

Structural bioinformatics

STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time

Deniz Dalli, Andreas Wilm, Indra Mainz and Gerhard Steger*

Heinrich-Heine-Universität Düsseldorf, Institut für Physikalische Biologie, D-40225 Düsseldorf, Germany

Received on February 10, 2006; revised and accepted on April 10, 2006

Advance Access publication April 13, 2006

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Alignment of RNA has a wide range of applications, for example in phylogeny inference, consensus structure prediction and homology searches. Yet aligning structural or non-coding RNAs (ncRNAs) correctly is notoriously difficult as these RNA sequences may evolve by compensatory mutations, which maintain base pairing but destroy sequence homology. Ideally, alignment programs would take RNA structure into account. The Sankoff algorithm for the simultaneous solution of RNA structure prediction and RNA sequence alignment was proposed 20 years ago but suffers from its exponential complexity. A number of programs implement lightweight versions of the Sankoff algorithm by restricting its application to a limited type of structure and/or only pairwise alignment. Thus, despite recent advances, the proper alignment of multiple structural RNA sequences remains a problem.

Results: Here we present STRAL, a heuristic method for alignment of ncRNA that reduces sequence–structure alignment to a two-dimensional problem similar to standard multiple sequence alignment. The scoring function takes into account sequence similarity as well as up- and downstream pairing probability. To test the robustness of the algorithm and the performance of the program, we scored alignments produced by STRAL against a large set of published reference alignments. The quality of alignments predicted by STRAL is far better than that obtained by standard sequence alignment programs, especially when sequence homologies drop below ~65%; nevertheless STRAL's runtime is comparable to that of CLUSTALW.

Availability: STRAL is implemented in C. Source code (under GNU public license) as well as a precompiled Debian package can be downloaded at <http://www.biophys.uni-duesseldorf.de/stral/>

Contact: stral@biophys.uni-duesseldorf.de

Supplementary information: Supplementary data available at *Bioinformatics* online.

1 INTRODUCTION

Non-coding RNAs (ncRNAs) are RNA molecules or elements that do not code for proteins but nevertheless are functional in biological processes, including localization, replication, translation, degradation and stabilization of biological macromolecules (for review and further references see Eddy, 2001; Storz, 2002; Winkler and Breaker, 2005; Fedor and Williamson, 2005; Gottesmann, 2005). Prominent examples are small nuclear RNAs, which are involved in

mRNA splicing, and riboswitches, which are located in non-translated regions of mRNAs, where they bind metabolites and control gene expression.

For analysis of ncRNA function, knowledge about their secondary and tertiary structure is crucial. Structure prediction for single sequences is performed by dynamic programming, which allows the thermodynamically optimal structure or structure ensemble to be found (Zuker, 2000, 2003; Hofacker, 2003; Rivas and Eddy, 1999). These algorithms rely on correctness of thermodynamic parameters, neglect the influence of kinetics on structure formation and are not able to take into account interactions with other macromolecules. The alternative method—called comparative sequence analysis—needs a set of homologous RNAs and predicts base–base interactions on the basis of compensatory mutations (Chiu and Kolodziejczak, 1991; Gautheret *et al.*, 1995; Lescoute *et al.*, 2005). Its reliability increases with the number and divergence of sequences, but it needs an alignment (nearly) perfect with respect to sequence and structure. Other approaches for consensus structure prediction based on RNA alignments include PFOLD (Knudsen and Hein, 2003) and RNAALFOLD (Hofacker *et al.*, 2002). Furthermore, RNA alignments are an essential basis for phylogeny inference (e.g. Hudelot *et al.*, 2003), homology searches (e.g. Gräf *et al.*, 2006; Eddy, 2002) and approaches to searching for new ncRNAs (e.g. Washietl *et al.*, 2005).

The structurally correct alignment of RNA sequences is, however, a difficult problem. An algorithm for simultaneously optimizing the sequence and structure of an RNA set was published by Sankoff in 1985; however, the algorithm is not employable due to its computational complexity $\mathcal{O}(n^{3m})$ and memory usage $\mathcal{O}(n^{2m})$ for m sequences of length n . Thus, several variants of this algorithm have been developed which are restricted to pairwise alignment only and implement other simplifications to make the calculation tractable (Mathews and Turner, 2002; Hofacker *et al.*, 2004; Havgaard *et al.*, 2005a; Holmes, 2005).

