BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S153–S160

Stochastic pairwise alignments

U. Mückstein, I.L. Hofacker and P.F. Stadler

Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien, Währingerstraße 17, Vienna, A-1090, Austria and The Santa Fe Institute, Santa Fe, New Mexico, USA

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Motivation: The level of sequence conservation between related nucleic acids or proteins often varies considerably along the sequence. Both regions with high variability (mutational hot-spots) and regions of almost perfect sequence identity may occur in the same pair of molecules. The reliability of an alignment therefore strongly depends on the level of local sequence similarity. Especially in regions of high variability, many alignments of almost equal quality exist, and the *optimal* alignment is highly arbitrary.

Results: We discuss two approaches which deal with the inherent ambiguity of the alignment problem based on the computation of the partition function over all canonical pairwise alignments. The ensemble of possible alignments can be described by the probabilities P_{ij} of a match between position i in the first and position j in the second sequence. Alternatively, we introduce a probabilistic backtracking procedure that generates ensembles of suboptimal alignments with correct statistical weights.

A comparison between structure based alignments and large samples of stochastic alignments shows that the ensemble contains correct alignments with significant probabilities even though the optimal alignment deviates significantly from the structural alignment. Ensembles of suboptimal alignments obtained by stochastic backtracking can be used as input to any bioinformatics method based on pairwise alignment in order to gain reliability information not available from a single optimal alignment.

Availability: The software described in this contribution is available for downloading at http://www.tbi.univie.ac.at/~ulim/probA/

Contact: ivo@tbi.univie.ac.at

INTRODUCTION

The optimal alignment of two sequences may become susceptible to small perturbations of the scoring parameters if the evolutionary relationship between two sequences becomes more distant (Vingron, 1996). In addition, the dynamic programming algorithms used to derive the 'optimal' alignment have an inherent ambiguity that arises from the non-uniqueness of optimal solutions and the particular scheme by which the search space is evaluated (Giegerich, 2000). As a consequence, the reliability of an alignment may vary considerable along the sequence. Several approaches dealing with this effect have been reported, starting with the investigation of suboptimal alignments by Vingron and Argos (1990) and Saqi and Sternberg (1991). The use of the partition function of all alignments was pioneered by Miyazawa (1994).

In this contribution we introduce a modified alignment algorithm that avoids the generation of solutions that are represented differently but are equivalent from a semantic point of view. Furthermore, we include a parameter governing the relative weight of alignment paths with different scores (Kschischo and Lassig, 2000) and extend previous approaches to stochastic pairwise alignments by a probabilistic backtracking procedure that can be used to obtain ensembles of suboptimal alignments with correct statistical weights.

In the following section we briefly review the theory of probabilistic alignments. Then we describe the stochastic backtracking procedure. A few applications of stochastic alignments are discussed in the Results section. In particular, we compare an ensemble of suboptimal alignments with a 'true' alignment of two proteins that is obtained from purely structural considerations. Finally, we briefly discuss potential further applications of stochastic alignments.

BACKGROUND

Pairwise alignments

We consider two sequences $\mathbf{a} = (a_1 a_2 \ldots a_m)$ and $\mathbf{b} = (b_1 b_2 \ldots b_n)$ taken from an alphabet $\mathfrak{A}$. An alignment $\mathcal{A}$ of $\mathbf{a}$ and $\mathbf{b}$ is the sequence of pairs (a^*_j, b^*_j), $1 \leq j \leq \ell \leq n + m$ such that

(i) $a^*_j, b^*_j \in \mathfrak{A} \cup \{\_\}$, where $\_ \notin \mathfrak{A}$ is the so-called 'gap character',

(ii) $(a^*_j, b^*_j) \neq (\_, \_)$

(iii) There are strictly monotone functions

$j' : \{1, \ldots, m\} \to \{1, \ldots, \ell\}$ and $j'' : \{1, \ldots, n\} \to \{1, \ldots, \ell\}$ such that there is $k \in \{1, \ldots, m\}$ with $j'(k) = j$ whenever $a^*_j \neq \_$ and $l \in \{1, \ldots, n\}$ with $j''(l) = j$ whenever $b^*_j \neq \_$.

Condition (iii) is just a fancy way of expressing the fact that $\mathbf{a}^*$ and $\mathbf{b}^*$ are obtained from $\mathbf{a}$ and $\mathbf{b}$ by inserting gaps without disturbing the linear order of the letters. As a consequence, alignments decompose at all (mis)matches in the following sense: If $a^*_j \neq \_$ and $b^*_j \neq \_$ we may write $(a^*_j, b^*_j) = (a^*_{j'(k)}, b^*_{j''(l)})$ for some $k \leq m$, $l \leq n$, and hence $\mathcal{A}$ is the concatenation of an alignment $\mathcal{A}'$ of the subsequences $\mathbf{a}[1..k-1]$ and $\mathbf{b}[1..l-1]$, the (mis)match (a_k, b_l) and an alignment $\mathcal{A}''$ of the subsequences $\mathbf{a}[k+1..m]$ and $\mathbf{b}[l+1..n]$.

