Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach

Matthias Höchsmann, Björn Voss, and Robert Giegerich

Abstract—In functional, noncoding RNA, structure is often essential to function. While the full 3D structure is very difficult to determine, the 2D structure of an RNA molecule gives good clues to its 3D structure, and for molecules of moderate length, it can be predicted with good reliability. Structure comparison is, in analogy to sequence comparison, the essential technique to infer related function. We provide a method for computing multiple alignments of RNA secondary structures under the tree alignment model, which is suitable to cluster RNA molecules purely on the structural level, i.e., sequence similarity is not required. We give a systematic generalization of the profile alignment method from strings to trees and forests. We introduce a tree profile representation of RNA secondary structure alignments which allows reasonable scoring in structure comparison. Besides the technical aspects, an RNA profile is a useful data structure to represent multiple structures of RNA sequences. Moreover, we propose a visualization of RNA consensus structures that is enriched by the full sequence information.

Index Terms—Alignment of trees, RNA secondary structures, noncoding RNAs.

1 Introduction

THE investigation of RNA secondary structures is a challenging task in molecular biology. RNA molecules have a variety of functions in the cell, which often depend on structural properties. A similar structure often implies a similar function, the converse conclusion is weaker. If there is a certain amount of conservation on the sequence level, a multiple sequence alignment of RNAs, having a similar function, can elucidate conserved structural regions among these sequences. However, functional RNA families such as tRNA, rRNA, and RNAse P RNA exhibit a highly conserved secondary structure but little sequence similarity. Therefore, it is of general interest to compare RNA secondary structures directly, i.e., without relying on sequence similarity [1], [2], [3]. These approaches are limited to a pairwise comparison. *Stemtrace* [4] is an interactive visualization tool for comparative RNA structure analysis. However, it requires manual intervention. An early approach for multiple structure comparison based on sequence alignments is provided in [5]. A multiple structure alignment strategy which builds upon pairwise structure alignment and multiple sequence alignment is implemented in the tool *MARNA* [6]. It calculates pairwise similarities of RNA structures, based on the algorithm in [3], which in turn are used to weight edges in the multiple sequence alignment tool *T-Coffee* [7]. As *T-Coffee* computes a sequence alignment, this step inherits the problem of pairwise sequence alignments of RNA structures. That is: A base-pair is not treated as a unit.

We provide a purely structure-based approach for comparing multiple RNA secondary structures. Our work is based on the tree alignment model [8] and we generalize the progressive profile alignment approach from sequences [9], [10] to trees. A tree profile representation of RNA secondary structure alignments makes the progressive approach applicable to multiple RNA structure comparison. As a tree is a fundamental data structure, our algorithm is useful beyond the scope of RNA structure comparison such as structured text databases, HTML Web pages, and image analysis [11], [12], [13], [14]. The algorithms presented here are implemented as an extension of the structure comparison tool *RNAforester* [15]. The source code distribution is part of the *Vienna RNA Package* [16], http://www.tbi.univie.ac.at/~ivo/RNA/. An online version of the software will be available at the *Bielefeld Bioinformatics Server*, http://bibiserv.techfak.uni-bielefeld.de.

• *The authors are with the International Graduate School in Bioinformatics and Genome Research, University of Bielefeld, D-33501 Bielefeld, Germany. E-mail: {mhoechsm, bvoss, robert}@techfak.uni-bielefeld.de.*

Manuscript received 7 June 2004; revised 2 Aug. 2004; accepted 3 Aug. 2004. For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0064-0604.

2 Preliminaries

Let Σ be a set of symbols, the *alphabet*. The *gap symbol* "-" not in Σ, will play the special role to indicate deletions. We define $\Sigma^{-} = \Sigma \cup \{\text{-}\}$ and the *tuple alphabet* $\Sigma^n = \{(a_1, \ldots, a_n) \mid a_i \in \Sigma^{-}\} \backslash \{(\text{-}, \ldots, \text{-})\}$.

We consider *rooted, ordered, node-labeled trees*, called *trees* for short. An (ordered) *forest* is a sequence of trees. A function *label* assigns a label from some alphabet Σ to each node in a tree or forest. We use $\mathcal{F}(\Sigma)$ for the set of Σ-*labeled* forests. Where convenient, we identify a tree with the forest containing only this tree. A *string* over Σ is an unary tree in $\mathcal{F}(\Sigma)$, where each inner node has exactly one descendant. This latter definition implies that and how the string case is embedded into our generalization to trees. Alternatively, a string can be considered as a forest wherein each tree is just a leaf, labeled by a character. No matter which embedding is chosen, our algorithm degenerates to the standard string alignment algorithm in this special case.

3 Alignments of Forests

We give the following definitions in terms of forests. As trees and strings are special cases thereof, the definitions apply to them as well.

In the tree edit model [17], deleting a tree node v means that the children of node v become the children of the parent node of v. Moreover, if v has any siblings, the deletion preserves the preorder relation of these nodes. If v is a root node, then its children have no common ancestor any more and they split up into a forest. Fig. 1 gives an example of the delete operation. We speak of deletions and insertions when editing T into T', but an insertion into T is nothing but a deletion from T' and, hence, requires no extra definition. In contrast to the common operational view of editing one structure into another, our notion of an alignment is a declarative model, a data structure rather than a process: Following [8], an *alignment* of two forests, with labels from some alphabet, is a forest with labels from the pair alphabet. Labels of the form $(a,b), (a,\text{-}), (\text{-},b)$ with $a, b \in \Sigma$ denote the edit operations $R(eplace)$, $D(elete)$, and $I(nsert)$, respectively. Here, this notion applies to (pairs of) strings, trees, and forests. Clearly, it generalizes to alignments of more than two items. We now formalize this view.

Let $A \in \mathcal{F}(\Sigma^n)$. Its componentwise projections $A|_1, \ldots, A|_n$ are elements of $\mathcal{F}\Sigma^{-}$. Let $F \in \mathcal{F}(\Sigma^{-})$. $\pi(F) \in \mathcal{F}(\Sigma)$ is the forest that results from successive deletion of nodes v with $label(v) = \text{-}$. It is easy to show that the order of node deletions is irrelevant and, thus, $\pi(F)$ is uniquely defined. Let $F_1, \ldots, F_n \in \mathcal{F}(\Sigma)$. $A \in \mathcal{F}(\Sigma^n)$ is an alignment of $F_1, \ldots, F_n$ iff $F_i = \pi(A|_i)$, for $1 \le i \le n$. Since strings and trees are special cases of forests, these definitions apply to them as well. An example of a pairwise tree alignment is given in Fig. 2.

