Bioinformatics Advance Access published April 8, 2004

Alignment of RNA Base Pairing Probability Matrices

Ivo L. Hofacker, Stephan H.F. Bernhart and Peter F. Stadler

Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, Vienna, A-1090, Austria and Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstrasse 7b, Leipzig, D-04103, Germany

ABSTRACT

Motivation: Many classes of functional RNA molecules are characterized by highly conserved secondary structures but little detectable sequence similarity. Reliable multiple alignments can therefore be constructed only when the shared structural features are taken into account. Since multiple alignments are used as input for many subsequent methods of data analysis, structure based alignments are an indispensable necessity in RNA bioinformatics.

Results: We present here a method to compute pairwise and progressive multiple alignments from the direct comparison of basepairing probability matrices. Instead of attempting to solve the folding and the alignment problem simultaneously as in the classical Sankoff algorithm we use McCaskill's approach to compute base pairing probability matrices which effectively incorporate the information on the energetics of each sequences. A novel, simplified variant of Sankoff's algorithms can then be employed to extract the maximum weight common secondary structure and an associated alignment.

Availability: The programs `pmcomp` and `pmmulti` described in this contribution are implemented in `Perl` and can be downloaded together with the example data sets from http://www.tbi.univie.ac.at/RNA/PMcomp/. A web server is available at http://rna.tbi.univie.ac.at/cgi-bin/pmcgi.pl,

Contact: Ivo L. Hofacker,
Tel: ++43 1 4277 52738, Fax: ++43 1 4277 52793,
ivo@tbi.univie.ac.at

INTRODUCTION

Many functional classes of RNA molecules, including tRNA, rRNA, RNAse P RNA, SRP RNA, exhibit a highly conserved secondary structure but little sequence homology. Reliable alignments thus have to take structural information into account.

Sankoff's algorithm (Sankoff, 1985) that simultaneously allows the solution of the structure prediction and alignment problem is computationally very expensive, $\mathcal{O}(n^6)$ in CPU time and $\mathcal{O}(n^4)$ in memory for a pair of sequences of length n. Currently available software packages such as `foldalign` (Gorodkin *et al.*, 1997) and `dynalign` (Mathews & Turner, 2002) therefore implement only restricted versions. A further complication is that one should ideally implement the full loop-based RNA energy model (Mathews *et al.*, 1999), this is currently only done by `dynalign`. Stochastic context free grammars (SCFGs) present an alternative avenue to approaching the structure-alignment problem, see e.g. Sakakibara *et al.* (1994); Holmes & Rubin (2002); Klein & Eddy (2003). SCFGs do not use energies, but instead rely on probabilities for production rules derived from a training set. This again limits SCFGs to relatively simple scoring schemes with few parameters.

In this contribution we describe a different approach. Instead of starting from sequences alone, we first compute base pairing probability matrices predicted by means of McCaskill's algorithm (McCaskill, 1990) (implemented in the `RNAfold` program of `Vienna RNA Package` (Hofacker *et al.*, 1994; Hofacker, 2003)). The problem then becomes the alignment of the base pairing probability matrices. This appears to be an even harder threading problem, which is known to be NP-complete in the general case (Lathrop, 1994). For RNA structure alignments, however, we shall see the threading problem remains tractable as long as we score the alignment based on the notion of a common secondary structure.

IMPLEMENTATION OF A BANDED SANKOFF ALGORITHM

Suppose we are given two sequences A and B with their pair probability matrices P^A and P^B, resp. A natural way of determining the similarities of P^A and P^B is to search for the secondary structure of maximal "weight" that P^A and P^B have in common. In other words, we have to find a list of matches between pairs (i,j) from A and (k,l) from B that form a secondary structure and satisfy

$$\sum_{\text{matches } (ij;kl)} \left(\Psi^A_{ij} + \Psi^B_{kl}\right) \to \max, \qquad (1)$$

