A fast structural multiple alignment method for long RNA sequences

Yasuo Tabei¹,², Hisanori Kiryu², Taishin Kin² and Kiyoshi Asai*¹,²

Address: ¹Graduate School of Frontier Science, University of Tokyo, CB04 Kiban-tou 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan and ²Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Email: Yasuo Tabei - tabei@cb.k.u-tokyo.ac.jp; Hisanori Kiryu - kiryu-h@aist.go.jp; Taishin Kin - kin-taishin@aist.go.jp; Kiyoshi Asai* - asai@k.u-tokyo.ac.jp

* Corresponding author

Abstract

Background: Aligning multiple RNA sequences is essential for analyzing non-coding RNAs. Although many alignment methods for non-coding RNAs, including Sankoff’s algorithm for strict structural alignments, have been proposed, they are either inaccurate or computationally too expensive. Faster methods with reasonable accuracies are required for genome-scale analyses.

Results: We propose a fast algorithm for multiple structural alignments of RNA sequences that is an extension of our pairwise structural alignment method (implemented in SCARNA). The accuracies of the implemented software, MXSCARNA, are at least as favorable as those of state-of-art algorithms that are computationally much more expensive in time and memory.

Conclusion: The proposed method for structural alignment of multiple RNA sequences is fast enough for large-scale analyses with accuracies at least comparable to those of existing algorithms. The source code of MXSCARNA and its web server are available at http://mxscarna.ncrna.org.

Background

Non-coding RNAs (ncRNAs) are transcribed RNA molecules that do not encode proteins. Their functions often depend on their 3D-structures rather than their primary sequences. The secondary structures of RNA sequences can be identified by various methods, including minimization of the free energy [1-3]. However, it is not always possible to obtain the accurate secondary structures. More reliable predictions of the secondary structures are possible if we have a set of RNA sequences with a common secondary structure. For consensus structure prediction, RNAalifold [4], Pfold [5], and McCaskill-MEA [6] are applicable only to sets of aligned RNA sequences. Multiple alignment tools that consider only sequence similarities, e.g. ClustalW [7], Dialign [8], and T-Coffee [9], however, have limited accuracy for RNA sequences with low similarity.

Simultaneous prediction of the common secondary structure and optimal alignment of RNA sequences is computationally quite expensive, even if pseudo-knotted structures are excluded. For example, the strict algorithm of Sankoff [10] requires O(L³N) in time and O(L²N) in memory for N sequences of length L. Its faster variants that restrict the distances of the base pairs in the primary sequences are proposed for pairwise alignments [11-14].

Published: 23 January 2008

BMC Bioinformatics 2008, 9:33 doi:10.1186/1471-2105-9-33

Received: 14 September 2007

Accepted: 23 January 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/33

© 2008 Tabei et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Although structural alignment of multiple RNA sequences with reasonable computational cost is difficult, several algorithms have been proposed. Hofacker et al. proposed a method for progressive multiple alignments by direct comparison of the base-pairing probability matrices [12], implemented in PMulti which was recently reimplemented in FoldalignM [13] and Locarna [16] by Torarinsson et al. and Will et al., respectively. In Stemloc, Holmes et al. incorporated a constraint approach that limits the range of structures and alignments to be considered by pre-processing the sequences [13,14]. Siebert et al. proposed an approach distantly related to Sankoff’s algorithm and implemented it in MARNA [17] that uses the structural information for pairwise alignments before combining them into multiple alignments with T-Coffee [9]. Dalli et al. developed a new scoring approach, StrAl, that takes into account sequence similarities as well as base-pairing probabilities [18]. Xu et al. proposed a new sampling based algorithm that finds the common structure between input sequences by probabilistically sampled aligned stems based on stem conservation calculated from intrasequence base pairing probabilities and intersequence base alignment probabilities, which was implemented in RNAsampler [19]. Bauer et al. developed a graph based representation which modeled sequence-structured alignment as an integer linear program (ILP), and implemented it in RNAAlara [20]. Kiryu et al. proposed a variant of Sankoff’s algorithm with marked reduction of computation, which was implemented in Murlet [21]. All of these methods, however, are still too slow to apply to the RNA sequences longer than 1000 bases. Seibel et al. developed an alignment tool with an editor, which uses the secondary structure information of individual sequences to align multiple RNA sequences with low time complexities (4SALE) [22]. In order to extract the common secondary structure, it is also possible to find the structural motifs without aligning the whole sequences. For structural motif finding, Yao et al. proposed an algorithm based on covariance models (CMfinder) [23], and Hamada et al. proposed a graph mining approach (RNAmine) [24].

Here we propose a method, implemented in MXSCARNA, for fast multiple alignments of RNA sequences. This method extends our previous work in pairwise alignments (SCARNA) [25] to progressive multiple alignments with improved score functions, and simultaneously construct multiple alignments and the associated common secondary structures. The pairwise alignment in this progressive alignment is an heuristic algorithm that separately aligns 5’ parts and 3’ parts of the stems with rough consistency considerations.

In benchmark experiments, our method was at least as accurate as currently available state-of-art multiple alignment methods, but unlike those methods, the computations were fast enough for large-scale analyses, though the accuracies for the alignments of long sequences have not yet been confirmed.

**Results and Discussion**

**Algorithm**

**Overview of the algorithm**

The proposed method, implemented in MXSCARNA, progressively aligns multiple RNA sequences, in an extension of the pairwise structural alignment algorithm (implemented in SCARNA) of our previous work [25].

First the guide tree for the progressive alignment is built by Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [26] by using the pairwise similarities of the RNA sequences. Second the base-pairing probability matrices are calculated for all the RNA sequences by McCaskill’s algorithm [27]. Those base-pairing probabilities are used for extracting the potential stems and for the matching scores in the Dynamic Programming (DP) of the alignments. Third the RNA sequences are progressively aligned along the guide tree using SCARNA’s pairwise alignment algorithm with improved score functions introduced in this paper.

At the first stage of the progressive alignment, which corresponds to the bottom level of the guide tree, the pairs of RNA sequences are aligned by engineered DP algorithm of SCARNA’s pairwise alignment. The pairwise alignment is very fast because the potential stems extracted from the base-pairing probability matrices are decomposed into 5’ part and 3’ part and those two parts are independently aligned. In each upper-level step of the progressive alignment according to the guide tree, potential stems for groups of RNA sequences are extracted from the averaged base-pairing probability matrices.

The DP algorithm of the pairwise alignment uses the approximated posterior probabilities as score functions. The approximation uses the product of the pairwise posterior probabilities of Maximum Expected Accuracy (MEA) alignments and the base-pairing probabilities of the sequences. MEA alignment maximizes the expected number of positions where the two nucleotides are correctly aligned. To yield robust alignments, the pairwise posterior probabilities of MEA alignments are modified by the probability consistency transformation.

**Definitions**

**Definition 1: Stem candidate**

Given a base-pairing probability matrix for an RNA sequence and a threshold \( r \) (0 < \( r < 1 \), stem candidate is a set of continuous base pairs of which the base-pairing probabilities are greater than \( r \).
A stem fragment (a pair of underlined positions) comprises four overlapping stem fragments. A fragment consists of a 5’ (left) component and a 3’ (right) component. \(X_i\) (blue box) and \(X_{ai}\) (red box) are 1-continuous stem components.

Definition 2: Stem fragment
Given a base-pairing probability matrix for an RNA sequence, a threshold \(\tau (0 < \tau < 1)\), and an integer \(W\), stem fragment is a set of continuous base pairs of length \(W\), of which the base-pairing probabilities are greater than \(\tau\).

A stem candidate longer than \(W\) is represented by a set of overlapping stem fragments of fixed-length \(W\) (Figure 1). Smaller values in \(W\) or \(\tau\) increase the sensitivity of the predictions of the stems and decrease the specificity of them. \(W\) and \(\tau\) are set to 2 and 0.01 respectively in all the computational experiments in this paper. For each stem fragment, the 5’ stem component and the 3’ stem component, which are representatives of the 5’ and 3’ portions of the stem fragment, respectively, are defined as follows.

Definition 3: Stem component
For each stem fragment, a stem component \(X_{ai}\) either a 5’ stem component or a 3’ stem component, is an object that has the following properties:

- \(p(X_{ai})\): position, the position of the leftmost base of the 5’ or 3’ part of the stem fragment.
- \(s(X_{ai})\): sequence, the nucleotide sequence of the 5’ or 3’ part of the stem fragment.
- \(c(X_{ai})\): partner component, the complementary (3’ or 5’) stem component.
- \(d(X_{ai})\): loop distance, the distance to the complementary (3’ or 5’) stem component.