We now turn to scoring alignments. Among all possible alignments, we are interested only in those that satisfy an optimality criterion. For distance problems, optimal means minimal, while for similarity problems optimal means maximal. Given a similarity function $\sigma : \Sigma^n \to \mathbb{R}$, the similarity score of an alignment $A \in \mathcal{F}(\Sigma^n)$ is defined by

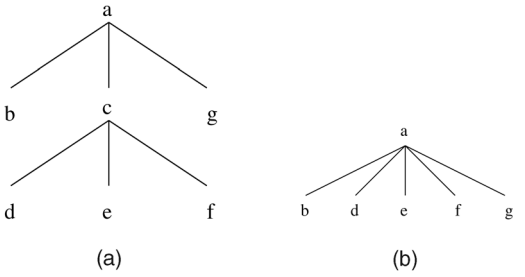

Fig. 1. (a) A tree with nodes $a, \ldots, g$. (b) The tree after node c has been deleted.

$$\sigma(A) = \sum_{v \text{ node in } A} \sigma(label(v)).$$

The similarity score of forests $F_1, \ldots, F_n$, written as $\sigma(F_1, \ldots, F_n)$, is the maximal score that can be obtained by an alignment of $F_1, \ldots, F_n$. An alignment of $F_1, \ldots, F_n$ is optimal if it achieves this score.

Problem 1 (Multiple Forest Alignment (MFA) Problem). *Compute $\sigma(F_1, \ldots, F_n)$ and an optimal alignment of $F_1, \ldots, F_n$.*

Computing $\sigma(F_1, \ldots, F_n)$ is an unsolved problem in the literature. The complexity of the problem of computing multiple sequence alignments depends on the scoring function. Among different scoring schemes, the sum-of-pairs score is the one that received most attention. Wang and Jiang in [18] showed that the problem of computing a multiple sequence alignment with optimal sum-of-pairs score is NP-complete; this remains true if the scoring scheme is a metric [19].

As the sequence alignment model is a special case of the forest alignment model, we cannot hope to solve this problem exactly within polynomial time. Inspired by the progressive sequence alignment method of *ClustalW* [9], we focus on generalizing this widely used heuristics to forest alignments. This is done in Section 6.

4 Forest Representation of RNA Secondary Structures

An RNA structure is denoted by an RNA sequence and the set of bases that form hydrogen bonds. Representing the bonds as arcs above the sequence (drawn as a straight line), an RNA structure is an *RNA secondary structure* iff the arcs are not crossing. A coarse grained representation of RNA secondary structures which uses the structural elements hairpin loop (H), bulge (B), interior loop (I), and multiloop (M) as its basic elements is proposed in [5]. This encoding produces small forests, but developing a reasonable scoring scheme on this level of abstraction is a difficult problem. Following [20], we represent RNA secondary structures as forests on the level of paired and unpaired bases. The parent and sibling relationship of the forest nodes is determined by the nesting of base pair bonds. The $5'$ to $3'$ nature of the RNA molecule imposes the order among sibling nodes. Fig. 3a shows a 2D plot of an RNA molecule and Fig. 3b depicts the corresponding forest representation.

Bases that pair in one structure can be unpaired in a related structure because the pair is not stable in terms of energy or a mutation of one base forbids a pairing. Accordingly, the bases that are involved in such events should be replaced by each other. The RNA representation in Fig. 3b is not yet suitable for creating an adequate scoring scheme for these basic events. Clearly, each node of an RNA forest is involved in exactly one edit operation in a forest alignment. Since a base pair is encoded as a single node, the score for deleting the pairing between bases a and u would be $\sigma((a,u),\text{-}) + \sigma(\text{-},a) + \sigma(\text{-},u)$. We extend the forest representation to allow explicit scoring of base pair deletions under the tree alignment model. Base-pairs are represented by three connected nodes: The *P-node* stands for the base pair bond and is labeled with P. Its children nodes are ordered according to the $5'$ to $3'$ ordering of the bases and the leftmost and rightmost child are the bases that pair. The children nodes of a P-node can be P-nodes, except for the leftmost or rightmost child. Fig. 3c gives an example of our *extended* RNA forest representation which is in flavor of parse trees for context free grammars that describe RNA secondary structures [22].

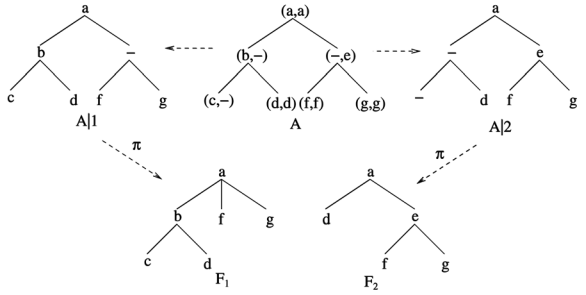

Fig. 2. A is an alignment of F_1 and F_2.

The alphabet of labels of our extended forest representation is $\Sigma = \{P, a, c, g, u\}$. For a simpler description, we ignore the primary sequence information and only compare structures.[1] This is facilitated by a reduced alphabet $\Sigma_{P,B} = \{P, B\}$, where B stands for an arbitrary base.[2] We use the following scoring scheme for given parameters b_m, b_d, p_m, p_d:

$$\begin{aligned}\sigma(B,B) &= b_m\\ \sigma(B,\text{-}) = \sigma(\text{-},B) &= b_d\\ \sigma(P,P) &= p_m\\ \sigma(P,\text{-}) = \sigma(\text{-},P) &= p_d\\ \sigma(P,B) = \sigma(B,P) &= -\infty.\end{aligned}$$

A meaningful scoring scheme should satisfy $b_m >= 0$, $b_d <= 0$, $p_m >= 0$, and $p_d <= 0$. A replacement of a P-node and a B-node is not defined in our model and the scoring contribution for this case must be $-\infty$.

5 RNA Secondary Structure Profiles

Multiple alignments of protein sequences are useful to relate proteins of similar functions to protein families. The identification of proteins that also belong to a certain family naturally gives rise to the question of finding a kind of representative sequence of a protein family. Such representations, that are well-known for multiple sequence alignments, are *profile*, *consensus sequence*, and *signature*. Here, we restrict our attention to the profile representation [23].

A *profile* for a multiple sequence alignment contains the frequency of characters of each row and is also known as a *weight matrix*. In analogy of our view of sequence alignments as sequences of edit operations, and tree alignments as trees labeled with edit operations (see Section 3), we consider a profile of a sequence alignment as a sequence of relative frequency vectors. Consequently, a profile of a forest alignment is a forest labeled with relative frequency vectors. We give the following definition of a *profile for a forest alignment*:

1. Our implementation of the algorithm considers sequence information.
2. A scoring scheme σ such that $\sigma(a_1,b_1) = \sigma(a_2,b_2)$ for $a_1,a_2,b_1,b_2 \in \{a,c,g,u\}$ has the same effect.

