

Sequence analysis

TOPAS: network-based structural alignment of RNA sequences

Chun-Chi Chen^{1,2}, Hyundoo Jeong³, Xiaoning Qian^{1,2} and Byung-Jun Yoon^{1,2,*}

¹Department of Electrical and Computer Engineering, ²TEES-AgriLife Center for Bioinformatics & Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA and ³Department of Electronic Engineering, Chosun University, Gwangju 61452, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 10, 2017; revised on December 7, 2018; editorial decision on December 29, 2018; accepted on January 4, 2019

Abstract

Motivation: For many RNA families, the secondary structure is known to be better conserved among the member RNAs compared to the primary sequence. For this reason, it is important to consider the underlying folding structures when aligning RNA sequences, especially for those with relatively low sequence identity. Given a set of RNAs with unknown structures, simultaneous RNA alignment and folding algorithms aim to accurately align the RNAs by jointly predicting their consensus secondary structure and the optimal sequence alignment. Despite the improved accuracy of the resulting alignment, the computational complexity of simultaneous alignment and folding for a pair of RNAs is $O(N^6)$, which is too costly to be used for large-scale analysis.

Results: In order to address this shortcoming, in this work, we propose a novel network-based scheme for pairwise structural alignment of RNAs. The proposed algorithm, TOPAS, builds on the concept of topological networks that provide structural maps of the RNAs to be aligned. For each RNA sequence, TOPAS first constructs a topological network based on the predicted folding structure, which consists of sequential edges and structural edges weighted by the base-pairing probabilities. The obtained networks can then be efficiently aligned by using probabilistic network alignment techniques, thereby yielding the structural alignment of the RNAs. The computational complexity of our proposed method is significantly lower than that of the Sankoff-style dynamic programming approach, while yielding favorable alignment results. Furthermore, another important advantage of the proposed algorithm is its capability of handling RNAs with pseudoknots while predicting the RNA structural alignment. We demonstrate that TOPAS generally outperforms previous RNA structural alignment methods on RNA benchmarks in terms of both speed and accuracy.

Availability and implementation: Source code of TOPAS and the benchmark data used in this paper are available at <https://github.com/bjyoontamu/TOPAS>.

Contact: bjyoon@ece.tamu.edu

1 Introduction

RNA sequence alignment methods play important roles in comparative genome analysis, especially for accelerating the discovery of novel noncoding RNAs (ncRNAs) as well as studying their functions and structures. Sequence alignment techniques provide effective means of quantitatively evaluating the similarity between different

RNA sequences, which can be used for computational identification of homologous RNAs that belong to the same functional family. As revealed in various comparative studies of RNAs, for many RNA families, the secondary structure of the RNAs tends to be better conserved among the members compared to their primary sequence (Freyhult *et al.*, 2006; Glotz *et al.*, 1981; Johnsson *et al.*, 2014; Raué

et al., 1988; Zwieb et al., 1981). As a consequence, it is critical for RNA sequence alignment techniques to sensibly incorporate the underlying RNA secondary structure into the alignment process in order to obtain accurate alignment results.

RNA is a single-stranded molecule that is composed of a chain of nucleotides with four different types of bases A, C, G and U. Due to the base-pairing interactions between different bases, an RNA may fold onto itself thereby forming a complicated structure. Although the prediction of the native three-dimensional structure of an RNA is challenging, the two-dimensional RNA secondary structure is amenable to mathematical analysis and computational prediction (Flamm et al., 2000; Greenleaf et al., 2008; Tinoco and Bustamante, 1999), thanks to the quasi-hierarchical nature of the folding structure. Canonical Watson-Crick pairs A–U and C–G are typically formed between bases and wobble pairs G–U are also frequently observed in RNA secondary structure. In a typical RNA secondary structure, the base-pairs appear in a nested manner such that two base-pairs (i_1, i_2) and (j_1, j_2) — i_k and j_k referring to base locations—either satisfy $i_1 < i_2 < j_1 < j_2$ or $i_1 < j_1 < j_2 < i_2$. An RNA secondary structure that contains non-nested crossing base-pairs is called a *pseudoknot*. In general, pseudoknots make computational analysis of RNAs—e.g. structure prediction and structural alignment—significantly more challenging.

Probably the first—and arguably also the most influential—method that has been proposed for structural alignment of RNAs with unknown structures is the algorithm proposed by Sankoff, which simultaneously solves the sequence alignment problem and the consensus RNA secondary structure prediction problem through a dynamic programming approach (Sankoff, 1985). Several different implementations of Sankoff-style algorithms exist to date for RNA structural alignment. For example, *Dynalign* and *Foldalign* are popular methods that use thermodynamic models to evaluate the free energy of a potential secondary structure and utilizes dynamic programming to find the structure with the lowest free energy that is common to the RNAs to be aligned (Fu et al., 2014; Harmanci et al., 2008; Havgaard et al., 2005; Mathews and Turner, 2002; Sundfeld et al., 2016). Another method, called *PARTS*, introduces a pseudo-free energy model based on the base-pairing and alignment probabilities to find the best structural alignment that maximizes the joint probability (Harmanci et al., 2008). Although Sankoff-style algorithms generally yield more accurate and reliable alignment results compared to alignment techniques that solely rely on sequence similarity, their main downside is the sharp increase in complexity. For example, the complexity of the original Sankoff algorithm for the structural alignment of two RNA sequences of length N is $O(N^6)$ in time and $O(N^4)$ in space (i.e. memory) (Hamada et al., 2009). The extremely high complexity of the original Sankoff algorithm makes it impractical for large-scale genome analysis, and a number of simplified variations of Sankoff-style algorithms have been developed to efficiently solve the RNA structural alignment problem (Gardner et al., 2005; Will et al., 2007, 2015). One such example is *PMcomp*, which uses base-pairing probabilities as a lightweight energy model and imposes restrictions on the matching base-pairs to reduce the overall computational complexity to $O(N^4)$ (Hofacker et al., 2004). *LocARNA* adopts a light-weight energy model like *PMcomp*, and it simplifies the dynamic programming approach by incorporating the sparse property of base-pairing (Will et al., 2007). *SPARSE* (Will et al., 2015) and *RAF* (Chuong et al., 2008) further improve the alignment speed achieving quadratic time complexity. In order to improve the alignment speed, *SPARSE* (Will et al., 2015) utilizes ensemble-based sparsification and *RAF* (Chuong et al., 2008) exploits the fact that the probable alignment