This definition of the pairwise alignment contains one important ambiguity: there is no way to distinguish between

```
A---XXXXB      and    AXXXX---B
AYYY----B             A----YYYB
```

Therefore, we restrict the definition of an alignment by excluding the second alternative. We call an alignment *canonical*, if, whenever gaps in one sequence immediately follow gaps in the other, the gaps in the first sequence $\mathbf{a}$ precede those in the second sequence $\mathbf{b}$. Note that canonical alignments are uniquely determined by their sequence of (mis)matches which in turn is equivalent to the *alignment path* (Durbin *et al.*, 1998; Yu and Hwa, 2001).

Partition function

In the probabilistic interpretation of the sequence alignment problem, see e.g. (Durbin *et al.*, 1998), the score $S(\mathcal{A})$ of an alignment $\mathcal{A}$ is given as a (possibly rescaled) log-odds ratio for obtaining the two aligned sequences from a common ancestor compared to a chance event. In particular, the score $s(a, b)$ of a match (a, a) or mismatch (a, b), $a \neq b$, of two aligned letters is obtained by the log-odds ratio

$$s(a, b) = k \log \frac{f_{ab}}{f_a f_b} \tag{1}$$

where k is an arbitrary positive constant, f_a and f_b are the frequencies of the letters a and b in a prescribed dataset and f_{ab} is the frequency of finding (a, b) in homologous positions. This framework is readily extended to affine gap functions of the form

$$\gamma(l_g) = g_o + g_{\text{ext}}(l_g - 1). \tag{2}$$

The *gap-open penalty* g_o and the *gap-extension* penalty g_{ext} satisfy $g_o \geq g_{\text{ext}}$. Under these assumptions the *alignment score function* $S(\mathcal{A})$ is additive

$$S(\mathcal{A}) = S(\mathcal{A}') + s(a_k, b_l) + S(\mathcal{A}''). \tag{3}$$

This observation is the basis of all dynamic programming algorithms for pairwise alignments (Needleman and Wunsch, 1970). Not surprisingly, it will play a crucial role in our discussion as well. The optimal alignment(s) can be obtained efficiently using the dynamic programming algorithm of Gotoh (1982).

It is not hard to verify that in this framework the probability of a particular alignment $\mathcal{A}$ satisfies

$$\text{Prob}(\mathcal{A}) \propto \exp\left(\frac{S(\mathcal{A})}{k}\right), \tag{4}$$

see e.g. (Yu and Hwa, 2001).

In a thermodynamic interpretation of the alignment problem (Miyazawa, 1994; Kschischo and Lassig, 2000), on the other hand, the score of the alignment $\mathcal{A}$, $S(\mathcal{A})$, is the analogue of (negative) energy. The constant k that depends on the definition of substitution scores corresponds to Boltzmann's constant. In addition, one considers a parameter T that is analogous to the thermodynamic temperature. The *partition function* for the alignment problem is *defined* in the familiar way as

$$Z(T) = \sum_{\mathcal{A}} e^{\frac{S(\mathcal{A})}{kT}} = \sum_{\mathcal{A}} \exp(\beta S(\mathcal{A})), \tag{5}$$

where $\beta = 1/(kT)$. Immediately, we see that

$$\text{Prob}(\mathcal{A}; T) = \frac{1}{Z(T)} \exp(\beta S(\mathcal{A})) \tag{6}$$

where for $T = 1$ we recover the 'true' probability. In the limit $T \to 0$, we have $\text{Prob}(\mathcal{A}; 0) = 0$ for all alignments with a score $S(\mathcal{A})$ less than the maximal score $S_0 = \max S(\mathcal{A})$. In the limit $T \to \infty$, on the other hand, all alignments have the same $\text{Prob}(\mathcal{A}; \infty) = 1/Z(\infty)$, where $Z(\infty) = \binom{m+n}{n}$ is the total number of possible canonical alignments. Thus, the temperature parameter T governs the relative 'importance' of the optimal alignment(s) just as thermodynamic temperature determines the occupation of the ground state (Kschischo and Lassig, 2000). In this sense, we can interpret T as a measure of our interest in suboptimal alignments.

Let $Z_{i,j}$ denote the partition function for the alignments of the subsequences $\mathbf{a}[1..i]$ and $\mathbf{b}[1..j]$. The values $Z_{i,j}$ can be computed by recursions analogous to the ones for the optimal alignment, see e.g. (Miyazawa, 1994; Bucher and Hoffmann, 1996; Yu and Hwa, 2001):

$$\begin{aligned}
Z^M_{i,j} &= \left(Z^M_{i-1,j-1} + Z^E_{i-1,j-1} + Z^F_{i-1,j-1}\right) e^{\beta s(a_i, b_j)} \\
Z^E_{i,j} &= Z^M_{i,j-1} e^{\beta g_o} + Z^E_{i,j-1} e^{\beta g_{\text{ext}}} \\
Z^F_{i,j} &= \left(Z^M_{i-1,j} + Z^E_{i-1,j}\right) e^{\beta g_o} + Z^F_{i-1,j} e^{\beta g_{\text{ext}}} \\
Z_{i,j} &= Z^M_{i,j} + Z^E_{i,j} + Z^F_{i,j}
\end{aligned} \tag{7}$$