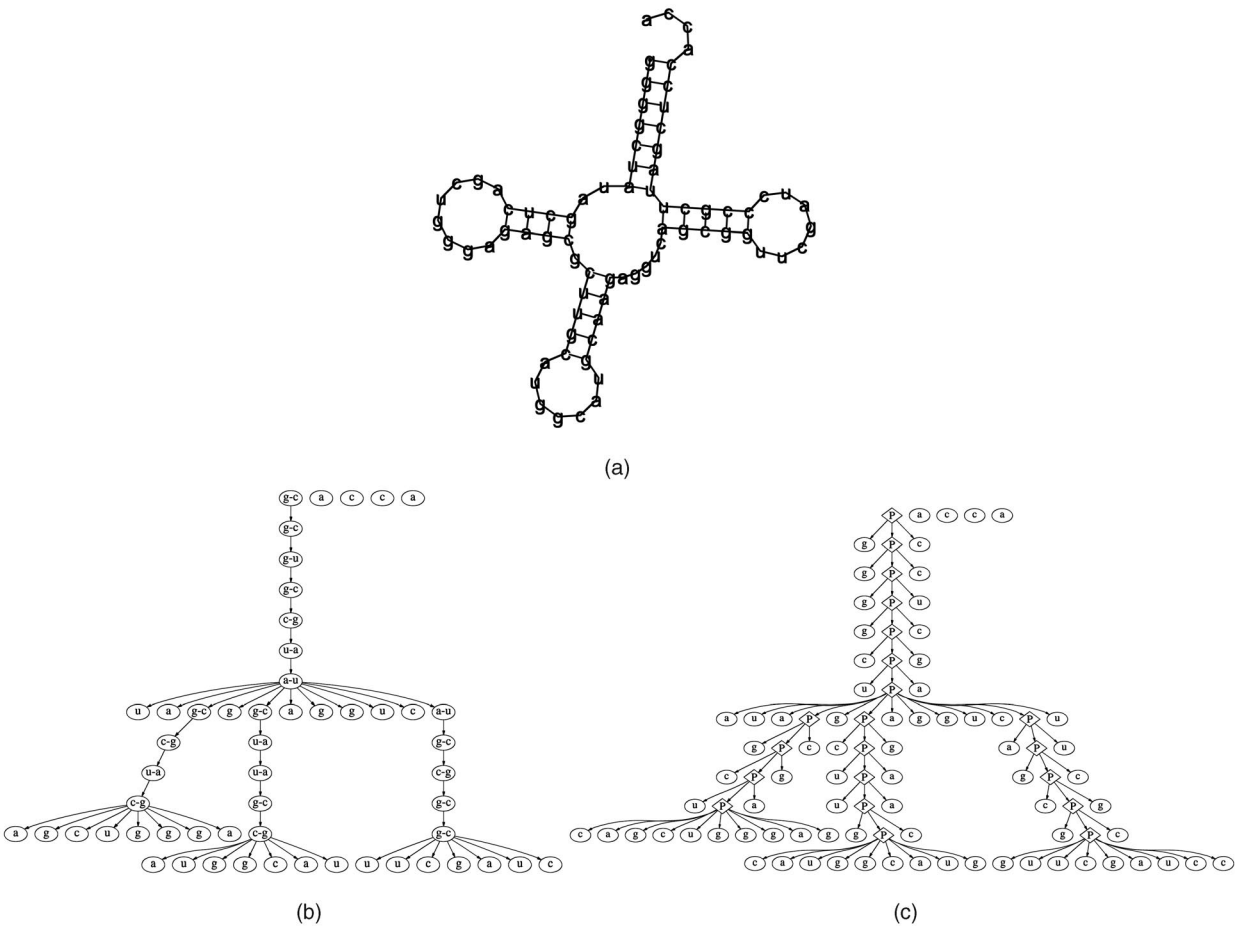

Fig. 3. (a) Secondary structure of the *E.coli* tRNA for alanine taken from the Genomic tRNA Database [21]. (b) Forest representation of (a). Base-pairs correspond to internal nodes labeled with the bases of that pair. Unpaired bases correspond to leaf nodes and their label is a single base. (c) Extended forest representation of (a). A base pair is represented explicitly by a P-node. The leftmost and rightmost child of a P-node are the bases of that pair.

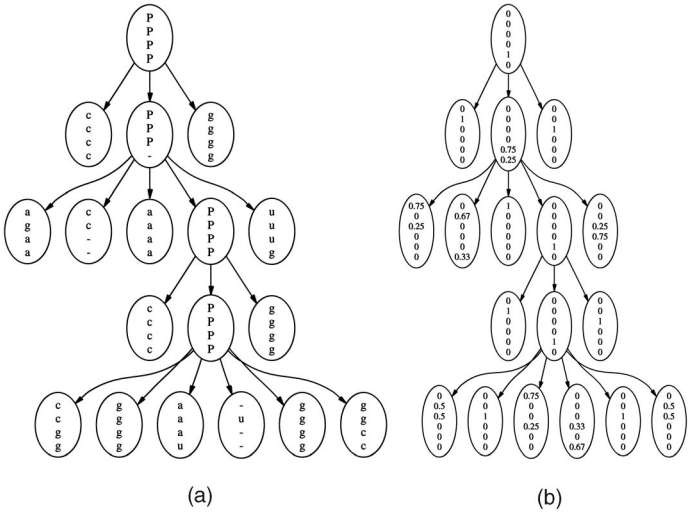

Fig. 4. (a) Shows a multiple tree alignment of RNA secondary structures and (b) its corresponding profile. The rows of the frequency vectors stand, from top to bottom, for the frequencies of the bases a, c, g, u, P, -. Note that the frequency of a base is zero iff the frequency of a basepair bond is greater zero.

Definition 1. *Let $k = |\Sigma^-|$. Given a multiple forest alignment $A \in \mathcal{F}(\Sigma^n)$, its profile alignment $P_A \in \mathcal{F}(\mathbb{R}^k)$ is obtained by converting each label in A to its relative frequency vector.*

An example of a multiple tree alignment and its corresponding profile is shown in Fig. 4. Since the profile of a forest alignment is also a forest, it is straightforward to define a *profile-forest to profile-forest alignment*. All that we need is a scoring function $\delta : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. The *sum-of-pairs score*, introduced in [24], defines the score of a column of a multiple sequence alignment as the sum of scores of all combinations of pairwise scores for the column. Adhering to the sequence alignment tradition, we use the analog scheme for frequency vectors:

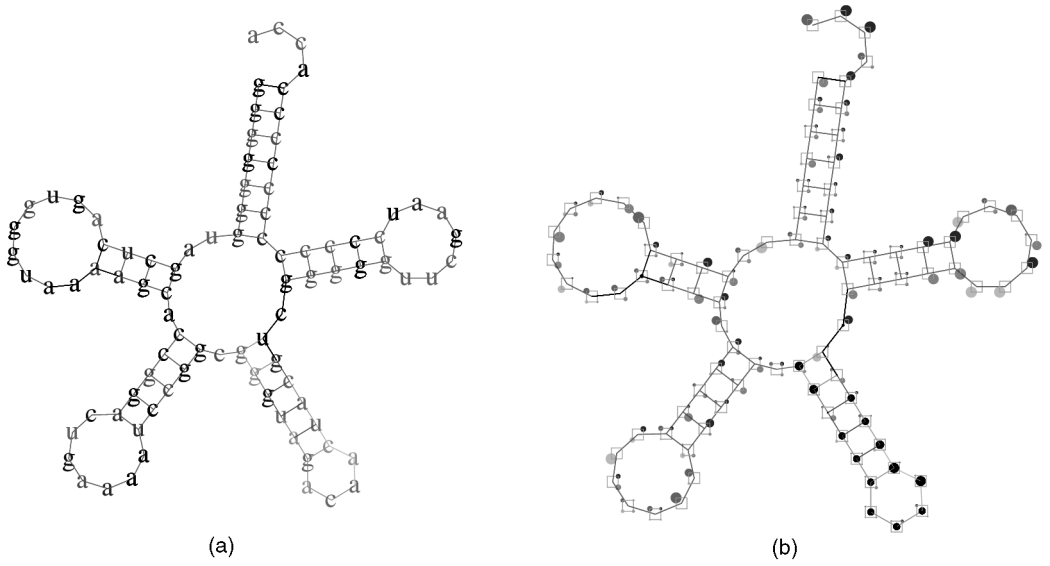

Fig. 5. 2D plots of a multiple alignment of 20 secondary structures of *E.coli* tRNAs. In both plots, the consensus structure is shown. The lighter a basepair bond is drawn, the less frequent does it exist in the structures. Bases or basepair bonds that have a frequency of 100 percent are drawn in red. (a) The most frequent base at each residue is printed with the base frequency indicated by gray scale. (b) The frequencies of the bases a, c, g, u are proportional to the radius of circles that are arranged clockwise on the corners of a square, starting at the upper left corner. Additionally, these circles are colored red, green, blue, and magenta for the bases a, c, g, u, respectively. The frequency of a gap is proportional to a black circle growing at the center of the square.

Definition 2. *The weighted sum-of-pairs score of two relative frequency vectors $p, q \in \mathbb{R}^{|\Sigma^-|}$ is defined as follows:*

$$\sigma(p,q) = \sum_{a,b \in \Sigma^-, (a,b) \neq (-,-)} p_a q_b \sigma(a,b).$$

Unlike for distances where the score of two equal forests is zero, the similarity value can be an arbitrary positive value. The similarity score of two equal forests of size n can be the same as for two different forests of size m, where $m > n$. Therefore, we introduce relative scores that are upper bounded by 1.

Definition 2. *The relative similarity score of forests F_1 and F_2 is defined by:*

$$\sigma_r(F_1, F_2) = \frac{2\sigma(F_1, F_2)}{\sigma(F_1, F_1) + \sigma(F_2, F_2)}.$$

The self-similarity score of a forest results from a perfect matching alignment for reasonable scoring schemes. This score can be computed, without self-aligning the forests, in $O(|F|)$. A new profile-forest can be constructed in $O(|F|)$ from an alignment of profiles by using the mean value of the aligned frequency vectors as the frequency vector of the profile.

6 Algorithm

Driven by the importance for multiple sequence comparison in the field of biology, several heuristics and approximation schemes have been developed that produce "good" alignments. A popular strategy is based on the idea of constructing a multiple alignment from optimal pairwise alignments in an iterative, also called *progressive*, fashion [10]. A widely used implementation of a profile-based, progressive multiple sequence alignment is *ClustalW* [9].

An efficient algorithm that solves the pairwise tree distance problem is introduced in [8]. A corresponding similarity version for forests is provided in [15]. We now describe how the pairwise case is extended to multiple alignments:

1. Convert $F_1, \ldots, F_n$ into single structure profiles $P_1, \ldots, P_n$.
2. Construct all $\frac{n(n-1)}{2}$ pairwise relative similarity scores of $P_1, \ldots, P_n$.
3. Choose P_i and P_j of maximal similarity, compute their alignment P_{ij}, and replace both by P_{ij}.
4. Compute the relative similarity score of P_{ij} with all others.
5. Iterate Steps 3 and 4 until only a single profile alignment $P_{1\ldots n}$ is left.

Step 4 is in contrast to *ClustalW*, which first constructs a guide tree solely based on the initial pairwise distances. We avoid the computation of a guide tree because it carries the danger of introducing a methodical cycle. Often, alignments are used to construct phylogenies. It has been observed (A. Dress, personal communication) that any such phylogeny tends to reproduce the guide tree, no matter how well this tree really suits the data. To

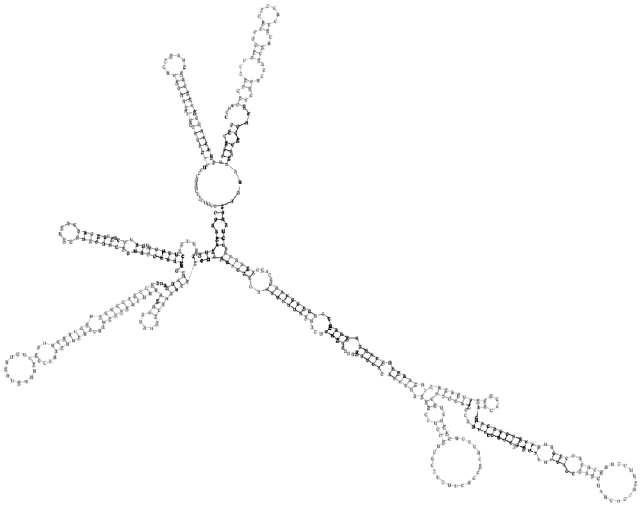

Fig. 6. Multiple alignment of the four 5'UTRs of human and mouse ferritin heavy chain mRNA (5HSA015337, 5MMU002159) and SLC11A3 iron-transporter mRNA (5HSA023193, 5MMU011005). The alignment clearly superposes the IRE elements, automatically marked red by our visualization.

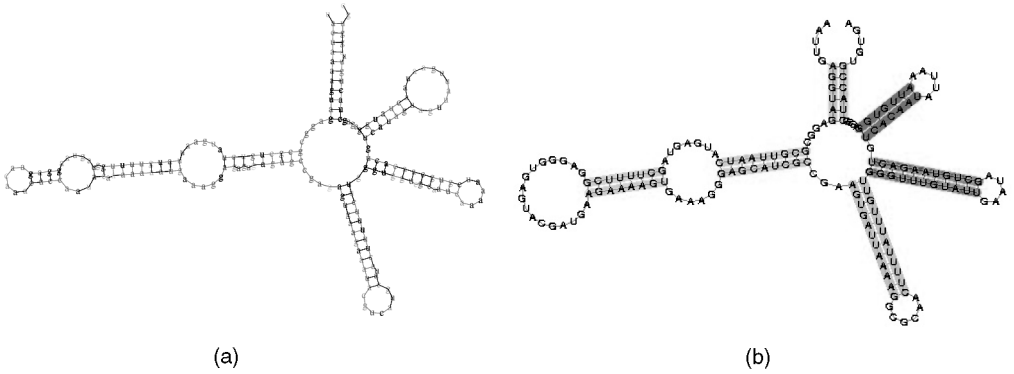

Fig. 7. (a) Consensus structure of Lysine riboswitches derived from the profile obtained by the calculation of the multiple alignment in the Appendix. (b) Consensus structure of Lysine riboswitch published in [38]. The consensus structure in (a) is calculated only from a subset (11 of the 48 seed sequences) of the structures that are incorporated in (b). This is one reason for a slight variation between these structures.