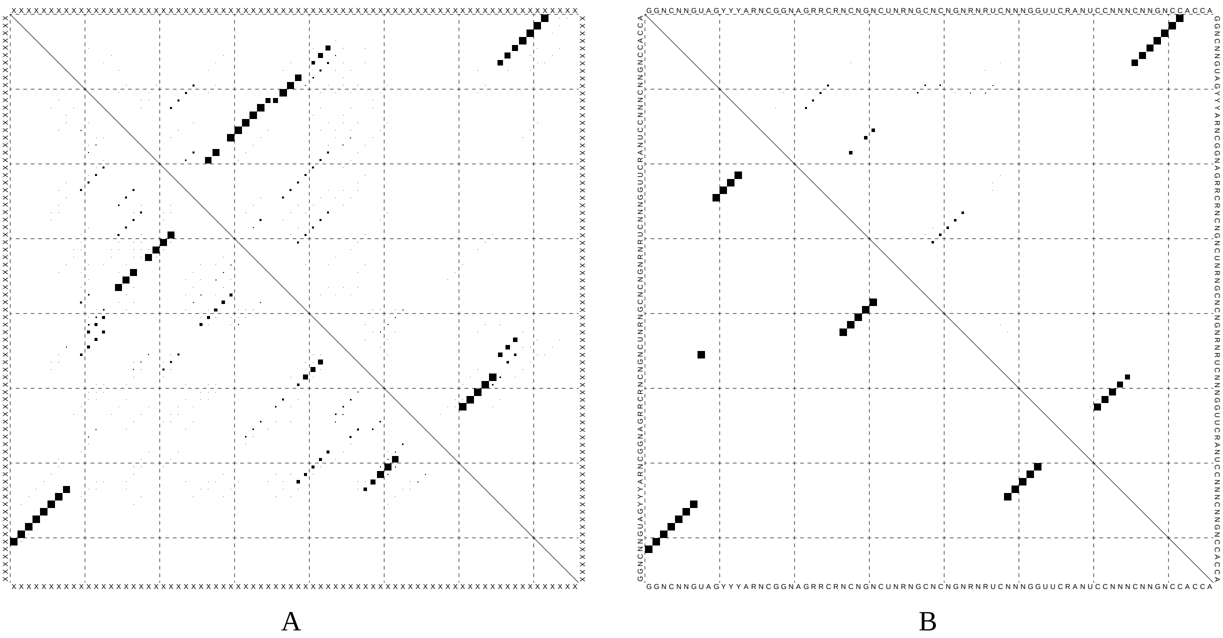


Fig. 1. A: Two base pairing probability matrices of tRNAs DF1140 (GAA from *Mycoplasma capricolum*, upper right) and DA0980 (TGC from *Thermoproteus tenax*, lower left) taken from M. Sprinzl's tRNA database (Sprinzl *et al.*, 1998).
B: pairwise alignment obtained based solely on the structural information using `pmcomp` with $\gamma = -5$. Note that the tRNA cloverleaf is not obvious from both individual structure predictions (A), while it is easily identified in the average of the aligned dot plots (B, upper right). The lower left triangle shows the base pairs of the consensus structure which, with the exception of the spurious isolated pair is identical to the consensus structure given in the database.

where Ψ^A_{ij} is the weight of pair (i,j) from sequence A. Here, we use $\Psi_{ij} = \log(P_{ij}/p_{\min})$ with $p_{\min}$ the minimum pair probability that is deemed significant. This form suggests itself since $\log(P_{ij})$ represents the mean free energy of the pair (i,j), however alternatives for the match score such as the product of pair probabilities $P^A_{ij} \cdot P^B_{kl}$ can be easily substituted.