This situation resembles that of (pure) sequence alignment: the algorithm for aligning two sequences is relatively cheap (Smith and Waterman, 1981), whereas the same approach cannot be applied to multiple sequence alignment due to its complexity of $\mathcal{O}(n^m)$ (Fuellen, G., 1997). This led to the development of very successful heuristic alignment methods, for example CLUSTAL (Thompson *et al.*, 1994a). Here we follow the same heuristic multiple sequence alignment approach but enhance it using a scoring function that emphasizes structural features. For this purpose, we project the structure features calculated by a thermodynamic approach (RNAFOLD; Hofacker, 2003) on top of the sequence. Similar

*To whom correspondence should be addressed.

approaches have already been proposed in the literature (Bonhoeffer *et al.*, 1993; Yang and Blanchette, 2004) but have not been implemented. We call our program STRAL.

To test the performance of STRAL, we compare it with two sequence alignment programs—CLUSTAL (Thompson *et al.*, 1994a) and PROALIGN (Löytynoja and Milinkovitch, 2003)—and three different structure alignment programs:

PMSTRING (PMMULTI in string-like alignment mode; see Hofacker *et al.*, 2004) follows a concept related to that of our approach: it uses strings of pairing probabilities obtained from Sankoff-like pairwise alignments to produce a guide tree for a multiple alignment. This procedure, however, ‘often produces misaligned [sequence] pairs’ (Hofacker *et al.*, 2004).

MARNA (Siebert and Backofen, 2005) uses a distantly related approach: it performs pairwise alignments of RNAs using sequence and structure information (based on RNAFOLD predictions) and combines the pairwise, weighted information into a multiple alignment with T-COFFEE (Notredame *et al.*, 2000). **STEMLOC** (Holmes, 2005) implements the pairwise Sankoff algorithm using a pair stochastic context-free grammar and a heuristic for multiple alignment. To reduce computing cost and memory usage, constraints are imposed by ‘fold and alignment envelopes’, which take into account, for example, the 1000 best single sequence structure predictions and the 100 best alignments.

In the following, we present our string-like alignment algorithm and test the performance of the respective program STRAL on several hundred sets of homologous RNAs. The quality of alignments produced by the Sankoff-like algorithm implemented in STEMLOC is clearly superior to that produced by STRAL; this significant difference is, however, accompanied by a huge factor in computing time and memory usage in favor of STRAL. The quality of alignments predicted by STRAL is similar to that of standard sequence alignment programs if sequence similarity is >65%, but is far better with lower sequence similarities; nevertheless STRAL’s runtime is close to that of CLUSTALW.

2 SYSTEMS AND METHODS

STRAL is implemented in C and should compile under any Unix system. We used the GNU C compiler (GCC) versions 3 and 4. The program was thoroughly tested on several Linux distributions, including Debian Version 3.1 and Red Hat Fedora Core 3. To facilitate installation, support for GNU autotools is built in. We also provide a precompiled Debian package. STRAL requires RNAfold’s RNALib (Hofacker *et al.*, 1994; Hofacker, 2003) version 1.5 and the squid library (Eddy, 2005) version 1.9g. Static versions of these libraries compiled for i386 are included in the package, which can be downloaded at <http://www.biophys.uni-duesseldorf.de/stral/>. All alignments have been computed using STRAL version 0.5.2. Runtime comparison was performed on a 1.8 GHz Dual-Opteron machine with 4 GB memory; STEMLOC computations had to be performed on 2.4 GHz Opterons with 16 GB memory. Programs have been compiled with optimization level 3 (GCC 4).

2.1 Reference alignments

As reference alignments we used dataset-1 from the RNA alignment benchmark database BRALibase (Gardner *et al.*, 2005). This dataset consists of 388 alignments of Group I introns, 5S rRNAs, tRNAs, and U5 spliceosomal RNAs with five sequences per alignment.

2.2 Scoring of alignments

As proposed by Gardner *et al.* (2005), we used two independent yet complementary scores to evaluate alignment quality: the widely used sum-of-pairs score (SPS) implemented in BaliScore (Thompson *et al.*, 1999) and the Structure Conservation Index (SCI; see Washietl *et al.*, 2005). The SPS measures the level of sequence consistency between a test and a reference alignment by comparing all possible character pairs per column between both alignments; it ranges from 0 to 1 (complete agreement). The SCI is a measure for structural conservation and works independently of a reference alignment. It ranges from 0 (no detectable conservation) to values slightly above 1.0 (sequence agreement and structure conservation). For statistical analysis of results we used R (<http://cran.r-project.org/>).

2.3 Parameters of alignment programs

The parameter choices for STRAL are described in Section 4.1. PROALIGN v0.5a3 was allowed to use up to 256 MB of memory and a bandwidth of 400 nt (`-Xmx256m -bandwidth=400`). PMMULTI v1.1 was used either in its slow and thorough variant or in string-like alignment mode (`--fast_progressive`). Other programs (CLUSTALW v1.83, MARNA, STEMLOC v0.19b) were used with default options.

