined in this study is the negative squared Euclidean distance (that is, $s(x_i, c_i) = -\|x_i - c_i\|^2$), if $x_i \neq c_i$. As to $x_i = c_i$, the value of $s(x_i, x_i)$ represents the *a priori preference* for $x_i$ to serve as an exemplar and, therefore, it is not calculated in the same way as $s(x_i, x_k)$, where $x_i \neq x_k$, because it does not represent an assignment similarity. As suggested in (5), the preference values can be set to a global (shared) value, or customized for particular data points. Particularly, moreover, high values of the preferences will cause the AP algorithm to find many exemplars (clusters), while low values will lead to a small number of exemplars. Here, we set a global value to $s(x_i, x_i)$ for all $1 \leq i \leq n$ such that a total of 9,267 non-terminal nucleotides on the $\eta$-$\theta$ plot is classified into 23 conformation clusters, as was illustrated in Figure 2. The 3D conformations of these 23 exemplar nucleotides are shown in Figure 3.

For our purpose of transforming RNA 3D structures into 1D sequences, we further assigned a letter to each of 23 clusters, as named in Table 1. We used the set of these 23 letters as a

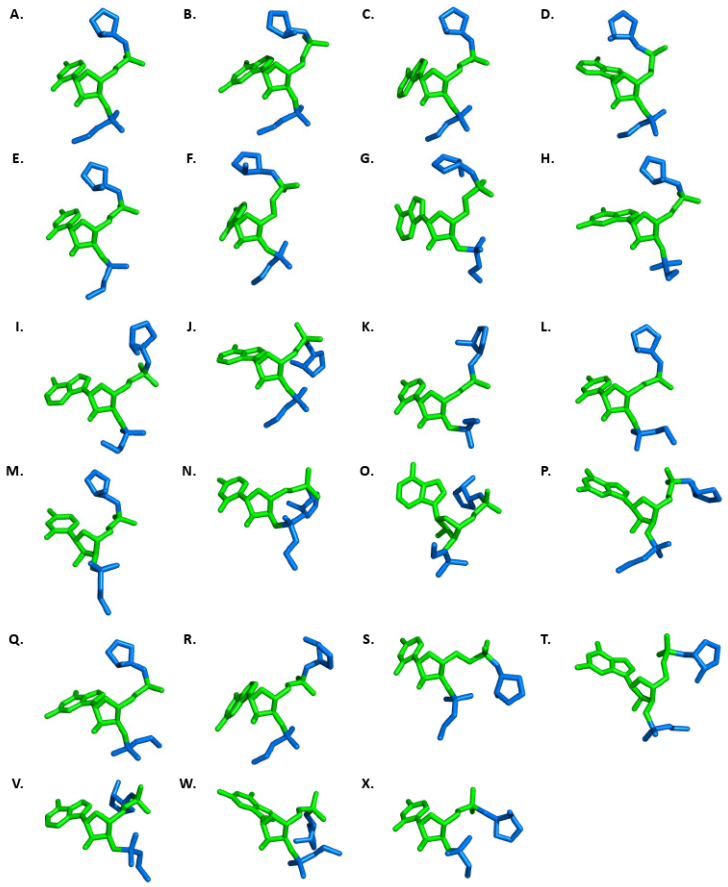


Figure 3: Three-dimensional conformations of 23 exemplar nucleotides, where the exemplar nucleotides are shown in green, whereas the portions of the previous and next nucleotides that affect the pseudo-torsions are shown in blue.