edges in the sequence alignment tend to be sparse. All the aforementioned Sankoff-style algorithms utilize energy models (or pseudo energy models based on base-pairing probability) and aim to find the optimal structural alignment through dynamic programming with various simplifications and constraints to reduce the overall complexity.

In contrast to the Sankoff-style algorithms, we propose a novel approach for RNA structural alignment by adopting the concept of *topological network* that integrates the sequence and structural information of the RNAs to be aligned. Topological networks provide convenient ways of concisely representing the complicated interactions and relationships among parts or entities that form a larger whole. Well known examples of such networks are the protein–protein interaction (PPI) networks and the co-expression networks. In a PPI network, nodes correspond to proteins and edges between nodes represent interactions between the corresponding proteins. In a co-expression network, nodes typically correspond to genes and the presence of an edge between two nodes imply that there exists a significant correlation between the expression levels of the connected genes. In recent years, there has been growing interest in developing efficient computational tools for comparative analysis of large-scale biological networks (Yoon et al., 2012), especially for the comparison and alignment of PPI networks (Jeong and Yoon, 2015; Jeong et al., 2016; Liao et al., 2009; Sahraeian and Yoon, 2013; Singh et al., 2008). By comparing the PPI networks (Gursoy et al., 2008) that capture the physical interactions among proteins in different species, PPI network alignment aims to predict the functional correspondence between proteins across networks and identify network modules that may be conserved in different species. In order to obtain accurate alignment results that are biologically meaningful, network alignment methods generally consider both the *sequence similarity* between proteins and the *topological similarity* between networks during the alignment process (Yoon et al., 2012).

In this paper, we propose a novel RNA structural alignment algorithm called TOPAS (TOPological network-based Alignment of Structural RNAs) that builds on the concepts of topological networks and network alignment. TOPAS first constructs a topological network for each RNA sequence such that the network captures the sequence and structural properties of the RNA. The constructed topological networks are then aligned by utilizing an efficient network alignment technique, which leads to an accurate structural alignment that seamlessly integrates the sequence similarity and the structural similarity between the given RNAs. The network-based approach that is adopted by TOPAS for representing and aligning RNAs makes the algorithm very flexible, allowing it to handle RNAs with arbitrary structures, including pseudoknots. We compare our proposed algorithm TOPAS with several well-known RNA structural alignment algorithms and show that TOPAS outperforms previous algorithms in term of speed and accuracy.

2 Materials and methods

RNA structural alignment aims to predict an accurate alignment of a given set of RNAs, such that their common folding structures are faithfully aligned to each other. For fast and accurate structural alignment of RNAs, we propose an innovative network-based approach. In the proposed approach, we first construct a topological network for each RNA that provides a graphical representation of its sequence composition as well as its potential secondary structure. Next, the constructed topological networks are efficiently aligned using a network alignment technique, where the resulting network

alignment gives rise to the structural alignment of the corresponding RNAs. Recent studies in comparative network analysis (Singh *et al.*, 2008; Yoon *et al.*, 2012) have shown that accurate network alignment results can be attained by sensibly integrating the similarity between nodes across networks as well as the topological similarity between the networks. In a similar way, network alignment techniques can be used to reliably align topological networks representing RNA sequences and their folding structures, thereby predicting an accurate structural alignment of the RNAs that incorporates both the sequence similarity and the structural similarity between the RNAs. In what follows, we discuss the two main steps of the proposed RNA structural alignment algorithm TOPAS—i.e. construction of the topological networks based on the given RNAs and finding the RNA structural alignment through the alignment topological networks—in more details.

2.1 Topological network construction from RNA sequences

For each of the RNAs to be aligned, we first construct a topological network that provides a graphical representation of the RNA sequence and its potential folding structure. The backbone of the topological network is formed based on the primary sequence of the RNA, where every nucleotide in the RNA is represented as a node in the topological network. Next, nodes that can form a base-pair in the RNA folding structure are also connected by a weighted edge, where the weight is determined by the corresponding base-pairing probability. The base-pairing probabilities can be estimated by using thermodynamic equilibrium models with experimentally determined parameters (Mathews, 2004; McCaskill, 1990; Turner and Mathews, 2010), which are widely used for RNA structure prediction. In order to keep the topological network sparse by keeping only the edges that correspond to reliable base-pairs, edges with base-pairing probabilities that are lower than a threshold P_{Th} are removed from the network. This has the effect of reducing the overall cost of network alignment and enhancing the accuracy of the final alignment results. The sequence similarity between nodes across different networks is estimated by using a pair hidden Markov model (pair-HMM) (Mount, 2009; Yoon, 2009). Alignment probabilities between nucleotides are estimated through the *forward-backward* algorithm based on the given pair-HMM, and their normalized bit scores are used as a measure of node similarity across networks, which incorporate the sequence similarity between the corresponding RNAs. The detailed network construction process and the proposed network-based RNA structural alignment algorithm TOPAS are elaborated in Section 2.2.