More generally, we look for an alignment of the sequences A and B with N_{gap} insertions or deletions together with a consensus secondary structure $\mathcal{S}$ such that

$$\sum_{(ij;kl)\in\mathcal{S}} \left(\Psi^A_{ij} + \Psi^B_{kl} + \tau(A_i, A_j; B_k, B_l)\right) + \gamma N_{\text{gap}} + \sum_{i\in A, k\in B \;\notin \mathcal{S}} \sigma(A_i, B_k) \to \max \tag{2}$$

Here $\gamma < 0$ is the gap penalty, its absolute value has to be estimated or optimized. In our implementation γ is given in units of the maximum attainable match score, experience from sequence alignment thus suggest values between -3 and -5. For simplicity we only consider simple linear gap costs here, the extension to affine gap costs (Gotoh, 1982) is straightforward and increases computational requirements only by a constant factor. The scores $\sigma(A_i, B_k)$ and $\tau(A_i, A_j; B_k, B_l)$ describe the substitution of unpaired bases and base pairs, respectively. In the simplest case we disregard sequence-specific components, setting $\sigma = \tau = 0$. However, τ can be used to include contributions based on covariation or substitution. For example, one might want to use parameters derived along the lines of the `ribosum` matrices Klein & Eddy (2003). Similarly, parameters from a standard sequence similarity alignment package can be used as substitution scores σ.

Ideally, the structure dependent part of the scoring scheme would reflect the evolutionary likelihood of *structural transitions* instead of measuring structural similarity; at present there do not seem to be sufficient data available to actually derive such a parameter set. Furthermore, it is by no means clear that such an evolutionary motivated scoring model can be formulated in terms of additive contributions of individual base pairs.

Let $S_{i,j;k,l}$ be the score of the best matching of the subsequences $A[i..j]$ and $B[k..l]$. Furthermore, let $S^M_{i,j;k,l}$ be the best match subject to the constraint that (i,j) and (k,l) are matched base pairs. With this definition one easily obtains dynamic programming recursions

$$S_{i,j;k,l} = \max \begin{cases} S_{i+1,j;k,l} + \gamma, \\ S_{i,j;k+1,l} + \gamma, \\ S_{i+1,j,k+1,l} + \sigma(A_i, B_k), \\ \max_{h\leq j, q\leq l} \left(S^M_{i,h;k,q} + S_{h+1,j;q+1,l}\right) \end{cases}$$

$$S^M_{i,j;k,l} = S_{i+1,j+1,k+1,l+1} + \Psi^A_{ij} + \Psi^B_{kl} + \tau(A_i, A_j; B_k, B_l) \qquad (3)$$

with the initialization $S_{i,j;k,l} = |(j-i)-(l-k)|\gamma$ for $j-i \leq M+1$ or $l-k \leq M+1$. Here M is the minimum size of a hairpin loop, usually $M = 3$. The first two terms in the upper line of equ.(3) account for gaps in one of the two sequences. The third term describes the extension of both sub-sequences with an unpaired position. The max-term, finally, describes a match of the pairs (i,h) in A with (k,q) in B. The expression for the score restricted to a match of (i,j) with (k,l) is straightforward. Recursion (3) requires $\mathcal{O}(n^4)$ memory to store the scores $S_{i,j;k,l}$ and requires $\mathcal{O}(n^6)$ operations.

This is the same as the (maximum circular matching version of the) Sankoff algorithm when we set $P^A_{ij} = 1$ if A_i and A_j can form a base pair and $P^A_{ij} = 0$ otherwise. The algorithm shown here is not restricted to canonical alignments, in which adjacent insertions and deletions occur only in one of the two possibles orders (Waterman, 2003). This restriction is important if suboptimal alignments are computed, an appropriate modification of the algorithm is straightforward.

Restricting the difference $\Delta = |(j-i)-(l-k)|$ in the "span" of matching base pairs $(i,j) \in A$ and $(k,l) \in B$ reduces the complexity to $\mathcal{O}(n^5)$ CPU usage; restricting this difference for *all* partial alignments reduces the computational effort to $\mathcal{O}(n^3)$ memory and $\mathcal{O}(n^4)$ CPU usage. Note that in the latter case Δ must be larger than the difference in sequence length. The bigger Δ, the smaller the reduction of the computational effort but also the smaller the chance of missing significant structural homologies.