The matrix $Z^M_{i,j}$ contains the partition function over all alignments that end in a (mis)match (a_i, b_j). Similarly, $Z^E_{i,j}$ contains the partition function over all alignments in which residue b_j is aligned to a gap (i.e. all alignments ending with a gap in sequence **a**) and $Z^F_{i,j}$ describes alignments ending with a gap in **b**. The boundary conditions are: $Z^M_{0,0} = Z^E_{0,0} = Z^F_{0,0} = 1$,
$Z^E_{0,1} = e^{\beta g_o}$, $Z^E_{0,j} = e^{\beta(g_o+(j-1)g_{\text{ext}})}$ for $j > 1$,
$Z^F_{1,0} = e^{\beta g_o}$, $Z^F_{i,0} = e^{\beta(g_o+(i-1)g_{\text{ext}})}$ for $i > 1$.
The values $Z^M_{i,0}$, $Z^M_{0,j}$, $Z^E_{i,0}$, and $Z^F_{0,j}$ for $i \geq 1$ and $j \geq 1$ may remain undefined. Note that the recursion for $Z^F_{i,j}$ differs from the 'usual' form in order to account for the asymmetric definition of canonical alignments.

Match probabilities

Using the partition function, the probability of each match (i, j) between the two sequences can be calculated (Miyazawa, 1994). A related approach based on Bayesian inference is discussed in (Zhu *et al.*, 1998). We define a class Ω of alignments that meet certain criteria. The probability to find an alignment that belongs to the class Ω is

$$\text{Prob}(\Omega) = \frac{1}{Z} \sum_{\mathcal{A} \in \Omega} e^{\beta S(\mathcal{A})} = \frac{Z(\Omega)}{Z} \tag{8}$$

The probability that i and j are matched is therefore given by $P_{ij} = \text{Prob}(\Omega_{i,j})$ where $\Omega_{i,j}$ is the class of alignments in which a_i is matched to b_j. For each $\mathcal{A} \in \Omega_{i,j}$ the score of the whole alignment $\mathcal{A}$ is the sum

$$S(\mathcal{A}) = S(\mathcal{A}^{i,j}_{1,1}) + S(\mathcal{A}^{m,n}_{i,j}) - s(a_i, b_j) \tag{9}$$

where $S(\mathcal{A}^{i,j}_{1,1})$ and $S(\mathcal{A}^{m,n}_{i,j})$ are the scores of the partial alignments. The probability of a (mis)match (a_i, b_j) is now obtained from

$$\begin{aligned} Z(\Omega_{i,j}) &= \sum_{\mathcal{A} \in \Omega_{i,j}} e^{\beta S(\mathcal{A}^{i,j}_{1,1}) + \beta S(\mathcal{A}^{m,n}_{i,j}) - \beta s(a_i, b_j)} \\ &= e^{-\beta\, s(a_i, b_j)} \times \underbrace{\sum_{\mathcal{A} \in \mathfrak{A}^{i,j}_{1,1}} e^{\beta S(\mathcal{A}^{i,j}_{1,1})}}_{Z^{\text{M}}_{ij}} \times \underbrace{\sum_{\mathcal{A} \in \mathfrak{A}^{m,n}_{i,j}} e^{\beta S(\mathcal{A}^{m,n}_{i,j})}}_{\widehat{Z}^{\text{M}}_{ij}} \end{aligned} \tag{10}$$

where $Z^{\text{M}}_{ij} = Z(\mathfrak{A}^{i,j}_{1,1})$ is the partition function of the set $\mathfrak{A}^{i,j}_{1,1}$ of all alignments of the partial sequences $\mathbf{a}[1..i]$ and $\mathbf{b}[1..j]$ that end with a (mis)match of (a_i, b_j). Analogously, $\widehat{Z}^{\text{M}}_{ij} = Z(\mathfrak{A}^{i,j}_{m,n})$ is the partition function of the set $\mathfrak{A}^{i,j}_{m,n}$ of all alignments of the partial sequences $\mathbf{a}[i..m]$ and $\mathbf{b}[j..n]$ that begin with a (mis)match (a_i, b_j).