A stem fragment is written as \([X_{ai}, X_{ai}']\) by using the mutually complementary stem components, 5’ stem component \(X_{ai}\) and 3’ stem component \(X_{ai}’\), which represent the 5’ and 3’ parts of the stem fragment. \(X_{ai}\) and \(X_{ai}’\) satisfy

\[X_{ai} = c(X_{ai}’)\] and \(X_{ai}’ = c(X_{ai})\).

The loop distance \(d(X_{ai})\) can be written as

\[d(X_{ai}) = p(c(X_{ai})) - p(X_{ai}) - W\]

Definition 4: stem component sequence (SCS)
A stem component sequence (SCS) is a sorted sequence of all the stem components of an RNA sequence, in order of their positions and, if the positions are the same, according to their loop distances.

For \(i < j\), a SCS \(X = X_1 X_2 ... X_m\) satisfies

\[p(X_{ai}) < p(X_{aj})\text{ or } p(X_{ai}) = p(X_{aj}) \& d(X_{ai}) < d(X_{aj})\]

Definition 5: relations of stem fragments without an overlap
Two stem fragments, \([X_{ai}, X_{ai}’]\) and \([X_{jb}, X_{jb}’]\) of an RNA sequence are, parallel if and only if

\[p(X_{ai}) < p(X_{ai}’) < p(X_{jb}) < p(X_{jb}’)\text{ or } p(X_{ai}) < p(X_{ai}’) < p(X_{jb}) < p(X_{jb}’)\]

nested if and only if \(p(X_{ai}) < p(X_{jb}) < p(X_{ai}’) < p(X_{jb}’)\text{ or } p(X_{ai}) < p(X_{ai}’) < p(X_{jb}) < p(X_{jb}’)\).

pseudo-knotted if and only if \(p(X_{ai}) < p(X_{ai}’) < p(X_{jb}) < p(X_{jb}’)\text{ or } p(X_{ai}) < p(X_{ai}’) < p(X_{jb}) < p(X_{jb}’)\).

Definition 6: relations of overlapping stem fragments
Two stem fragments, \([X_{ai}, X_{ai}’]\) and \([X_{jb}, X_{jb}’]\) of an RNA sequence are, r-continuous if and only if

\[r = p(X_{ai}) - p(X_{aj}) = p(X_{ai}’) - p(X_{aj}’)\]

ill-continuous if and only if \(X_{ai}\) overlaps \(X_{aj}\) and \(X_{ai}’\) overlaps \(X_{aj}’\), and

\[p(X_{ai}) \cdot p(X_{aj}) \neq p(X_{ai}’) \cdot p(X_{aj}’)\]

contradictory if and only if only one side, either 5’ part or 3’ part, of the stem fragments overlap.

The three possible relationships between stem fragments without an overlap: parallel, nested, and pseudo-knotted, may exist in the same secondary structure of an RNA sequence. However, among the three possible relationships between overlapping stem fragments, only r-continuous stem fragments may coexist in the same secondary structure of an RNA sequence. 1-continuous, a special case of r-continuous, means that the two stem fragments are adjacent in the RNA sequence and a part of a stem candidate with a length of \(W + 1\) (Figure 1). As described later, two overlapping stem components in the alignment are controlled to belong to two r-continuous stem fragments in DP.
Building stem component sequences
In a base-pairing probability matrix, which is calculated by McCaskill's algorithm [27], a potential stem is located in two symmetry locations as continuous counterdiagonal positions which have high base-pairing probabilities. Therefore, the stem components for each RNA sequence defined in previous section are extracted by scanning counterdiagonal windows of length W in the base-pairing probability matrix and selecting the windows whose elements are greater than \( \tau \). Smaller value in \( W \) or \( \tau \) increase the specificity of them. \( W \) and \( \tau \) are set to 2 and 0.01 respectively in all the computational experiments in this paper.

The stem components are sorted in order of their positions and loop distances to construct a stem component sequence (SCS).

For each group alignment in the progressive alignment, the average of the base-pairing probability matrices is calculated directly according to the alignment of the group of RNA sequences. The stem components for the group are extracted from the averaged matrix, and the SCS is constructed by sorting the stem components.

Alignment of stem component sequences
Before the pairwise alignments or group alignments, RNA sequences or groups of RNA sequences are represented by their stem component sequences (SCSs). Those two SCSs are aligned by SCARNA's pairwise DP algorithm in each stage of the progressive alignment. The alignment of the two SCSs uses two DP matrices, \( M(i, j) \) and \( G(i, j) \). For two SCSs, \( \{X_i\} \) and \( \{Y_j\} \), \( M(i, j) \) is the best score of the alignment of the pair \( X_i \) and \( Y_j \), given that \( X_i \) matches \( Y_j \) and \( G(i, j) \) is the best score given that \( X_i \) mismatches \( Y_j \). The recursions for \( M(i, j) \) and \( G(i, j) \) are written as:

\[
M(i, j) = \max \begin{cases} M(\alpha, \beta) + \delta(i, j) \\ M(p, q) + s(i, j) \\ G(p, q) + s(i, j) \end{cases}
\]

(1)

\[
G(i, j) = \max \begin{cases} M(i - 1, j) \\ M(i, j - 1) \\ G(i - 1, j) \\ G(i, j - 1) \end{cases}
\]

(2)

with the initial conditions; \( M(0, 0) = 0, M(\cdot, 0) = M(0, \cdot) \) = \( G(0, 0) = G(\cdot, 0) = G(0, \cdot) = -\infty \).

Figure 2
An example of relations of indices of stem components in SCS alignment. \( \alpha/\beta \) and \( p/q \) in equation (1) are the indices (smaller than \( ij \)) of stem components of \( X/Y \). The red lines separate the stem components into groups which have the same positions in RNA sequences. When \( X_i \) and \( Y_j \) match in DP of SCS alignment, the stem components of the adjacent previous match must be either non-overlapping or 1-continuous with \( X/Y \). \( X_\alpha \), \( Y_\beta \) are 1-continuous with \( X/Y \). \( X_p \), \( Y_q \) are the nearest stem components that do not overlap with \( X/Y \).
The first term of equation (1) controls the 1-continuous case where the continuous matches of two overlapping stem components form a match of the corresponding stem longer than \( W \). \( \alpha_i/\beta_j \) are the indices (smaller than \( ij \)) of the components that are 1-continuous with \( X_i/Y_j \). The positions of \( X_{\alpha_i} \), \( Y_{\beta_j} \) are adjacent to \( X_i \) and \( Y_j \) in the nucleotide sequences (Figures 1 and 2), i.e.

\[
\begin{align*}
p(X_i) &= p(X_{\alpha_i}) + 1, \quad p(c(X_i)) = p(c(X_{\alpha_i})) - 1, \\
p(Y_j) &= p(Y_{\beta_j}) + 1, \quad p(c(Y_j)) = p(c(Y_{\beta_j})) - 1
\end{align*}
\]

\( \delta(i, j) \) corresponds to the incremental score for the match of the overlapping stem components, which is discussed in the next section.

The second and third terms of equation (1) keep the stem components in the adjacent DP match from overlapping in the nucleotide sequences. \( p_f/q_f \) are the indices (smaller than \( ij \)) of the nearest components that do not overlap with \( X_i/Y_j \) (Figure 2). \( s(i, j) \) is a match score for \( X_i \) and \( Y_j \) which is discussed in the next section.

Equation (2) refers only adjacent positions in DP matrix because overlaps of \( X_i \) and \( Y_j \) with the other stem components are permitted. Because the 5’ stem components and the 3’ stem components are handled independently, there is no term for bifurcation in secondary structures in equations (1) and (2).

The traceback pointer keeps the triplets, indices of \( X, Y \) and the selection of \( M \) or \( G \), in the recursion (1) and (2). The first term of the triplet, the index of \( X \), can be either \( \alpha_i \), \( p_f \), \( i \), or \( i - 1 \) and the second term of the triplet, the index of \( Y \), can be either \( \beta_j \), \( q_f \), \( j \), or \( j - 1 \). In the traceback of the DP, \( M(i, j) \) and \( G(i, j) \) are used jointly to obtain the optimal path and to select \( M \) or \( G \), which gives the maximum score of the alignment. The alignments of SCs are constructed by selecting the stem components that appear in the path with the selected \( M \). All of the mismatched stem components are excluded from the alignment. The algorithm makes the adjacent DP matches of stem components either not overlapping in the nucleotide sequences or consistently overlapping (1-continuous) as a match of the stems longer than \( W \). Pairwise alignment of the SCs requires only \( O(|X||Y|) \) in time and in memory. That computational complexities are evaluated as \( L^2 \) for two RNA sequences of length \( L \) because the number of the stem components is regarded as a linear function of the nucleotide sequence [25].