2.2 RNA structural alignment based on topological networks

Let $G_n = (V_n, E_n)$ be the n th topological network. V_n is the set of nodes in the network, where each node corresponds to a nucleotide in the n th sequence. E_n is the set of weighted edges between the nodes, where each edge reflects that the connected nodes may form a base-pair in the RNA with a base-pairing probability exceeding the threshold P_{Th} . Given two topological networks G_1 and G_2 , we aim to accurately align the networks by integrating their node similarity and topological similarity, thereby predicting an accurate structural alignment of the RNAs represented by the networks. Let R be the overall similarity between the two networks, where the element $R(a, b)$ is the overall similarity score between two nodes $a \in V_1$ and $b \in V_2$. To compute the overall similarity R , we integrate the following three types of similarities: (i) *structural similarity*

R_S between the underlying secondary structures of the two RNAs; (ii) *connected similarity* R_C for consecutive node (nucleotide) alignment; and (iii) *sequence similarity* R_E for nucleotide-level sequence resemblance. The structural similarity R_S and the connected similarity R_C reflect the topological similarity between the networks G_1 and G_2 , while R_E reflects the similarity between nodes in the two networks (i.e. the sequence-level similarity between the corresponding RNAs).

In order to compute R , we adopt a similar approach that was originally used in the *IsoRank* network alignment algorithm (Singh *et al.*, 2008). In *IsoRank*, two nodes in different networks are likely to be matched (or aligned) to each other if their neighbors are also well matched to one another. This gives rise to a similarity diffusion scheme that can be iteratively applied until convergence, thereby computing the overall similarity scores. Following similar principles, we compute the structural similarity $R_S(a, b)$ and connected similarity $R_C(a, b)$ by

$$R_S(a, b) = \sum_{\substack{c \in N_{G_1}(a) \\ d \in N_{G_2}(b)}} \frac{P_{S_1}(a, c)P_{S_2}(b, d)}{D(c)D(d)} R(c, d) \quad (1)$$

and

$$R_C(a, b) = \frac{1}{2} (R(a-1, b-1) + R(a+1, b+1)) \quad (2)$$

where $N_{G_n}(x)$ is defined as the set of connected neighbors of the node x in the topological network G_n . $P_{S_1}(a, c)$ is the base-pairing probability for the node pair at (a, c) in the network G_1 , and $P_{S_2}(b, d)$ is the base-pairing probability for the node pair at (b, d) in the network G_2 . $D(c) = \sum_{u \in N_{G_1}(c)} P_{S_1}(u, c)$ and $D(d) = \sum_{v \in N_{G_2}(d)} P_{S_2}(v, d)$ are the weighted degrees of nodes c and d , respectively. These are illustrated in Figure 1.

The structural similarity R_S measures the topological similarity between nodes in different topological networks based on the base-pairing probabilities in the respective RNAs, such that nodes (nucleotides) involved in conserved base-pairs are likely to be aligned in the network alignment (hence the RNA structural alignment). Next, the connected similarity R_C is inspired by the message-passing based sequence alignment scheme proposed in Yoon (2014). R_C is computed based on the principle that two nucleotides in two RNA sequences are likely to be aligned if their neighboring nucleotides are also aligned in the RNA sequence alignment. As mentioned before, both R_S and R_C attempt to estimate the topological

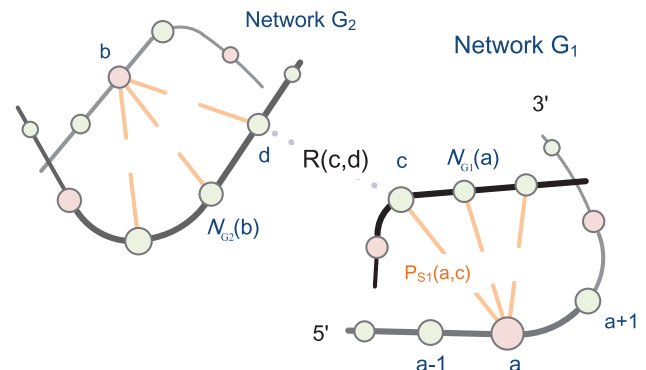


Fig. 1. Illustration of topological networks for RNA structural alignment. $R(c, d)$ denotes the pairwise similarity between nodes at position c in network G_1 and position d in network G_2 . $P_{S_1}(a, c)$ is the base-pairing probability for nodes at position (a, c) in network G_1 . $N_{G_1}(a)$ denotes the set of the neighbors of node at position a if there exists the base-pairing interaction in network G_1 .

similarity between the given networks by capturing the similarity between the neighborhoods of two nodes that belong to different networks.