Vingron and Argos (1990) observed that (3) can be utilized to determine the score of an optimal alignment that contains the (mis)match (a_k, b_l) by computing an *optimal* alignment of the sub-sequences $\mathbf{a}[1..k-1]$ and $\mathbf{b}[1..l-1]$ by 'forward' recursion and an optimal alignment of the subsequences $\mathbf{a}[k+1..m]$ and $\mathbf{b}[l+1..n]$ by a 'backward' recursion in which the sequences are simply read in the opposite direction. The same approach can be exploited to compute the matrices $\widehat{Z}^M$, $\widehat{Z}^F$, and $\widehat{Z}^E$ by means of backward recursions that are analogous to the forward recursions in Equation (7). Thus, we obtain the match probabilities as follows:

$$P_{ij} = \frac{Z^{\text{M}}_{ij} \widehat{Z}^{\text{M}}_{ij}}{Z} \exp(-\beta s(a_i, b_j)) \tag{11}$$

Equation (11) was first derived by Miyazawa (1994) who then proceeds with a greedy method to extract a single 'locally most probable' alignment from the match probability matrix (P_{ij}).

Match probabilities can be conveniently visualized in dot plots where each possible match is represented by a box with area P_{ij}, plotted on a rectangular $m \times n$ grid indexed by i and j. Such dot plots provide an excellent overview of likely alignment alternatives. An example in the context of local alignments is shown below, see Figure 1.

Local alignments

The partition function version of the local alignment problem *sensu Smith–Waterman* considers the ensemble of all alignments of the sub-sequence $\mathbf{a}[i..k]$ with the sub-sequence $\mathbf{b}[j..l]$. The corresponding recursions are

$$Z^M_{i,j} = \left(Z^M_{i-1,j-1} + Z^E_{i-1,j-1} + Z^F_{i-1,j-1} + 1\right) e^{\beta s(a_i, b_j)} \tag{12}$$

where the additional term 1 has to be included to describe the case that (i, j) is the first pair in the aligned part of the sequences. The recursions for Z^F and Z^E are the same as in Equation (7). The boundary conditions are $Z^M_{i,0} = Z^M_{0,j} = Z^E_{i,0} = Z^E_{0,j} = Z^F_{i,0} = Z^F_{0,j} = 1$. The analogue of this recursion for the case of linear gap functions, $g_o = g_{\text{ext}}$ is discussed in Kschischo and Lassig (2000).

The important difference to the global alignment problem is that now the only valid alignments are those that *end* with a (mis)match (k, l). The 'empty alignment', i.e., the case in which **a** and **b** are totally unrelated is also a valid local alignment, thus

$$Z = 1 + \sum_{k,l} Z^M_{k,l}. \tag{13}$$

The pairing probabilities for this model are again obtained from Equation (11), see Figure 1 for an example.

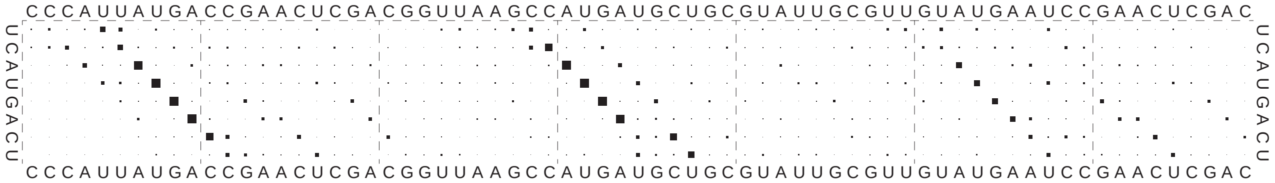

Fig. 1. Example of a local alignment. The parameters are $s(a, a) = 3$, $s(a, b) = -2$ if $a \neq b$, $g_o = -3$, and $g_{\text{ext}} = -2$, $\beta = 0.5$.

STOCHASTIC BACKTRACKING

While the matrix of match probabilities provides an accurate description of the ensemble of possible alignments, most sequence analysis methods require alignments as input. In such cases, one would like to represent the ensemble by a properly weighted sample of stochastic pairwise alignments. Such a sample can indeed be generated from the partition function matrices by a stochastic version of the backtracking procedure, as shown below.

The stochastic backtracking algorithm, just like ordinary backtracking, constructs an alignment path starting at the final position, (m, n), of the Z matrices. At each step the path is extended by a match, a gap in sequence **a**, or a gap in sequence **b**. The probability for each choice can be calculated from the Z matrices as follows.

In the first step, one of the three possible states of the alignment is selected depending on a random number r, $0 \leq r < 1$: Residue a_m is matched to residue b_n if $r < p(\text{match}) = Z^M_{m,n}/Z_{m,n}$, a gap is introduced in sequence **a** with probability $p(\text{gap in } \mathbf{a}) = Z^E_{m,n}/Z_{m,n}$ if the random number r satisfies $p(\text{match}) \leq r < p(\text{match}) + p(\text{gap in } \mathbf{a})$, otherwise a gap is inserted in sequence **b**.

In the following steps of the stochastic backtracking, the probability of each state is dependent on the previous choice. If the previous state of the alignment was a gap in sequence **a**, there are only two possibilities to extend the alignment, either return to the match state or to add another gap in sequence **a**. The probability to introduce a gap in sequence **b**, $p(\text{gap in } \mathbf{b})$ is zero, because the algorithm is designed to arrange gaps in order gaps in **a** $\Rightarrow$ gaps in **b**. The complete set of transition probabilities is listed in Table 1.