The pairwise alignment of SCs allows some inconsistent matches by ignoring strict treatments of the complementary components. For two stem fragments, \( [X_{n'}, X_{p}] \) and \( [Y_{m'}, Y_{l}] \), if \( X_{n'} \) matches \( Y_m \) in the SCs alignment, \( X_{n'} \) should match \( Y_{l} \). Let us define such a match as left-right consistent. Because 5’ stem components and 3’ stem components are aligned independently, left-right consistency is not guaranteed in general. Any match which is not left-right consistent is removed as a post process. If any two of the stem components of a same SCs appear in the SCs alignment and their complementary components overlap (i.e. contradictory in Definition 6), those complementary components do not appear together in the alignment because the alignment of complementary components are controlled to be either nonoverlapping or r-continuous. Therefore, the post process also guarantees that no pair of contradictory stem fragments appears in the alignment [25].

The score function using the MEA alignment

In our previous work [25], a function of the RIBOSUM [28] score, loop distance, base-pairing probabilities, and the stacking energy were used as the score \( s(i, j) \) in recursion (1). In MXSCARNA, the score function is replaced by an approximated posterior probability according to the principle of Maximum Expected Accuracy (MEA). Recent studies have shown that the accuracy of the resulting sequence alignment and secondary structure predictions is better than that of predictions made by the conventional maximum likelihood estimation (MLE) algorithms [21,29-32].

In the following, for nucleotide sequences \( x \) and \( y \), \( x_i \sim y_j \) means that \( x_i \in x \) and \( y_j \in y \) are aligned on the same column in the alignment, and \( x_i \sim x_j \) means that \( x_i \) and \( x_j \) form a base pair. For two RNA sequences, \( x \), \( y \) and \( k \), \( l \) \( \in \{ 1, \cdot \cdot \cdot, |x| \} \), \( m \), \( n \) \( \in \{ 1, \cdot \cdot \cdot, |y| \} \), let \( P(x_m \sim y_n, x_k \sim y_l, x_{m'} \sim y_{n'}, \sim x_j, y_{j'} \) be the posterior probability, i.e. the sum of the probabilities that two positions of the sequences, \( x_m \) and \( y_n \), \( x_{m'} \) and \( y_{n'} \), \( x_k \) and \( y_l \) are aligned in the alignment, and that two pairs of the nucleotides, \( x_k \) and \( x_{m'} \), \( y_j \) and \( y_l \) form base pairs in the secondary structures as well; this is computed by the inside-outside algorithm of the pair Stochastic Context Free Grammar (pair SCFG) [5] for structural pairwise alignments of RNA sequences. We wanted to use posterior probability as the score function \( s(i, j) \), but the computational costs, \( O(L^6) \) in time and \( O(L^4) \) in memory for sequences of length \( L \) are impractical. We instead used the following approximated posterior probability introduced by Kiryu et al. [21].

\[
P(x_k \sim y_m, x_j \sim y_{m'}, x_k \sim y_{l}, y_{n}, \sim x_j, y_{j'}) = P(x_k \sim y_m | x, y) P(x_j \sim y_n | x, y) P(y_m \sim y_n | x, y) P(y_{m'} \sim y_{n'} | x, y)
\]

\( P(x_k \sim y_l | x, y) \) and \( P(y_m \sim y_n | x, y) \) are the base-pairing probabilities that the particular positions \( x_k \) and \( x_{m'} \), \( y_m \) and \( y_{n'} \).
A pair-HMM for pairwise sequence alignment. A pair-HMM is used for alignment of loop regions and calculation of the posterior probabilities in score function. The state $M$ has emission probability distribution $p_{x,y}$ for emitting an aligned pair $x$ and $y$. The state $l_x$ has distributions $q_{x_{i}}$ for emitting symbol $x_i$ against a gap. The state $l_y$ has distributions $q_{y_{j}}$ for emitting symbol $y_j$ against a gap. The parameters $\delta$ and $\varepsilon$ are the state transition probabilities.

respectively, form base pairs; these probabilities are computed by McCaskill’s algorithm \[32\].

$$P(x_k \sim y_m | x, y)$$ and $$P(x_k \sim y_n | x, y)$$ are the posterior probabilities modified by probability consistency transformation \[32\], which is computed as follows.

$$\tilde{P}(x_k \sim y_m | x, y) = \frac{1}{|S|} \sum_{z \in S} \sum_{i=1}^{1-L} P(x_k \sim z_i | x, z)P(z_i \sim y_m | z, y). \quad (3)$$

where $S$ is the set of RNA sequences to be aligned. In this transformation, the probability of specific nucleotides of two sequences being aligned are replaced by the average over the products of probabilities that the two nucleotides are aligned to the same nucleotides in arbitrary third sequences. This calculation requires $O(N^3L^3)$ in time and $O(N^2L^2)$ in memory. The probability consistency transformations are applied twice in current implementation.

$P(x_k \sim z_i | x, z)$ is the posterior probability, i.e. the sum of the probabilities that particular positions of the two sequences, $x_k$ and $z_i$, are aligned in some alignment; this is computed by the forward-backward algorithm of the pair Hidden Markov Model (pair HMM) \[31\] for pairwise alignment of the sequences. Our new matching scores in (1) are defined as follows.

$$s(i, j) = \sum_{0 \leq w \leq W} P(x_{p(X_i)+w} \sim y_{p(Y_j)+w}, x_{p(X_i)+W-1-w} \sim y_{p(Y_j)+W-1-w} ) \quad (4)$$

$$\delta_s(i, j) = P(x_{p(X_i)+W-1} \sim y_{p(Y_j)+W-1}, x_{p(X_i)} \sim y_{p(Y_j)} ) \quad (5)$$

where $X'_i/Y'_j$ are the complementary stem components of $X_i/Y_j$.

The sum of the probabilities, not the logarithms of the probabilities, is used for the matching score, in an effort to maximize the number of correctly aligned bases including the implicit prediction of the base pairs (MEA principle).

Alignment of loop region

The remaining loop regions (except the selected common stems) are aligned by using the consistency-transferred posterior probabilities, $\tilde{P}(x_k \sim y_m | x, y)$, as the matching scores. The probabilities, not the logarithms of the probabilities, again are used, according to the MEA principle. The recursion is shown following.

$$A(k, m) = \max \begin{cases} A(k-1, m-1) + \tilde{P}(x_k \sim y_m | x, y) \\ A(k-1, m) \\ A(k, m-1) \end{cases}$$

Emission and transition probabilities for the pair HMM in MXSCARNA (Figure 3) were trained via Expectation-Maximization (EM) on a set of unaligned sequences that is extracted from the Rfam database and that do not overlap the sequences of the dataset for subsequent experiments.

Computational Experiments

Datasets

To test the empirical performance of MXSCARNA, we used three datasets for the benchmark multiple alignments: an original multiple alignment dataset, the BRAli-basell multiple alignment dataset \[33\], and Kiryu et al.’s multiple alignment dataset \[21\].

Our original dataset comprised 1669 multiple alignments of 5 sequences, the secondary structures of which have been published, obtained from the Rfam 7.0 database \[34\]. There are 27 families of RNA sequences in the dataset and the sequence identities varied from 35% to 100%. Sequences that included bases other than A, C, G, and U
were removed because some of the alignment programs were unable to align them. The BRAliBaseII benchmark dataset included 481 multiple alignments of 5 sequences. The sequences of each multiple alignment were extracted from tRNA, Intron,gpII, 5S_rRNA, and U5 families in the Rfam 5.0 database and the signal recognition particle RNA family (SRP) in the SRPDB database [35]. Because the dataset did not include consensus secondary structure annotations to the alignments, we used the secondary structure annotations recovered by Kiryu et al. [21]. Kiryu et al.’s multiple alignment benchmark dataset was generated from selected seed alignments in the Rfam 7.0 database that have published consensus structures [21]. For each sequence family, as many as 1000 random combinations of 10 sequences were generated. The alignments whose mean pairwise sequence identity exceeded 95% and whose gap characters accounted for more than 30% of the total number of characters aligned were removed. As such, this dataset consisted of 85 multiple alignments of 10 sequences, generated from 17 sequence families, with five alignments for each. The dataset was reasonably divergent, and its mean length varied from 54 to 291 bases, and mean pairwise sequence identities varied from 40% to 94%.

