Stochastic Pairwise Alignments

Ulrike Mückstein[a], Ivo L. Hofacker[a,*], and Peter F. Stadler[a,b]

[a]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien,
Währingerstraße 17, A-1090 Vienna, Austria
Email: {ulim,ivo,studla}tbi.univie.ac.at,
[b]The Santa Fe Institute, Santa Fe, New Mexico, USA.
[*]Author for correspondence: Phone: ++43 1 4277 52738, Fax: ++43 1 4277 52793

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.
Results: The probability P_{ij} of a match between position i in the first and position j in the second sequence is computed using the the partition function over all canonical pairwise alignments. A probabilistic backtracking procedure can then be used to generate 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, or the match probability matrices themselves, are therefore promising starting points for improved iterative multiple alignment procedures. In particular, it should be possible to overcome the problem of fixating an incorrect pairwise alignment in an early iteration.
Availability: The software described in this contribution is available for downloading at `http://www.tbi.univie.ac.at/~ulim/probA/`
Contact: Ivo L. Hofacker,
Tel: ++43 1 4277 52738, Fax: ++43 1 4277 52793,
Email: ivotbi.univie.ac.at

1 INTRODUCTION

The optimal alignment of two sequences may become susceptible to small perturbations of the scoring parameters when 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 to 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. In section 3.2 we compare an ensemble of suboptimal alignments with a "true" alignment of two proteins that is obtained based on purely structural considerations. Finally we briefly discuss potential further applications of stochastic alignments.

2 THEORY

2.1 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 $j' : \{1, \dots, m\} \to \{1, \dots, \ell\}$ and $j'' : \{1, \dots, n\} \to \{1, \dots, \ell\}$ such that there is $k \in \{1, \dots, m\}$ with $j'(k) = j$ whenever $a^*_j \neq \_$ and $l \in \{1, \dots, n\}$ 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 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 further by excluding the second alternative. We say that an alignment is *canonical*, if, between two consecutive matches we always first have the gaps (if any) in the first sequence $\mathbf{a}$ and then 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).

2.2 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 $g_o > g_{\text{ext}}$. 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 algorithm 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

$$\text{Prob}(\mathcal{A}) \propto e^{\frac{S(\mathcal{A})}{k}}\,, \tag{4}$$

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 usual way as

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

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

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

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. The temperature parameter T thus governs the relative "importance" of the optimal alignment(s) just as thermodynamic temperature determines the occupation of the ground state. 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):

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

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.

2.3 Match Probabilities

Using the partition function, the probability of each match (i, j) between the two sequences can be calculated. 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

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

The probability that i and j are matched is therefore given by

$$P_{ij} = \text{Prob}(\Omega_{i,j}) \qquad (9)$$

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

$$S(\mathcal{A}) = S(\mathcal{A}^{i,j}_{1,1}) + S(\mathcal{A}^{m,n}_{i,j}) - s(a_i, b_j) \qquad (10)$$

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 an (a_i, b_j) (mis)match is now obtained from

$$\begin{aligned} Z(\Omega_{i,j}) &= \sum_{\mathcal{A}\in\Omega_{i,j}} e^{\beta S(\mathcal{A}^{i,j}_{1,1}) + \beta S(\mathcal{A}^{m,n}_{i,j}) - \beta s(a_i,b_j)} \\ &= e^{-\beta\, s(a_i,b_j)} \times \\ &\qquad \underbrace{\sum_{\mathcal{A}\in\mathfrak{A}^{i,j}_{1,1}} e^{\beta S(\mathcal{A}^{i,j}_{1,1})}}_{Z^{\mathrm{M}}_{ij}} \times \underbrace{\sum_{\mathcal{A}\in\mathfrak{A}^{m,n}_{i,j}} e^{\beta S(\mathcal{A}^{m,n}_{i,j})}}_{\widehat{Z}^{\mathrm{M}}_{ij}} \qquad (11) \\ &= Z^{\mathrm{M}}_{ij} \widehat{Z}^{\mathrm{M}}_{ij} e^{-\beta s(a_i,b_j)} \end{aligned}$$

where $Z^{\mathrm{M}}_{ij} = Z(\mathfrak{A}^{i,j}_{1,1})$ 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 $\widehat{Z}^{\mathrm{M}}_{ij} = Z(\mathfrak{A}^{i,j}_{m,n})$ 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).