At each step of the backtracking process the selection of the next alignment state is done stochastically, so that an alignment $\mathcal{A}$ is generated with probability $\exp(\beta S(\mathcal{A}))/Z$. Repeated application of this procedure yields an equilibrium sample of alignments.

Stochastic backtracking can be applied to the local alignment problem as well. The difference is the need to choose start and end points of the local alignment fairly. We assume the local alignment starts and ends with a match. The probability of choosing i, j as start point is then given by $Z^M_{i,j}/Z$. We then proceed as shown above, except that after each match i, j the alignment terminates with probability $e^{\beta s(a_i,b_j)}/Z^M_{i,j}$.

RESULTS

Well-definedness of the alignment

Stochastic backtracking provides an ensemble of stochastic alignments, which are distributed according to the probability of each alignment. If the alignment is well defined, the ensemble will be dominated by the optimal, most likely, alignment. A simple, entropy-like measure for the diversity of alignments is thus the probability of the optimal alignment, or equivalently the difference between the score of the optimal alignment $S(\mathcal{A}_{\text{opt}})$ and the analogue of the free energy of the ensemble

$$\Delta S^{\text{ensemble}} = \beta S(\mathcal{A}_{\text{opt}}) - \ln Z = \ln \text{Prob}(\mathcal{A}_{opt}) \qquad (14)$$

In order to quantify the importance of suboptimal alignments we computed this entropy measure for random nucleic acid and protein sequences of different length as well as several examples of real RNA sequences.

The entropy distribution of different pairwise alignments of a set of functionally identical nucleic acids provides a good measure of the relatedness of the members of this set. The RNAse P RNA sequences, for example, comprise a set of functionally identical sequences with no conservation at the sequence level. The entropies of pairwise alignments between different RNAse P RNA sequences correspond to entropies of random nucleic acid sequences. In the case of 16S RNA, on the other hand, the relationship on the sequence level is significantly higher (data not shown).

Comparison with structure-based alignments

As a demonstration of the program we consider the alignment between leghaemoglobin from yellow lupin, *Lupinus luteus*, PDB entry 1GDJ (Harutyunyan *et al.*, 1995), and chain A of human deoxyhaemoglobin, PDB entry 2HHB_A (Fermi *et al.*, 1984). The proteins 1GDJ and 2HHB_A are dissimilar in sequence (pairwise identity 14%), but have quite similar structures. Therefore we can use an alignment of their 3D structures as the standard to which purely sequence-based alignments have to be

Table 1. Complete set of transition probabilities. The probability of each state depends on the previous choice

next state	next indices		previous state			
	$i \leftarrow$	$j \leftarrow$	start	matched	gap in sequence **a**	gap in sequence **b**
match	$i-1$	$j-1$	$\frac{Z^M_{m,n}}{Z_{m,n}}$	$\frac{Z^M_{i-1,j-1}}{Z_{i,j}} \quad e^{\beta s(a_i,b_j)}$	$\frac{Z^M_{i,j-1}}{Z^E_{i,j}} \quad e^{\beta g_o}$	$\frac{Z^M_{i-1,j}}{Z^F_{i,j}} \quad e^{\beta g_o}$
gap in sequence **a**	i	$j-1$	$\frac{Z^E_{m,n}}{Z_{m,n}}$	$\frac{Z^E_{i-1,j-1}}{Z_{i,j}} \quad e^{\beta s(a_i,b_j)}$	$\frac{Z^E_{i,j-1}}{Z^E_{i,j}} \quad e^{\beta g_\mathrm{ext}}$	$\frac{Z^E_{i-1,j}}{Z^F_{i,j}} \quad e^{\beta g_o}$
gap in sequence **b**	$i-1$	j	$\frac{Z^F_{m,n}}{Z_{m,n}}$	$\frac{Z^F_{i-1,j-1}}{Z_{i,j}} \quad e^{\beta s(a_i,b_j)}$	0	$\frac{Z^F_{i,j-1}}{Z^F_{i,j}} \quad e^{\beta g_\mathrm{ext}}$

compared. For the analysis of the stochastic ensemble one million stochastic alignments between 1GDJ and 2HHB_A were generated.

The global 3D alignment of two proteins has been characterized as NP hard (Lathrop, 1994), thus one has to rely on heuristics to find good solutions. To reduce the influence of the particular heuristic used, we employed different web-accessible structure alignment programs to extract reliably aligned regions. The underlying assumption is that a region which is identically aligned by various methods is, in fact, reliably aligned.

The programs we used are based on different principles: Combinatorial Extension, CE[†] (Shindyalov and Bourne, 1998), uses similarity in local geometry of C_α; TOP[‡] (Lu, 2000, 1996) and COMPARER[§] (Sali and Blundell, 1990), both utilize topological features for the computation of 3D structure alignments; SARF2[¶] (Alexandrov and Fischer, 1996) and MATRAS[‖] (Kawabata and Nishikawa, 2000), employ secondary structure information to generate structure alignments.