Finally, the overall similarity score R is computed by combining the structural similarity R_S , the connected similarity R_C , and the sequence similarity R_E as follows

$$R = (\alpha \cdot R_S + \beta \cdot R_C + (1 - \alpha - \beta) \cdot R_E), \quad (3)$$

where α and β are weighting parameters that control the contribution from R_S and that from R_C such that $0 \leq \alpha, \beta, \alpha + \beta \leq 1$. The Equation (3) can be rewritten in a matrix form as $R = AR$, where the matrix A represents the linear combination of the three types of similarities (R_S, R_C, R_E) according to Equations (1), (2) and (3). We can efficiently compute the overall similarity R by using the power method as follows:

$$R^{(k+1)} \leftarrow AR^{(k)} / |AR^{(k)}|, \quad (4)$$

where $R^{(k+1)}$ is the estimation of the similarity score matrix R in the $(k+1)$ th iteration, and the initial similarity $R^{(0)}$ is set to a random vector with unit L_1 -norm and nonnegative elements. The convergence rate of the power method is dominated by the second largest eigenvalue of the matrix A , but the number of iterations can be limited to a fixed number N_{It} or the iteration could be stopped if the residual is lower than a predefined tolerance. Based on the estimated node-to-node similarity scores in R , we can now find the optimal network alignment through dynamic programming. To be more specific, the estimated scores that measure the similarities between nodes that belong to different topological networks (which represent different RNAs) can be used to find the best pairwise alignment between the networks that maximize the sum of the similarity scores of the aligned nodes. Since the nodes in the topological networks correspond to nucleotides in the corresponding RNAs, the structural alignment of the RNAs can be readily obtained from the resulting network alignment. The pseudo-code of the proposed network-based RNA structural alignment algorithm TOPAS is shown in Figure 2.

The computational complexity of TOPAS is dominated by the estimation of the overall similarity R . Typically, the matrix A is very sparse, which allows efficient computation of R . The overall computational complexity will be $O(kd_1d_2N^2)$, where k is the number of iterations in the power method, d_1 is the number of base-pairing interaction edges in the network G_1 , and d_2 is the number of base-pairing interaction edges in G_2 . For typical RNAs, we have $kd_1d_2 \ll N^2$. Additionally, the space complexity of TOPAS is $O(N^2)$ which is much lower than $O(N^4)$ required by the traditional Sankoff algorithm. It is worth noting that LocARNA and RAF also have the same low space complexity $O(N^2)$.

3 Results

3.1 Construction of topological networks

Given a pair of RNA sequences, TOPAS constructs topological networks for the respective RNAs based on the base-pairing probabilities estimated using the *RNAstructure* package (version 5.8). *RNAstructure* is a software package for RNA secondary structure analysis, which also includes a tool for single RNA structure prediction based on the nearest-neighbor thermodynamic model and the sequence alignment derived from a pair-HMM (Harmanci et al., 2008; Reuter and Mathews, 2010). Previously, the PARTS algorithm utilized precomputed base-pairing and alignment probabilities to evaluate the pseudo-free energy, and in a similar way, TOPAS

RNA structural alignment based on topological networks:

Output: Structural alignment (\hat{S}_1, \hat{S}_2) .

Input: RNA sequences (S_1, S_2) , probabilistic model (P_{S_1}, P_{S_2})

Parameters $(\alpha, \beta, N_{It}, P_{Th})$.

```

1. Construct topological networks
for  $n = 1$  to 2 do
    Construct  $G_n = (V_n, E_n)$  from the sequence data
     $(S_n, P_{S_n}, P_{Th})$ 
end

2. Run power method to estimate similarity  $R$ 
Initialize the similarity vector  $R^{(0)}$  with a nonzero random unit
vector.

for  $k = 1$  to  $N_{It}$  do
    Initialize  $R_S, R_C$  to 0
    for  $a = 1$  to  $\text{length}(V_1)$  do
        for  $b = 1$  to  $\text{length}(V_2)$  do
            Update structure similarity
            foreach  $(c, d) \in (N_{G_1}(a), N_{G_2}(b))$  do
                 $R_S(a, b) + =$ 
                 $R^{(k-1)}(c, d)[P_{S_1}(a, c)P_{S_2}(b, d)/D(c)D(d)]$ 
            end
            Update connected similarity
            if  $\text{Exist } R(a-1, b-1)$  then
                 $R_C(a, b) + = \frac{1}{2}R^{(k-1)}(a-1, b-1)$ 
            end
            if  $\text{Exist } R(a+1, b+1)$  then
                 $R_C(a, b) + = \frac{1}{2}R^{(k-1)}(a+1, b+1)$ 
            end
            Update overall similarity
             $R_A^{(k)}(a, b) = \alpha R_S(a, b) + \beta R_C(a, b) + (1 - \alpha - \beta)R_E(a, b)$ 
        end
    end
    Normalize and update the overall similarity
     $R^{(k)} = R_A^{(k)} / |R_A^{(k)}|$ 
    Stop criterion
    if  $|R^{(k)} - R^{(k-1)}| < \text{Tolerance}$  then
        break
    end
end

3. Run dynamic programming (Needleman-Wunch) to maximize
the overall similarity  $R(\hat{S}_1, \hat{S}_2)$ 
4. Output the corresponding RNA structural alignment  $(\hat{S}_1, \hat{S}_2)$ 

```

Fig. 2. Pseudocode of the proposed RNA structural alignment algorithm

utilizes the probabilistic models in *RNAstructure* for predicting the RNA structural alignment based on topological networks.