3 ALGORITHM

The steps of the algorithm include

- (1) a pairwise alignment of all sequences of a set of homologous ncRNAs,
- (2) production of a guide tree using the alignment scores from step 1, and
- (3) a progressive alignment of the sequences, guided by the tree.

This strategy follows that of CLUSTAL (Thompson *et al.*, 1994a).

3.1 Pairwise alignment

We use the RNAFOLD library (Hofacker *et al.*, 1994; Hofacker, 2003) to compute the partition function and matrix of base-pairing probabilities P_{ij} of base i with base j for a single sequence. This probability matrix is then condensed into three linear vectors (Bonhoeffer *et al.*, 1993), holding for each base i the probabilities of being paired downstream $p_i^1 = \sum_{j>i} P_{ij}$, paired upstream $p_i^2 = \sum_{j<i} P_{ij}$ or unpaired $p_i^0 = 1 - (p_i^1 + p_i^2)$, respectively. Thus we lose the specific pairing information but we can apply an alignment method that is cheap in terms of computing costs while still using thermodynamic information. Next, we choose a particular combination of these vectors—a structural part $S_{\text{struct}} = f(p^1, p^2)$ and a sequence part $S_{\text{seq}} = f(p^0)$ —as a similarity score

$$\begin{aligned} s_{i,k} &= \alpha(S_{\text{struct}}) + S_{\text{seq}} \\ &= \alpha\left(\sqrt{p_{A_i}^1 p_{B_k}^1} + \sqrt{p_{A_i}^2 p_{B_k}^2}\right) + \sqrt{p_{A_i}^0 p_{B_k}^0} \cdot d(A_i, B_k) \end{aligned} \quad (1)$$

for matching bases i and k from different sequences A and B , respectively. The idea for this score is to favor structural alignment of paired nucleotides from both sequences as well as sequence alignment of unpaired nucleotides. The factor α gives the ratio of structure over sequence similarity. The positive 4×4 single nucleotide substitution matrix d for aligning single-stranded regions is adapted either from Klein and Eddy (2003) (RIBOSUM85-60) or from Gotoh (1999) (see Section 4.1.2).

Let $V_{i,k}$ be the value of the optimal alignment of prefixes $A[1..i]$ and $B[1..k]$ with base conditions $V_{0,0} = 0$, $V_{i,0} = E_{i,0} = -g_o - i \cdot g_e$, $V_{0,k} = F_{0,k} = -g_o - k \cdot g_e$ and an affine gap weight model. That is, a single gap of length q is given by weight $g_o + q \cdot g_e$; g_o and g_e denote gap-open and gap-extension values, respectively. If we do not charge end gaps, which is the default setting, set $V_{i,0} = V_{0,k} = 0$. Then follow the dynamic programming recurrences (Gusfield, 1999; Gotoh, 1999) for aligning two sequences:

$$E_{i,k} = \max \{E_{i,k-1}, V_{i,k-1} - g_o\} - g_e \quad (2a)$$

$$F_{i,k} = \max \{F_{i-1,k}, V_{i-1,k} - g_o\} - g_e \quad (2b)$$

$$G_{i,k} = V_{i-1,k-1} + s_{i,k} \quad (2c)$$

$$V_{i,k} = \max \{E_{i,k}, F_{i,k}, G_{i,k}\} \quad (2d)$$

$G(i, k)$ is the maximum value of an alignment of $A[1..i]$ and $B[1..k]$, where $A(i)$ and $B(k)$ are aligned opposite each other. $E(i, k)$ aligns $A(i)$ to the left of $B(k)$; thus, the alignment ends in a gap in A . $F(i, k)$ aligns $A(i)$ to the right of $B(k)$; thus, the alignment ends in a gap in B . $V(i, k)$ is defined as the maximum value of the three terms $E(i, k)$, $F(i, k)$ and $G(i, k)$. This pairwise alignment is carried out for all $m(m-1)/2$ pairs of the m sequences $S = S(1), \dots, S(m)$ of the ncRNA set.

3.2 Guide tree

The $m(m-1)/2$ scores from the previous step are converted into a distance matrix

$$D_{S_p, S_q} = -\log \left(\frac{V_{S_p, S_q} - V_{\min} + 1}{V_{\max} - V_{\min}} \right)$$

with $V_{\min} = \min_{1 \leq p < q \leq m} (V_{S_p, S_q})$ and

$$V_{\max} = \max_{1 \leq p < q \leq m} (V_{S_p, S_q}).$$

This is used as the starting point for the construction of a guide tree. We implemented four different methods: UPGMA (Sokal and Michener, 1958), neighbor joining (NJ; Saitou and Nei, 1987; Felsenstein, 1989, 1997), weighted neighbor joining (Weighbor; Bruno *et al.*, 2000) and BIONJ (Gascuel, 1997); the default is UPGMA. In the case of an unrooted tree, it is possible to root the tree using the midpoint method (Thompson *et al.*, 1994b).