Evaluation measures

The qualities of the alignments were evaluated by the Sum-of-Pairs Score (SPS) for the accuracy of the alignments and by the Matthews Correlation Coefficient (MCC) [36] for the accuracy of the secondary structure predictions. The SPS and MCC of the alignment to be evaluated (named as a test alignment) for the reference alignment were defined as follows. The SPS was defined as the proportion of correctly aligned nucleotide pairs:

\[
SPS = \frac{\sum_{i=1}^{I} SP_{i}}{\sum_{j=1}^{J} SP_{j}},
\]

where \( I \) is the number of columns in the test alignment, \( J \) is the number of columns in the reference alignment, on column \( i \) in the test alignment \( SP_{i} \) is the total number of "correct" nucleotide pairs which also appear in the reference alignment, on column \( j \) in the reference alignment \( SP_{j} \) is the total number of nucleotide pairs. The MCC was defined as

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},
\]

\[
FP' = FP - \xi,
\]

where TP indicates the number of correctly predicted base pairs, TN the number of base pairs that were correctly predicted as unpaired, FP the number of incorrectly predicted base pairs, and FN the number of true base pairs that were not predicted. The term \( \xi \) accounts for predicted base pairs that were not present in the reference structure but were compatible with it. Compatible base pairs are not true positives but have to be neither inconsistent (one or both nucleotides being a part of a different base pair in the reference structure) nor pseudo-knotted with respect to the reference structure [37]. In order to calculate MCC for each test alignment, the reference alignment and the "correct" consensus secondary structure are taken from the database. In order to compare the accuracies of the alignments in terms of the implicitly predicted common secondary structures, the common secondary structures for each test alignment by the alignment programs were predicted by the Pfold program [5].
Table 2: Accuracies for the original multiple alignment dataset. SPS and MCC values (%) for the original multiple alignment dataset are presented. Each family has 5 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the average SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences/total no. of sequences.

| SPS:Family | SPSCARNA | Murlet | ProbCons | MAFFT | ClustalW | StrAl | RNASampler |
|------------|----------|--------|----------|--------|---------|------|-----------|
| IRE        | 62       | 29     | 98       | 96     | 93      | 89   | 77        | 83 92     |
| s2m        | 78       | 43     | 96       | 96     | 96      | 96   | 96        | 95 82     |
| UnaL2      | 78       | 54     | 95       | 98     | 98      | 96   | 92        | 94 82     |
| Hammerhead_3| 71       | 55     | 96       | 94     | 93      | 89   | 83        | 89 91     |
| SECIS       | 42       | 64     | 65       | 66     | 65      | 52   | 45        | 56 46     |
| sno_14q_I_II| 71      | 74     | 90       | 95     | 96      | 93   | 93        | 87 71     |
| tRNA       | 48       | 76     | 87       | 87     | 87      | 84   | 76        | 82 87     |
| ctRNA_pGA1 | 74       | 80     | 88       | 86     | 85      | 88   | 84        | 86 89     |
| Tymo_tRNA  | 70       | 83     | 84       | 83     | 82      | 75   | 75        | 79 77     |
| Y          | 64       | 95     | 71       | 72     | 72      | 73   | 65        | 64 65     |
| SRP_bact   | 52       | 95     | 71       | 71     | 70      | 66   | 70        | 62 66     |
| Purine     | 55       | 100    | 74       | 77     | 77      | 78   | 75        | 81 69     |
| 5S_rRNA    | 60       | 117    | 88       | 89     | 89      | 86   | 86        | 86 83     |
| S_box      | 66       | 130    | 79       | 80     | 78      | 78   | 68        | 72 67     |
| U4         | 67       | 141    | 79       | 81     | 80      | 79   | 79        | 79 69     |
| RFN        | 65       | 150    | 86       | 87     | 87      | 88   | 81        | 80 81     |
| 5_8S_rRNA  | 67       | 154    | 91       | 93     | 93      | 90   | 88        | 89 78     |
| U1         | 60       | 158    | 81       | 81     | 79      | 79   | 79        | 76 74     |
| Telomerase_cil | 56      | 171    | 48       | 50     | 49      | 43   | 38        | 41 37     |
| Lysine     | 50       | 180    | 78       | 81     | 79      | 72   | 70        | 76 72     |
| U2         | 66       | 185    | 76       | 76     | 75      | 71   | 71        | 73 71     |
| U17        | 75       | 214    | 91       | 93     | 93      | 90   | 89        | 89 87     |
| U3         | 51       | 246    | 43       | 44     | 44      | 47   | 41        | 40 34     |
| SRP_euk_arch | 46      | 294    | 50       | 56     | 47      | 42   | 42        | 48 44     |
| tmRNA      | 46       | 373    | 48       | 50     | 50      | 47   | 46        | 39 42     |
| RnaseP_bact_b | 64      | 387    | 82       | 80     | 79      | 78   | 74        | 74 66     |
| Telomerase_vert | 66     | 463    | 69       | 70     | 69      | 69   | 66        | 64 65     |

Average(all) 78 79 78 75 72 73 70
Average(sub) 85 85 85 82 78 80 78

| SPS:Family | RNAAlara | Locarna | FoldalignM(1) | FoldalignM(2) | MARNA | PMmulti | Stemloc |
|------------|-----------|---------|---------------|---------------|-------|---------|--------|
| IRE        | 85       | 88 (100/100) | 87 (99/101) | 85 (97/100) | 92   | 90 (101/101) | 87 (101/101) |
| s2m        | 96       | 99 (43/50) | 95 (50/50) | 95 (47/50) | 94   | 98 (44/50) | 93 (50/50) |
| UnaL2      | 93       | 93 (78/89) | 89 (75/89) | 86 (87/89) | 88   | 87 (75/89) | 75 (89/89) |
| Hammerhead_3| 88       | 80 (100/100) | 79 (99/100) | 77 (97/100) | 89   | 87 (77/100) | 94 (100/100) |
| SECIS       | 48       | 47 (62/63) | 46 (63/63) | 44 (63/63) | 49   | 39 (63/63) | 60 (60/63) |
| sno_14q_I_II| 78       | 76 (98/98) | 73 (98/98) | 61 (97/98) | 67   | 73 (97/98) | 89 (97/98) |
| tRNA       | 91       | 82 (103/103) | 90 (100/103) | 78 (97/103) | 59   | 86 (103/103) | 88 (103/103) |
| ctRNA_pGA1 | 86       | 74 (20/28) | 75 (28/28) | 74 (27/28) | 79   | 72 (20/28) | 84 (28/28) |
| Tymo_tRNA  | 79       | 78 (49/59) | 68 (59/59) | 68 (56/59) | 66   | 71 (49/59) | 57 (57/59) |
| Y          | 56       | 49 (21/24) | 50 (24/24) | 47 (24/24) | 60   | 43 (11/24) | 68 (23/24) |
| SRP_bact   | 63       | 62 (70/70) | 64 (70/70) | 56 (67/70) | 56   | 61 (63/70) | 60 (67/70) |
| Purine     | 64       | 65 (45/45) | 64 (45/45) | 62 (45/45) | 64   | 65 (45/45) | 37 (45/45) |
| 5S_rRNA    | 85       | 57       | 71       | 78       | 81   | 74       | 32       |
| S_box      | 66       | 463 | 69       | 70       | 69   | 66       | 64 65     |
| U4         | 78       | 85       | 85       | 82       | 78   | 80       | 78 78     |
Table 2: Accuracies for the original multiple alignment dataset. SPS and MCC values (%) for the original multiple alignment dataset are presented. Each family has 5 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the average SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences aligned/total no. of sequences.

FoldalignM consists of two modes: FoldalignM_FoldalignM and FoldalignM_McCaskill, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively. (Continued)

| Lysine | 59 |
| U2    | 71 |
| U17   | 79 |
| U3    | 32 |
| SRP_euk_arch | 42 |
| tmRNA | 34 |
| RnaseP_bact_b | 66 |
| Telomerase_vert | 51 |