The upper part of Figure 2 shows the 3D structure alignments computed by the specified on-line programs. The 3D alignments display three regions which can be aligned without ambiguity. A region is accepted as reliably aligned if all structure alignment methods agree on the alignment of this region. About 65% of the consensus structure alignment can be classified as reliably aligned. These positions exhibit significantly higher match probabilities than segments for which no consistent structure alignment exists. Furthermore, at least one of this reliable aligned segments is found in the vast majority of stochastic alignments as can be seen from the lower panels in Figure 2.

The correspondence between reliably aligned residues and matches with higher match probabilities is not absolute. Whereas nearly all matches with $P \geq 0.7$ are part of reliably aligned regions, matches with lower match probabilities are found both inside and outside the consensus region, see Figure 2. The optimal alignment, on the other hand, which includes per definition the highest possible number of matches with high match probabilities, contains only about 87% of the positions that are classified as reliably aligned in the structure consensus. On the other hand, some 3% of the suboptimal alignments include all of the reliable aligned positions of the structure consensus. The retrieval of a biological correct alignment is therefore dependent upon the inclusion of the information provided by the suboptimal alignments.

As can be seen in the lower panel of Figure 2, correctness increases with the alignment score. However, stochastic backtracking finds a substantial number of suboptimal alignments that are more correct than the optimal alignment.

We tested our method on two other pairs of structurally similar alignments. In this case we measured the correctness of the optimal alignment and a set of 1 million stochastic alignments by comparison with the structural alignment returned by CE. In this case 'correctness' is the percentage of matches from the reference alignment present in the stochastic alignment. The results are summarized in the table below, data are for $T = 1$.

PDB entry	length	%corr		% at least as
		optimal	stochastic	good as opt
1YCC/1CTJ	108/89	51.3	50.3	41.7
4HHB/1GDJ	141/153	67.2	63.5	38.9
1SAC/1C4R	204/182	44.4	38.4	21.3

As can be seen from the data, the quality of the average stochastic alignments is almost as good as the optimal alignment. As shown in the last column a significant fraction of stochastic alignments is as correct or better than the optimal alignment.

[†] http://cl.sdsc.edu/ce.html

[‡] http://bioinfo1.mbfys.lu.se/TOP/webtop.html

[§] http://www-cryst.bioc.cam.ac.uk/∼robert/cpgs/COMPARER/comparer.html

[¶] http://123d.ncifcrf.gov/run2.html

[‖] http://bongo.lab.nig.ac.jp/∼takawaba/Matras.html

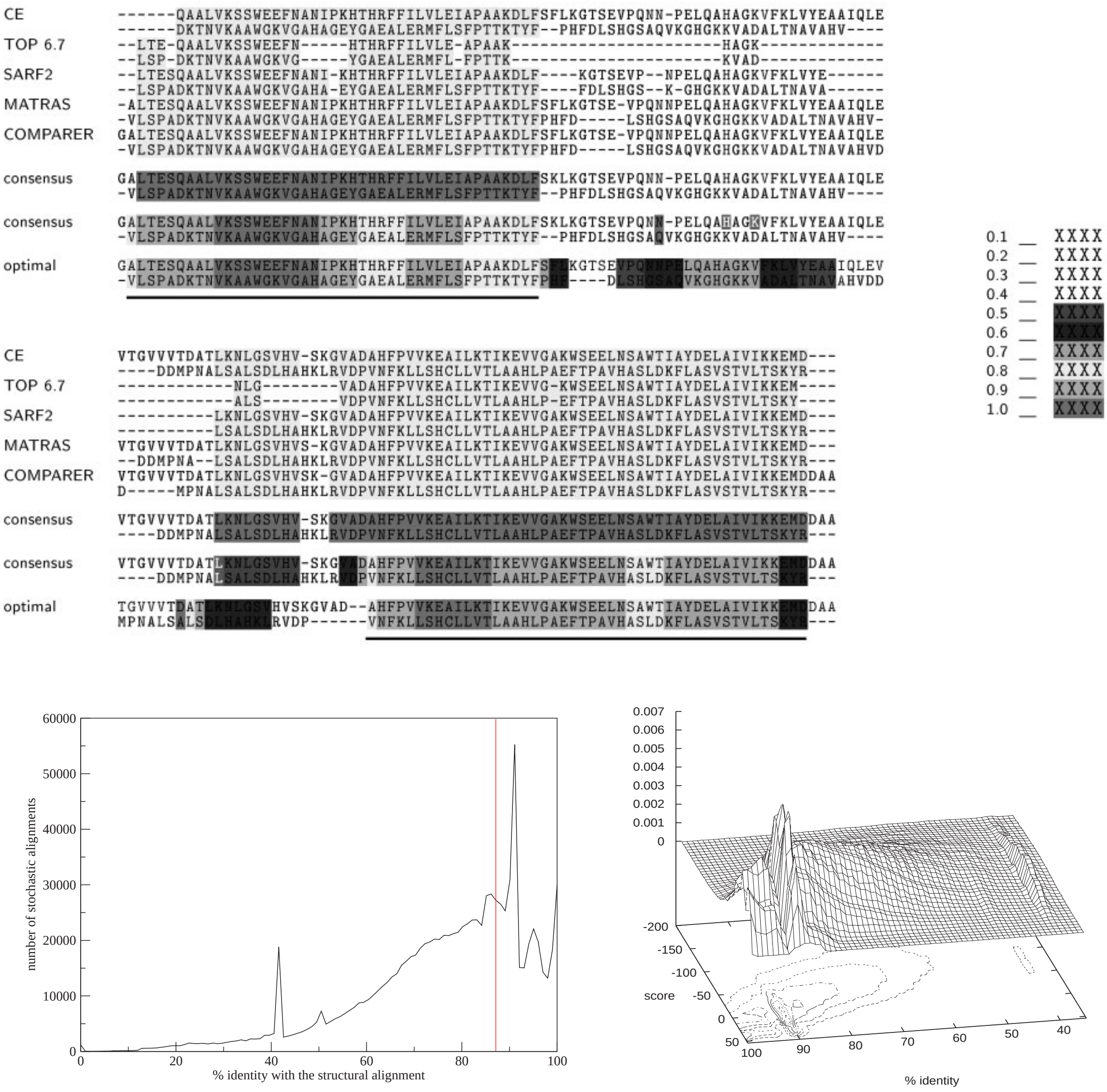

Fig. 2. Upper panel: Reliably aligned region were extracted from 5 pairwise structure alignments computed with different on-line 3D structure alignment programs. For each pairwise alignment the upper sequence is 1GDJ, the lower 2HHB_A. The structure alignment program used is indicated in the first column. The second column displays the structure alignments, reliably aligned regions are indicated in yellow. In the consensus, which directly follows the different 3D structure alignments reliably aligned regions are colored red. The alignment in the regions were no consistent alignment between the different structure alignment exists is like in the structure alignment generated with CE.

Match probabilities in the 3D structure alignments and in the optimal alignment are indicated below in different colors. The color code for match probabilities is specified on the right side of the figure. The figure was prepared using the Texshade package http://homepages.uni-tuebingen.de/beitz/txe.html. Regions with high match probabilities that are included in the optimal alignment as well as in the structure alignment consensus are indicated by a black line below the alignments.

Lower Panels: L.h.s.: Distribution of the fraction of stochastic alignments with different fractions of matches with the structure alignment. The optimal alignment is indicated by the red line. The three sharp peaks in the curve correspond to stochastic alignments containing one or more of the reliably aligned regions shown in figure 2. The first sharp peak at 41.6% corresponds to alignments that include the N-terminal block; the second peak indicates alignments that have both of the longer reliable sections at the N- and C-termini, while the third sharp peak at 100% corresponds to the correct alignments.

R.h.s.: Joint probability of finding an alignment with given fraction of matches in the structural consensus and a given alignment score.

The stochastic alignments used the Gonnet matrix with a PAM distance of 300 and the default temperature $T = 1$.

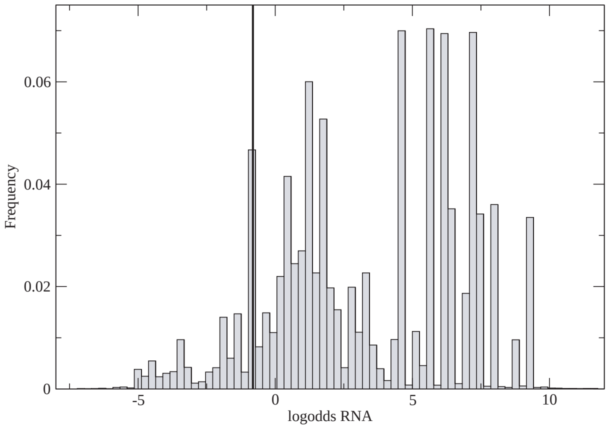

Fig. 3. Distribution of log-odds scores for ‘structural RNA’ as computed by `qrna` for a sample of 40 000 stochastic alignments of the TAR sequence from HIVMAL and HIVELI. Alignments with positive log-odds score (about 84%) are classified as structural RNA (the correct classification). The optimal alignment chosen by our program has a score of -0.814 (indicated by the vertical line) and is thus mis-qualified. Parameters used were $s(a, a) = 1$, $s(a, b) = 0$ if $a \neq b$, $g_o = -3$, and $g_{\text{ext}} = -1$, $\beta = 4/3$.

Using stochastic backtracking with other methods

Pairwise alignments are used ubiquitously as input for other bioinformatics methods. Usually, a single optimal alignment is used as input and thus there is no way to estimate the sensitivity of the method with respect to alignment errors and ambiguities. Stochastic backtracking allows to run the method repeatedly with different, yet plausible, alignments and thus generate statistically sound results.

As a simple example we use Rivas and Eddy’s `qrna` program that, given a pairwise alignment, classifies the aligned sequences as either coding sequences, structural RNA with common structure, or neither (Rivas and Eddy, 2001).