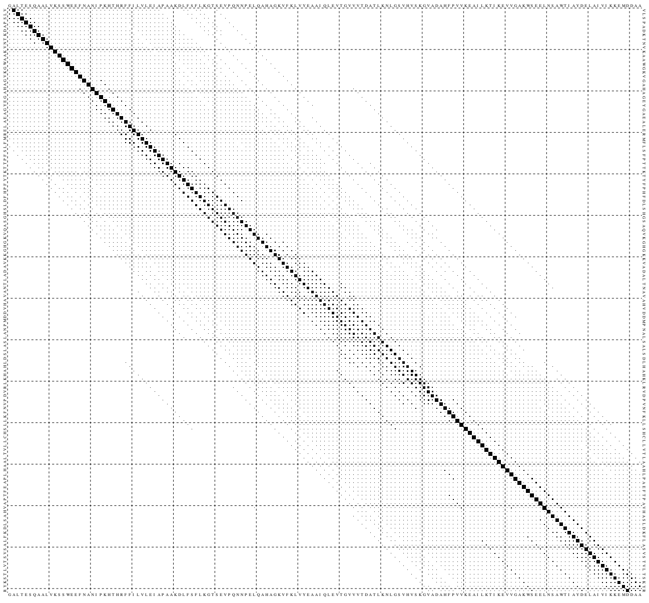


Figure 1: Dot plot of the alignment between leghaemoglobin from yellow lupin, *Lupinus luteus*, and chain A of human deoxyhaemoglobin. The horizontal axis of the dot plot is labeled by *Lupinus luteus* leghaemoglobin, the vertical axes by chain A of human deoxyhaemoglobin. Regions of high sequence similarity are illustrated by large black dots representing high match probabilities. In regions of lower sequence similarity many different possibilities to align the two sequences exits. The different possibilities to align these regions are depicted by a multitude of small dots, each representing a match of low probability. The alignment was prepared using the score matrix Gonnet 120, $T = 0.4$.

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 equ.(7). Thus we obtain the match probabilities as follows:

$$P_{ij} = \frac{Z^{\mathrm{M}}_{ij} \widehat{Z}^{\mathrm{M}}_{ij}}{Z} e^{-\beta s(a_i,b_j)} \qquad (12)$$

Equ.(12) 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, see figure 1.

2.4 Stochastic Backtracking

For every position (i, j) of an alignment the probability for matching residues a_i and b_j, for introducing a gap in sequence **a** and for introducing a gap in sequence **b** can be calculated. The stochastic backtracking algorithm just like ordinary backtracking starts at final positions, (m, n), of the Z matrix.

The probabilities of the three possible alignment states are obtained by the following reasoning: The probability of the match (a_m, b_n) is simply the fraction

$$\begin{array}{llll} \text{match} & p = \dfrac{Z^M_{m,n}}{Z_{m,n}} & i \leftarrow m-1 & j \leftarrow n-1 \\ \text{gap in } \mathbf{a} & p = \dfrac{Z^E_{m,n}}{Z_{m,n}} & i \leftarrow m & j \leftarrow n-1 \\ \text{gap in } \mathbf{b} & p = \dfrac{Z^F_{m,n}}{Z_{m,n}} & i \leftarrow m-1 & j \leftarrow n \end{array}$$

One of the three possible states of the alignment is selected in the following way depending on a random number r, $0 \le r < 1$: Residue a_m is matched to residue b_n if $r < p(\text{match})$, a gap is introduced in sequence **a**, if the random number is $p(\text{match}) \le r < (p(\text{match}) + p(\text{gap in } \mathbf{a}))$, 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 positions were matched in the previous step we have the probabilities

$$\begin{array}{llll} & & i \leftarrow & j \leftarrow \\ \text{match} & p = \dfrac{Z^M_{i-1,j-1} e^{\beta s(a_i,b_j)}}{Z_{i,j}} & i-1 & j-1 \\ \text{gap in } \mathbf{a} & p = \dfrac{Z^E_{i-1,j-1} e^{\beta s(a_i,b_j)}}{Z_{i,j}} & i & j-1 \\ \text{gap in } \mathbf{b} & p = \dfrac{Z^F_{i-1,j-1} e^{\beta s(a_i,b_j)}}{Z_{i,j}} & i-1 & j \end{array}$$