Average(all) 68

| Average(sub) | 77 | 74 | 73 | 69 | 72 | 73 | 74 |

| MCC:Family | %id | length | MXSCARNA | Murlet | ProbCons | MAFFT | ClustalW | StrAl | RNAsampler |
|------------|-----|--------|----------|--------|----------|--------|----------|-------|------------|
| IRE        | 62  | 29     | 90       | 91     | 75       | 72     | 65       | 43    | 89         |
| S2m        | 78  | 43     | 84       | 83     | 84       | 84     | 84       | 81    | 81         |
| UnaL2      | 78  | 54     | 52       | 70     | 51       | 53     | 51       | 50    | 51         |
| Hammerhead_3 | 71  | 55     | 99       | 95     | 93       | 86     | 72       | 79    | 96         |
| SECIS      | 42  | 64     | 76       | 60     | 55       | 35     | 23       | 37    | 73         |
| sno_14q_I_II | 71 | 74     | 93       | 98     | 93       | 93     | 91       | 91    | 95         |
| rRNA       | 48  | 76     | 91       | 89     | 86       | 84     | 76       | 83    | 95         |
| ctRNA_pGA1 | 74  | 80     | 96       | 93     | 88       | 89     | 81       | 92    | 94         |
| Tymo_rRNA  | 70  | 83     | 87       | 85     | 75       | 73     | 72       | 80    | 90         |
| Y         | 64  | 95     | 95       | 85     | 86       | 83     | 67       | 83    | 94         |
| SRP_bact   | 52  | 95     | 81       | 72     | 54       | 50     | 58       | 57    | 83         |
| Purine     | 55  | 100    | 90       | 94     | 90       | 86     | 80       | 84    | 91         |
| 5S_rRNA    | 60  | 117    | 75       | 79     | 69       | 70     | 69       | 70    | 70         |
| S_box      | 66  | 130    | 90       | 87     | 86       | 81     | 76       | 81    | 86         |
| U4         | 67  | 141    | 75       | 71     | 62       | 62     | 54       | 65    | 67         |
| RFN        | 65  | 150    | 84       | 83     | 84       | 84     | 82       | 83    | 82         |
| 5_8S_rRNA  | 67  | 154    | 58       | 51     | 47       | 45     | 41       | 46    | 52         |
| U1         | 60  | 158    | 70       | 68     | 61       | 56     | 60       | 57    | 71         |
| Telomerase  | 56  | 171    | 65       | 41     | 28       | 21     | 24       | 31    | 60         |
| Lysine     | 50  | 180    | 87       | 90     | 76       | 66     | 63       | 71    | 89         |
| U2         | 66  | 185    | 73       | 76     | 58       | 51     | 62       | 59    | 77         |
| U17        | 75  | 214    | 79       | 80     | 78       | 76     | 75       | 72    | 72         |
| U3         | 51  | 246    | 46       | 26     | 22       | 19     | 46       | 21    | 39         |
| SRP_euk_arch | 46 | 294    | 72       | 75     | 46       | 37     | 35       | 49    | 72         |
| tmRNA      | 46  | 373    | 51       | 54     | 50       | 49     | 43       | 42    | 49         |
| RnaseP_bact_b | 64 | 387    | 73       | 58     | 63       | 58     | 53       | 60    | 37         |
| Telomerase  | 66  | 463    | 64       | 51     | 47       | 44     | 40       | 36    | 53         |

Average(all) 78 | 74 | 67 | 63 | 61 | 63 | 74

Average(sub) 86 | 85 | 77 | 74 | 68 | 72 | 86

| MCC:Family | RNAlama | Locarna | FoldalignM(1) | FoldalignM(2) | MARNA | PMmulti | Stemloc |
|------------|---------|---------|---------------|---------------|-------|---------|---------|
| IRE        | 81      | 89      | 85            | 85            | 82    | 89      | 81      |
| s2m        | 84      | 83      | 85            | 86            | 79    | 85      | 84      |
| UnaL2      | 53      | 53      | 53            | 51            | 51    | 50      | 45      |
| Hammerhead_3 | 95   | 94      | 87            | 91            | 95    | 91      | 98      |
| SECIS      | 63      | 74      | 74            | 75            | 57    | 67      | 78      |
| sno_14q_I_II | 92   | 87      | 92            | 84            | 81    | 93      | 85      |
| tRNA       | 95      | 87      | 95            | 87            | 67    | 90      | 95      |
Table 2: Accuracies for the original multiple alignment dataset. SPS and MCC values (%) for the original multiple alignment dataset are presented. Each family has 5 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the average SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences aligned/total no. of sequences.

FoldalignM consists of two modes: FoldalignM_FoldalignM and FoldalignM_McCaskill, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively. (Continued)

| Family          | SPS 95 | SPS 96 | SPS 97 | SPS 98 | MCC 95 | MCC 96 | MCC 97 | MCC 98 |
|-----------------|--------|--------|--------|--------|--------|--------|--------|--------|
| ctRNA_pGa1      | 95     | 95     | 97     | 95     | 95     | 95     | 96     | 96     |
| Tymo_tRNA       | 85     | 87     | 87     | 82     | 62     | 82     | 83     | 83     |
| Y               | 86     | 87     | 94     | 89     | 87     | 83     | 83     | 93     |
| SRP_bact        | 65     | 87     | 86     | 83     | 63     | 80     | 69     |        |
| Purine          | 72     | 89     | 88     | 88     | 82     | 86     | 89     |        |
| S5_rRNA         |        |        |        |        |        |        |        |        |
| S_box           | 72     |        |        |        |        |        |        |        |
| U4              | 72     |        |        |        |        |        |        |        |
| RFN             | 79     |        |        |        |        |        |        |        |
| 5_8S_rRNA       | 46     |        |        |        |        |        |        |        |
| U1              | 60     |        |        |        |        |        |        |        |
| Telomerase      |        |        |        |        |        |        |        |        |
| Lysine          | 72     |        |        |        |        |        |        |        |
| U2              | 72     |        |        |        |        |        |        |        |
| U17             | 72     |        |        |        |        |        |        |        |
| U3              | 72     |        |        |        |        |        |        |        |
| SRP_euk_arch    | 72     |        |        |        |        |        |        |        |
| tmRNA           | 72     |        |        |        |        |        |        |        |
| RNaseP_bact     | 72     |        |        |        |        |        |        |        |
| Telomerase      | 46     |        |        |        |        |        |        |        |
| Average(all)    | 65     |        |        |        |        |        |        |        |
| Average(sub)    | 81     | 83     | 85     | 85     | 75     | 83     | 83     |        |

Comparison of accuracies with those of other aligners

To compare the accuracies of the alignment methods we used a Linux machine with an AMD Opteron processor (2 GHz and 4 GB RAM).

We compared the performance of MXSCARNA with that of Murlet [21], ProbCons [32], MAFFT [38], ClustalW [7], StrAl [18], MARNa [17], RNAsampler [19], RNRAlara [20], FoldalignM [15], Locarna [16], PMMulti [12], and Stemloc [13] on the three datasets described earlier. Whereas ProbCons, MAFFT, and ClustalW align RNA sequences on the basis of sequence similarities only, StrAl, MARNa, RNAsampler, RNRAlara, FoldalignM, Locarna, PMMulti, Stemloc, and Murlet weigh both sequence similarities and secondary structures. The command line options for the programs in the experiments are shown in Table 1. The results for the original dataset are shown in Table 2. Because MARNa, Locarna, FoldalignM, PMMulti, and Stemloc impose high time and memory demands, those programs were executed only on families of which the average sequence lengths were less than or equal to 100 bases. The SPS of MXSCARNA was comparable to those of Murlet and ProbCons, which currently are the best performing aligners [21]. In addition, the MCC of MXSCARNA was one of the highest among aligners. In particular, the MCC of MXSCARNA is similar to that of Stemloc, which aligns only short sequences that have simple secondary structures.

The results from the BRAlibaseII benchmark multiple alignment dataset are shown in Table 3. Because of their prohibitive requirements for memory and time, Stemloc, FoldalignM, PMmulti, and MARNa were not applied to the SRP family data. Again, MXSCARNA was comparable to Murlet and ProbCons in terms of SPS and one of the best performers among multiple aligners according to the MCC. These trends continue in Table 4, which contains the results from Kiryu et al.’s benchmark dataset comprising 10 sequences for each alignment.

All results are summarized in Table 5.

Evaluation of new score function

In order to evaluate the performance of our new score function (4), we compared it in pairwise alignment with the previous score function of SCARNA, which is a linear combination of RIBOSUM score, stacking energy, loop-distance penalty, base-pairing probability. Dowell’s dataset [39], which consists of R100 dataset and percid dataset, are used for the evaluation. R100 is a dataset which consists of 100 pairwise alignments chosen randomly from tRNA and 5SrRNA families in Rfam 7.0 database.
Table 3: Accuracies for the BRAlibaseII multiple alignment dataset. SPS and MCC values (%) for the BRAlibaseII multiple alignment dataset are presented. Each family has 5 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the results of the average value of the SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences aligned/total no. of sequences. FoldalignM consists of two modes: FoldalignM_FoldalignM and FoldalignM_McCaskill, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively.