3.3 Progressive alignment

During the progressive alignment, the sequences of the set $S = S(1), \dots, S(m)$ are aligned according to their position in the tree, starting with the two sequences with lowest distance, resulting in a group of aligned sequences. To this group, a further sequence or a group of sequences is added, which makes necessary modifications of equations (1) and (2), and the base conditions. We measure the SPS, so $V_{i,0} = E_{i,0} = n \cdot (-g_o - i \cdot g_e)$ and $V_{0,k} = F_{0,k} = n \cdot (-g_o - k \cdot g_e)$, with n being the number of all pairwise alignments of groups A and B . In the case of free end gaps, $V_{i,0} = V_{0,k} = 0$.

For (2c) the scoring function (1) has to be modified. Let C be the desired alignment of groups A and B ($C = A \cup B$); then

$$s_{i,k} = \sum_{S(t) \in C} \sum_{\substack{S(u) \in C \\ 1 \leq t < u \leq |C|}} \left[\alpha \left(\sqrt{p_{S(t),i}^1 p_{S(u),k}^1} + \sqrt{p_{S(t),i}^2 p_{S(u),k}^2} \right) + \sqrt{p_{S(t),i}^0 p_{S(u),k}^0} \cdot d(S(t)_i, S(u)_k) \right].$$

To speed up calculation, the scores of individual groups are stored in a double linked list. For (2a) and (2b) the gap values have to be modified according to the size of groups A and B :

$$g_{\text{open}} = \left(|A| - \sum_{i=1}^{|A|} \delta_i \right) \cdot g_o \quad \text{with } \delta_i = \begin{cases} 1 & \text{if } A_i \text{ is gap,} \\ 0 & \text{else} \end{cases}$$

$$g_{\text{ext}} = |A| \cdot g_e.$$

These equations accordingly hold for sequences in B . We have not tried to reduce computing costs by introducing true profiles (Griboskov *et al.*, 1990); with our scoring function, this would force a grouping of individual scores ($s_{i,k} \in \mathbb{R}^{0,+}$) and gap sizes, respectively, into small numbers of classes (Gotoh, 1993).

4 RESULTS

4.1 Choice of parameters

For parameter optimization we used the RNA alignment benchmark datasets published by Gardner *et al.* (2005) (see also Section 2). The determination of ‘correct’ parameters was quite difficult; for example, the ratio α of structure over sequence similarity and gap values g_o and g_e depend on each other. So we examined a rather huge parameter space. Contrary to our expectations, however, modification of the final parameters by a factor of 2 leads to only marginally different alignments.

4.1.1 Scoring function The scoring function (1) implies a large influence of sequence on the alignment only in predicted loop regions and not in predicted helical regions. A major improvement in alignment quality arose by skipping this restriction; that is, formally we set the probability of a base to be unpaired $p_i^0 = 1$. Consequently, sequence information is used for both structured and unstructured regions.

4.1.2 Substitution matrix d The 4×4 single nucleotide substitution matrix given by Gotoh (1999) emphasizes identity substitutions [$d(X, X) = 4$] and allows for pyrimidine to pyrimidine and purine to purine substitutions only [$d(Y, Y) = d(R, R) = 1$]. The RIBOSUM85-60 matrix from Klein and Eddy (2003) overall contains higher values [$3.51 \leq d(X, X) \leq 4.70$, $d(Y, Y) = 1.43$, $d(R, R) = 1.02$]; only $d(G, C) = 0$. The former performed better for RNA sets of high similarity; the latter yielded a slightly better performance over the full similarity range.

4.1.3 Structure over sequence ratio α Overall, scoring values were relatively constant in an α range from 3 to 11. A factor $\alpha = 7$ produced the best results (see Table 1) when end gaps were free of costs. Nevertheless, scoring values improved when a higher α was applied to sequence sets containing fewer than five sequences and/or a sequence similarity lower than $\sim 50\%$. In contrast, α values lower than 2 did not lead to an improvement on high-similarity sets.

4.1.4 Gap values To determine appropriate values for gap costs, we aligned the complete dataset-1 several times with different gap opening, gap extension and α values. The alignments were then scored by the product of SPS and SCI. These values were averaged over all 388 alignments for each parameter combination.