3.2 Parameters for network-based structural alignment using TOPAS

Equation (3) estimates the overall similarity between nodes (which correspond to bases) across networks (which represent the RNA sequences to be aligned), where the parameter α weights the topological similarity R_S and the parameter β weights the connected similarity R_C . In addition, the sequence similarity R_E should be included to avoid symmetric structural ambiguity (i.e. $\alpha + \beta < 1$), but the contribution from the sequence similarity should be kept at a relatively low level so that it does not dominate the final alignment result when analyzing sequences with low sequence identity (SI). We illustrate the effect of the weight parameters (α, β) on the structural alignment accuracy based on two pairs of tRNAs obtained from the

Rfam database (Griffiths-Jones *et al.*, 2003): (i) the first tRNA pair (X14835.1/6927-7002, M32222.1/12777-1363) has been selected to illustrate the high sequence identity case ($SI=0.77$) and (ii) the other tRNA pair (X14835.1/6927-7002, M86496.1/1024-1089) has been selected to illustrate the low sequence identity case ($SI=0.24$). The respective secondary structures of these three tRNAs are shown

in Figure 3(a–c), which have been drawn using VARNA (Darty *et al.*, 2009). The accuracy of the structural alignment algorithm is assessed in terms of sensitivity ($SEN = \frac{TP}{TP+FN}$) and positive predictive value ($PPV = \frac{TP}{TP+FP}$). TP, FP and FN are the number of true positives, false positives and false negatives, respectively, and they are calculated by comparing the predicted alignment edges with those in

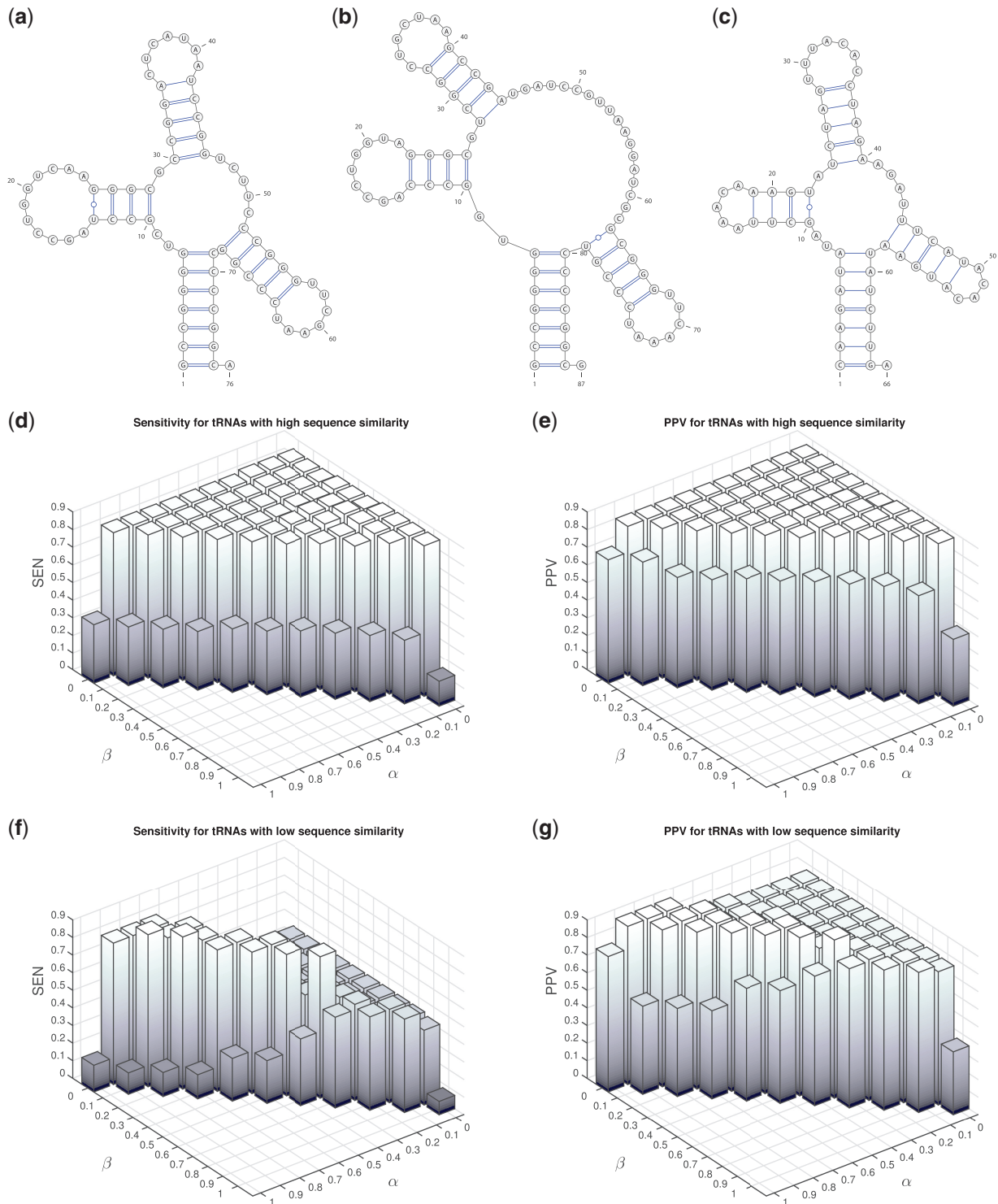


Fig. 3. Illustration of the effect of the parameters α and β on the alignment accuracy. (a) Secondary structure of the tRNA X14835.1/6927-7002. (b) Secondary structure of the tRNA M32222.1/12777-1363. (c) Secondary structure of the tRNA M86496.1/1024-1089. (d) Sensitivity (SEN) for a tRNA pair with high sequence identity. (e) Positive predictive value (PPV) for a tRNA pair with high sequence identity. (f) Sensitivity for a tRNA pair with low sequence identity. (g) Positive predictive value for a tRNA pair with low sequence identity.

the true alignment. The F-Score = $2 / (\frac{1}{SEN} + \frac{1}{PPV})$ is also measured for performance evaluation and comparison.