*structural alphabet* (SA) and encoded RNA 3D structures as 1D sequences of SA letters by using the *nearest neighbor rule*, by which each nucleotide in an RNA molecule is assigned with the letter of the cluster whose exemplar (center) is nearest to the nucleotide being encoded. Next, like ordinary nucleotide (or amino acid) sequences, we applied methods of classical sequence alignments, such as global (8), semiglobal (9), local (10) and normalized local (11) alignments, to the SA-encoded 1D sequences of two RNA 3D structures for determining their structural similarities. In this study, we chose 23 as the number of the clusters on the $\eta$-$\theta$ plot based on the following two reasons. First, over 60% of nucleotides on the $\eta$-$\theta$ plot fall within the helical region (defined by the intersection of the two perpendicular gray regions in Figure 1). As illustrated in Figure 2, the helical region is partitioned into four clusters when $N = 23$. However, if $N = 46$, then an overpartitioning (with more than 10 clusters) in this helical region can be observed. This overpartitioning results was actually due to the fact that the helical region is so highly populated in the dataset of currently collected RNA structures that any clustering algorithm may tend to divide it into a lot of clusters. In fact, according to our experiments (data not shown), the value of the AUC obtained using our testing dataset with $N = 46$ is not better than that with $N = 23$. Second, choosing $N = 23$ will allows one to apply BLAST, the most widely used tool of sequence homology search,

Table 1: The structural alphabet of 23 conformational clusters classified by the AP algorithm with their associated letters and the $\eta$ and $\theta$ pseudo-torsion angles of their exemplars.

| No. | Letter | $(\eta, \theta)$ | No. | Letter | $(\eta, \theta)$ | No. | Letter | $(\eta, \theta)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | A | (168.7, 221.4) | 9 | I | (208.5, 167.9) | 17 | Q | (162.5, 1.4) |
| 2 | B | (169.1, 205.7) | 10 | J | (23.1, 228.9) | 18 | R | (248.7, 218.9) |
| 3 | C | (167.3, 235.1) | 11 | K | (229.4, 104.9) | 19 | S | (318.9, 127.7) |
| 4 | D | (163.7, 257.1) | 12 | L | (179.8, 71.4) | 20 | T | (299.4, 3.2) |
| 5 | E | (169.4, 179.5) | 13 | M | (203.8, 307.5) | 21 | V | (88.3, 292.5) |
| 6 | F | (139.7, 216.6) | 14 | N | (92.5, 232.2) | 22 | W | (48.3, 52.5) |
| 7 | G | (194.1, 227.2) | 15 | O | (69.6, 153.8) | 23 | X | (5.9, 314.3) |
| 8 | H | (173.3, 125.9) | 16 | P | (310.6, 220.1) | | | |

for efficiently performing the structurally similar search on the database consisting of the SA-encoded sequences of RNA 3D structures.

For the accuracy of aligning two SA-encoded sequences, we derived a $23 \times 23$ log-odds matrix for SA-letter substitution using the statistical method proposed by Henikoff and Henikoff (12). Let $\{a_1, a_2, \ldots, a_{23}\}$ denote the structural alphabet of 23 SA letters and $f_{ij}$ be the observed substitution frequency of SA-letter pair $(a_i, a_j)$. Then the relative frequency $q_{ij}$ of an SA-letter pair $(a_i, a_j)$ is $q_{ij} = \dfrac{f_{ij}}{\sum_{k=1}^{23} \sum_{l=1}^{k} f_{kl}}$, and the frequency of occurrence of SA letter $a_i$ in an SA-letter pair $(a_i, a_j)$ is $p_i = q_{ii} + \dfrac{\sum_{k=1, k\neq i}^{23} q_{ik}}{2}$. The expected frequency $e_{ij}$ for a substitution between two SA-letters $(a_i, a_j)$ is $p_i p_j$ for $i = j$ and $p_i p_j + p_j p_i = 2 p_i p_j$ for $i \neq j$. The logarithm of the odds matrix is finally calculated by $score(a_i, a_j) = \lambda \log_2 \left( \frac{q_{ij}}{e_{ij}} \right)$, where $\lambda$ is a positive scale factor.