Here $Z_{i,j} = (Z^M_{i-1,j-1} + Z^E_{i-1,j-1} + Z^F_{i-1,j-1}) e^{\beta s(a_i,b_j)}$. 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 $\mathbf{a} \Rightarrow$ gaps in **b**. Thus

$$\begin{array}{llll} & & i \leftarrow & j \leftarrow \\ \text{match} & p = \dfrac{Z^M_{i,j-1} e^{\beta g_o}}{Z^E_{i,j}} & i-1 & j-1 \\ \text{gap in } \mathbf{a} & p = \dfrac{Z^E_{i,j-1} e^{\beta g_{\text{ext}}}}{Z^E_{i,j}} & i & j-1 \end{array}$$

where $Z^E_{i,j} = Z^M_{i,j-1} e^{\beta g_o} + Z^E_{i,j-1} e^{\beta g_{\text{ext}}}$ is the sum over all alignments up to position (i, j) that end with a gap in sequence **a**.

In the case that the previous state of the alignment was a gap in sequence **b**, three possible states to continue the alignment are available: return to the match state, switch to a gap in sequence **a**, or continue the alignment with another gap in sequence **b**. Thus

$$\begin{array}{llll} & & i \leftarrow & j \leftarrow \\ \text{match} & p = \dfrac{Z^M_{i-1,j} e^{\beta g_o}}{Z^F_{i,j}} & i-1 & j-1 \\ \text{gap in } \mathbf{a} & p = \dfrac{Z^E_{i-1,j} e^{\beta g_o}}{Z^F_{i,j}} & i & j-1 \\ \text{gap in } \mathbf{b} & p = \dfrac{Z^F_{i,j-1} e^{\beta g_{\text{ext}}}}{Z^F_{i,j}} & i-1 & j \end{array}$$

where $Z^F_{i,j} = (Z^M_{i-1,j} + Z^E_{i-1,j}) e^{\beta g_o} + Z^F_{i-1,j} e^{\beta g_{\text{ext}}}$ is the sum of all possible alignments up to position (i, j) that end with a gap in sequence **b**.

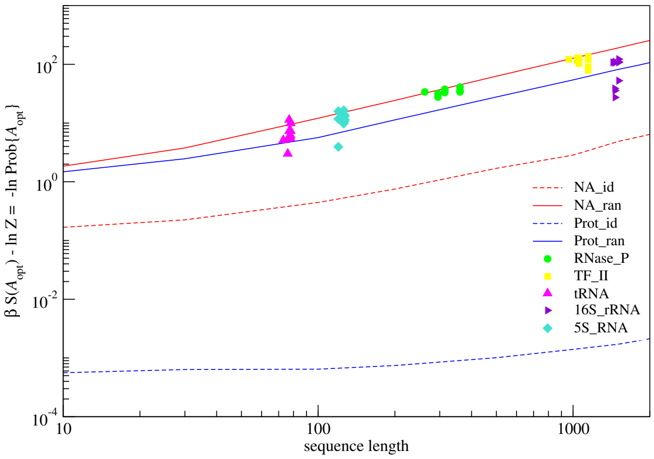


Figure 2: Well-definedness of an ensemble of stochastic alignments, measured by the probability of the optimal alignment. The blue lines correspond to the mean log-probabilities of protein sequences, the red curves are for nucleic acids. In each case, the upper (full) curve refers to alignments of random sequences, while the lower (dotted) curve shows the entropy values for alignments of two identical sequences. The symbols indicate different classes of biological RNA sequences.