For the first pair with high SI, the performance of the proposed network-based structural alignment is not very sensitive to the choice of the parameters (α, β) as can be seen in Figure 3(d) and (e). In this case, sequence similarity provides sufficient clues for predicting a relatively accurate alignment, although the alignment quality can be further improved by considering the structural similarity between the RNAs. However, for RNAs with relatively low SI, the sequence similarity between the RNAs is not sufficient per se for finding an accurate alignment. This is illustrated in Figure 3(f) and (g) based on the tRNA pair with low SI. As we can see from these plots, in this case, placing a larger weight on topological similarity generally leads to a higher SEN and a higher PPV. In practice, the weight parameters α and β can be estimated through a grid search based on the available training data, to strike a balance between structural similarity and sequence similarity for reliable and accurate prediction of RNA structural alignments.

3.3 Performance evaluation

In order to evaluate the performance of our proposed structural alignment method TOPAS, we used the sequence pairs in the *BRALiBase* 2.1 dataset K2 (Wilm et al., 2006) as the benchmark for performance assessment and comparison. In the dataset, there were 389 sequence pairs that contained unknown bases. These sequence pairs were excluded in our performance assessment, and the final size of the test dataset was about 95.67% of the size of the original *BRALiBase* 2.1 dataset K2. The benchmark consists of RNA sequences from 36 RNA structural families, including 8587 RNA sequence pairs with an average length of 109 bases and an average sequence identity of 0.67. For comparison, we also assessed the performance of several widely used Sankoff-style structural alignment algorithms based on the same benchmark. Table 1 lists the structural alignment algorithms that were considered in our performance evaluation and comparison.

The performance evaluation results based on *BRALiBase* 2.1 dataset K2 are summarized in Table 2. The parameters of TOPAS were set to $(\alpha, \beta, N_{IT}, P_{Th}) = (0.40, 0.56, 30, 0.01)$. All experiments were performed on an iMac (3.5GHz CPU, 32 GB RAM, OS X 10.9.5) and the computational time was measured in seconds for all algorithms. The overall computation time of TOPAS consists of two major parts: the time needed for computing the base-pairing probabilities using the RNAstructure package (Reuter and Mathews, 2010) and the computation time for constructing the topological networks and predicting the structural alignment of RNAs based on the constructed networks. The base-pairing probabilities used as the input for the TOPAS algorithm can also be computed by other RNA folding packages, such as the popular ViennaRNA package

(Hofacker, 2009), based on one's preference. In Tables 2 and 3, the computation time shown for TOPAS corresponds to the time needed for the network-based structural alignment and it does not include the time for computing the input base-pairing probabilities using RNAstructure.

The computation time of TOPAS for network-based structural alignment depends on the length of the RNA sequences to be aligned and the number of probabilistic interaction edges inferred by the probabilistic model for secondary structure prediction. As we can see in Table 2, TOPAS yields highly accurate structural alignment results, outperforming previous structural alignment algorithms in terms of accuracy. In terms of alignment speed, TOPAS was also among the fastest among the compared algorithms. The total computation time of TOPAS for aligning all sequence pairs in the benchmark was comparable to that of SPARSE, which was the fastest among all algorithms. However, SPARSE resulted in the lowest SEN and PPV as a trade-off.

In order to find out how the sequence similarity of the RNAs affects the alignment accuracy of different structural alignment algorithms, we grouped the RNA pairs in the benchmark based on their sequence identity (SI). Figure 4 shows the alignment accuracy (i.e. SEN and PPV) as a function of SI (RNA pairs have been grouped based on their rounded SI). As we can see in Figure 4(a) and (b), TOPAS consistently outperforms other structural alignment algorithms at most SI levels. For sequences with very low SI (20–30%), the alignment accuracy of TOPAS tended to degrade and TOPAS did not perform as well as some of the other Sankoff-style algorithms like *FoldAlign*. The structural alignment predicted by TOPAS relies on effective estimation of the topological similarity. We suspect that the degradation of alignment accuracy for low SI sequence pairs is likely due to the quality degradation of the topological similarity estimated by the probabilistic models utilized by TOPAS.

Table 2. Performance evaluation results based on the *BRALiBase* 2.1 K2 dataset

	SEN	PPV	F-Score	Log ₁₀ (Time)
TOPAS	0.878	0.938	0.907	3.349
<i>PARTS</i>	0.860	0.931	0.894	5.625
<i>Foldalign</i>	0.860	0.923	0.891	5.657
<i>Dynalign2</i>	0.706	0.914	0.797	5.803
<i>LocaRNA</i>	0.862	0.922	0.891	4.128
<i>SPARSE</i>	0.848	0.931	0.888	3.653
<i>RAF</i>	0.865	0.938	0.900	3.200

Note: Accuracy is measured by comparing the predicted alignment edges against the true edges in the benchmark. Total computation time was measured for completing the structural alignment of all sequence pairs in the benchmark (in seconds) Best performance is shown in bold.