Table 1. SPS and SCI dependence upon structure-over-sequence ratio α

α	SPS	SCI	SPS-SCI	Sign.
0	0.838	0.741	0.621	*
1	0.849	0.773	0.656	—
2	0.852	0.794	0.677	—
3	0.855	0.800	0.684	—
4	0.854	0.809	0.691	—
5	0.855	0.818	0.700	—
6	0.855	0.820	0.701	—
7	0.854	0.821	0.701	—
8	0.851	0.817	0.696	—
9	0.847	0.814	0.690	—
10	0.844	0.813	0.686	—
11	0.839	0.809	0.678	—
12	0.836	0.803	0.671	—
13	0.834	0.800	0.667	—
14	0.834	0.792	0.657	—
15	0.824	0.786	0.648	—

Values for gap-opening g_o and gap extension g_e were fixed at 8.0 and 0.5, respectively, with free end gaps. The significance of SPS-SCI obtained using the different α values was determined by Friedman tests against $\alpha = 7$; only $\alpha = 0$ was significantly different (worse) than higher values.

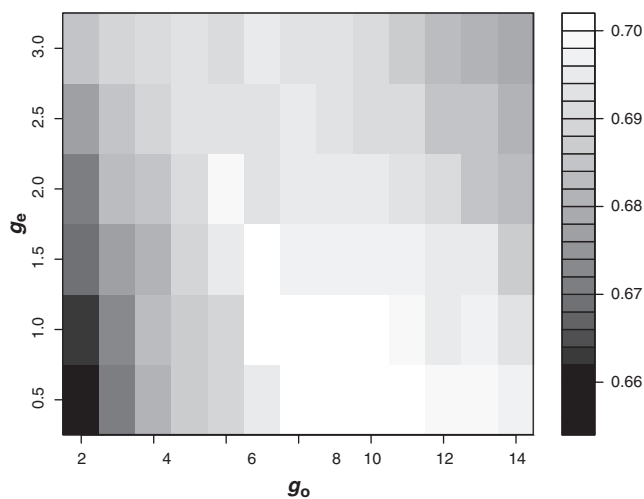


Fig. 1. Alignment quality measured by SPS-SCI averaged over all 388 alignments is shown for several gap parameter combinations. The bar at the right gives averaged SPS-SCI values from 0.654 to 0.702 in grey scale. The structure-over-sequence ratio α is fixed to a value of 7 in this example. Note that absolute values change only slightly; that is alignment quality is rather robust to the choice of gap parameters.

An example with $\alpha = 7$ is shown in Figure 1. Several parameter combinations are near optimal. We implemented a gap opening value of $g_o = 8$ and a gap extension value of $g_e = 0.5$ as default, as this combination produced high-scoring alignments for other values of α , too (data not shown). Overall alignment quality, however, does not vary significantly upon 2-fold changes of either gap opening or gap extension values.

4.1.5 Guide tree Alignments based on guide trees derived from UPGMA showed the highest accuracy. Trees produced by the other methods showed similar alignment accuracy, except for the trees derived from the midpoint method (data not shown).

4.2 Benchmark

To demonstrate the power of STRAL we compared its performance with that of CLUSTAL (Thompson *et al.*, (1994a), PROALIGN (Löytynoja and Milinkovitch, 2003) (the best-performing programs according to the evaluation by Gardner *et al.*, 2005), PMMULTI in string-like alignment mode (PMSTRING; see Hofacker *et al.*, 2004, MARNA (Siebert and Backofen, 2005) and STEMLOC (Holmes, 2005). For the comparison we used the multiple RNA sequence set (dataset-1) from the above-mentioned benchmark. Aligning all 388 RNA sets (with five sequences each) from this data set took ~ 106 s using STRAL. This time reduced to only ~ 9 s when probability matrices were precomputed. CLUSTALW, PROALIGN, PMSTRING, MARNA, and STEMLOC needed ~ 8 s, ~ 15 min, ~ 1 h, ~ 5 h and nearly 2 days, respectively. Note that PMSTRING is only prototyped in Perl; for STEMLOC calculations a (faster) machine with more memory had to be used (see Section 2). MARNA was not able to align those 46 sequence sets containing ambiguity code, whereas STEMLOC failed to align 5 sequence sets for unknown reasons.

Program performance was measured using SPS and SCI and plotted as a function of the sequence similarity (see Fig. 2). The structure and sequence alignment program STEMLOC performed best, but at the cost of long time and high memory usage. With the exception of STEMLOC, STRAL clearly outperformed all other programs: alignments produced by STRAL not only showed higher structural conservation (Fig. 2B) than the structure alignment programs MARNA and PMSTRING but also achieved more conservation on the sequence level (see Fig. 2A) than the pure sequence alignment programs CLUSTAL and PROALIGN. The performance difference became drastic when sequence similarity dropped below 60%. Here, the performance of STRAL is clearly better as the alignment process is guided by structural features, too, whereas pure sequence alignment programs cannot handle these sets, which are only poorly conserved on the sequence level.