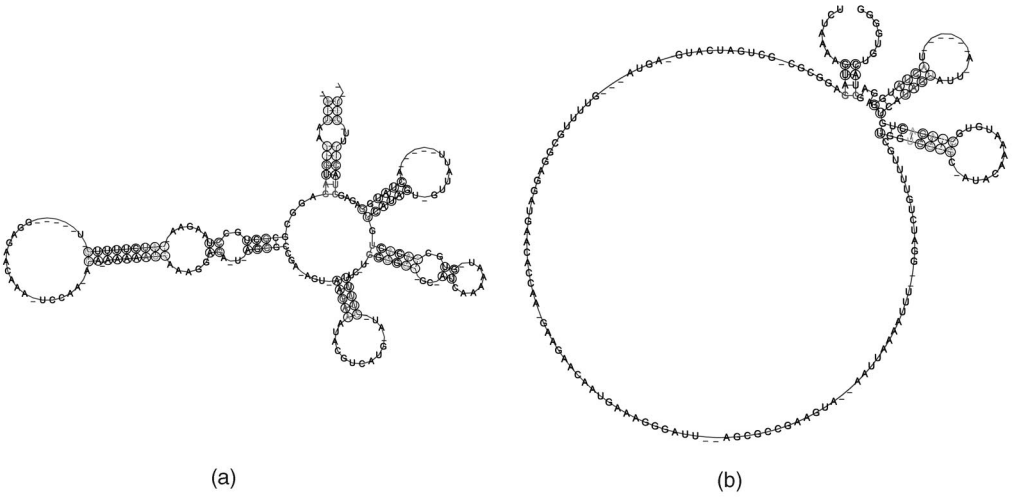

Fig. 8. *RNAalifold* consensus structure predictions of (a) the *RNAforester* derived sequence alignment (see the Appendix) and (b) the *ClustalW* alignment (see the Appendix).

facilitate clustering of forests, joining alignments in Step 3 can be restricted to a minimal cutoff value c. If the best alignment score of P_i and P_j in Step 3 is below c, these profiles are put into the result list of clusters which are not aligned further.

6.1 Efficiency Analysis

The asymptotic efficiency of this approach is as follows: Let there be n structures of average size s, measured in terms of nodes in the tree representation. Let d be the maximum degree of a tree node. The pairwise algorithm has time efficiency $O(s^2d^2)$ and space efficiency $O(s^2d)$, see [8]. In Step 2, this algorithm is called for all pairs of tree profiles yielding time efficiency $O(s^2d^2n^2)$. Both, Step 3 and 4 are repeated $n-1$ times. In the ith iteration, Step 4 computes $n-i$ pairwise alignment scores. Consequently, the overall runtime of Steps 3 to 5 is in $O(s^2d^2n^2)$. Thus, the runtime of our algorithm is in $O(s^2d^2n^2)$. In Step 1, n forests are stored, requiring $O(ns)$ space. The allocated space of a pairwise alignment can be freed after the alignment score is calculated. The scores are stored in a table of size n^2. In Step 3, the optimal alignment is obtained by recalculating the alignment matrix and a backtracking procedure in $O(s^2d^2)$ time and $O(s^2d)$ space. Thus, the overall space requirement of our algorithm is bounded by $O(ns+n^2+s^2d)$.

In case of RNA structures, d corresponds to the maximal loop size, which can be considered as a constant for thermodynamic reasons [25]. Hence, RNA structure alignments under the tree alignment model can be calculated with the same asymptotic efficiency as sequence alignments. Note, this does not hold if RNA structure alignments are calculated under the tree edit model [1].[3] The worst-case runtime of the tree edit algorithm by Zhang and Sasha [26], and tree alignment algorithm by Jiang et al. [8] are both in $O(s^4)$, but they depend on different parameters of the trees. While the degree of an RNA structure tree is constant, the number of leaves and the depth of a tree grow with the length of an RNA molecule. The latter two influence the complexity of the tree edit algorithm. Algorithmic improvements of the tree edit algorithm are proposed in [27], [28], but those do not achieve quadratic runtime for trees of bounded degree. The average runtime of Zasha and Zhang's algorithm for RNA secondary structures turned out to be in $O(s^3)$ [29].

7 Consensus Structure and Visualization

An alignment of RNA structures naturally gives rise to the question of a consensus structure. Since bases paired in one structure can be aligned to bases unpaired in another structure, this leaves some choice. We use the final profile that is calculated by the algorithm presented in the previous section to infer a consensus structure.