For the purpose of constructing this BLOSUM-like matrix, a dataset of structurally similar RNA pairs was obtained from the DARTS database (13), which used an automated method to classify 1,333 RNA tertiary structures into 244 groups of highly identical structures, and the SCOR database (14, 15), which organized many RNA structural motifs in a hierarchical classification system similar to the SCOP database for protein domains. From the initial dataset of 1,333 high-resolution RNA 3D structures, the DARTS database first selected 244 representative structures based on RNA sequence and 3D structure resemblances and then marked each of the remaining structures as either a highly identical structure or a highly identical fragment of a representative structure. A highly identical structure is defined as a structure that is globally almost identical (i.e., with at least 90% sequence or 3D structure identity) to some other structure of similar size (i.e., size ratio[1] is between 1 and 1.5), while a highly identical fragment is defined as a structure that is almost identical to only a small substructure of a larger structure (i.e., size ratio is greater than 1.5). Note that 101 out of 244 representative structures have no highly identical structure. For our purpose, we used only the remaining 143 representative structures and their highly identical structures to construct our BLOSUM-like matrix. In addition, a set of structurally similar RNA motif pairs was obtained from the SCOR database based on the following criteria: (1) motifs must belong

[1]The size ratio is defined as the number of nucleotides of the bigger structure divided by the number of nucleotides of the smaller structure.

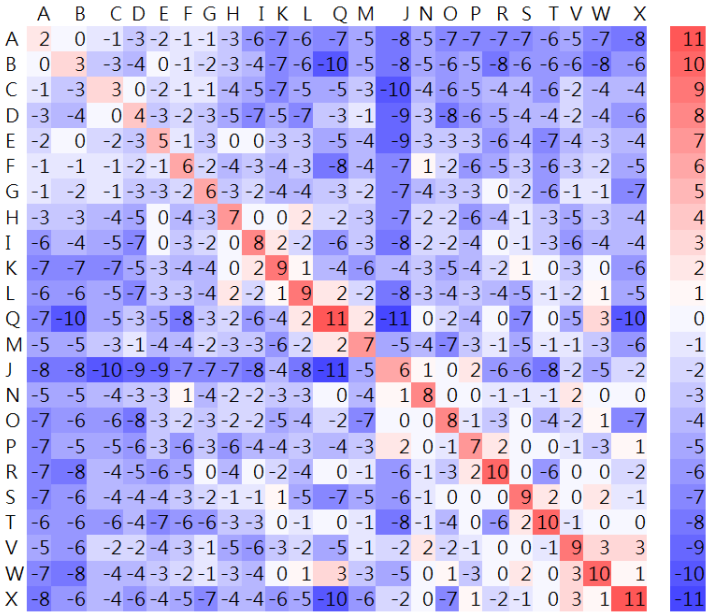


Figure 4: The BLOSUM-like substitution matrix for the 23 SA-letters we derived in this study.

to a structural family, (2) motifs must have length greater than 3 nt, (3) motifs must have specified starting and ending positions in the chain, and (4) motif pairs must have no 100% sequence identity. In total, 3,391 RNA structural alignment pairs from 143 DARTS groups of 686 high-resolution RNA 3D structures and 430,628 RNA motif pairs from 334 SCOR classes of 6,220 structural motifs were analyzed, which together accounted for 8,500,322 SA-letter pairs. The $\lambda$ value used in this study was set to 1.6 for the best performance, by testing various values ranging from 1 to 2. Figure 4 illustrates the BLOSUM-like substitution matrix for the 23 SA-letters we derived in this study.

As we done before (4), we implemented four different types of pairwise alignments that are global (8), semiglobal (9), local (10) and normalized local (11) alignments for a variety of practical applications. Moreover, a grid-like search procedure was performed to optimize the parameters of open and extension gap penalties by varying the open gap penalty from $-15$ to $-1$ in steps of 1 and the extension gap penalty from $-3$ to $-0.5$ in steps of 0.5. It is worth mentioning here again that the Smith-Waterman algorithm (10) for the local alignment was originally designed to remove non-similar initial and terminal fragments but not non-similar internal fragments in a sequence alignment, resulting in a so-called *mosaic effect* by including poor internal fragments in a local alignment (11). Such a mosaic effect still can be observed in local alignments of two RNA 3D structures, as demonstrated on the help page of our iPARTS server. To eliminate this mosaic effect, we implemented the algorithm proposed by Arslan *et al.* (11) to solve the so-called *normalized local alignment problem*, which aims to find the subsequences, say $I$ and $J$, of two given sequences that maximizes $S(I,J)/(|I|+|J|)$ among all subsequences $I$ and $J$ with $|I|+|J| \geq T$, where $S(I,J)$ is the alignment score between $I$ and $J$, and $T$ is a threshold for the minimal overall length of $I$ and $J$. Usually, an alignment should be sufficiently long to be biologically meaningful. Therefore, the above length constraint of $|I|+|J| \geq T$ is necessary, since length normalization in the normalized local alignment problem favors short local alignments. The user can vary the value of $T$ to control the result of optimal normalized local alignment. If $T$ is small, the optimal normalized local alignment tends to