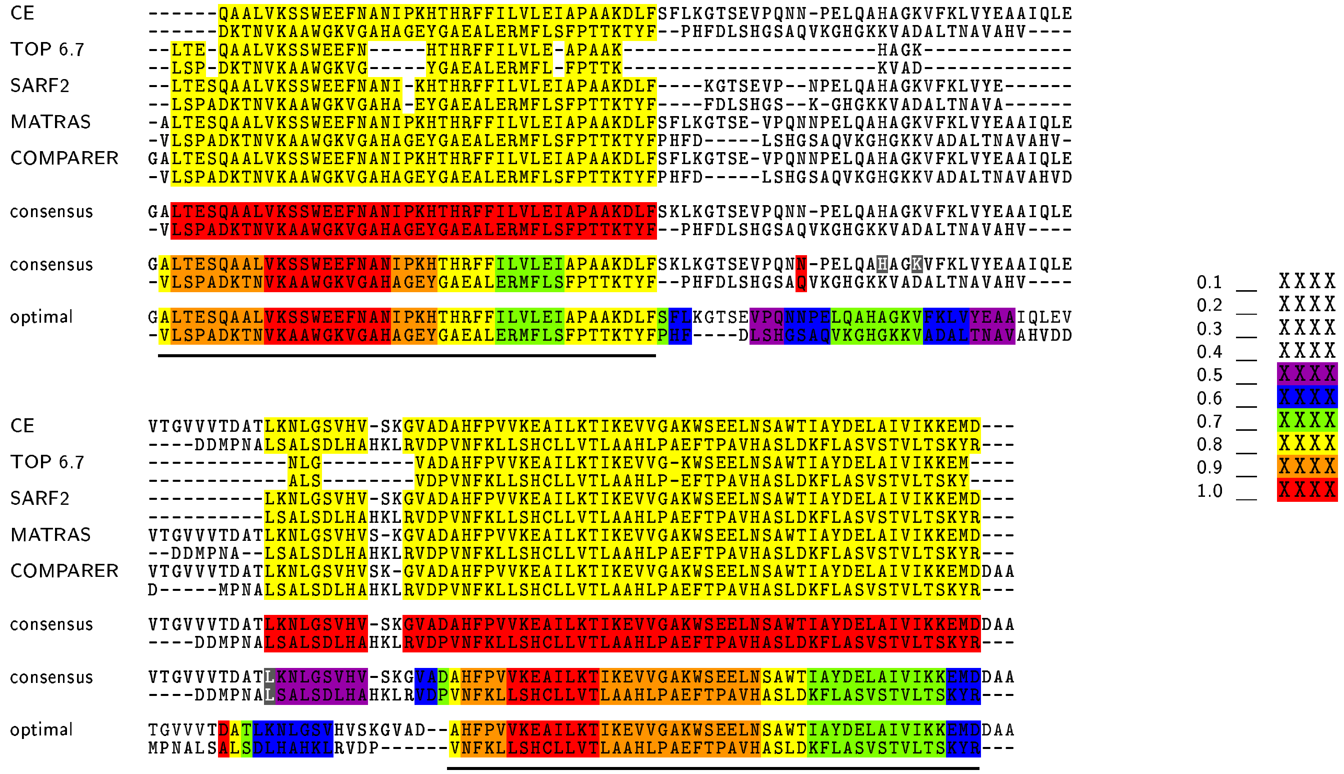
At each step of the backtracking process the selection of the next alignment states is done stochastically. Repeated application of this procedure yields an equilibrium sample of alignments.

3 RESULTS

3.1 Alignment Well-definedness

Stochastic backtracking provides an ensemble of stochastic alignments 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