3. The tree edit model is different from the tree alignment model, i.e., tree alignments form a subset of tree edits.

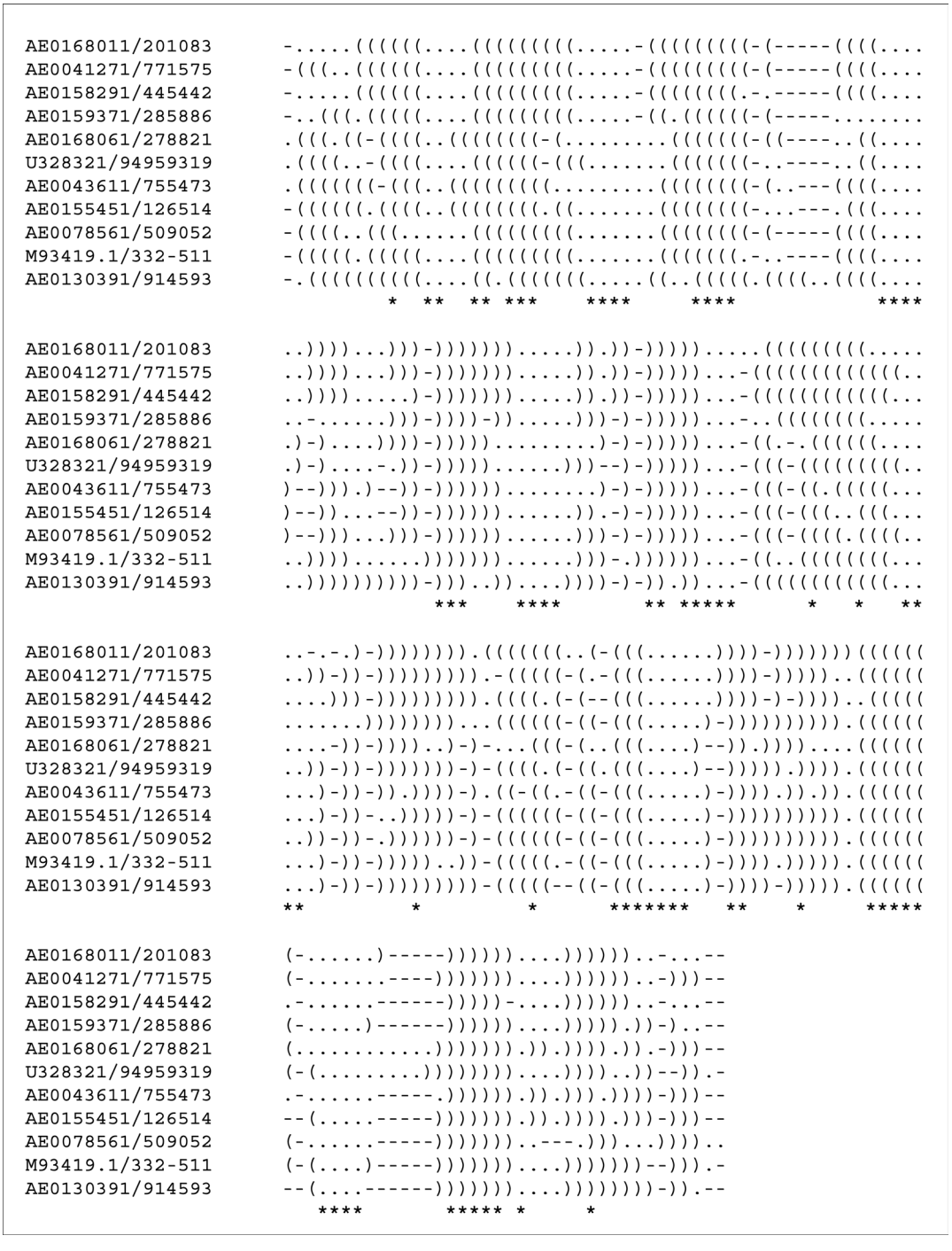

Fig. 9. ASCII representation of pure structure alignment.

An RNA profile is *compatible* to a single structure if the leftmost and rightmost child of each P-node is a B-node. A compatible profile can be obtained by an arbitraty profile by deleting (see Section 3) P-nodes from the profile. An optimal consensus structure corresponds to a compatible profile that maximizes the sum of P-node frequencies. This can be computed efficiently by a dynamic programming algorithm.

RNA secondary structures are commonly represented graphically as circle plots, dot plots, mountain plots, or 2D plots. These techniques were also applied to the visualization of consensus structures of RNAs [30], [31]. We concentrate on the 2D plot of RNA consensus structures and refine it to include full sequence information. We print the consensus structure in two forms that differ only in the way sequence information is included. Both express the frequency of base-pairs and the presence (or absence) of gaps as a gradient from light gray to black. With respect to sequence information, we provide either the most frequent base (respectively, gap) at each residue, or we indicate frequencies of all bases and gaps in an arrangement of colored dots. A basepair is only drawn if it is present in at least 50 percent of the structures. Figs. 5a and 5b show an example of a multiple structure alignment that was calculated by *RNAforester*.

8 Experiments

Multiple alignment of RNA secondary structures is helpful to reveal a common structural property of RNAs. We present two applications for multiple structure alignment, motif discovery and

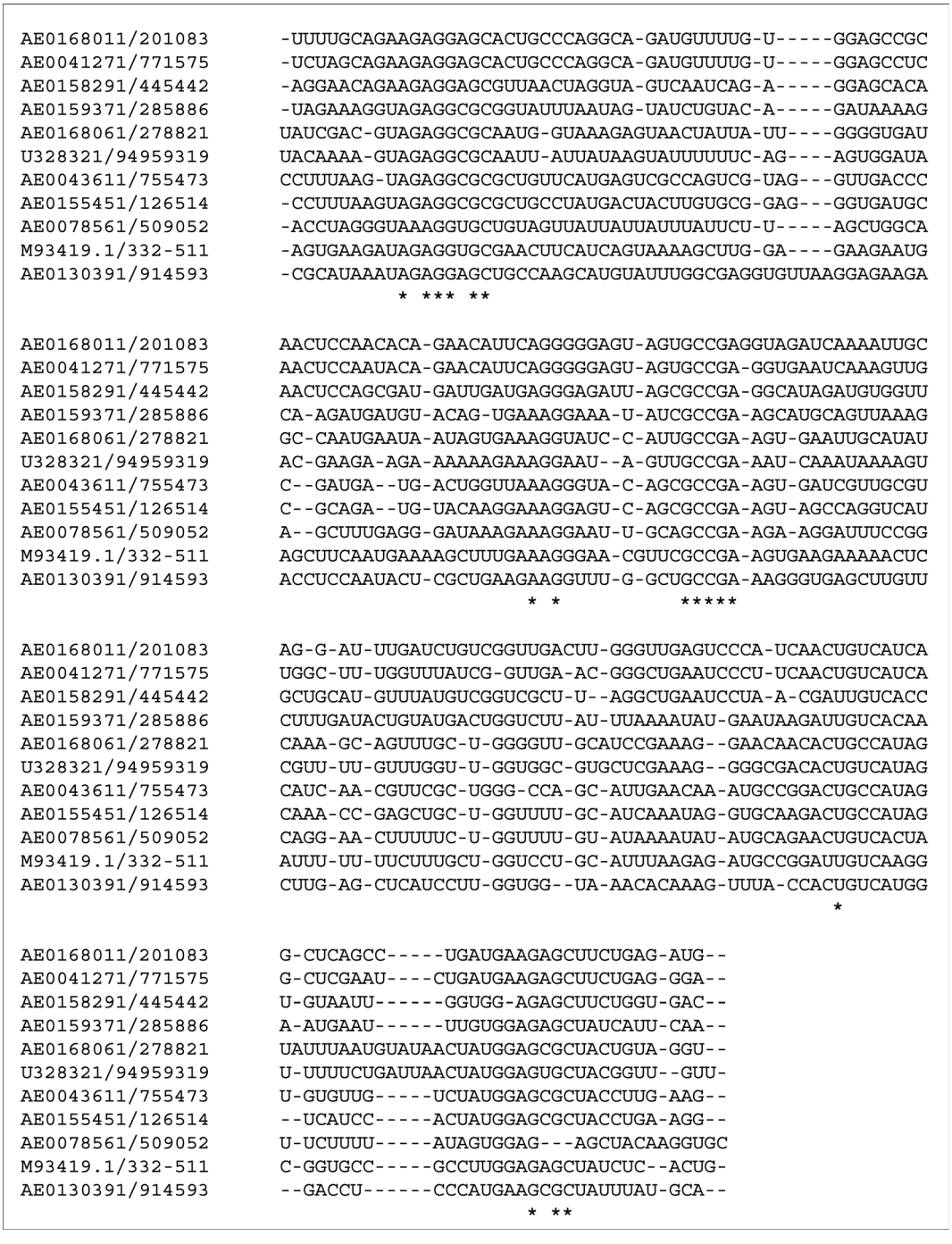

Fig. 10. Derived multiple sequence alignment.

consensus structure prediction. The latter application is initially based on sequence information where no verified structures are given. If a set of trusted structures is available, a multiple structure alignment is, in analogy to multiple sequence alignment, a basic tool which can be embedded in a larger framework of structure analysis.

8.1 Motif Discovery

Multiple structure alignment can be used for searching regulatory structural motifs common to several RNAs. One of the best investigated regulatory motifs is the iron responsive element (IRE), which is a specific stem-loop structure and can be found in the untranslated terminal regions (UTRs) of many mRNAs. It regulates, for example, the translational efficiency of these mRNAs according to the amount of iron in the cell [10]. The 5'UTRs of human and mouse ferritin heavy chain mRNA and SLC11A3 iron-transporter mRNA were taken from the UTR database [32]. These UTRs are known to contain iron responsive elements. Their secondary structures were predicted with *mfold* (Version 3.1) [33] and a multiple structure alignment of the UTRs was calculated using *RNAforester*. In Fig. 6, the resulting alignment is displayed. The red colored stem-loop shows the conserved iron responsive element that occurs in all structures. All other structural elements shown in black or gray can only be found in some of the structures. Thus, the described approach is useful for the detection of common structural motifs in a set of RNA secondary structures. This example works well because the element of interest resides in similar positions in the globally aligned structures. Should this

Fig. 11. Multiple RNA sequence alignment calculated by *ClustalW* (1.82).

position vary, a local similarity comparison can be employed [15]. Unfortunately, this is restricted to pairwise comparisons.

8.2 Consensus Structure Prediction

A multiple sequence alignment is often the first step in the determination of a consensus structure. Homologous sequence regions are matched in the alignment and force regions of sequence diversity between them to be aligned to each other. Among a family of RNAs, bases that are only involved in stacking regions of an RNA are expected to be less sequence conserved in an alignment than unpaired bases. Thus, regions of diversity in an alignment are subject to structural observations. It is obvious that such a strategy requires a considerable amount of sequence conservation to be successful.

We use the multiple structure alignment to predict consensus structures the other way around. First, the structure of an RNA is predicted by *RNAfold* [16], [34]. In the second step, a pure multiple structure alignment is computed by the algorithm presented in Section 6. A pure structure alignment means that sequence conservation is not favored by the scoring scheme. In contrast to the sequence-based approach, structural conserved regions are the anchor regions of the alignment. Clearly, our approach is sensitive to a correct prediction of RNA structures. On the other hand, it is useful to identify correct predictions of structures by the following hypothesis: *The more structurally conserved and the less sequence conserved a multiple alignment is, the better are the predicted structures*. This is motivated by the observations that random structures do

rarely fold into similar structures and that the mutual information content of a sequence alignment of a family of structural RNAs is regarded as a measure of its quality [35], [36], [37]. Obviously, our hypothesis gets the stronger the more structures a multiple alignment includes. A theoretical statistical analysis of our approach is not provided here, but a comparison of real data to random data has been done.

We exemplify our approach by the RNAs belonging to the family of *Lysine riboswitches* taken from *Rfam* database [39] (Accession number: RF00168), originally described in [38]. We use the 48 seed sequences of the database and predict their structures with *RNAfold* (default energy parameters and with option *-noLP*). A multiple structure alignment of these structures is computed, using the following parameters: $p_m = 1$, $p_d = -1$, $b_m = 0$, and $b_d = -2$. The clustering parameters are $c = 0$. The resulting clusters have the following sizes where the number in brackets is the number of clusters: $1(8), 3(3), 4(2), 5(1), 7(1), 11(1)$. The alignment of cluster of size 11 shows little sequence conservation. The consensus structure of this cluster is shown in Fig. 7a and the consensus structure published in *Rfam* is shown in Fig. 7b. Both are in good correspondence. Other clusters that have minor sequence conservation are also close to the published consensus structure. Clusters that are highly sequence conserved (almost perfect) do not result in consensus structures that are close to the published one. We shall return to the lessons to be learned from this in the conclusions. Experiments on random structures show that the expected cluster size for our experiment, under the above parameters, is close to 1.

The approach is limited by the size of the structures. The reason is not due to computational time constraints, but to the fact that the results of structure prediction become more unreliable. The longer the sequences are, the more sequences are necessary to form clusters. For sequences of families up to 300 bases in the *Rfam* database, our approach always produces consensus structures close to the published ones.

8.3 Structure versus Sequence Alignment

A pure structure alignment produces a sequence alignment as a coproduct. Here, we compare the quality of multiple sequence alignment which is obtained by structure comparison with a *ClustalW* multiple sequence alignment. To measure the quality, we use the tool *RNAalifold* [40] which computes the consensus structure of aligned RNA sequences taking into account both thermodynamic stability and sequence covariation.

We derive a sequence alignment from the structure alignment studied in Section 8.2 (see the Appendix). For the same sequences, we calculate a multiple sequence alignment with *ClustalW* (see the Appendix). The consensus structure predictions of *RNAalifold* are shown in Fig. 8. Clearly, the *RNAforester* alignment produces a better consensus structure in terms of sequence covariation and is closer to the published one in Fig. 7b. Naturally, if the structure of an RNA is not important for its function, a classical sequence alignment should produce the best result and be the method of choice.

9 Conclusion and Future Work

Our approach is a faithful generalization of established techniques used in sequence comparison. All the experience that has accumulated with using multiple sequence alignments; therefore, carries over now to RNA secondary structures. The experiments in Section 8 clearly show the applicability of multiple structure alignment to current tasks in bioinformatics.

Alignments of predicted minimal free energy structures can rightfully be criticized, because structure prediction may produce "optimal" structures quite different to the (suboptimal) native structure. The use of sequence similarity, if sufficient, is advocated as a means to avoid this dilemma. However, our experiments contribute two new considerations to this issue:

- They demonstrate an effect that, at the first sight, is paradoxical: Strong sequence similarity can mislead the determination of the consensus structure. This happens because very similar sequences tend to fold into a similar structure, be it wrong or right.
- They demonstrate that a multiple structure alignment, when applying the cutoff value in the clustering step, may produce meaningful alignments even in the presence of incorrect predictions.

As a consequence, a new approach to consensus construction becomes feasible, where first a good candidate consensus (or several) is constructed and, subsequently, sequences that do not fall into a consensus cluster are refolded, given the candidate consensus as a target structure. Algorithms for the later task have already been suggested by [41], but their combination with the method presented here has yet to be explored.

Appendix A

Multiple RNA Structure Alignment Calculated by *RNAforester*

An ASCII representation of pure structure alignment is shown in Fig. 9. A derived multiple sequence alignment is shown in Fig. 10.

Appendix B

Multiple RNA Sequence Alignment Calculated by *ClustalW* (1.82)

See Fig. 11.

References

[1] B.A. Shapiro and K. Zhang, "Comparing Multiple RNA Secondary Structures Using Tree Comparisons," *Computer Applications in the Biosciences,* vol. 6, no. 4, pp. 309-318, 1990.

[2] F. Corpet and B. Michot, "RNAlign Program: Alignment of RNA Sequences Using Both Primary and Secondary Structures," *Computer Applications in the Biosciences,* vol. 10, no. 4, pp. 389-399, 1994.

[3] T. Jiang, G. Lin, B. Ma, and K. Zhang, "A General Edit Distance between RNA Structures," *J. Computational Biology,* vol. 9, no. 2, pp. 371-388, 2002.

[4] K. Wojciech and B. Shapiro, "Stem Trace: An Interactive Visual Tool for Comparative RNA Structure Analysis. " *Bioinformatics,* vol. 15, no. 1, pp. 16-31, 1999.

[5] B. Shapiro, "An Algorithm for Comparing Multiple RNA Secondary Stuctures," *Computer Applications in the Biosciences,* vol. 4, no. 3, pp. 387-393, 1988.

[6] S. Siebert and R. Backofen, "MARNA: A Server for Multiple Alignment of Rnas," *Proc. German Conf. Bioinformatics,* pp. 135-140, 2003.

[7] C. Notredame, D. Higgins, and J. Heringa, "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment," *J. Molecular Biology,* vol. 302, pp. 205-217, 2000.

[8] T. Jiang, J. Wang, and K. Zhang, "Alignment of Trees–An Alternative to Tree Edit," *Theoretical Computer Science,* vol. 143, no. 1, pp. 137-148, 1995.

[9] J. Thompson, D. Higgins, and T. Gibson, "*CLUSTAL W*: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice," *Nucleic Acids Research,* vol. 22, pp. 4673-4680, 1994.

[10] D. Feng and F. Doolittle, "Progressive Sequence Alignment as a Prerequisite to Correct to Correct Phylogenetic Trees," *J. Molecular Evolution,* vol. 25, pp. 351-360, 1987.

[11] P. Kilpeläinen and H. Mannila, "Ordered and Unordered Tree Inclusion," *SIAM J. Computing,* vol. 24, pp. 340-356, 1995.

[12] D. Reis, P. Golgher, A. Silva, and A. Laender, "Automatic Web News Extraction Using Tree Edit Distance," *Proc. 13th Conf. World Wide Web,* pp. 502-511, 2004.

[13] B. Rössler, J. Zhang, and M. Höchsmann, "Visual Guided Grasping and Generalization Using Self-Valuing Learning," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems,* 2002.
[14] T. Sebastian, P. Klein, B. Kimia, "Recognition of Shapes by Editing Their Shock Graphs," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, pp. 550-571, 2004.
[15] M. Höchsmann, T. Toller, R. Giegerich, and S. Kurtz, "Local Similarity in RNA Secondary Structures," *Proc. IEEE Computational Systems Bioinformatics Conf.,* pp. 159-168, 2003.
[16] I. Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast Folding and Comparison of RNA Secondary Structures," *Monatshefte f. Chemie,* vol. 125, pp. 167-188, 1994.
[17] K. Tai, "The Tree-to-Tree Correction Problem," *J. ACM,* vol. 26, pp. 422-433, 1979.
[18] L. Wang and T. Jiang, "On the Complexity of Multiple Sequence Alignment," *J. Computational Biology,* vol. 1, pp. 337-348, 1994.
[19] P. Bonizzoni and G. Della Vedova, "The Complexity of Multiple Sequence Alignment with Sp-Score that is a Metric," *Theoretical Computer Science,* vol. 259, no. 1, pp. 63-79, 2001.
[20] S. Le, R. Nussinov, and J. Mazel, "Tree Graphs of RNA Secondary Structures and Their Comparison," *Computational Biomedical Research,* vol. 22, pp. 461-473, 1989.
[21] T. Lowe and S. Eddy, "tRNAscan-se: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence," *Nucleic Acids Research,* vol. 25, pp. 955-964, 1997.
[22] D. Dowell and S. Eddy, "Evaluation of Several Lightweight Stochastic Context-Free Grammars for RNA Secondary Structure Prediction," *BMC Bioinformatics,* vol. 5, no. 71, 2004.
[23] M. Gribskov, A. McLachlan, D. Eisenberg, "Profile Analysis Detection of Distantly Related Proteins," *Proc. Nat'l. Academy of Science,* no. 88, vol. 43, pp. 55-58, 1987.
[24] H. Carillo and D. Lipman, "The Multiple Sequence Alignment Problem in Biology," *SIAM J. Applied Math.,* vol. 48, no. 5, pp. 1073-1082, 1988.
[25] M. Zuker and P. Stiegler, "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information," *Nucleic Acids Research,* vol. 9, pp. 133-148, 1981.
[26] K. Zhang and D. Shasha, "Simple Fast Algorithms for the Editing Distance between Trees and Related Problems," *SIAM J. Computing,* vol. 18, no. 6, pp. 1245-1262, 1989.
[27] P.N. Klein, "Computing the Edit-Distance between Unrooted Ordered Trees," *Proc. Sixth Ann. European Symp.,* no. 1461, pp. 91-102, 1998.
[28] S. Dulucq and H. Touzet, "Analysis of Tree Edit Distance Algorithms," *Proc. 14th Ann. Symp. Combinatorial Pattern Matching,* pp. 83-95, 2003.
[29] S. Dulucq and L. Tichit, "RNA Secondary Structure Comparison: Exact Analysis of the Zhang-Sasha Tree Edit-Algorithm," *Theoretical Computer Science,* vol. 306, pp. 471-484, 2003.
[30] D. Konings and P. Hogeweg, "Pattern Analysis of RNA Secondary Structure: Similarity and Consensus of Minimal-Energy Folding," *J. Molecular Biology,* vol. 207, pp. 597-614, 1989.
[31] F. Chetouani, P. Monestie, P. Thebault, C. Gaspin, and B. Michot, "ESSA: An Integrated and Interactive Computer Tool for Analysing RNA Secondary Structure," *Nucleic Acids Research,* vol. 25, no. 17, pp. 3514-3522, 1997.
[32] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, and C. Saccone, "Utrdb and Utrsite: Specialized Database of Sequences and Functional Elements of 5′ and 3′ Untranslated Regions of Eukaryotic mRNAs," *Nucleic Acids Research,* vol. 30, pp. 335-340, 2002.
[33] M. Zuker, "Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction," *Nucleic Acids Research,* vol. 31, no. 13, pp. 3406-3415, 2003.
[34] I. Hofacker, "Vienna RNA Secondary Structure Server," *Nucleic Acids Research,* vol. 31, no. 13, pp. 3429-3431, 2003.
[35] A. Wong and D. Chiu, "An Event-Covering Method for Effective Probabilistic Inference," *Pattern Recognition,* vol. 20, pp. 245-255, 1987.
[36] D. Chiu and T. Kolodziejczak, "Inferring Consensus Structure from Nucleic Acid Sequences," *Computer Applications in the Biosciences,* vol. 7, pp. 347-352, 1991.
[37] J. Gorodkin, L. Heyer, S. Brunak, and G. Stormo, "Displaying the Information Contents of Structural RNA Alignments," *Bioinformatics,* vol. 13, pp. 583-586, 1997.
[38] M. Mandal, B. Boese, J. Barrick, W. Winkler, and R. Breaker, "Riboswitches Control Fundamental Biochemical Pathways in Bacillus Subtilis and Other Bacteria," *Cell,* vol. 113, no. 5, pp. 577-586, 2003.
[39] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. Eddy, "Rfam: An RNA Family Database," *Nucleic Acids Research,* vol. 31, no. 1, pp. 439-441, 2003.
[40] I. Hofacker, M. Fekete, and P. Stadler, "Secondary Structure Prediction for Aligned RNA Sequences," *J. Molecular Biology,* vol. 319, pp. 1059-1066, 2002.
[41] K. Zhang, "Computing Similarity between RNA Secondary Structures," *Proc. IEEE Int'l Joint Symp. Intelligence and Systems,* pp. 126-132, 1998.