A rather *ad hoc* approach to measuring RNA alignment quality is to compare consensus structures predicted from calculated alignments. Figure 3 shows such consensus structures predicted by RNAALIFOLD (Hofacker *et al.*, 2002) on the basis of different alignments of aphto- and cardiovirus IRES regions (see also Hofacker *et al.*, 2004). CLUSTAL clearly failed to create a correct alignment which comprised enough conserved structure. The consensus structures predicted from alignments computed by STRAL and PMMULTI (slow variant) are almost identical. A more detailed comparison of the predicted structures is given in Supplementary Table S1.

To give a visual impression of the differences in alignment quality obtained using CLUSTALW and STRAL, alignments of 14 selenocysteine insertion sequences (SECIS; taken from Kryukov and Gladyshev, 2004) from methanogenic organisms are shown in Figure 4. In the CLUSTAL alignment (Fig. 4C) the thermodynamically predicted stem-loop structures are not superimposed (top right triangle) and no statistically significant base pairs are predicted (lower left triangle); the ‘sequence alignment’ has a

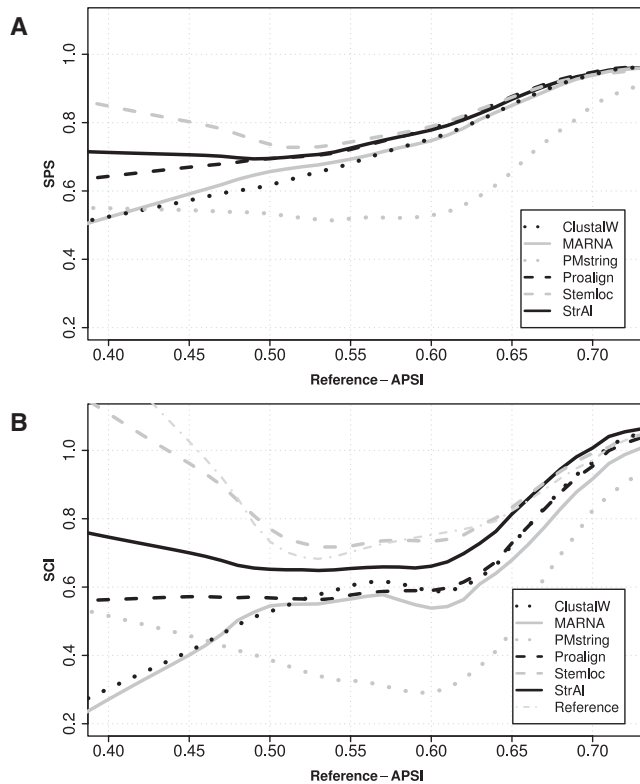


Fig. 2. Performance of alignment programs on the BRAliBase benchmark: all 388 RNA sets from dataset-1 (Gardner *et al.*, 2005) were aligned and each alignment scored. The performance is plotted against sequence similarity measured as mean pairwise sequence identity. Curves are fitted using Lowess smoothing. **A:** Performance measured as SPS. **B:** Performance measured as SCI. See Section 2.2 for an explanation of these scores; for a color version of this figure, see Supplementary Figure S1.

length of 44 nt. In contrast, the alignment computed by means of STRAL (Fig. 4B) is close to the ‘correct’ one (Fig. 4A), which was manually refined by means of CONSTRUCT (Lück *et al.*, 1999): the thermodynamically predicted stem-loop structures are perfectly superimposed (except for a single sequence), the highly conserved internal loop nucleotides are aligned and alignment length is only 41 nucleotides.

5 DISCUSSION

We have implemented a multiple RNA alignment program named STRAL that combines structural and sequence information in a ‘cheap’ dynamic programming approach. That is, when pairing vectors are precomputed, STRAL is nearly as fast as CLUSTALW. STRAL requires computational resources similar to other sequence alignment programs with $\mathcal{O}(k^2n^2)$ time and $\mathcal{O}(n^2)$ memory cost, whereas true structure alignment programs such as DYNALIGN (Mathews, 2005), FOLDALIGN v. 2 (Havgaard *et al.*, 2005a,b); PMCOMP (Hofacker *et al.*, 2004) and STEMLOC (Holmes, 2005), have costs of at least $\mathcal{O}(n^4)$ for pairwise alignment. Nevertheless,