$$\Delta S^{\text{ensemble}} = \beta S(\mathcal{A}_{\text{opt}}) - \ln Z = \ln \text{Prob}(\mathcal{A}_{opt}) \quad (13)$$



```
CE          ------QAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNN-PELQAHAGKVFKLVYEAAIQLE
            ------DKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF--PHFDLSHGSAQVKGHGKKVADALTNAVAHV----
TOP 6.7     --LTE-QAALVKSSWEEFN-----HTHRFFILVLE-APAAK----------------------HAGK------------
            --LSP-DKTNVKAAWGKVG-----YGAEALERMFL-FPTTK----------------------KVAD------------
SARF2       --LTESQAALVKSSWEEFNANI-KHTHRFFILVLEIAPAAKDLF----KGTSEVP--NPELQAHAGKVFKLVYE------
            --LSPADKTNVKAAWGKVGAHA-EYGAEALERMFLSFPTTKTYF----FDLSHGS--K-GHGKKVADALTNAVA------
MATRAS      -ALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE-VPQNNPELQAHAGKVFKLVYEAAIQLE
            -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD-----LSHGSAQVKGHGKKVADALTNAVAHV-
COMPARER    GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE-VPQNNPELQAHAGKVFKLVYEAAIQLE
            -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD-----LSHGSAQVKGHGKKVADALTNAVAHVD

consensus   GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSKLKGTSEVPQNN-PELQAHAGKVFKLVYEAAIQLE
            -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF--PHFDLSHGSAQVKGHGKKVADALTNAVAHV----

consensus   GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSKLKGTSEVPQNN-PELQAHAGKVFKLVYEAAIQLE
            -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF--PHFDLSHGSAQVKGHGKKVADALTNAVAHV----

optimal     GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEV
            -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF----DLSHGSAQVKGHGKKVADALTNAVAHVDD


CE          VTGVVVTDATLKNLGSVHV-SKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMD---
            ----DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
TOP 6.7     ------------NLG--------VADAHFPVVKEAILKTIKEVVG-KWSEELNSAWTIAYDELAIVIKKEM----
            ------------ALS--------VDPVNFKLLSHCLLVTLAAHLP-EFTPAVHASLDKFLASVSTVLTSKY----
SARF2       ----------LKNLGSVHV-SKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMD---
            ----------LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
MATRAS      VTGVVVTDATLKNLGSVHVS-KGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMD---
            --DDMPNA--LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
COMPARER    VTGVVVTDATLKNLGSVHVSK-GVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
            D-----MPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---

consensus   VTGVVVTDATLKNLGSVHV-SKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
            ----DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---

consensus   VTGVVVTDATLKNLGSVHV-SKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
            ----DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---

optimal     TGVVVTDATLKNLGSVHVSKGVAD--AHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
            MPNALSALSDLHAHKLRVDP------VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
```

```
0.1 __ XXXX
0.2 __ XXXX
0.3 __ XXXX
0.4 __ XXXX
0.5 __ XXXX
0.6 __ XXXX
0.7 __ XXXX
0.8 __ XXXX
0.9 __ XXXX
1.0 __ XXXX
```

Figure 3: 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.

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.

Figure 2 shows that nucleic acid sequence alignments are less well defined than protein alignments and the values for biological nucleic acid sequences lie only slightly below those for completely random nucleic acid 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.

3.2 A Structure-based Alignment

As an application of the algorithm 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 must be 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 assump-

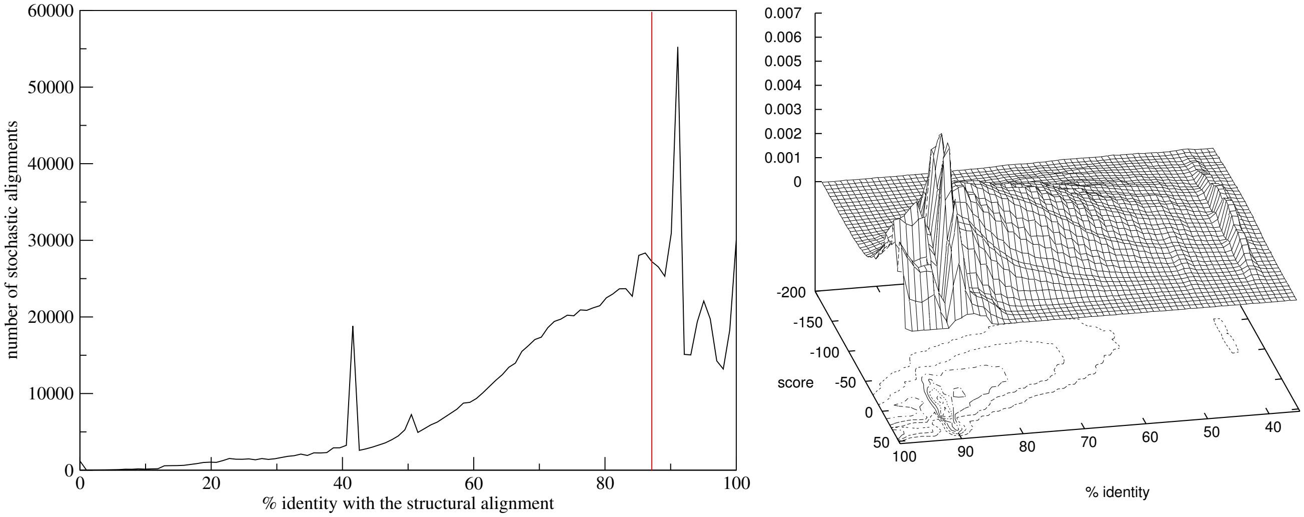


Figure 4: L.h.s.: Distribution of the fraction of stochastic alignments with different fractions of matches with the structure alignment. The optimal alignment is indicated by the red line. The three sharp peaks in the curve correspond to stochastic alignments containing one or more of the reliably aligned regions shown in figure 3. 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.

tion is that a region that is identically aligned by various methods is, in fact, reliably aligned.