Standard backtracking can be used to retrieve the matched sequence positions. Here we have two kinds: matches of a base pair $(i,j) \in A$ with $(k,l) \in B$ and matches of unpaired bases corresponding to the third term in the first line of equ.(3). Note that matches of unpaired bases do not contribute to the score in the simple scoring scheme with $\sigma = \tau = 0$. Thus the exact positions of gaps within a stretch of unpaired bases is arbitrary in this case.

MULTIPLE ALIGNMENTS

Given the alignment we can define a "consensus" or "average" base pair probability matrix as

$$P^{AB}_{p,q} = \begin{cases} \sqrt{P^A_{i_p,j_q} P^B_{k_p,l_q}} & \text{for matches} \\ 0 & \text{for gaps} \end{cases} \qquad (4)$$

where i_p is the positions in sequence A corresponding to the position p of the alignment. As a consequence we can easily extend the method to progressive alignments of more than two sequences by using $P^{AB}_{p,q}$ to compute Ψ. Using the geometric mean for the "consensus" probability matrix together with a match score $\Psi \sim \log(P)$ is roughly equivalent to the usual sum of pair score for multiple alignments. For other scoring schemes the arithmetic mean $(P^A_{i_p,j_q} + P^B_{k_p,l_q})/2$ may be more appropriate.

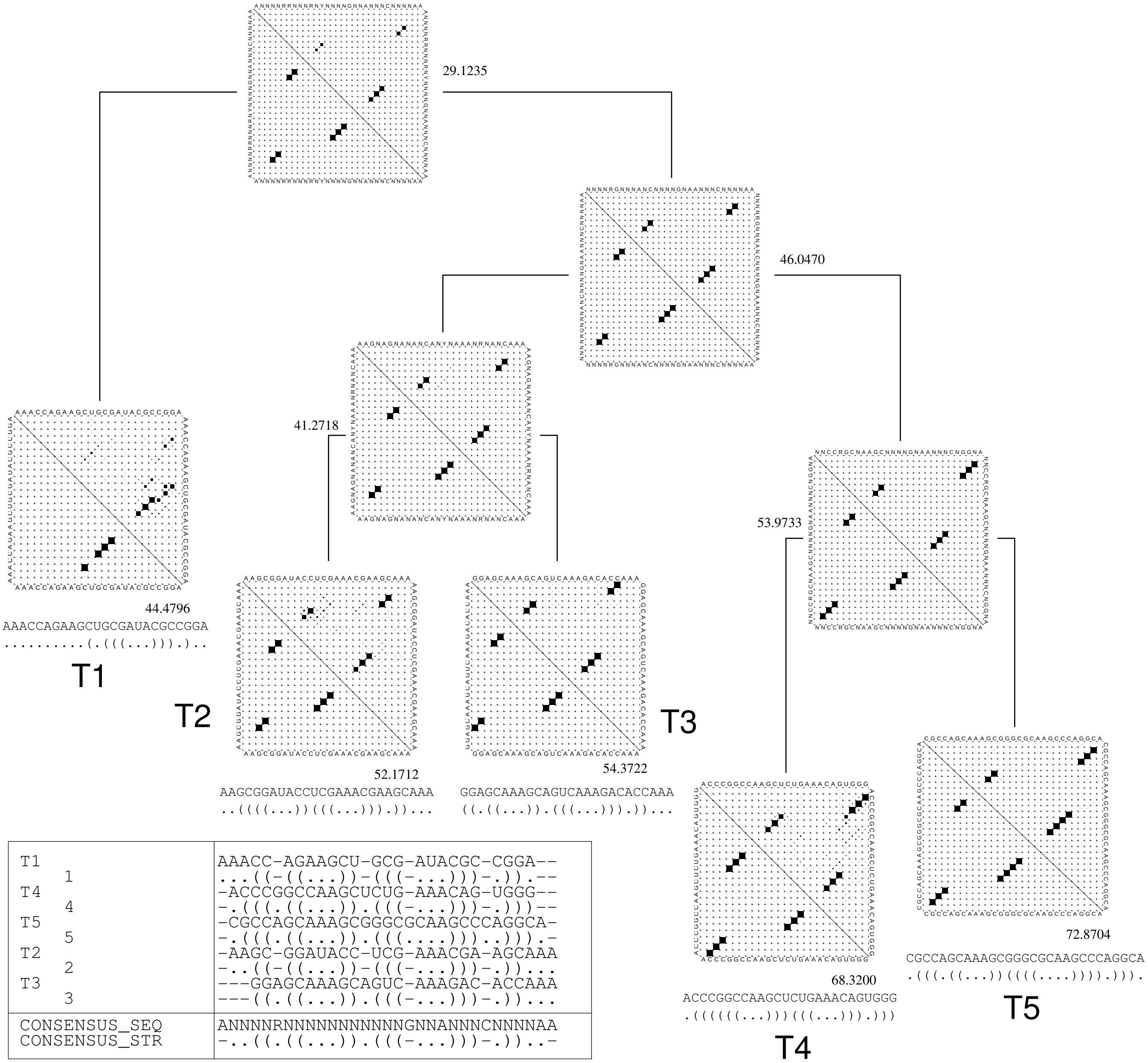
In the current implementation a script `pmmulti` first calls `pmcomp` to compute all pairwise alignments, then takes the similarity scores to produce a guide tree using the weighted pairgroup clustering method and assembles the multiple alignment. For the construction of multiple alignments it is not necessarily desirable to use canonical alignments since the correct order of insertions/deletions in a pairwise alignment is in general determined by the other sequences in the multiple alignment.

A coarse, but *much* faster, method to compare pair probabilities was already introduced in Bonhoeffer *et al.* (1993). From the pairing probabilities of base i we construct a vector containing the probabilities of being paired upstream $p^{<}(i) = \sum_{j>i} P_{ij}$, downstream $p^{>}(i) = \sum_{j<i} P_{ji}$, or unpaired $p^{\circ}(i) = 1 - p^{<}(i) - p^{>}(i)$. The resulting profiles can be aligned by means of a standard string/profile alignment algorithm in $\mathcal{O}(n^2)$ time using

$$\rho = \sqrt{p^{>}_A p^{>}_B} + \sqrt{p^{<}_A p^{<}_B} + \sqrt{p^{\circ}_A p^{\circ}_B} \qquad (5)$$

as the match score. While this fast method, to which we will refer as the *"string-like alignment"*, often produces misaligned pairs, we have found the quality to be sufficient for the pairwise alignments used to construct the guide tree. Thus for a multiple alignment of N RNAs, we can use a fast approximate algorithm to compute the $N(N-1)/2$ pairwise alignments and restrict the expensive Sankoff algorithm to the $N-1$ progressive alignments along the guide trees.

As an example for the quality of `pmmulti` alignments we compare the predicted conserved structure of the IRES Ib element of Aphtovirus and Cardiovirus, two genera of the family picornaviridae in Fig. 3. The two approximate methods for structure-enhanced alignments, `MARNA` (Siebert & Backofen, 2003) and an alignment computed by `pmmulti` in the "string-like" alignment mode (labeled `PMstring` in the Figure) produce acceptable results. In



```
T1             AAACC-AGAAGCU-GCG-AUACGC-CGGA--
     1         ...((-((...))-(((-...)))-.)).--
T4             -ACCCGGCCAAGCUCUG-AAACAG-UGGG--
     4         -.(((.((...)).(((-...)))-.)))--
T5             -CGCCAGCAAAGCGGGCGCAAGCCCAGGCA-
     5         -.(((.((...)).((....)))..))).-
T2             -AAGC-GGAUACC-UCG-AAACGA-AGCAAA
     2         -..((-((...))-(((-...)))-.))...
T3             ---GGAGCAAAGCAGUC-AAAGAC-ACCAAA
     3         ---((.((...)).(((-...)))-.))...
CONSENSUS_SEQ  ANNNNRNNNNNNNNNNNGNNANNNCNNNNAA
CONSENSUS_STR  -..((.((...)).(((-...)))-.))..-
```

Fig. 2. A toy example for a multiple alignment of five RNA sequences aligned solely based on their base pairing probability matrices. For each sequence `T1`-`T5` we give its minimum free energy structure and the self-similarity score of their base pairing probability matrices. The guide tree is constructed based on the scores of pairwise alignments using $\gamma = -3$. The inset contains the final multiple alignment.

contrast, `ClustalW` (Thompson *et al.*, 1994) yields a low quality alignment in structural terms, showing many inconsistent mutations. The Sankoff algorithm (`pmmulti`), finally, slightly improves the manual alignment, finding one additional compensatory mutation.

DISCUSSION

The dynamic programming algorithm (3) is a variant of Sankoff's algorithm (Sankoff, 1985) for simultaneously aligning and folding two RNA sequences. The main idea of our implementation is that we avoid implementing the sophisticated loop-based thermodynamic energy model for RNA folding (Mathews *et al.*, 1999), instead relying

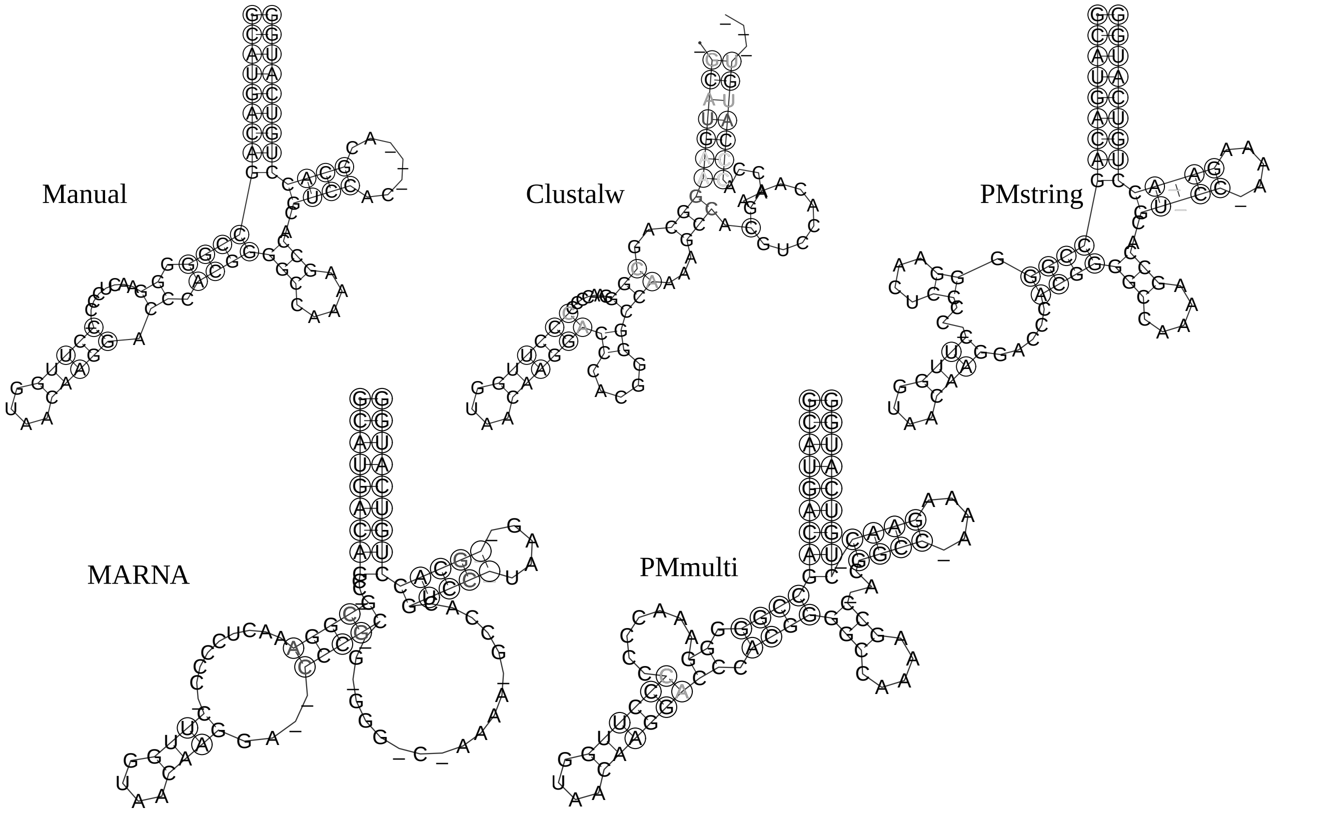


Fig. 3. Consensus structure of the IRES Ib region predicted from four Aphtovirus (FDI251473, FMDVALF, PIFMDV2, FAN133359) and three Cardiovirus sequences (MNGPOLY, EMCBCG, TMEPP), to be found e.g. at http://rna.tbi.univie.ac.at/cgi-bin/virusdb.cgi. All structures were predicted by the `RNAalifold` program (Hofacker *et al.*, 2002) from multiple alignments produced by different methods as input. Upper left: manual alignment taken from Witwer *et al.* (2001), upper center: `ClustalW` alignment, upper right: `pmmulti` in string-like alignment mode, lower left: `MARNA` (Siebert & Backofen, 2003) alignment, lower center: `pmmulti`. Circles indicate consistent and compensatory mutations while gray letters mark inconsistent mutations.
While the approximate methods `MARNA` and `PMstring` (`pmmulti` in string-like alignment mode) produce acceptable results, purely string-based approaches `ClustalW` yields a low quality alignment in structural terms, showing many inconsistent mutations. The two approximate methods for structure-enhanced alignments, `MARNA` and `PMstring` produce acceptable results but do not reach the number of compensatory mutations obtained in the manually edited alignment (upper left). The Sankoff algorithm (`pmmulti`), on the other hand, slightly improves the manual alignment, finding one additional compensatory mutation.

on a much simpler analogue of Nussinov's weighted circular matching problem (Nussinov *et al.*, 1978). Nevertheless, our approach does include thermodynamic information about the RNA molecules via the base pair probability matrices P^A and P^B that are used as input. Another advantage over combined alignment-and-folding programs such as `dynalign` (Mathews & Turner, 2002) and `FoldAlign` (Gorodkin *et al.*, 1997) is the fact that the input pairing matrices can be computed independently. For example, one might want to use the results of kinetic folding simulations (Flamm *et al.*, 2000) that can differ significantly from the equilibrium thermodynamics results for some molecules.

Multiple structural alignments can be analyzed further by familiar techniques. For example, a parsimony program can be used to extract the phylogenetic relationships from structural information. Since the alignment is constructed such that only base pairs (i.e., matching pairs of parentheses) and unpaired positions are aligned with each other, the original version of the Fitch algorithm can be used to obtain meaningful parsimony scores directly from the aligned dot-parenthesis strings. This effectively weights base pairs with double weight compared to unpaired positions. For the example in Fig. 2 the unique most parsimonious tree is `((T1)(T2)((T3)((T4)(T5))))` with a score of 13, compared to the guide tree shown in the figure, `((T1)((T2)(T3))((T4)(T5)))`, with a score of 14. More elaborate scoring schemes and even associated maximum likelihood techniques acting directly on RNA secondary structures are algorithmically unproblematic but will require detailed knowledge of the dynamics of RNA structure evolution, while at present even RNA sequence evolution in the presence of (partially) conserved structures is understood only partially, see e.g. (Knudsen & Hein, 1999; Savill *et al.*, 2001; Otsuka & Sugaya, 2003).

Alignments of base pairing probability matrices could also be employed as part of a structure search procedure. For example, the locally stable secondary structure motifs in a large RNA (as computed e.g. using the `RNALfold` algorithm (Hofacker *et al.*, 2004).) could be compared with a the base pairing probability matrices of one or a collection of query sequences. Due to high computational cost of the Sankoff algorithm it will probably by necessary to pre-select promising candidate sequences using a pattern search procedure. Tools such as `HyPa` (Gräf *et al.*, 2001), `rnaforester` (Höchsmann *et al.*, 2003), or the Algebraic Dynamic Programming approach advocated in (Meyer & Giegerich, 2002) could be used to scan for candidate sequences, that are then folded by McCaskill's algorithm (e.g. using `RNAfold -p`) and compared to either each individual or the consensus structure of a multiple alignment of query structures.

Acknowledgments

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, by the Austrian Gen-AU bioinformatics integration network sponsored by BM-BWK, and by the DFG Bioinformatics Initiative BIZ-6/1-2.

REFERENCES

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. & Schuster, P. (1993). RNA multi-structure landscapes. a study based on temperature dependent partition functions. *Eur. Biophys. J.*, **22**, 13–24.

Flamm, C., Fontana, W., Hofacker, I. & Schuster, P. (2000). RNA folding kinetics at elementary step resolution. *RNA*, **6**, 325–338.

Gorodkin, J., Heyer, L. J. & Stormo, G. D. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucl. Acids Res.*, **25**, 3724–3732.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gräf, S., Strothmann, D., Kurtz, S. & Steger, G. (2001). `HyPaLib`: a database of RNAs and RNA structural elements defined by hybrid patterns. *Nucleic Acids Res.*, **29**, 196–198.

Höchsmann, M., Töller, T., Giegerich, R. & Kurtz, S. (2003). Local similarity in RNA secondary structures. In *Proc of the Computational Systems Bioinformatics Conference, Stanford, CA, August 2003 (CSB 2003)*. pp. 159–168.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucl. Acids Res.*, **31**, 3429–3431.

Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, **125**, 167–188.

Hofacker, I. L., Priwitzer, B. & Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 191–198.

Holmes, I. & Rubin, G. M. (2002). Pairwise RNA structure comparison using stochastic context-free grammars. In Altman, R. B., Dunker, K. A., Hunter, L. & Klein, T. E., (eds.) *Pacific Symposium on Biocomputing (PSB 2002)*. World Scientific, Singapore, pp. 163–174.

Klein, R. J. & Eddy, S. R. (2003). `RSEARCH`: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44 (16 pages).

Knudsen, B. & Hein, J. J. (1999). Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics*, **15**, 446–454.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Eng.*, **7**, 1059–1068.

Mathews, D., Sabina, J., Zuker, M. & Turner, H. (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews, D. H. & Turner, D. H. (2002). Dynalign: An algorithm for finding secondary structures common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Meyer, C. & Giegerich, R. (2002). Matching and significance evaluation of combined sequence-structure motifs in RNA. *Z. Phys. Chem.*, **216**, 193–216.

Nussinov, R., Piecznik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithms for loop matching. *SIAM J. Appl. Math.*, **35**, 68–82.

Otsuka, J. & Sugaya, N. (2003). Advanced formulation of base pair changes in the stem regions of ribosomal RNAs; its application to mitochondrial rRNAs for resolving the phylogeny of animals. *J. Theor. Biol.*, **222**, 447–460.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C. & Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, **22**, 5112–5120.

Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.

Savill, N. J., Hoyle, D. C. & Higgs, P. G. (2001). RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics*, **157**, 399–411.

Siebert, S. & Backofen, R. (2003). MARNA: A server for multiple alignment of RNAs. In Mewes, H.-W., Heun, V., Frishman, D. & Kramer, S., (eds.) *Proceedings of the German Conference on Bioinformatics. GCB 2003*, volume 1. belleville Verlag Michael Farin, München, D, pp. 135–140.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, **26**, 148–153.

Thompson, J. D., Higgs, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.

Waterman, M. S. (2003). Introduction of Computational Biology. Chapman & Hall, Boca Raton, FL.

Witwer, C., Rauscher, S., Hofacker, I. L. & Stadler, P. F. (2001). Conserved RNA secondary structures in picornaviridae genomes. *Nucl. Acids Res.*, **29**, 5079–5089.