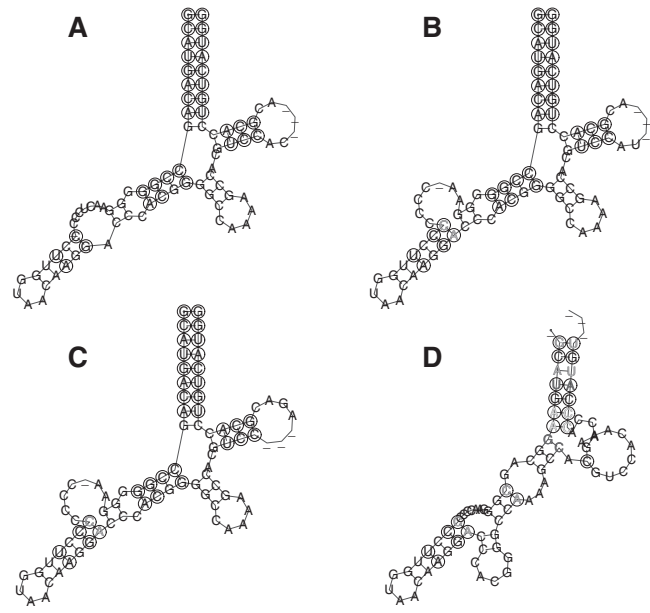


Fig. 3. IRES consensus structures predicted by RNAALFOLD on the basis of a manually constructed reference alignment (A; see Hofacker *et al.*, 2004, and references therein) and of alignments predicted by STRAL (with NJ tree) (B), PMMULTI (slow, thorough variant) (C) and CLUSTALW (D), respectively. Consistent and compensatory mutations are indicated by circles. Grey letters indicate inconsistent mutations. The CLUSTALW prediction shows several inconsistent mutations and the overall structure is different from the reference. Predictions made by PMMULTI and STRAL are almost identical and share only one inconsistent mutation. For further details, see Supplementary Table S1.

it has been shown that these structural alignment programs do not necessarily produce high-quality alignments (Gardner *et al.*, 2005).

The parameters used in the algorithm of STRAL have been optimized using the benchmark data set BRAliBase (Gardner *et al.*, 2005). Clearly, the inclusion of sequence and structure into the scoring function improves predictions in comparison with both pure sequence alignment and pure structure alignment (as, for example, in PMSTRING). With respect to all parameters, STRAL’s performance is quite robust to modifications by a factor of ~ 2 from the default parameters. It is likely, however, that the performance of other programs will also be improved by such a parameter optimization (e.g. cf. Katoh *et al.*, 2005).

The use of a condensed vector representation of the pairing probabilities instead of the full pairing matrix—as done in PMMULTI in pairwise mode—allows for a very fast pairwise alignment computation during every alignment step. Yet a future improvement could be the use of such ‘perfect’ pairwise alignments in combination with STRAL for the multiple alignment step(s). A further improvement of STRAL will be the inclusion of a recursive step in addition to the purely progressive approach (for review, see Gotoh, 1999).

Our approach offers a fast and reliable compromise between the computationally very demanding true structural alignment and pure sequence alignment.

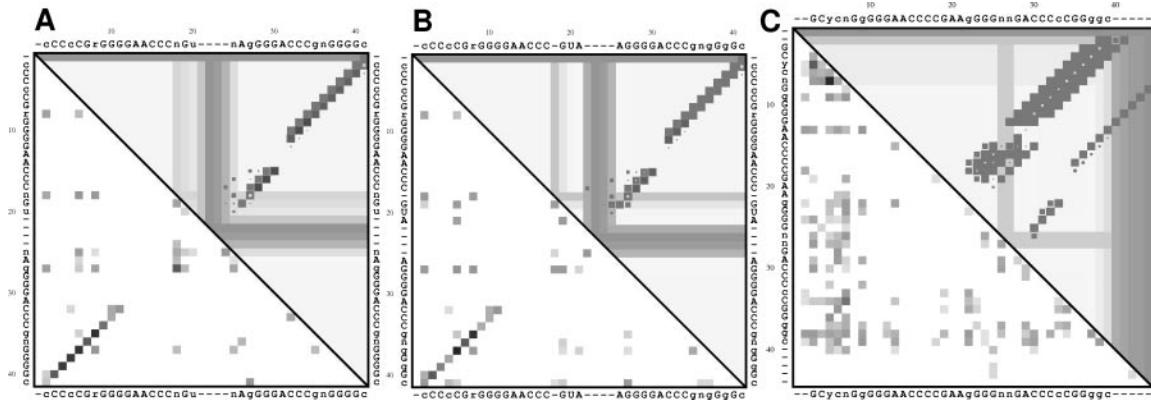


Fig. 4. Visualization as base-pairing dotplots, drawn by CONSTRUCT (Lück *et al.*, 1999), of alignments produced manually (A) or by programs STRAL (B) and CLUSTAL (C), respectively. Each top-right triangle shows the thermodynamic base-pairing probability of individual sequences; the horizontal and vertical bars denote gaps; the lower-left triangle shows mutual information content normalized by pair entropy (Martin *et al.*, 2005). For further details see Supplementary Figure S2.