be short; otherwise, it tends to be a long local alignment that may contain some non-similar internal fragments.

# References


1. Duarte, C. M. and Pyle, A. M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *Journal of Molecular Biology,* **284**, 1465–1478 [PubMed:9878364] [doi:10.1006/jmbi.1998.2233].

2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Research,* **28**, 235–242 [PubMed:10592235] [PubMed Central:PMC102472] [doi:10.1093/nar/28.1.235].

3. Wadley, L. M., Keating, K. S., Duarte, C. M., and Pyle, A. M. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *Journal of Molecular Biology,* **372**, 942–957 [PubMed:17707400] [PubMed Central:PMC2720064] [doi:10.1016/j.jmb.2007.06.058].

4. Chang, Y.-F., Huang, Y.-L., and Lu, C. L. (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Research,* **36**, W19–W24 [PubMed:18502774] [PubMed Central:PMC2447761] [doi:10.1093/nar/gkn327].

5. Frey, B. J. and Dueck, D. (2007) Clustering by passing messages between data points. *Science,* **315**, 972–976 [PubMed:17218491] [doi:10.1126/science.1136800].

6. Xu, R. and Wunsch, D., I. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks,* **16**, 645–678 [PubMed:15940994] [doi:10.1109/TNN.2005.845141].

7. Charikar, M., Guha, S., Tardos, E., and Shmoys, D. B. (2002) A constant-factor approximation algorithm for the $k$-median problem. *Journal of Computer and System Sciences,* **65**, 129–149 [doi:10.1145/301250.301257].

8. Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Evolution,* **48**, 443–453 [PubMed:5420325] [doi:10.1016/0022-2836(70)90057-4].

9. Setubal, J. and Meidanis, J. (1997) *Introduction to Computational Molecular Biology*, PWS Publishing Company, Boston.

10. Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology,* **147**, 195–197 [PubMed:7265238] [doi:10.1016/0022-2836(81)90087-5].

11. Arslan, A. N., Eğecioğlu, O., and Pevzner, P. A. (2001) A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics,* **17**, 327–337 [PubMed:11301301] [doi:10.1093/bioinformatics/17.4.327].

12. Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America,* **89**, 10915–10919 [PubMed:1438297] [PubMed Central:PMC50453] [doi:10.1073/pnas.89.22.10915].

13. Abraham, M., Dror, O., Nussinov, R., and Wolfson, H. J. (2008) Analysis and classification of RNA tertiary structures. *RNA,* **14**, 2274–2289 [PubMed:18824509] [PubMed Central:PMC2578864] [doi:10.1261/rna.853208].

14. Klosterman, P. S., Tamura, M., Holbrook, S. R., and Brenner, S. E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Research,* **30**, 392–394 [PubMed:11752346] [PubMed Central:PMC99131] [doi:10.1093/nar/30.1.392].

15. Tamura, M., Hendrix, D. K., Klosterman, P. S., Schimmelman, N. R., Brenner, S. E., and Holbrook, S. R. (2004) SCOR: structural classification of RNA, version 2.0. *Nucleic Acids Research,* **32**, D182–D184 [PubMed:14681389] [PubMed Central:PMC308814] [doi:10.1093/nar/gkh080].