| SPS:Family | tRNA | Intron_gpl | 5S_rRNA | U5 | SRP |
|------------|------|------------|---------|----|-----|
| %id        | 69   | 64         | 70      | 72 | 67  |
| length     | 76   | 80         | 117     | 118| 300 |
| MXSCARNA   | 91   | 91         | 89      | 74 | 88  |
| Murlet     | 91   | 90         | 90      | 75 | 88  |
| ProbCons   | 91   | 80         | 90      | 76 | 88  |
| MAFFT      | 89   | 77         | 90      | 72 | 87  |
| ClustalW   | 85   | 75         | 89      | 72 | 86  |
| StrAl      | 89   | 75         | 89      | 73 | 86  |
| RNAsampler | 92   | 79         | 90      | 78 | 82  |
| Average(all)| 84 | 85 | 85 | 83 | 81 |
| Average(sub)| 83 | 84 | 84 | 82 | 80 |

| MCC:Family | tRNA | Intron_gpl | 5S_rRNA | U5 | SRP |
|------------|------|------------|---------|----|-----|
| %id        | 69   | 64         | 70      | 72 | 67  |
| length     | 76   | 80         | 117     | 118| 300 |
| MXSCARNA   | 94   | 90         | 92      | 69 | 72  |
| Murlet     | 92   | 90         | 91      | 72 | 71  |
| ProbCons   | 91   | 80         | 80      | 75 | 72  |
| MAFFT      | 90   | 77         | 91      | 72 | 68  |
| ClustalW   | 83   | 76         | 80      | 77 | 68  |
| StrAl      | 88   | 74         | 76      | 74 | 67  |
| RNAsampler | 94   | 83         | 84      | 80 | 76  |
| Average(all)| 80 | 80 | 78 | 75 | 73 |
| Average(sub)| 80 | 79 | 76 | 75 | 73 |

| SPS:Family | tRNA | Intron_gpl | 5S_rRNA | U5 | SRP |
|------------|------|------------|---------|----|-----|
| %id        | 69   | 64         | 70      | 72 | 67  |
| length     | 76   | 80         | 117     | 118| 300 |
| MXSCARNA   | 94   | 90         | 92      | 69 | 72  |
| Murlet     | 92   | 90         | 91      | 72 | 71  |
| ProbCons   | 91   | 80         | 80      | 75 | 72  |
| MAFFT      | 90   | 77         | 91      | 72 | 68  |
| ClustalW   | 83   | 76         | 80      | 77 | 68  |
| StrAl      | 88   | 74         | 76      | 74 | 67  |
| RNAsampler | 94   | 83         | 84      | 80 | 76  |
| Average(all)| 80 | 80 | 78 | 75 | 73 |
| Average(sub)| 80 | 79 | 76 | 75 | 73 |

| MCC:Family | tRNA | Intron_gpl | 5S_rRNA | U5 | SRP |
|------------|------|------------|---------|----|-----|
| %id        | 69   | 64         | 70      | 72 | 67  |
| length     | 76   | 80         | 117     | 118| 300 |
| MXSCARNA   | 94   | 90         | 92      | 69 | 72  |
| Murlet     | 92   | 90         | 91      | 72 | 71  |
| ProbCons   | 91   | 80         | 80      | 75 | 72  |
| MAFFT      | 90   | 77         | 91      | 72 | 68  |
| ClustalW   | 83   | 76         | 80      | 77 | 68  |
| StrAl      | 88   | 74         | 76      | 74 | 67  |
| RNAsampler | 94   | 83         | 84      | 80 | 76  |
| Average(all)| 80 | 80 | 78 | 75 | 73 |
| Average(sub)| 80 | 79 | 76 | 75 | 73 |

| SPS:Family | tRNA | Intron_gpl | 5S_rRNA | U5 | SRP |
|------------|------|------------|---------|----|-----|
| %id        | 69   | 64         | 70      | 72 | 67  |
| length     | 76   | 80         | 117     | 118| 300 |
| MXSCARNA   | 94   | 90         | 92      | 69 | 72  |
| Murlet     | 92   | 90         | 91      | 72 | 71  |
| ProbCons   | 91   | 80         | 80      | 75 | 72  |
| MAFFT      | 90   | 77         | 91      | 72 | 68  |
| ClustalW   | 83   | 76         | 80      | 77 | 68  |
| StrAl      | 88   | 74         | 76      | 74 | 67  |
| RNAsampler | 94   | 83         | 84      | 80 | 76  |
| Average(all)| 80 | 80 | 78 | 75 | 73 |
| Average(sub)| 80 | 79 | 76 | 75 | 73 | 79
Table 4: Accuracies for Kiryu et al.'s dataset. SPS and MCC values (%) for Kiryu et al.'s dataset are presented. Each family has 10 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the results of the average value of the SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences aligned/total no. of sequences. FoldalignM consists of two modes: FoldalignM熟悉 FoldalignM and FoldalignM McCaskill, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively.

| Family     | %id | length(nt) | MXSCARNA | Murlet | ProbCons | MAFFT | ClustalW | StrAl | RNAAlara |
|------------|-----|------------|----------|--------|----------|-------|----------|-------|----------|
| SPS:Family |     |            |          |        |          |       |          |       |          |
| SPS:Family | RNASampler | Locarna | FoldalignM(1) | FoldalignM(2) | MARNA | PMmulti | Stemloc |
| UnaL2      | 72  | 88         | 68       | 60     | 83 (5/5) | 69 (5/5) | 82 (5/5) |
| SECIS      | 54  | 49         | 42       | 41     | 47 (5/5) | 35 (5/5) | 82 (5/5) |
| tRNA       | 82  | 79         | 84       | 66     | 54 (5/5) | 69 (5/5) | 91 (5/5) |
| sno_14q_I_II | 64  | 87         | 54       | 50     | 49 (5/5) | 39 (3/5) | 75 (7/5) |
| SRP_bact   | 54  | 52         | 55       | 51     | 43 (5/5) | 36 (4/5) | 47 (3/5) |
| THI        | 68  | 65         | 65       | 62     | 62 (4/5) | 58 (5/5) | 71 (5/5) |
| S_box      | 87  | 57         | 57       | 78 (5) | 57 (5) | 57 (5) | 77 (3/5) |
| SRP_euk_arch | 96  | 95         | 91       | 91     | 96 (5/5) | 87 (5/5) | 75 (5/5) |
| RFN        | 82  | 72         | 73       | 63     | 77 (5/5) | 58 (4/5) | 80 (5/5) |
| U1         | 63  | 68         | 50       | 50     | 50       | 50     | 50     |
| Lysine     | 71  | 55         | 58       | 58     | 58       | 58     | 58     |
| U2         | 65  | 64         | 65       | 65     | 65       | 65     | 65     |
| T-box      | 32  | 15         | 22       | 22     | 22       | 22     | 22     |
| IRES_HCV   | 93  | 75         | 92       | 92     | 92       | 92     | 92     |
| SRP_euk_arch | 33  | 40         | 37       | 37     | 37       | 37     | 37     |
| Average(all)| 68  | 64         | 62       | 62     | 62       | 62     | 62     |
| Average(sub)| 72  | 70         | 64       | 57     | 52       | 60     | 76     |

| MCC:Family | %id | length(nt) | MXSCARNA | Murlet | ProbCons | MAFFT | ClustalW | StrAl | RNAAlara |
|------------|-----|------------|----------|--------|----------|-------|----------|-------|----------|
| MCC:Family |     |            |          |        |          |       |          |       |          |
| MCC:Family | RNASampler | Locarna | FoldalignM(1) | FoldalignM(2) | MARNA | PMmulti | Stemloc |
| UnaL2      | 73  | 54         | 42       | 41     | 46       | 36    | 24       | 32    | 44       |
| SECIS      | 41  | 64         | 78       | 78     | 59       | 20    | 23       | 45    | 70       |
| tRNA       | 45  | 73         | 93       | 97     | 91       | 85    | 65       | 85    | 97       |
| sno_14q_I_II | 64  | 75         | 87       | 91     | 91       | 91    | 66       | 75    | 87       |
Table 4: Accuracies for Kiryu et al.’s dataset. SPS and MCC values (%) for Kiryu et al.’s dataset are presented. Each family has 10 RNA sequences. Family: Rfam family name. %id: average sequence identity. length: average sequence length (bases) in each family. Average(all): the results of the average value of the SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families with an average sequence length of less than or equal to 100 bases. Because PMmulti and Stemloc were unable to align all data, the proportion of data that was aligned is given in parentheses as no. of sequences aligned/total no. of sequences. FoldalignM consists of two modes: FoldalignM_FoldalignM and FoldalignM McCaskill, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively. (Continued)

|        | RNASampler | Locarna | FoldalignM(1) | FoldalignM(2) | MARNa | PMmulti | Stemloc |
|--------|------------|---------|---------------|---------------|-------|---------|---------|
| UnaL2  | 51         | 42      | 57            | 50            | 36    | 18      | 39      |
| tRNA   | 78         | 80      | 75            | 80            | 40    | 64      | 77      |
| sno_144 | 94        | 96      | 95            | 89            | 51    | 91      | 98      |
| SRP_bact| 73         | 74      | 80            | 75            | 47    | 44      | 64      |
| THI    | 72         | 60      | 50            | 51            | 59    | 41      | 77      |
| S_box  | 86         | 82      | 83            | 87            | 83    | 55      | 88      |
| S_rRNA | 62         | 75      | 72            | 73            | 59    | 58      | 66      |
| Retroviral_psi | 86 | 88 | 76 | 89 | 84 | 73 | 87 |
| RFN    | 69         | 69      | 67            | 70            | 65    | 58      | 72      |
| S_8S_rRNA | 38     | 40      | 43            | 41            | 21    | 14      | 34      |
| U1     | 64         | 73      | 45            |              |       |         |         |
| Lysine | 86         | 80      | 66            |              |       |         |         |
| U2     | 79         | 65      | 66            |              |       |         |         |
| T-box  | 44         | 1       | 2             |              |       |         |         |
| IRES_HCV | 63     | 47      | 61            |              |       |         |         |
| SRP_euk_arch | 62 | 64   | 52            |              |       |         |         |

Average(all) 72 71 65 54 48 54 61

Average(sub) 72 72 68 60 52 59 69

|        | MCC:Family | RNA | Sampler | Locarna | FoldalignM(1) | FoldalignM(2) | MARNa | PMmulti | Stemloc |
|--------|------------|-----|---------|---------|---------------|---------------|-------|---------|---------|
| SPS    |            |     | RFN     |         |               |               |       |         |         |
| Kiryu  |            |     |         |         |               |               |       |         |         |

Table 6: Accuracies for the different scorers. SPS and MCC values (%) for the SPS and MCC of the different scorers are presented. The scorers considered are RNASampler, Locarna, FoldalignM(1), FoldalignM(2), MARNa, PMmulti, and Stemloc. The SPS and MCC are shown in Table 6. It is observed that the new score function of MXSCARNA outperformed the previous score function of SCARNA.