Table 1. List of RNA structural alignment algorithms that were considered in this work for performance comparison with TOPAS

Program	Version/Package	Command ^a (Configure file)	Reference
<i>PARTS</i>	RNAstructure 5.8	parts default.conf	Harmanci et al. (2008)
<i>Dynalign2</i>	RNAstructure 5.8	dynalign_ii default.conf	Fu et al. (2014)
<i>Foldalign</i>	2.1.0	foldalign -global seq_files	Havgaard et al. (2005) and Sundfeld et al. (2016)
<i>LocaRNA</i>	LocaRNA 1.9.2	locarna seq_files	Will et al. (2007)
<i>SPARSE</i>	LocaRNA 1.9.2	sparse seq_files	Will et al. (2015)
<i>RAF</i>	1.0.0	raf predict seq_files	Chuong et al. (2008)

^aNote that the 'Command (Configure file)' column describes the command that was used to run the algorithm. In all cases, default configurations were used for performance evaluation. Here, the performance of Foldalign 2.1 is compared as the accuracy of Foldalign 2.1 is better than Foldalign 2.5 for the test dataset.

Table 3. Performance evaluation results for RNA families with pseudoknots

	wcaG RNA				Downstream-peptide RNA			
	SEN	PPV	F-Score	Log ₁₀ (Time)	SEN	PPV	F-Score	Log ₁₀ (Time)
TOPAS	0.847	0.911	0.878	2.410	0.861	0.899	0.880	1.908
TOPAS (PK)	0.854	0.912	0.882	2.401	0.866	0.901	0.883	1.903
PARTS	0.839	0.908	0.872	4.401	0.827	0.895	0.860	3.879
Foldalign	0.834	0.905	0.868	3.381	0.805	0.890	0.845	2.725
Dynalign2	0.413	0.806	0.546	3.979	0.438	0.797	0.565	3.266
LocaRNA	0.824	0.902	0.861	2.816	0.827	0.897	0.861	2.190
SPARSE	0.766	0.903	0.828	2.732	0.854	0.907	0.880	2.140
RAF	0.841	0.913	0.876	2.322	0.821	0.900	0.859	2.201

Note: Accuracy is measured by comparing the predicted alignment edges against the true edges. Total computation time was measured for completing the structural alignment of all sequence pairs in the given family (in seconds). Best performance is shown in bold.

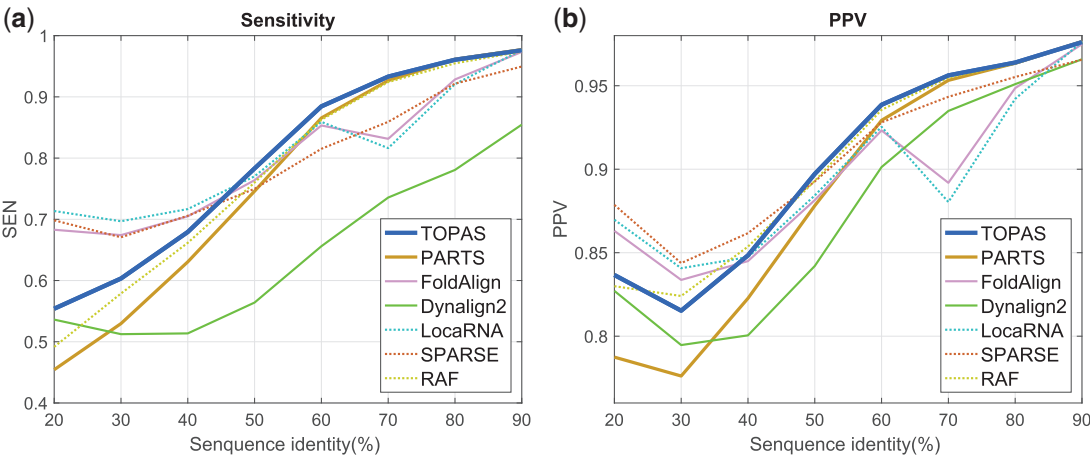


Fig. 4. Performance evaluation results based on the *BRAliBase* 2.1 K2 dataset. (a) Sensitivity (SEN) of different algorithms is shown as a function of sequence identity (SI). (b) Positive predictive value (PPV) of different algorithms are shown as a function of sequence identity (SI)

In order to evaluate the performance of structural alignment for RNAs with pseudoknots, we used sequences from two RNA families—Downstream-peptide RNA and wcaG RNA—in the *Rfam* database. For each family, 2000 pairs were randomly selected for performance assessment. Table 3 summarizes the alignment results for the two RNA families with pseudoknots. We can again see from Table 3 that TOPAS generally outperforms other structural alignment algorithms in terms alignment speed and accuracy.

The performance of TOPAS can be further improved if the base-pairing probabilities for RNAs with pseudoknots could be better estimated. In Table 3, TOPAS (PK) shows the results obtained by TOPAS when the minimum interleaved base pairs of pseudoknots in the RNA structure are corrected. This experiment was carried out to verify the potential improvement that could be attained through better estimation of base-pairing probabilities for crossing base-pairs in RNAs with pseudoknots. There are six interleaved base-pairs in wcaG RNA and five interleaved base-pairs in Downstream-peptide RNA that are corrected to test the improvement for TOPAS (PK). Currently, most RNA secondary structure prediction packages exclude pseudoknots, as allowing secondary structures with crossing base-pairs would lead to a sharp increase in computational cost and memory requirement. More accurate estimation of the base-pairing probabilities for RNA pseudoknots would improve the quality of the topological networks, and TOPAS could take direct advantage of such improvement as the network-based approach adopted by TOPAS is not restricted to nested RNA secondary structures.

4 Conclusions

Various methods have been developed for RNA structural alignment to date, where Sankoff-style algorithms that simultaneously predict the optimal alignment and folding have been especially popular. Although such Sankoff-style algorithms are known to yield accurate alignment results, especially for RNAs with relatively low sequence similarity, they typically suffer from high complexity in time and space. In this paper, we proposed TOPAS, a novel algorithm for pairwise structural alignment of RNAs based on an innovative network-based approach. Given two RNAs with unknown structure, TOPAS first constructs topological networks for the respective RNAs by incorporating their structural information extracted through probabilistic base-pairing models. The resulting networks are then aligned through an efficient network alignment technique, thereby predicting the best structural alignment of the given RNAs in a way that sensibly integrates their sequence similarity as well as their structural similarity. As shown by extensive performance evaluation based on several RNA families and the *BRAliBase* 2.1 K2 dataset, the proposed algorithm TOPAS outperforms popular Sankoff-style RNA structural alignment algorithms in many cases, resulting in comparable or higher alignment accuracy at a significantly reduced computational cost. Moreover, owing to the flexibility of the network-based alignment approach adopted by TOPAS, the proposed RNA structural alignment algorithm is not restricted to nested folding structures and it can effectively align RNAs with pseudoknots. To the best of our knowledge, TOPAS is the first RNA

structural alignment algorithm that explicitly adopts a network-based approach. As we have shown in this paper, the topological networks constructed by TOPAS lead to accurate alignment results. However, we would like to note that the approach introduced in our paper is by no means the only—and not necessarily the optimal—way of constructing such networks. We expect that the overall accuracy of the RNA structural alignment may be further improved in the future by constructing topological networks that are further enriched with additional information that may be useful in predicting the RNA alignment. The scheme that was adopted by TOPAS for computing the overall similarity R between nodes across different networks can be viewed as performing a random walk with restart. In fact, random walk based models have been shown to be useful for comparative network analysis, and several different models have been proposed to date (Jeong and Yoon, 2015; Jeong et al., 2016; Sahraeian and Yoon, 2013; Singh et al., 2008). Devising and incorporating novel random walk models that are optimized for network-based structural alignment of RNAs may potentially enhance the speed and accuracy of the RNA structural alignment algorithm even further.

Funding

This work was supported by the National Science Foundation Awards CCF-1149544, CCF-1447235; United States Department of Agriculture National Institute of Food and Agriculture competitive grant USDA-NIFASCR-2017-51181-26834 through the National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University; and by the TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering.

Conflict of Interest: none declared.

References

- Chuong, B.D. et al. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, i68–i76.
- Darty, K. et al. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
- Flamm, C. et al. (2000) RNA folding at elementary step resolution. *RNA*, **6**, 325–338.
- Freyhult, E.K. et al. (2006) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.
- Fu, Y. et al. (2014) Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res.*, **42**, 13939–13948.
- Gardner, P.P. et al. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
- Glotz, C. et al. (1981) Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, *Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. *Nucleic Acids Res.*, **9**, 3287–3306.
- Greenleaf, W.J. et al. (2008) Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, **319**, 630–633.
- Griffiths-Jones, S. et al. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
- Gursoy, A. et al. (2008) Topological properties of protein interaction networks from a structural perspective. *Biochem. Soc. Trans.*, **36**, 1398–1403.
- Hamada, M. et al. (2009) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*, **25**, 3236–3243.
- Harmanci, A.O. et al. (2008) PARTS: probabilistic alignment for RNA joint secondary structure prediction. *Nucleic Acids Res.*, **36**, 2406–2417.
- Havgaard, J.H. et al. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.
- Hofacker, I.L. (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, **26**, 12–12.
- Hofacker, I.L. et al. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
- Jeong, H. and Yoon, B.-J. (2015) Accurate multiple network alignment through context-sensitive random walk. *BMC Syst. Biol.*, **9**, S7.
- Jeong, H. et al. (2016) Effective comparative analysis of protein–protein interaction networks by measuring the steady-state network flow using a Markov model. *BMC Bioinformatics*, **17**, 395.
- Johansson, P. et al. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta*, **1840**, 1063–1071.
- Liao, C.-S. et al. (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253–i258.
- Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
- Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
- McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
- Mount, D.W. (2009) Using hidden Markov models to align multiple sequences. *Cold Spring Harb. Protoc.*, 2009, pdb-top41.
- Raué, H. et al. (1988) Evolutionary conservation of structure and function of high molecular weight ribosomal RNA. *Progress Biophys. Mol. Biol.*, **51**, 77–129.
- Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 1.
- Sahraeian, S.M.E. and Yoon, B.-J. (2013) SMETANA: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLoS One*, **8**, e67995.
- Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
- Singh, R. et al. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. USA*, **105**, 12763–12768.
- Sundfeld, D. et al. (2016) Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, **32**, 1238–1240.
- Tinoco, I. and Bustamante, C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
- Turner, D.H. and Mathews, D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38** (Database issue), D280–D282.
- Will, S. et al. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
- Will, S. et al. (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.
- Wilm, A. et al. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 1.
- Yoon, B.-J. (2009) Hidden Markov models and their applications in biological sequence analysis. *Curr. Genomics*, **10**, 402–415.
- Yoon, B.-J. (2014) Sequence alignment by passing messages. *BMC Genomics*, **15**, 1.
- Yoon, B.-J. et al. (2012) Comparative analysis of biological networks: hidden Markov model and Markov chain-based approach. *IEEE Signal Process. Mag.*, **29**, 22–34.
- Zwieb, C. et al. (1981) Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species. *Nucleic Acids Res.*, **9**, 3621–3640.