The programs we used are based on different principles: Combinatorial Extension, CE[1] (Shindyalov and Bourne, 1998), uses similarity in local geometry of C_α; TOP[2] (Lu, 2000, 1996) and COMPARER[3] (Šali and Blundell, 1990) both utilize topological features for the computation of 3D structure alignments; SARF2[4] (Alexandrov and Fischer, 1996) and MATRAS[5], (Kawabata and Nishikawa, 2000) employ secondary structure information to generate structure alignments.

Figure 3 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 structue alignment consensus are classified as reliably aligned, this 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 Figure 4.

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 3. 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.

The correlation between the score of a suboptimal alignment and the percentage of reliably aligned positions in the structure consensus it includes can be seen most easily from a plot of the joint probability of obtaining an alignment with a given percentage of matches that are contained in the structural consensus and a given alignment score. The r.h.s. of Figure 4 shows that this correlation is weak, especially for alignments with near optimal score. Suboptimal alignments, that include all of the reliably aligned positions of the consensus, can have an alignment score nearly as high as the optimal alignment, that is a score of approx. 50, on the other hand their scores can be as low as -100. This highlights the fact that the score of an suboptimal alignment cannot be regarded as a reliable measure of its accuracy.

[1] http://cl.sdsc.edu/ce.html

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

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

[4] http://123d.ncifcrf.gov/run2.html

[5] http://bongo.lab.nig.ac.jp/~takawaba/Matras.html

4 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. The algorithm has been implemented in `C`. In addition to computing probabilities for individual matches our software produces correctly weighted samples of alignments by means of stochastic backtracking. The software package `probA` is available for download[6].

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. It will be desirable therefore 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.

The stochastic pairwise alignments are useful also in a completely different context. Many tools in bioinformatics require pairwise or multiple 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 varying the temperature parameter T). These can be used to investigate the sensitivity of method to realistic variations of the input alignments.

Acknowledgments

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

[6] `http://www.tbi.univie.ac.at/~ulim/probA/`

REFERENCES

Alexandrov, N. N. and Fischer, D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins: Struct. Funct. Genet.* **25**, 354–65 (1996).

Bucher, P. and Hoffmann, K. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In: *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96)* (States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds.), pp. 44–50. Menlo Park, CA: AAAI Press (1996).

Durbin, R., Eddy, S. R., Krogh, A. and G., M. *Biological sequence analysis*. Cambridge: Cambridge University Press (1998).

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

Giegerich, R. *Explaining and Controlling Ambiguity in Dynamic Programming*. Tech. rep., Faculty of Technology, Bielefeld University, Bielefeld, Germany (2000). Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching.

Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162** (3), 705–708 (1982).

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. The structure of deoxy- and oxy-leghaemoglobin from lupin. *J. Mol. Biol.* **251**, 104–115 (1995).

Kawabata, T. and Nishikawa, K. Protein structure comparison using the Markov transition model of evolution. *Protein structure* **41** (1), 108–122 (2000).

Kschischo, M. and Lassig, M. Finite-temperature sequence alignment. *Pacific Symposium Biocomputing* **1**, 624–35 (2000).

Lathrop, R. H. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7** (9), 1059–1068 (1994).

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

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

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

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48** (3), 443–453 (1970).

Šali, A. and Blundell, T. L. Definition of general topological equivalence in protein struc tures. A procedure involving comparison of properties and relationships throug h simulated annealing and dynamic programming. *J. Mol. Biol.* **212** (2), 403–428 (1990).

Saqi, M. A. and Sternberg, M. J. A simple method to generate non-trivial alternate alignments o f protein sequences. *J. Mol. Biol.* **219** (4), 727–732 (1991).

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

Vingron, M. Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.* **6** (3), 346–352 (1996).

Vingron, M. and Argos, P. Determination of reliable regions in protein sequence alignmen ts. *Protein Eng.* **3** (7), 565–569 (1990).

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

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