Time and memory
The computational complexities of the proposed method for $N$ sequences of length $L$ were evaluated as follows. The construction of the guide tree using the alignments of all pairs of the sequences required $O(N^2L^2)$ in time and $O(L^2 + N^2)$ in memory. The calculation of base-pairing probability matrices for $N$ sequences by McCaskill’s algorithm [27] required $O(NL^3)$ in time and $O(NL^2)$ in memory. The probability consistency transformation (see (3) in Method) required $O(N^3L^3)$ in time and $O(N^2L^2)$ in memory. Pairwise alignment of stem component sequences required $O(N^2L^2)$ in time and memory as is explained in Method. Therefore, the total computational complexities were $O(N^3L^3)$ in time and $O(N^2L^2)$ in memory. For the base-pairing probabilities, the computational time for

|        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|
| SRP_bact| 47     | 93     | 66     | 56     | 46     | 49     | 54     |
| THI    | 55     | 105    | 71     | 70     | 70     | 62     | 38     | 48     | 58     |
| S_box  | 66     | 107    | 90     | 89     | 87     | 79     | 77     | 75     | 76     |
| S_8S_rRNA | 57   | 116    | 75     | 67     | 62     | 64     | 53     | 66     | 69     |
| Retroviral_psi | 92 | 117    | 86     | 86     | 86     | 84     | 86     | 86     |
| RFN    | 66     | 140    | 67     | 71     | 72     | 73     | 71     | 60     | 70     |
| S_8S_rRNA | 61   | 154    | 38     | 43     | 35     | 16     | 14     | 26     | 33     |
| U1     | 59     | 157    | 69     | 61     | 57     | 56     | 61     | 52     | 56     |
| Lysine | 49     | 181    | 83     | 81     | 71     | 33     | 52     | 61     | 64     |
| U2     | 62     | 182    | 74     | 71     | 71     | 56     | 38     | 39     | 58     | 68     |
| T-box  | 45     | 244    | 72     | 78     | 80     | 51     | 41     | 26     | 0      | 0      |
| IRES_HCV | 94 | 261    | 63     | 62     | 62     | 63     | 26     | 34     | 63     |
| SRP_euk_arch | 40 | 291    | 70     | 63     | 40     | 21     | 23     | 38     | 33     |

Average(all) 72 71 65 54 48 54 61

Average(sub) 72 72 68 60 52 59 69

|        | RNA | Sampler | Locarna | FoldalignM(1) | FoldalignM(2) | MARNa | PMmulti | Stemloc |
|--------|-----|---------|---------|---------------|---------------|-------|---------|---------|
| UnaL2  | 51  | 42      | 57      | 50            | 36            | 18    | 39      |         |
| SECIS  | 78  | 80      | 75      | 80            | 40            | 64    | 77      |         |
| tRNA   | 94  | 96      | 95      | 89            | 51            | 91    | 98      |         |
| sno_144L_II | 84 | 95      | 95      | 86            | 79            | 77    | 87      |         |
| SRP_bact| 73  | 74      | 80      | 75            | 47            | 44    | 64      |         |
| THI    | 72  | 60      | 50      | 51            | 59            | 41    | 77      |         |
| S_box  | 86  | 82      | 83      | 87            | 83            | 55    | 88      |         |
| S_rRNA | 62  | 75      | 72      | 73            | 59            | 58    | 66      |         |
| Retroviral_psi | 86 | 88      | 76      | 89            | 84            | 73    | 87      |         |
| RFN    | 69  | 69      | 67      | 70            | 65            | 58    | 72      |         |
| S_8S_rRNA | 38 | 40      | 43      | 41            | 21            | 14    | 34      |         |
| U1     | 64  | 73      | 45      |               |               |       |         |         |
| Lysine | 86  | 80      | 66      |               |               |       |         |         |
| U2     | 79  | 65      | 66      |               |               |       |         |         |
| T-box  | 44  | 1       | 2       |               |               |       |         |         |
| IRES_HCV | 63 | 47      | 61      |               |               |       |         |         |
| SRP_euk_arch | 62 | 64      | 52      |               |               |       |         |         |

Average(all) 70 66 54

Average(sub) 72 73 72 72 60 54 72
each sequence can be reduced to $O(L^2W^2)$ by restricting the maximum distance of the base pairs to a fixed constant $W$ \[^{40}\]. The computation of probability consistency transformation for a pair of sequences can also be calculated in $O(L^2)$ time by restricting the effective width of transformation to a fixed value. Those reductions reduce total time complexity to $O(N^3L^2)$. We will address those improvements in future work.

Comparisons of alignment tools in regard to execution time for nucleotide sequences of various lengths are presented in Figures 4 and 5. Randomly generated sequences were allocated into groups of the same lengths and were used for alignment. Stemloc aligned sequences of not more than 100 bases; FoldalignM and Locarna were faster than Stemloc and aligned sequences of 500 bases or less. Because the lengths of the sequences were the same in each alignment task, the banded Dynamic Programming (DP) technique of these methods was effective. Although the Murlet program returned results for sequences as long as 4000 bases in the best case, it was much slower than MXSCARNA. MXSCARNA required only 17 seconds to align 5 sequences of 500 bases and returns alignments for sequences as long as 5000 bases, though the accuracies for sequences longer than 500 bases have not yet been evaluated. Similar comparisons for various numbers of the sequences are presented in Figure 6. The execution time of MXSCARNA is acceptable even for 50 sequences.

### Table 5: Summary of accuracies for all three datasets

The summary of SPS and MCC values (%) for all three multiple alignment datasets are presented. Average(all): the results of the average value of the SPS or MCC for all families. Average(sub): the average SPS or MCC for the subset of families. FoldalignM consists of two modes: FoldalignM_McCaskill and FoldalignM, which are separately evaluated and indicated as FoldalignM(1) and FoldalignM(2) respectively.

| Dataset               | MXSCARNA | Murlet | ProbCons | MAFFT | ClustalW | StrAl | RNAla |
|-----------------------|----------|--------|----------|-------|----------|-------|-------|
| original dataset      | Average(all) 78/78 | 79/74 | 78/67 | 75/63 | 72/61 | 73/63 | 70/74 |
|                       | Average(sub) 85/86 | 85/85 | 85/77 | 82/74 | 78/68 | 80/72 | 78/86 |
| BRAlibase III         | Average(all) 84/80 | 85/78 | 84/75 | 83/73 | 81/72 | 83/73 | 83/78 |
|                       | Average(sub) 83/82 | 84/79 | 84/76 | 82/75 | 80/72 | 82/75 | 83/80 |
| Kiryu et al.’s dataset | Average(all) 77/72 | 80/71 | 79/65 | 74/54 | 68/48 | 69/54 | 68/70 |
|                       | Average(sub) 82/72 | 86/72 | 85/68 | 80/60 | 73/52 | 75/59 | 72/72 |

### Table 6: Accuracy of new score function

The comparison of new score function of MXSCARNA and the old one which was used in SCARNA in terms of pairwise alignment. The SPS and MCC values (%) are used as accuracy measure for alignments. R100 is a dataset which consists of 100 pairwise alignments chosen randomly from tRNA and 5SrRNA families in Rfam 7.0 database \[^{34}\] and percid is a sequence identity balanced dataset which also consists of 100 pairwise alignments from these families.

| dataset | score function | SPS | MCC |
|---------|----------------|-----|-----|
| R100    | MXSCARNA       | 90  | 77  |
| SCARNA  | 84             | 74  |
| percid  | MXSCARNA       | 79  | 71  |
| SCARNA  | 78             | 69  |
Comparison of various multiple alignment tools in execution time and the number of sequences

Figure 4
Comparison of multiple alignment tools in execution time for various lengths of the sequences (maximal sequence length, 500 bases). The relationships between the length of the sequences (maximum, 500 bases) and the execution time for several multiple alignment tools are plotted. A set of randomly generated sequences of the same length is used for each alignment.