HIV viruses carry a prominent secondary structure motif, the so-called TAR hairpin, which occurs in the first 60 nucleotides of the 5′-UTR of their genomic RNA (see e.g. Huynen and Konings (1998) and the references therein). We used the first 60nt of the HIV-1 sequences HIVMAL and HIVELI and generated 40 000 pairwise alignments as input for `qrna`. While the optimal alignment picked by our program (one out of 17 with optimal score) was mis-qualified as ‘other’, 84% of the stochastic alignments were correctly classified as ‘structural RNA’, see Figure 3. The use of stochastic alignments allows to obtain both a more reliable average result and an error estimate.

DISCUSSION

The stochastic version of Gotoh’s pairwise sequence alignment algorithm described in this contribution computes the probability of each possible match in the alignment. Thus, it provides an internal measure of an alignments reliability not only globally but also locally. In addition, we present an algorithm that produces correctly weighted samples of alignments by means of stochastic backtracking. The algorithms have been implemented in `C` and are available in the software package `probA` that can be downloaded from the internet[**].

The stochastic pairwise alignments are useful in many different contexts: Numerous tools in bioinformatics require pairwise sequence alignments as input data. The program `probA` 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.

A comparison between structure based alignments and large samples of stochastic alignments shows that the ensemble contains correct alignments with significant probabilities even though the optimal alignment deviates significantly from the structural alignment. Such deviations occur even in those regions where the structural alignment appears to be very reliable.

This observation indicates that iterative multiple alignment programs are likely to be trapped in optimal pairwise alignments that may differ considerably from the true alignment. Thus, it will be desirable to develop multiple alignment tools that are explicitly based on either the match probability matrices P of the pairwise alignments or that use ensembles of pairwise alignments. It is important to notice, however, that the restriction to canonical alignments is inappropriate in a multiple alignment context. Considering the two situations

```
A---XXXXB      and    AXXXX---B
AYYY----B             A----YYYB
CXXXXYYYC             CXXXXYYYC
```

which correspond to the same canonical alignment of the first two sequences we see that only the second alternative, which is the one excluded by our definition of the canonical alignments, can be extended to the correct alignment of all three sequences.

ACKNOWLEDGMENTS

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P-13545-MAT.

[**] http://www.tbi.univie.ac.at/∼ulim/probA

REFERENCES

Alexandrov,N.N. and Fischer,D. (1996) Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins: Struct. Funct. Genet.*, **25**, 354–365.

Bucher,P. and Hoffmann,K. (1996) 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. and Smith,R.F. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96)*. AAAI Press, Menlo Park, CA, pp. 44–50.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Fermi,G., Perutz,M., Shaanan,B. and Fourme,R. (1984) The crystal structure of human deoxyhaemoglobin at 1.74å resolution. *J. Mol. Biol.*, **175**, 159.

Giegerich,R. (2000) Explaining and controlling ambiguity in dynamic programming. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, Technical report, Faculty of Technology, Bielefeld University, Bielefeld.

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

Harutyunyan,E.H., Safonova,T.N., Kuranova,I.P., Popov,A.N., Teplyakov,A.V., Obmolova,G.V., Rusakov,A.A., Vainshtein,B.K., Dodson,G.G., Wilson,J.C. and Perutz,M.F. (1995) The structure of deoxy- and oxy-leghaemoglobin from lupin. *J. Mol. Biol.*, **251**, 104–115.

Huynen,M. and Konings,D. (1998) Questions about RNA structures in HIV and HPV. In Myers,G.L. (ed.), *Viral Regulatory Structures and Their Degeneracy*, Vol. XXVIII of Santa Fe Institute Studies in the Sciences of Complexity, Addison Wesley Longman, Reading, MA, pp. 69–82.

Kawabata,T. and Nishikawa,K. (2000) Protein structure comparison using the Markov transition model of evolution. *Protein Struct.*, **41**, 108–122.

Kschischo,M. and Lassig,M. (2000) Finite-temperature sequence alignment. *Pac. Symp. Biocomput.*, **1**, 624–635.

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

Lu,G. (1996) A WWW service system for automatic comparison of protein structures. *Protein Data Bank Quarterly Newsletter*, **78**, 10–11.

Lu,G. (2000) A new method for protein structure comparisons and similarity searches. *J. Appl. Cryst.*, **33**, 176–183.

Miyazawa,S. (1994) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, **8**, 999–1009.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, pp. 19.

Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.

Saqi,M.A. and Sternberg,M.J. (1991) A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.*, **219**, 727–732.

Shindyalov,I. and Bourne,P. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Vingron,M. (1996) Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.*, **6**, 346–352.

Vingron,M. and Argos,P. (1990) Determination of reliable regions in protein sequence alignments. *Protein Eng.*, **3**, 565–569.

Yu,Y.K. and Hwa,T. (2001) Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *J. Comp. Biol.*, **8**, 249–282.

Zhu,J., Liu,J.S. and Lawrence,C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.