ACKNOWLEDGEMENTS

Acknowledgement is made to Dr M. Schmitz for critical reading of the manuscript. We thank Drs I. Hofacker and P. Stadler for providing the IRES reference alignment. A.W. was supported by a grant from the German National Academic Foundation.

Conflict of Interest: none declared.

REFERENCES

- Bonhoeffer, S. *et al.* (1993) RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.*, **22**, 13–24.
- Bruno, W. *et al.* (2000) Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, **17**, 189–197.
- Chiu, D. and Kolodziejczak, T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**, 347–352.
- Eddy, S. (2001) Noncoding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
- Eddy, S. (2005) SQUID—C function library for sequence analysis. <http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#squid>.
- Eddy, S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
- Fedor, M. and Williamson, J. (2005) The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell. Biol.*, **6**, 399–412.
- Felsenstein, J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
- Felsenstein, J. (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst. Biol.*, **46**, 101–111.
- Fuellen, G. (1997) A gentle guide to multiple alignment. *Complexity International*, **4**, <http://www.csu.edu.au/ci/vol04/mulali/mulali.html>.
- Gardner, P.P. *et al.* (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
- Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.
- Gautheret, D. *et al.* (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.*, **248**, 27–43.
- Gotoh, O. (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.*, **9**, 361–370.
- Gotoh, O. (1999) Multiple sequence alignment: algorithms and applications. *Adv. Biophys.*, **36**, 159–206.
- Gottesmann, S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.*, **21**, 399–404.
- Gräf, S. *et al.* (2006) A computational approach to search for non-coding RNAs in large genomic data. In Nellen, W. and Hammann, C. (eds), *Small RNAs: Analysis and Regulatory Functions*, volume 17 of *Nucleic Acids and Molecular Biology*. Springer Verlag, pp. 57–74.
- Gribnikov, M. *et al.* (1990) Profile analysis. *Methods Enzymol.*, **183**, 146–159.
- Gusfield, D. (1999) *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*. Cambridge University Press, Cambridge.
- Havgaard, J.H. *et al.* (2005a) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.
- Havgaard, J.H. *et al.* (2005b) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33**, W650–W653.
- Hofacker, I. *et al.* (1994) Fast folding and comparison of RNA structures. *Monatsh. Chem.*, **125**, 167–188.
- Hofacker, I. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
- Hofacker, I. *et al.* (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
- Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
- Holmes, I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**, 73.
- Hudlot, C. *et al.* (2003) RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol. Phylogenet. Evol.*, **28**, 241–252.
- Katoh, K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
- Klein, R. and Eddy, S. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
- Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
- Kryukov, G.V. and Gladyshev, V.N. (2004) The prokaryotic selenoproteome. *EMBO Rep.*, **5**, 538–543.
- Lescoute, A. *et al.* (2005) Recurrent structural RNA motifs, isothercity matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
- Löytynoja, A. and Milinkovitch, M.C. (2003) A hidden Markov model for progressive multiple alignment. *Bioinformatics*, **19**, 1505–1513.
- Lück, R. *et al.* (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.*, **27**, 4208–4217.
- Martin, L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
- Mathews, D. and Turner, D. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
- Mathews, D.H. (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*, **21**, 2246–2253.
- Notredame, C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
- Rivas, E. and Eddy, S. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

- Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and proto-sequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
- Siebert,S. and Backofen,R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352–3359.
- Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Sokal,R. and Michener,C. (1958) A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin*, **38**, 1409–1438.
- Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
- Thompson,J. *et al.* (1994a) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Thompson,J. *et al.* (1994b) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.
- Thompson,J. *et al.* (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
- Washietl,S. *et al.* (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 2454–2459.
- Winkler,W. and Breaker,R. (2005) Regulation of bacterial gene expression by riboswitches. *Ann. Rev. Microbiol.*, **59**, 487–517.
- Yang,Q. and Blanchette,M. (2004) STRUCTMINER: a tool for alignment and detection of conserved secondary structure. *Genome Inform Ser Workshop Genome Inform.*, **15**, 102–111.
- Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
- Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.