The Partition Function Variant of Sankoff's Algorithm

Ivo L. Hofacker[1] and Peter F. Stadler[1,2]

[1] Institut für Theoretische Chemie und Molkulare Strukturbiologie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria
`http://www.tbi.univie.ac.at/~ivo`

[2] Bioinformatics, Department of Computer Science, University of Leipzig, Kreuzstrasse 7b, D-04103 Leipzig, Germany
`http://www.bioinf.uni-leipzig.de/~studla`

Abstract. Many classes of functional RNA molcules 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. Sankoff's algorithm can be used to construct such structure-based alignments of RNA sequences in polynomial time. Here we extend the approach to a probabilistic one by explicitly computing the partition function of all pairwisely aligned sequences with a common set of base pairs. Stochastic backtracking can then be used to compute e.g. the probability that a prescribed sequence-structure pattern is conserved between two RNA sequences. The reliability of the alignment itself can be assessed in terms of the probabilities of each possible match.

1 Introduction

Sankoff's algorithm [1] simulateneously predicts a consensus structure for two (or, in its general version, more) RNA secondary structure and at the same time constructs their alignment. It is quite expensive in both CPU and memory requirements, $\mathcal{O}(N^6)$ and $\mathcal{O}(N^4)$, respectively. A further complication is that it requires the implementation of the full loop-based RNA energy model [2]. Currently available software packages such as `foldalign` [3] and `dynalign` [4] therefore implement only restricted versions. A complementary approach is taken in the `pmmatch` program [5]. Instead of attempting to solve the alignment and the structure prediction problem simultaneously, `pmmatch` utilizes the base pairing probability matrices predicted by means of McCaskill's algorithm [6] (implemented in the `RNAfold` program of `Vienna RNA Package` [7, 8]). The problem then becomes the alignment of the base pairing probability matrices. This appears to be an even harder threading problem, which in general is known to be NP-complete [9]. In the RNA case, the threading problem remains tractable as long as we score the alignment based on the notion of a common secondary structure. In fact, it reduces to a variant of the Sankoff algorithm in which the energy model for structure prediction part is reduced to base weights on the base pairs.

Suppose we are given two sequences A and B (of length $n = |A|$ and $m = |B|$) together 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. 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 obtains dynamic programming recursions

$$S_{i,j;k,l} = \max \begin{cases} S_{i+1,j;k,l} + \gamma, \\ S_{i,j;k+1,l} + \gamma, \\ S_{i+1,j,k+1,l} + \alpha(A_i, B_k), \\ \max_{h \le j, q \le l} \left(S^M_{i,h;k,q} + S_{h+1,j;q+1,l} \right) \end{cases} \tag{1}$$
$$S^M_{i,j;k,l} = S_{i+1,j+1,k+1,l+1} + \tau(P^A_{ij}, A_i, A_j; P^B_{kl}, B_k, B_l)$$

with the initialization $S_{i,j;k,l} = |(j - i) - (l - k)|\gamma$ for $j - i \le M + 1$ or $l - k \le M + 1$, where M is the minimum size of a hairpin loop, usually $M = 3$. The constant $\gamma < 0$ is a gap penalty. The scores $\alpha_{ik} = \alpha(A_i, B_k)$ and $\tau_{ij,kl} = \tau(P^A_{ij}, A_i, A_j; P^B_{ij}, B_k, B_l)$ describe the substitution of unpaired bases and base pairs, respectively. The latter term my depends on both the structures and the underlying sequences. Backtracking can be used to retrieve both the common secondary structure and the associated sequence alignment [5].

For both RNA folding and sequence alignment it is possible to compute partitions functions instead of optimal scores with essentially the same resources. In a second step probabilistic versions of the optimal structure of alignment can be constructed; see [6] for RNA folding and [10–14].

In this contribution we describe a "partition function version" of the Sankoff algorithm that computes the probabilities of matches in the structure-based alignments of two RNA molecules, thereby providing an intrinsic measure of the *local* quality of the structure-based alignments.

In the thermodynamic interpretation of the simultaneous folding and alignment problem a state θ is a pair $\theta = (\mathfrak{S}, \mathcal{A})$ of secondary structure $\mathfrak{S}$ consisting of all matched base pairs $(ij; kl)$, where (A_i, A_j) is a base pair in structure A and (B_k, B_l) is a base pair in structure B, and an alignment $\mathcal{A}$ of the underlying sequences A and B such that A_iB_k and A_jB_l are matches. Note that the alignment $\mathcal{A}$ in general contains further matches corresponding to unpaired nucleotides. The probability of a particular state is then

$$\text{Prob}[\theta] = Z^{-1} \exp(-\sigma(\theta)) \tag{2}$$

where the score is given explicitly in the form

$$\sigma(\theta) = \sum_{(ij;kl) \in \mathfrak{S}} \tau_{ij,kl} + \sum_{i \in A, k \in B \ \notin \mathfrak{S}} \alpha_{ik} + \gamma\big(m + n - 2|\mathcal{A}|\big)\,. \tag{3}$$

In the last term, $N_{\rm gap} = n + m - 2|\mathcal{A}|$ is the number of gaps in the alignment. The normalization constant

$$Z = \sum_{\theta} \exp(-\sigma(\theta)) \tag{4}$$

is the partition function of the model. The probability of a feature Ω can now be computed as the sum of the probabilites of all states $\theta \in \Omega$. In particular, we are interested in $\Omega^{(p,q)}$, the set of all states in which $A_p B_q$ is a match in the alignment.

2 Recursions

We first observe that equ.(1) can easily be transformed into a recursion for the partition function $Z_{ij;kl}$ of the model restricted to the subsequences $A[i..j]$ and $B[k..l]$. Explicitly, we obtain

$$\begin{aligned} Z_{ij;kl} = {} & Z_{i+1,j;kl} e^{\gamma} + Z_{ij;k+1,l} e^{\gamma} + Z_{i+1,j;k+1,l} e^{\alpha_{ik}} \\ & + \sum_{\substack{(k,q) \text{ paired in } B \\ (i,p) \text{ paired in } A}} Z_{i+1,p-1;k+1,q-1} Z_{p+1,j;q+1,l} e^{\tau_{ij,pq}} \end{aligned} \tag{5}$$

Let us now consider all states that contain the match $A_x B_y$. We have to distinguish four cases: (i) there is no matched base pair in $\mathfrak{S}$, (ii) $(i, x; j, y) \in \mathfrak{S}$, (iii) $(x, k; y, l) \in \mathfrak{S}$, and (iv) $A_x B_y$ is "immediately interior" to a matched pair $(i, j; k, l) \in \mathfrak{S}$ in the sense that $i < x < j$, $k < y < l$ and there no other pair $(i', j', k', l') \in \mathfrak{S}$ such that $i < i' < x < j' < j$ and $k < k' < y < l' < l$. Fig. 1 gives a graphical description: Clearly, these four cases are pairwise disjoint and cover all possibilities. We can therefore write the partition function Q^{xy} of all states that contain the match $A_x B_y$ in the alignment as follows:

$$\begin{aligned} Q^{xy} = {} & Z_{1,x-1;1,y-1} e^{\alpha_{ij}} Z_{x+1,n;y+1,n} + \\ & \sum_{i<x,\, k<y} Z_{i+1,x-1;k+1,y-1} e^{\tau_{i,x;k,y}} \widehat{Z}_{i,x;k,y} + \\ & \sum_{j>x,\, l>y} Z_{x+1,j-1;y+1,l-1} e^{\tau_{x,j;y,l}} \widehat{Z}_{x,j;y,l} + \\ & \sum_{i<x,\, j>x\; k<y,\, l>y} e^{\tau_{i,j;k,l}} Z_{i+1,x-1;k+1,y-1} e^{\alpha_{xy}} Z_{x+1,j-1;y+1,l-1} \widehat{Z}_{i,j;k,l} \end{aligned} \tag{6}$$

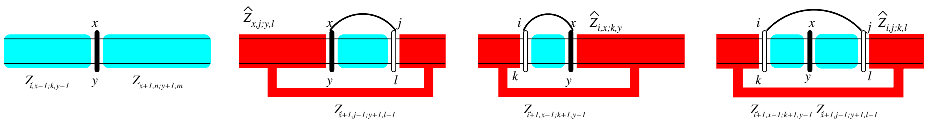


Fig. 1. Decomposition of the restricted partition function Q_{xy} into unconstrained partition functions $Z_{\dots}$ and $\widehat{Z}_{\dots}$ of sub-problems. For details see text.

where $\widehat{Z}_{i,j;k,l}$ denotes the partition function over all partial states *outside* the aligned interval $[i,j][k,l]$, i.e., excluding the positions i, j, k, and l. This corresponds to the states of the sub-problem with $A' = A[1..i-1]A[j+1..n]$ and $B' = B[1..k-1]B[l+1,m]$. We can easily find recursions for computing $\widehat{Z}_{i,j;k,l}$ from shorter subproblems (i.e., those with a larger "missing" interval) and the values of $Z_{i,j;k,l}$:

$$\begin{aligned}\widehat{Z}_{i,j;k,l} =& \widehat{Z}_{i,j+1;k,l}e^{\gamma} + \widehat{Z}_{i,j;k,l+1}e^{\gamma} + \widehat{Z}_{i,j+1;k,l+1}e^{\alpha_{j+1,l+1}} + \\ & \sum_{p<i,\, q<k} \widehat{Z}_{p,j+1;q,l+1}e^{\tau_{p,j+1;q,l+1}} Z_{p+1,i-1;q+1,k-1} + \\ & \sum_{p>j+1,\, q>l+1} \widehat{Z}_{i,p;k,q}e^{\tau_{j+1,p;l+1,q}} Z_{j+2,p-1;l+2,q-1}\end{aligned} \tag{7}$$

The probability for the match $A_x B_y$ given the input data and scoring scheme ist simply

$$P^{xy} = Q^{xy}/Z \tag{8}$$

Tabulating the $\mathcal{O}(n^4)$ entries of the partition functions $Z_{ij;kl}$ requires $\mathcal{O}(n^6)$ operation, just as the solution of the optimization problem. Then $\widehat{Z}_{ij;kl}$ can be computed also in $\mathcal{O}(n^6)$ operations. Given these two tables, recursion 6 can be evaluated in $\mathcal{O}(n^4)$ steps for each value x and y. The matrix of matching probabilites can therefore be computed in $\mathcal{O}(n^4)$ memory and $\mathcal{O}(n^6)$ CPU. Just as in the case of sequence alignements and secondary prediction, the partition function version is therefore not more expensive than the associated optimization problem.

3 Stochastic Backtracking

As described in [5], backtracking in the recursions (1) can be performed in $\mathcal{O}(n^3)$ to obtain a score-optimal alignment. When the partition functions $Z_{i,j;k,l}$ for the sub-problems are known, it is possible sample from the distribution of the alignments by means of "stochastic backtracking". This approach has recently be implemented for pairwise sequence alignment [14] and for RNA structure prediction in the latest release of the Vienna RNA Package [8, 15], see also [16, 17] and [18], where the idea was used to generate random RNA structures with uniform distribution. This method generalizes in a straightforward way to the Sankoff algorithm: From equ.(5) we obtain immediately that the subalignment

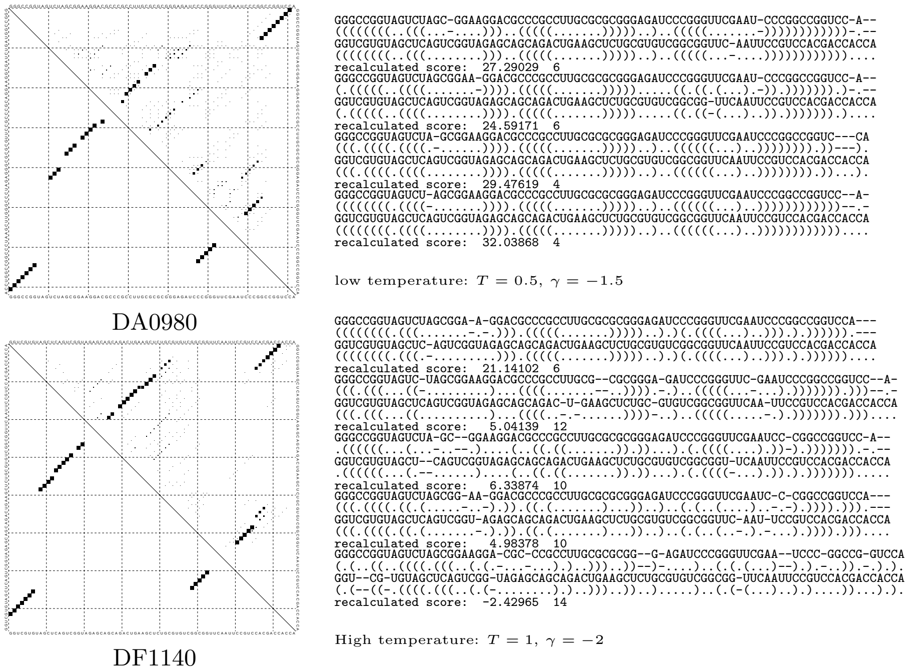


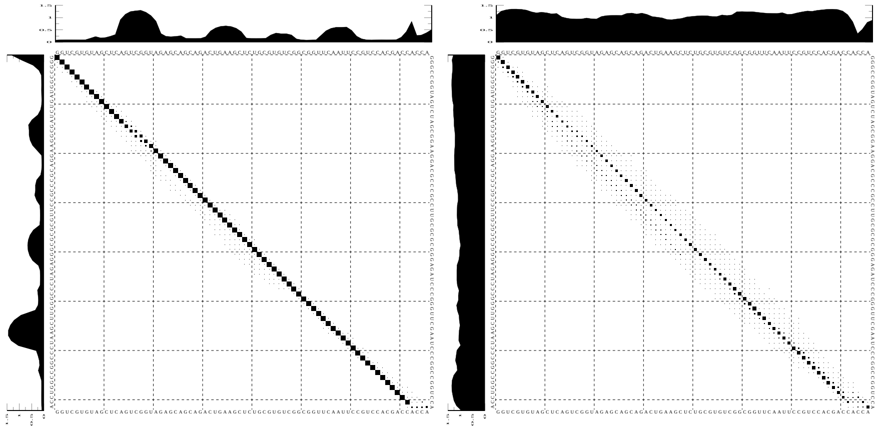
Fig. 2. Left: Two base pairing probability matrices of tRNAs taken from M. Sprinzl's tRNA database [19]: DA0980 (TGC from *Thermoproteus tenax* and DF1140 (GAA from *Mycoplasma capricolum*). Right: examples of pairwise alignments generated with two different parameter sets. The number of gaps (second column of numbers below the alignment) increases with temperature T even though $-\gamma/T$ decreases.

of $x[i..j]$ with $y[k..]$ is the from of one four types with the probabilities p listed below:

Deletion of x_i	$p = Z_{i+1,j;kl} e^{\gamma} / Z_{ij;kl}$
Deletion of y_k	$p = Z_{ij;k+1,l} e^{\gamma} / Z_{ij;kl}$
Unpaired Match of x_i and y_j	$p = Z_{i+1,j;k+1,l} e^{\alpha_{ik}} / Z_{ij;kl}$
Matched pair (x_i, x_p), $(y_k - y_q)$	$p = Z_{i+1,p-1;k+1,q-1} Z_{p+1,j;q+1,l} e^{\tau_{ij,pq}} / Z_{ij;kl}$

Chosing in each step of the backtracking procedure one of these alternatives with the correct probability results again in an $\mathcal{O}(n^3)$ algorithm that produces an alignment with the probability $p = \exp(-\text{score})$, Fig. 2.

The advantage of this procedure is that an emsemble of on the order of n^3 sample alignments can be computed economically (since we need $\mathcal{O}(n^6)$ time for the forward recursion and only $\mathcal{O}(n^3)$ for backtracking a single alignment). This samples can then be used to estimate the probabilities of features such as particular multiloops or non-local sequence-structure combinations.



$T = 0.5$, $\gamma = -1.5$ $\quad$ $T = 1$, $\gamma = -2$

Fig. 3. Match probabilities for the pairwise alignments of the two tRNAs DA0980 and DF1140 from Fig. 2. The area of the squares at position x, y is proportional to P^{xy}. The small panels along the axes show the position-wise entropies relative to each sequence.

4 An Example

As an example we consider here the alignment of two rather disparate tRNA sequences, Figs. 2 and 3. We use here

$$\tau(P^A_{ij}, A_i, A_j; P^B_{kl}, B_k, B_l) = 2 \ln n + \ln P^A_{ij} + \ln P^B_{kl} \tag{9}$$

for the pair score and neglect sequence similarity altogether, i.e., $\alpha_{ik} = 0$.

Note that for both sequences the predicted optimal secondary structures is not the clover-leaf, as shown in the l.h.s. of Fig. 2. Nevertheless, most stochastic backtrackings retrieve the clover-leaf as consensus structure of the two molecules. Since sequence similarity was not used in the scoring, the exact position of gaps within loop regions is arbitrary. For low temperatures (upper right panel in Fig. 2) alignments differ almost exclusively in the D-loop and at the 3' end of the tRNAs.

The local reliability of the alignment can be measured by the entropy of the match probabilities

$$S(x) = -\sum_y P^{xy} \ln P^{xy} - p^0(x) \ln p^0(x) \tag{10}$$

where $p^0(x) = 1 - \sum_y P^{xy}$ is the probability the position x is unmatched (i.e., opposite to a gap in the alignment). As can be seen in Fig.3, the alignment is typically much more well-defined in paired regions. For large values of the temperature T this difference disappears, however.

5 Concluding Remarks

We have introduced here a partition function version of the Sankoff algorithm. The algorithm is quite expensive both in memory and CPU time; The resource requirement is, however, essentially the same as for the "classical" version that computes the optimal alignment only. From the partition functions we can, in addition to the optimal alignment, also descriminate reliable from unreliable parts of a structure-based alignment of RNA molecules.

Stochastic pairwise alignments are useful in many different contexts: Numerous tools in bioinformatics require pairwise sequence alignments as input data. The present approach thus provides a tool that can be used to produce alignments with realistically distributed errors and varying overall quality (by choosing the temperature parameter T). These can be used to investigate the sensitivity of the method with respect to realistic variations of the input alignments. In particular, used as an input of a multiple alignment methods such as `t-coffee` [20] it can be used to produce multiple alignments together with estimates of local alignment quality.

While the Sankoff algorithm is too slow to scan large portions of a genome for conserved RNAs, it is still useful to post-process candidates for structurally conserved RNA detected by other methods, e.g. `qrna` [21].

The current implementation uses simple linear gap costs. A generalization to affine gap costs is straighforward along the lines of Gotoh's algorithm [22] for sequence alignments and should improve the placement of scattered gaps.

Acknowledgements

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893 and the DFG Bioinformatics Initiative BIZ-6/1-2.

References

1. Sankoff, D.: Simultaneous solution of the RNA folding, alignment, and protosequence problems. SIAM J. Appl. Math. **45** (1985) 810–825
2. Mathews, D., Sabina, J., Zuker, M., Turner, H.: Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J. Mol. Biol. **288** (1999) 911–940
3. Gorodkin, J., Heyer, L.J., Stormo, G.D.: Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucl. Acids Res. **25** (1997) 3724–3732
4. Mathews, D.H., Turner, D.H.: Dynalign: An algorithm for finding secondary structures common to two rna sequences. J. Mol. Biol. **317** (2002) 191–203
5. Hofacker, I.L., Berhart, S., Stadler, P.F.: Alignment of rna base pairing probability matrices. Bioinformatics (2003) submitted.
6. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29** (1990) 1105–1119

7. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chemie **125** (1994) 167–188
8. Hofacker, I.L.: Vienna RNA secondary structure server. Nucl. Acids Res. **31** (2003) 3429–3431
9. Lathrop, R.H.: The protein threading problem with sequence amino acid interaction preferences is np-complete. Protein Eng. **7** (1994) 1059–1068
10. Bucher, P., Hoffmann, K.: A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States, D.J., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R.F., eds.: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96), Menlo Park, CA, AAAI Press (1996) 44–50
11. Kschischo, M., Lassig, M.: Finite-temperature sequence alignment. Pacific Symposium Biocomputing **1** (2000) 624–35
12. Miyazawa, S.: A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng. **8** (1994) 999–1009
13. Yu, Y.K., Hwa, T.: Statistical significance of probabilistic sequence alignment and related local hidden markov models. J. Comp. Biol. **8** (2001) 249–282
14. Mückstein, U., Hofacker, I.L., Stadler, P.F.: Stochastic pairwise alignments. Bioinformatics **S153-S160** (2002) 18 ECCB 2002.
15. Flamm, C., Hofacker, I.L., Stadler, P.F.: Computational chemistry with RNA secondary structures. Kemija u industriji (2004) Proceedings CECM-2, Varaždin, June 19-21, 2003.
16. Ding, Y., Lawrence, C.E.: Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res. **29** (2001) 1034–1046
17. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31** (2003) 7280–7301
18. Tacker, M., Stadler, P.F., Bornberg-Bauer, E.G., Hofacker, I.L., Schuster, P.: Algorithm independent properties of RNA structure prediction. Eur. Biophy. J. **25** (1996) 115–130
19. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S.: Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. **26** (1998) 148–153
20. Notredame, C., Higgins, D., Heringa, J.: T-coffee: A novel method for multiple sequence alignments. J. Mol. Biol. **302** (2000) 205–217
21. Rivas, E., Eddy, S.R.: Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics **2** (2001) 19 pages
22. Gotoh, O.: An improved algorithm for matching biological sequences. J. Mol. Biol. **162** (1982) 705–708