Figure 5
Comparison of multiple alignment tools in execution time for various lengths of the sequences (maximal sequence length, 5000 bases). The relationships between the length of the sequences (maximum, 5000 bases) and the execution time for MXSCARNA, Murlet and Stemloc are plotted. A set of randomly generated sequences of a same length is used for each alignment. The number of the sequences used for the alignment is indicated after the names of the tools. The accuracies for the sequences longer than 500 bases have not been evaluated.

Figure 6
Comparison of various multiple alignment tools in execution time and the number of sequences. The relationships between the number of the sequences and the execution time for MXSCARNA, Murlet and Stemloc are plotted. A set of randomly generated sequences of a same length is used for each alignment. The lengths of the sequences used for the alignment is indicated after the names of the tools.

Figure 7
Relationship between sequence similarities and SPS. The relationship between sequence similarities and accuracies according to sum-of-pairs score (SPS) is shown. Lines are smoothed by local weighted regression.
aligners (as seen in the previous section), achieved robust accuracies in MCC but failed to compete among the other aligners in regard to SPS for moderate sequence identities. MXSCARNA, which incorporates information on Maximum Expected Accuracy (MEA) alignment in its structural alignments, yielded robust accuracies in terms of both SPS and MCC throughout the tested range of sequence similarities.

**Availability and requirements**

Project name: ncRNA.org project;

Project home page: [http://www.ncrna.org/](http://www.ncrna.org/);

Operating systems: Linux with gcc 3.0 and Cygwin with gcc 3.4;

Programming language: C++;

License: free software, except for inclusion to comertical software;

The source code of MXSCARNA and its web server, the dataset and its references are available at [http://mxxscarna.ncrna.org](http://mxxscarna.ncrna.org). On the web server W and r correspond to "SCSLENGTH" and "BASEPROBTHRESHHOLD" respectively, and "BASEPAIRSCORECONST" is a parameter of McCaskill-MEA [6] used for the secondary structure prediction, which controls the sensitivity and the specificity of the prediction ($\alpha$ in equation 4 in [6]).

**Conclusion**

We have developed MXSCARNA, a new structural multiple aligner of RNA sequences, which progressively applies the pairwise alignment algorithm used in SCARNA. The accuracies of MXSCARNA in terms of SPS and MCC were evaluated for three datasets: an original dataset, the BRAlibasell benchmark multiple alignment dataset, and Kiryu et al.'s multiple alignment dataset. MXSCARNA's accuracies were at least comparable to those of current state-of-art aligners. In addition, the accuracies of MXSCARNA were robust over a broad range of sequence similarities, whereas the other aligners tested showed reductions in SPS or MCC. The computational complexities of MXSCARNA were evaluated as $O(N^3L^3)$ in time and $O(N^2L^2)$ in memory for N sequences of length L. In the comparison of execution time for benchmark datasets, MXSCARNA was by far the fastest among the structural aligners and was fast enough for large-scale analyses. MXSCARNA aligns even 5000-base RNA sequences with acceptable computational costs though the accuracies of alignments for long sequences are not yet confirmed. The source code of MXSCARNA and its web server are available at the web site [41].

**Authors' contributions**

YT and KA developed the algorithm, and together they wrote the manuscript with the help of TK. YT implemented the algorithm into the software (MXSCARNA) and executed all of the computational experiments. HK contributed to the design of the new score function and closely collaborated in the computational experiments. TK helped to design the computational experiments and to write the manuscript. With the help of YT, KA organized the development of the web server. All the authors have read and approved the final version of the manuscript.

**Acknowledgements**

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas “Comparative Genomics” from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the “Functional RNA Project” funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan. The authors thank the Japan Biological Informatics Consortium (JBIC) for its support through the "Functional RNA Project" and Michiaki Hamada, Kengo Sato, and colleagues in the Computational Biology Research Center (CBRC) for useful discussions.

**References**

1. Mathews DH, Sabina J, Zuker M, Turner DH: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999, 288(5):911-940.
2. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ: Algorithms for loop matchings. SIAM J Appl Math 1978, 35:68-82.
3. Zuker M, Stiegler P: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl Acids Research 1981, 9:133-148.
4. Hofacker I, Fekete M, Stadler P: Secondary structure prediction for aligned RNA sequences. J Mol Biol 2002, 319:1059-1066.
5. Knudsen B, Hein J: FOLD: RNA secondary structure prediction using stochastic context-free grammars. Nucl Acids Res 2003, 31:3423-3428.
6. Kiry H, Kim T, Asai K: Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics 2006, 22:434-441.
7. Thompson J: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1999, 27:2185-2195.
8. Morgenstern B: DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 1998, 14:290-294.
9. Notredame C, Higgins DG, Heringa J: T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. Journal of Molecular Biology 2000, 302:205-217.
10. Bankoff D: Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. SIAM J Appl Math 1985, 45:810-825.
11. Hugavard JH, Lyngjs RB, Stormo GD, Gorodkin J: Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 2005, 21(9):1815-1824.
12. Hofacker I, Bernhart S, Stadler P: Alignment of RNA base pairing probability matrices. Bioinformatics 2004, 20:2222-2227.
13. Holmes I: A probabilistic model for the evolution of RNA structure. BMC Bioinformatics 2004, 5(166).
14. Holmes I, Rubin GM: Pairwise RNA structure comparison with stochastic context-free grammars. Pacific Symposium on Biocomputing 2002:163-174.
15. Torarinsson E, Hugavard JH, Gorodkin J: Multiple structural alignment and clustering of RNA sequences. Bioinformatics 2007, 23:926-932.
16. Will S, Reiche K, Hofacker I, Stadler P, Backofen R: Inferring Non-coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. PLoS Computational Biology 2007, 3(4)e45.
17. Siebert S, Backofen R: MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics 2005, 21:3352-3359.
18. Dalli D, Wilm A, Mains I, Steger G: STRAL: Progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. Bioinformatics 2006, 22(13):1593-1599.
19. Xu X, Ji Y, Stormo GD: RNASampler: a new sampling based algorithm for common RNA secondary structure prediction and structure alignment. Bioinformatics 2007, 23:1883-1891(15).
20. Bauer M, Klau GW, Reinert K: Accurate multiple sequence-strucure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics 2007, 8.
21. Kiry H, Tabei Y, Taishin K, Asai K, Murlet: a practical multiple alignment tool for structural RNA sequences. Bioinformatics 2007, 23:1588-1598.
22. Seibel PN, Muller T, Dandelar T, Schulz J, Wolf M: 4SALE - A tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics 2006, 7:498.
23. Yao Z, Weinberg Z, Ruzzo W: CMfinder - a covariance model based RNA motif finding algorithm. Bioinformatics 2006, 22:445-452.
24. Hamada M, Tsuda K, Kudo T, Kim T, Asai K: Mining frequent stem patterns from unaligned RNA sequences. Bioinformatics 2006, 22:2480-2487.
25. Tabei Y, Tsuda K, Taishin K, Asai K: SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. Bioinformatics 2006, 22:1723-1729.
26. Sokal RR, Michener CD: A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 1958, 28:1409-1438.
27. McCaskill J: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990, 29:1105-1119.
28. Klein R, Eddy S: RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 2005, 6(44).
29. Miyazawa S: A reliable sequence alignment method based on probabilities of residue correspondences. Protein Engineering 1995, 8:999-1009.
30. Holmes I, Durbin R: Dynamic programming alignment accuracy. J Comput Biol 1998, 5:493-504.
31. Eddy R, Durbin AKSR, Mitchison G: Biological Sequence Analysis Cambridge: UK: Cambridge University Press; 1998.
32. Do CB, Mahabhashayam MS, Burdno M, Batzoglou S: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 2005, 15(2):330-340.
33. Gardner P, Wilm A, Washed S: A benchmark of multiple sequence alignment programs upon structural RNAs. Nucl Acids Res 2005, 33(8):2433-2439.
34. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy S: Rfam: an RNA family database. Nucl Acids Res 2003, 31:449-451.
35. Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C, Samuelsson T: SRPDB: Signal Recognition Particle Database. Nucl Acids Res 2003, 31:363-364.
36. Matthews B: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta 1975, 405:442-451.
37. Gardner PP, Giegerich R: A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 2004, 5(140).
38. Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid and accurate multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002, 30:3059-3066.
39. Dowell RD, Eddy SR: Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. BMC Bioinformatics 2007, 8.
40. Kiry H, Kim T, Asai K: Rfold: An exact algorithm for computing local base pairing probabilities. Bioinformatics Advance Access 2007 in press.
41. MXSCARNA [http://mxscarna.ncrna.org/]

Publish with BioMed Central and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp