BIOINFORMATICS

Vol. 18 no. 6 2002
Pages 847–854

A novel approach to local reliability of sequence alignments

Maximilian Schlosshauer and Mattias Ohlsson

Complex Systems Division, Department of Theoretical Physics, University of Lund,
Sölvegatan 14A, S–223 62 Lund, Sweden

Received on August 14, 2001; revised on January 14, 2002; accepted on January 21, 2002

ABSTRACT

Motivation: The pairwise alignment of biological sequences obtained from an algorithm will in general contain both correct and incorrect parts. Hence, to allow for a valid interpretation of the alignment, the local trustworthiness of the alignment has to be quantified.

Results: We present a novel approach that attributes a reliability index to every pair of residues, including gapped regions, in the optimal alignment of two protein sequences. The method is based on a fuzzy recast of the dynamic programming algorithm for sequence alignment in terms of mean field annealing. An extensive evaluation with structural reference alignments not only shows that the probability for a pair of residues to be correctly aligned grows consistently with increasing reliability index, but moreover demonstrates that the value of the reliability index can directly be translated into an estimate of the probability for a correct alignment.

Contact: mattias@thep.lu.se

INTRODUCTION

Comparing protein sequences by means of an alignment algorithm has become an ubiquitous task of modern molecular biology. From such an alignment, evolutionary, structural and functional relationships between proteins can be delineated. Standard algorithms for global (Needleman and Wunsch, 1970) and local (Smith and Waterman, 1981) sequence alignment maximize a score function that favors matching of more similar residues over the pairing of more dissimilar residues and the insertion of gaps. Measures of similarity between residues are commonly derived from the observation of mutational probabilities; insertion of gaps is typically penalized by a score linearly growing with the length of the gap.

This implies that the optimal alignment obtained from such an algorithm is essentially nothing but the product of an optimization procedure that in turn is based on a predefined scoring scheme, namely similarity scores and gap penalties. The frequently subtle relationships between proteins can, however, not always be detected by this method. Alternative suboptimal alignments that score slightly lower than the optimal alignment and are therefore disregarded by the algorithm, but might actually pinpoint conserved regions between the proteins better. Furthermore, a change in the scoring parameters will often result in drastic alterations of the resulting alignment, in particular in the case of sequences of low similarity (Barton and Sternberg, 1987; Vingron and Waterman, 1994).

In general, we can therefore not expect the optimal alignment to be a mirror of biological truth in all its parts. Hence, a method is desired that can assess the local reliability of the alignment by attributing probabilities for a correct alignment to every region in the alignment, down to individual pairs.

Starting with the pioneering work by Vingron and Argos (1990), local reliability has been commonly deduced from an implicit or explicit study of suboptimal alignments competing with the optimal solution. Vingron and Argos demonstrated that regions in the optimal (global) alignment that remain unaltered among a large set of suboptimal alignments exhibit stronger agreement with the structural reference than regions that are represented only in a few close-to-optimal alignments.

An application of this idea also to local Smith–Waterman alignments has been described by Zuker (1991). Saqi and Sternberg (1991) calculated explicitely a limited set of suboptimal alignments that differ nontrivially from each other and can be used to identify larger parts of the correct alignment. A detailed study of suboptimal alignments was performed by Naor and Brutlag (1994). Chao *et al.* (1993) introduced an algorithm that allows to quantify the reliability of each individual residue pair in the optimal alignment; a modified and extensively evaluated version of this method has been described by Mevissen and Vingron (1996). Alternative approaches have been proposed in the context of a probabilistic interpretation of the alignment score (Kschischo and Lässing, 2000; Miyazawa, 1995), from which probabilities for individual residue–residue pairings are derived that can be interpreted as reliabilities.

The available methods so far, however, exhibit certain drawbacks. In many cases, the reliability measure does not evolve naturally from the task of sequence alignment itself but entails more or less complicated additional algorithms (as in Chao *et al.*, 1993; Mevissen and Vingron, 1996). Moreover, translating the algorithmically obtained reliability index into a probability for correct alignment may require the analysis of external databases with reference alignments (Mevissen and Vingron, 1996). Among the presented methods, only the algorithm by Chao *et al.* (1993) is able to assign reliability to gapped regions in the alignment, however at the price of involving an high degree of algorithmical sophistication.

We present a novel approach for quantifying the local reliability of sequence alignments that aims at a resolution of these problems. A 'fuzzy' implementation of the dynamic programming algorithm for sequence alignment in terms of mean field annealing enables us to study explicitely the local dynamics of the optimization task of sequence alignment. Quantifying these dynamics provides a measure for the presence of locally alternative solutions and can therefore be used to deduce a reliability index for each residue–residue and residue–gap pair in the optimal alignment.

In what follows, we shall for the sake of simplicity restrict ourselves to global Needleman–Wunsch alignments. Our method, however, possesses sufficient generality to be applied to local Smith–Waterman alignments (Smith and Waterman, 1981) and the like as well.

METHODS

Review of global sequence alignment

As our method is based on a reformulation of the Needleman–Wunsch algorithm for global sequence alignment (Needleman and Wunsch, 1970), we shall briefly review this algorithm in an implementation that will allow for a straightforward introduction of our algorithm.

Let $\mathbf{A} = (A_1 A_2 \dots A_M)$ and $\mathbf{B} = (B_1 B_2 \dots B_N)$ denote the two sequence strings containing M and N residues, respectively. We introduce a $(M+1) \times (N+1)$ alignment matrix such that we can represent every possible alignment of the two sequences by a directed path in this matrix (Figure 1). The matrix element, or node, (i, j) has (with obvious restrictions at the left and top margin of the matrix) three possible predecessors along the alignment path, specified by directions of propagation $k = 1, 2, 3$ (see Figure 2).

The score $\mathcal{S}(i, j)$ for the optimal alignment of the sequence prefix $(A_1 A_2 \dots A_i)$ of the whole sequence $\mathbf{A}$ with the sequence prefix $(B_1 B_2 \dots B_j)$ of the whole sequence $\mathbf{B}$ is then

$$\mathcal{S}(i, j) = \max_k \{\widetilde{\mathcal{S}}(i, j; k)\}. \tag{1}$$

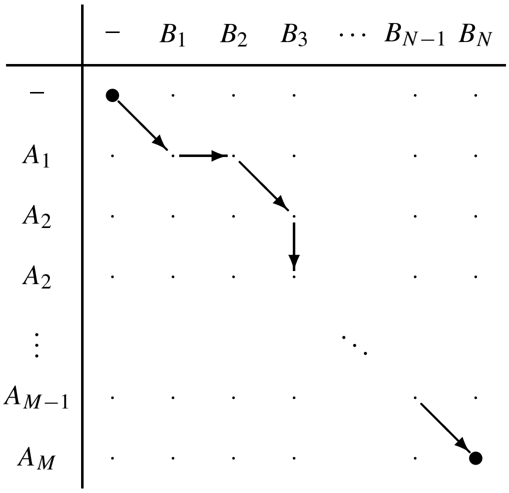


Fig. 1. The alignment matrix representation for an alignment of the two sequences $\mathbf{A} = (A_1 A_2 \dots A_M)$ and $\mathbf{B} = (B_1 B_2 \dots B_N)$.

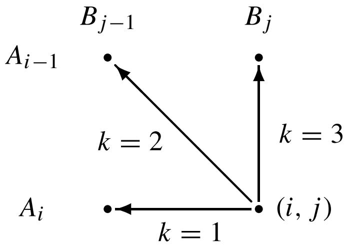


Fig. 2. The integer variable k specifies the possible predecessors to the node (i, j).

Here, $\widetilde{\mathcal{S}}(i, j; k)$ is the (optimal) score in (i, j) when the alignment path is forced to pass through the preceding node specified by the direction k, and it is recursively given by

$$\begin{aligned} \widetilde{\mathcal{S}}(i, j; k = 1) &= \max_{0 \le b < j} \{\mathcal{S}(i, b) - g(j - b)\}, \\ \widetilde{\mathcal{S}}(i, j; k = 2) &= \mathcal{S}(i-1, j-1) + \sigma(A_i, B_j), \\ \widetilde{\mathcal{S}}(i, j; k = 3) &= \max_{0 \le a < i} \{\mathcal{S}(a, j) - g(i - a)\}, \end{aligned} \tag{2}$$

where $\sigma(A_i, B_j)$ is the similarity score for aligning A_i with B_j and $g(l)$ the penalty for a gap of length l.

A directed sweep through the alignment matrix to recursively calculate $\mathcal{S}(i, j)$ for each (i, j) will finally yield $\mathcal{S}(M, N)$, the score for the optimal alignment of $\mathbf{A}$ with $\mathbf{B}$. To actually retrace the optimal alignment, we record, for each node (i, j), which of the three $\widetilde{\mathcal{S}}(i, j; k)$ in Equation (1) has yielded the optimal score by introducing binary variables $s_{ij,k}$ as

$$s_{ij,k} = \begin{cases} 1 & \text{if } \widetilde{\mathcal{S}}(i,j;k) = \max_{k'}\{\widetilde{\mathcal{S}}(i,j;k')\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then, we can replace the 'max' function in Equation (1) by the sum

$$\mathcal{S}(i,j) = \sum_k s_{ij,k}\,\widetilde{\mathcal{S}}(i,j;k). \tag{4}$$

Since for nodes in the first column or row of the alignment matrix only at most one directly preceding node exists, we can initialize as constants

$$\begin{aligned} s_{00,k} &= 0,\ k = 1,2,3, \\ s_{i0,1} = s_{i0,2} = 0,\ s_{i0,3} &= 1,\ i = 1,\dots,M, \\ s_{0j,2} = s_{0j,3} = 0,\ s_{0j,1} &= 1,\ j = 1,\dots,N, \end{aligned} \tag{5}$$

and

$$\begin{aligned} \mathcal{S}(i,0) &= g(i), \qquad i = 0,\dots,M, \\ \mathcal{S}(0,j) &= g(j), \qquad j = 0,\dots,N. \end{aligned} \tag{6}$$

For a affine linear gap penalty $g(l) = g_{\text{open}} + (l-1)g_{\text{ext}}$ (to which we shall restrict ourselves in the following), Equations (2) simplify to

$$\begin{aligned} \widetilde{\mathcal{S}}(i,j;k=1) &= \mathcal{S}(i,j-1) - g_{\text{open}}\,(1 - s_{i,j-1,1}) \\ &\quad - g_{\text{ext}}\,s_{i,j-1,1}, \\ \widetilde{\mathcal{S}}(i,j;k=2) &= \mathcal{S}(i-1,j-1) + \sigma(A_i,B_j), \\ \widetilde{\mathcal{S}}(i,j;k=3) &= \mathcal{S}(i-1,j) - g_{\text{open}}\,(1 - s_{i-1,j,3}) \\ &\quad - g_{\text{ext}}\,s_{i-1,j,3}. \end{aligned} \tag{7}$$

Introducing fuzzy alignment paths

The $s_{ij,k}$ as introduced above for the original Needleman–Wunsch algorithm are strictly binary 'winner-takes-all' variables. As such, they encode only a single alignment, namely the optimal alignment. To account for suboptimal deviations from this rigid alignment path, we propose a fuzzy recast of the Needleman–Wunsch algorithm by replacing the $s_{ij,k}$ by continuous 'winner-takes-most' variables $v_{ij,k}$ (Häkkinen *et al.*, 1998),

$$s_{ij,k} \to v_{ij,k} = \frac{e^{\widetilde{\mathcal{S}}(i,j;k)/T}}{\sum_{k'} e^{\widetilde{\mathcal{S}}(i,j;k')/T}}. \tag{8}$$

For $T > 0$, this can be viewed as a soft implementation of the 'max' function occurring in Equations (1) and (3); in the limit $T \to 0$, $v_{ij,k} \to s_{ij,k}$ and proper Needleman–Wunsch is recovered.

Accordingly, Equations (4) and (7) must now be expressed in terms of the $v_{ij,k}$ instead of the $s_{ij,k}$. Since by construction $\sum_k v_{ij,k} = 1$, the value of $v_{ij,k}$ can be interpreted as a probability that an optimal alignment path that passes through (i, j) contains the node specified by the direction k.

Relation to Mean Field Annealing

We note the similarity of Equation (8) to a Boltzmann probability when interpreting $\widetilde{\mathcal{S}}(i,j;k)$ as a negative energy and T as a fictitious temperature. In fact, it can be shown that the $v_{ij,k}$ are *mean field* (MF) approximations (Peterson and Söderberg, 1989) of the thermal averages $\langle s_{ij,k}\rangle_T$ of the original binary $s_{ij,k}$ for a suitably chosen energy function whose minimization corresponds to our optimization goal of finding the maximum-score alignment. Such MF approximations together with a stepwise lowering of T (*simulated annealing*; Kirkpatrick *et al.*, 1983 have in the past successfully been used to solve various combinatorial optimization problems (*mean field annealing*; for a review see e.g. Peterson and Söderberg, 1998).

Applied to our optimization problem of sequence alignment, solving the MF equations (8) iteratively under annealing T will yield a solution to the alignment problem which, due to the relative simplicity of this optimization task, is likely to coincide with the optimal alignment obtained from the Needleman–Wunsch algorithm. Therefore, we may use this MF annealing algorithm as an alternative method to solve the alignment problem.

The motivation for this alleged detour lies in the fact that it allows for an explicit study of the dynamics of the optimization procedure, represented by the evolution of the MF variables $v_{ij,k}$, Equations (8), under annealing T. Typically, variables $v_{ij,k}$ that exhibit only a late (i.e. low-T when $T \to 0$) onset of 'decision' towards their limiting values of one or zero, possibly accompanied by oscillatory fluctuations, indicate the presence of a local multiplicity of possible solutions represented by these $v_{ij,k}$ that makes fast convergence to a unique solution difficult (see Figure 3).

Obtaining a reliability index

With every $v_{ij,k}$ on the alignment path representing the alignment of a single pair, we can therefore monitor the evolution of the $v_{ij,k}$ under annealing T to estimate the presence of suboptimal alternatives to this aligned pair and thus to quantify its reliability. As it can be seen from Figure 3, the area that the curve $v_{ij,k}$ encloses with the T axis can be used to measure how fast and smoothly convergence is achieved.

We shall therefore define the local reliability $r(m)$ of the mth pair in the optimal alignment as

$$r(m) = \frac{1}{T_0}\int_0^{T_0} v^{(m)}\,\mathrm{d}T, \tag{9}$$

where $v^{(m)}$ denotes the $v_{ij,k}$ corresponding to the mth pair in the optimal alignment and T_0 the initial 'temperature' for the annealing process $T \to 0$. Thus, $r(m)$ measures the ratio between the area that the curve $v^{(m)}$ encloses with the T axis and the largest possible area that can be enclosed

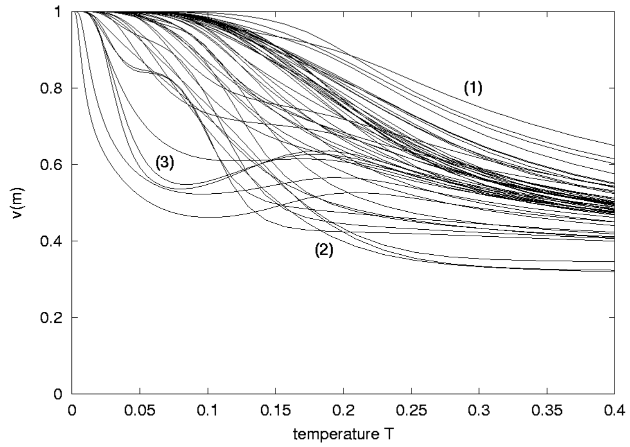


Fig. 3. A typical evolution of MF variables $v^{(m)} \equiv v_{ij,k}$, Equation (8), corresponding to the mth pair in the optimal alignment (thus all $v^{(m)} \to 1$ as $T \to 0$), under annealing T. Some $v^{(m)}$ exhibit a fast and smooth decision towards the limiting value of one (1), representing a dominating, non-ambiguous local solution to the optimization problem. Others show only late convergence at low T (2), sometimes with additional oscillations (3), which indicates the presence of competing alternative solutions.

(namely, by a curve with a constant value of one). We shall report $r(m)$ as a *reliability index* taking values between 0 (corresponding to $0 \leq r(m) < 0.1$) and 9 (corresponding to $0.9 \leq r(m) \leq 1.0$).

Note that since all $v_{0j,k}$ and $v_{i0,k}$ are required to retain their initial values, Equations (5), no meaningful reliability index can be assigned to initial end gaps. Since an attribution of reliability to end gaps is, however, of anyhow doubtful biological interpretation, this is no drawback; instead, we shall consistently refrain from assigning reliability indices both to initial and to terminal end gaps.

Implementation

Since the alignment obtained from the (approximative) MF annealing algorithm may occasionally differ from the exact Needleman–Wunsch alignment, we first calculate the optimal alignment with binary $s_{ij,k}$, Equation (3), corresponding to the zero-T limit and thus recovering the exact Needleman–Wunsch algorithm.

Consistency between the exact and the MF solution is then achieved by taking the resulting set of $s_{ij,k}$ and scores $\mathcal{S}(i, j)$ as the initial $T = 0$ configuration for an iteration of the MF Equations (8) under now *increasing* T (*inverse annealing*), $T \to \epsilon T$ with $\epsilon > 1$ (we used $\epsilon = 1.1$); the evolution of the $v^{(m)}$ is recorded by a series of discretized integration steps to compute the reliability index, Equation (9), for each aligned pair. This approach is justified by the observation that relevant features in the evolution of the $v_{ij,k}$ are present in both directions of changing T.

Instead of using a fixed value for T_0, cf. Equation (9), the procedure of inverse annealing is dynamically terminated when the $v^{(m)}$ have attained a certain average value $\overline{v} = v_0$ that ensures sufficient convergence towards the infinite-T value of 1/3 (we used $v_0 = 0.55$). All computations were performed with a modified PAM-250 matrix (Gonnet *et al.*, 1992) and a affine linear gap penalty $g(l) = 10 + l$ with free end gaps.

The complete algorithm for computing the optimal alignment with reliability indices assigned to each pair is summarized in Figure 4.

RESULTS

Test database

To demonstrate that our reliability index, Equation (9), serves in fact as a meaningful measure for the local trustworthiness of the optimal alignment, we need to evaluate whether pairs that receive high reliability indices are in fact more likely to be correctly aligned than pairs with lower reliability indices.

This requires the availability of a sufficiently large set of alignments that can be taken as a reference for the evaluation of the correctness of the Needleman–Wunsch alignments. As such a 'standard of truth', we used the *3D_ali* database (Pascarella and Argos, 1992) which provides a broad collection of multiple sequence alignments, organized into protein families, that have been obtained from thoroughly checked structural superpositions and sequence alignments. From this database, pairwise alignments and the corresponding sequences were extracted to be used as a test set for our algorithm.

To elucidate differences in the performance of our algorithm that are due to the degree of similarity between the two sequences, we divided the data set into three similarity classes (as in Mevissen and Vingron, 1996) of 25–30%, 30–40% and 40–50% sequence identity of residue pairs in the optimal Needleman–Wunsch alignment. Sequence pairs of lower similarity do usually not share relevant structural similarities, and algorithms normally fail to detect any possible relations. On the other hand, algorithms align in general sequences above 50% similarity correctly in most of the residues. Accordingly, sequence pairs with similarity below 25% or above 50% were excluded from our analysis.

The number of contained sequences varies strongly among different families in the 3D_ali database. However, many alignments within one family are usually very similar. Hence, to avoid biasing the results towards more highly populated families, the results obtained from the evaluation of one family were normalized by the number of sequence pairs taken into account. Furthermore, we

```
1. Perform a global (exact) Needleman–Wunsch alignment of the two given sequences:
   (a) For $i = 0, \ldots, M$ and $j = 0, \ldots, N$, initialize $s_{i0,k}$ and $s_{0j,k}$ according to Equations (5), and $\mathcal{S}(i, 0)$ and $\mathcal{S}(0, j)$ according to Equation (6).
   (b) By proceeding row by row $i \to i + 1$, $i = 1, \ldots, M$, and in each row column by column, $j \to j + 1$, $j = 1, \ldots, N$, calculate recursively for each node $(i, j)$ in the given order:
       i. $\widetilde{\mathcal{S}}(i, j; k)$ from Equations (7);
       ii. (binary) $s_{ij,k}$ from Equation (3);
       iii. $\mathcal{S}(i, j)$ from Equation (4).
   (c) When finished, retrace the optimal alignment by starting from node $(M, N)$ and following the directions encoded in the $s_{ij,k}$ until the initial node (0, 0) is reached.
2. Compute a reliability index for every pair in the optimal alignment using MF annealing:
   (a) Take the $s_{ij,k}$ and $\mathcal{S}(i, j)$ calculated from the Needleman–Wunsch algorithm in step 1 as initial configuration, with the binary $s_{ij,k}$, Equation (3), replaced throughout by continuous MF $v_{ij,k}$, Equation (8). Initialize $T$ close to zero, and all reliability indices as $r(m) = 0$.
   (b) Compute iteratively for each node $(i, j)$ until convergence:
       i. $\widetilde{\mathcal{S}}(i, j; k)$ from Equations (7) (with binary $s_{ij,k}$ replaced by MF $v_{ij,k}$);
       ii. (fuzzy) $v_{ij,k}$ from Equation (8);
       iii. $\mathcal{S}(i, j)$ from Equation (4) (with binary $s_{ij,k}$ replaced by MF $v_{ij,k}$).
   (c) Using the values of the variables $v^{(m)}$ on the optimal path (excluding end gaps),
       i. perform a single (discrete) integration step, e.g. via $r(m) \to r(m) + (\epsilon - 1)T\, v^{(m)}$, cf. Equation (9);
       ii. calculate the arithmetic average $\overline{v}$ of the $v^{(m)}$.
   (d) If $\overline{v} > v_0$, increase $T \to \epsilon T$ and go back to step 2b.
   (e) Normalize $r(m) \to r(m)/T$.
3. Output the optimal alignment together with the $r(m)$ (binned into reliability indices) for each aligned pair.
```

Fig. 4. Summary of the algorithm for the computation of the optimal alignment with reliability indices assigned to every aligned pair.

evaluated at most 40 alignments from each family and similarity class to limit computation time.

A total number of 5234 3D_ali alignments was analyzed, namely, 1193 alignments from 55 families in the 25–30% similarity class, 1873 alignments from 78 families in the 30–40% similarity class and 2168 alignments from 95 families in the 40–50% similarity class.

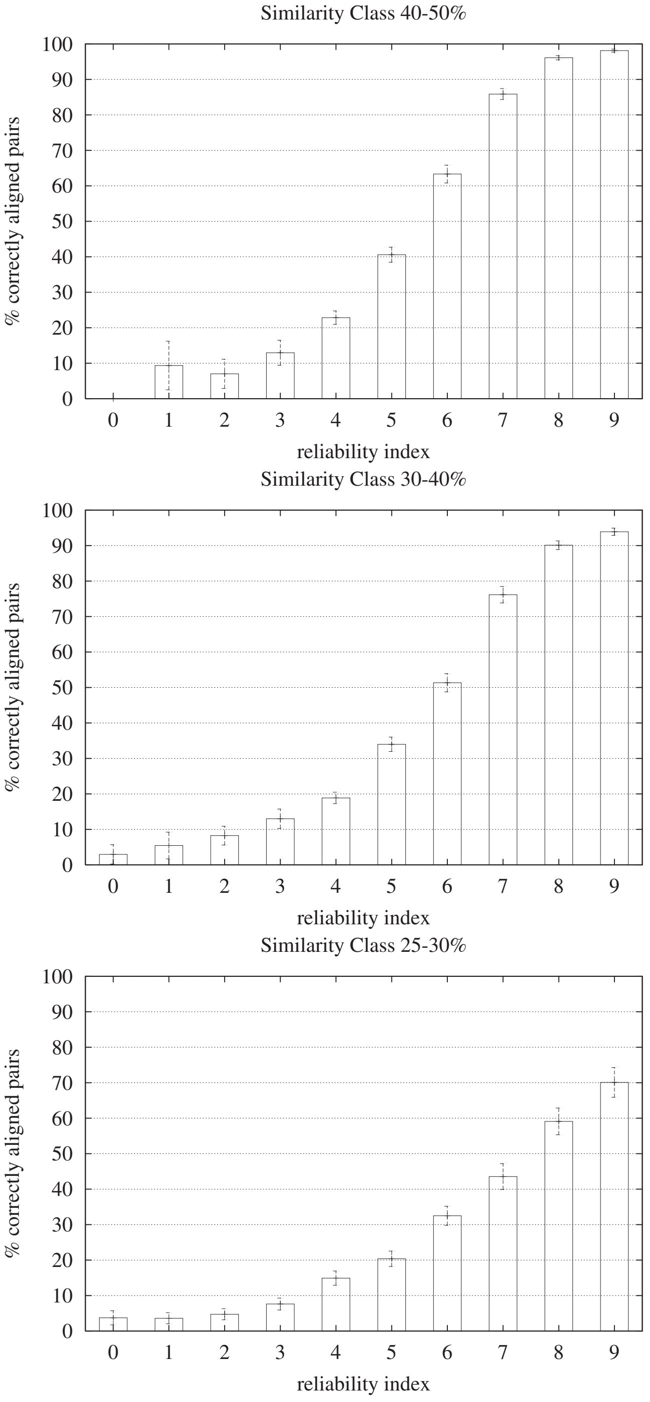
Reliability index and correct alignment

The statistics acquired from the analysis of the 3D_ali alignments are shown in Figure 5. The histograms depict, separately for the three similarity classes, the relationship between the percentage of correctly aligned residue–residue and residue–gap pairs and the assigned reliability index.

It can readily be seen that over the full range of reliability indices the percentage of correctly aligned pairs grows consistently with increasing reliability index. Since we have evaluated a large set of alignments, we can interpret this percentage as a probability for a correct alignment of the respective pair, which justifies our reliability index as a meaningful indicator of the local trustworthiness of the optimal alignment for all three similarity classes.

Moreover, the growth of the percentage of correctly aligned pairs with reliability index is observed to proceed roughly linearly (with some flattening out towards low and high reliability indices). This allows for a direct and intuitive translation of the value of the reliability index into an estimate of the probability for a correct alignment of the corresponding pair. For sequences with similarities above approximately 30%, the histograms demonstrate that the numerical value of $r(m)$, Equation (9), provides a good estimate of the probability for a correct alignment, particularly in the region of higher reliability indices. Only in the lowest similarity class, this direct numerical correspondence is not fully retained as even pairs with a reliability index of 9 are correctly aligned with only 70% probability.

To quantify the trustworthiness of our results, we estimated the standard errors of the percentages of correctly aligned pairs with bootstrap point estimates, shown as error bars in Figure 5. The standard errors are seen to be consistently small regardless of reliability index and sim-


Similarity Class 40-50%

Similarity Class 30-40%

Similarity Class 25-30%

Fig. 5. Percentage of correctly aligned pairs with standard errors versus assigned reliability index for the three different similarity classes.

ilarity class. This ensures that the translation of our reliability index into a probability for a correct alignment is in fact a meaningful and trustworthy procedure.

Our reliability index is explicitly illustrated in Figure 6. In this example *cytochrome C2* is aligned against *cytochrome C* using the Needleman–Wunsch algorithm. The reliability index is shown in the upper part of the figure together with the structural alignment according to the 3d_ali database in the lower part. It shows nicely the correspondence between the reliability index and the correct alignment—the incorrect aligned parts (without the '+' in the row 'correct') receive lower indices than the rest.

Computational effort

The computational effort scales as $\mathcal{O}(N^2 \cdot \ell)$, where N is the sequence length and ℓ the number of annealing steps required to attain the average convergence $\overline{v} = v_0$. However, shorter sequences are in general easier to align, and therefore $\overline{v}$ will for these sequences decrease more *slowly* to v_0 under increasing T than in the case of longer sequences (cf. Figure 3). This implies that ℓ increases with decreasing sequence lengths. In the range of typical sequence lengths up to $N = 1,000$ residues, these two opposite trends are found to lead to an effective scaling of $\mathcal{O}(N^\gamma)$ with $\gamma \approx 1.3$ (see Figure 7). For longer sequences, ℓ has attained its possible minimum and will therefore remain roughly constant at this value, which yields an $\mathcal{O}(N^2)$ scaling.

DISCUSSION AND CONCLUSION

A novel method for assigning reliability indices to both residue–residue and residue–gap pairs (excluding end gaps) in the optimal alignment of two protein sequences has been presented and evaluated. A fuzzy recast of the Needleman–Wunsch algorithm for sequence alignment together with simulated annealing allows for a natural estimate of the presence of suboptimal alternatives to each aligned pair which in turn provides a measure for the local reliability of the pair.

A thorough evaluation of our algorithm has shown that our reliability index in fact deserves its name. Furthermore, we could demonstrate that the value of the reliability index can directly be interpreted as an estimator of the probability for a correct alignment, at least for sequence pairs with similarities above the twilight zone of 25–30% identity. This is superior to previous approaches (e.g. Mevissen and Vingron, 1996) where a lack of distinction in the probabilities among different reliability indices requires external reference data to translate the value of the reliability index into a probability for correct alignment.

The overall decreased percentages of correctly aligned pairs for sequences of very low similarity (below 30% identity), in particular in the range of higher reliability indices, can be attributed to the fact that here the alignment obtained from a simple score optimization (as implemented by the alignment algorithm) frequently disagrees widely with the structural reference, especially in gapped regions; moreover, in this similarity class even the struc-

```
Sequence 1 = cytochrome C2
Sequence 2 = cytochrome C

 90 - 100% :                                                               *
 80 -  90% :               *  ******** *************       ******        ****** *  *********                             ***
 70 -  80% :  ***********  **************************************       ********************            ***    *      ********
 60 -  70% : ************  **************************************      *********************            ***  ******************
 50 -  60% : ************ ****************************************   *********************** **      ******  ******************
 40 -  50% : *************************************************************************************  ***************************
 30 -  40% : ******************************************************************************************************************
 20 -  30% : ******************************************************************************************************************
 10 -  20% : ******************************************************************************************************************
  0 -  10% : ******************************************************************************************************************

[Optimal   : EGDAAAGEKVSK-KCLACHTFDQGGANKVGPNLFGVFENTAAHKDNYAYSESYTEMKAKGLTWTEANLAAYVKNPKAFVLEKSGDPKAKSKMTF-KLTKDDEIENVIAYLKTLK-
alignment] : AGDSKKGANLFKTRCAQCHTLEEGGGNKIGPALHGLFGRKTGSVDGYAYTDA---NKQKGITWDENTLFEYLENPKKYI--------PGTKMAFGGLKKDKDRNDIITFMKEATA

[correct]  : ++++++++++++  ++++++++++++++++++++++++++++++++++++++      +++++++++++++++++++++          +++++  +++++++++++++++++++

[3d_ali]   : EGDAAAGEKVSKK-CLACHTFDQGGANKVGPNLFGVFENTAAHKDNYAYSESYTEMKAKGLTWTEANLAAYVKNPKAFVLEKSGDPKAKSKMTFK-LTKDDEIENVIAYLKTLK-
           : AGDSKKGANLFKTRCAQCHTLEEGGGNKIGPALHGLFGRKTGSVDGYAYTDANKQ---KGITWDENTLFEYLENPKKYIP-------G-TKMAFGGLKKDKDRNDIITFMKEATA
```

Fig. 6. The reliability index for the alignment of *cytochrome C2* and *cytochrome C*. The optimal alignment, using the Needleman–Wunsch algorithm and the structural alignment according to the 3d_ali database is shown in the lower part of the figure. A '+' in the row 'correct' indicates correct aligned residues.

tural alignment itself is of sometimes doubtful validity. Such systematic disagreements between the optimal and the structural alignment can obviously not always be reflected in the dynamics of the optimization process, which may result in a small number of pairs that receive higher reliability indices but are predominantly misaligned with respect to the structural reference.

It should be stressed that our method does not require the introduction of an algorithm distinct from the procedure of sequence alignment itself. Instead, we study directly the dynamics of the optimization problem of finding the optimal alignment to deduce our reliability measure. This allows also for a natural assignment of reliability indices to gapped regions, since both residue–gap and residue–residue pairs in the optimal alignment are the result of a common optimization procedure.

Despite its independence of additional external data, our method is not fully parameter free. In particular, the normalization factor in front of the integral in Equation (9) will have the most dominating influence on the obtained reliability index for a given pair. This normalization is a natural choice that has been shown to yield very satisfying results; no particular constraint, however, is imposed on the normalization, and small changes may be found to further improve our results.

ACKNOWLEDGMENTS

We would like to thank Carsten Peterson for providing the initial idea to this work and for many helpful discussions. This work was in part supported by the Swedish Natural Science Research Council and the Swedish Foundation for Strategic Research.

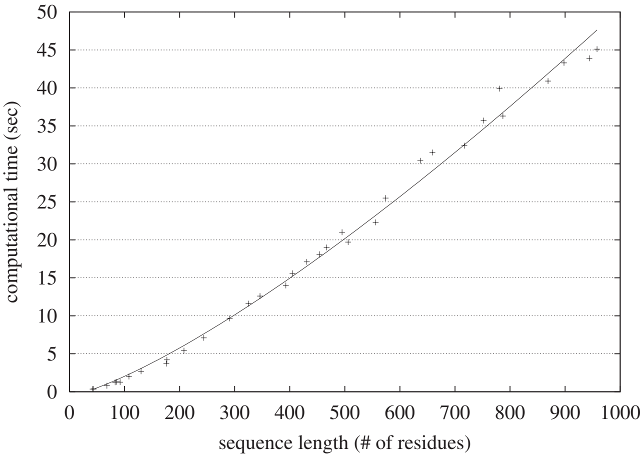


Fig. 7. Scaling of the computational effort for the calculation of the optimal alignment with assigned reliability indices as a function of sequence lengths, with a fitting curve representing an $\mathcal{O}(N^{\gamma})$ scaling where $\gamma \approx 1.3$. (Computational times measured on an 800 MHz Pentium III.)

REFERENCES

Barton,G.J. and Sternberg,M.J.E. (1987) Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.*, **1**, 89–94.

Chao,K.–M., Hardison,R.C. and Miller,W. (1993) Locating well-conserved regions within a pairwise alignment. *Comput. Appl. Biosci.*, **4**, 387–396.

Gonnet,G.H., Cohen,M.A. and Brenner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.

Häkkinen,J., Lagerholm,M., Peterson,C. and Söderberg,B. (1998) A Potts neuron approach to communication routing. *Neural Comput.*, **10**, 1587–1599.

Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kschischo,M. and Lässing,M. (2000) Finite-temperature sequence alignment. *Pac. Symp. Biocomput.*, **5**, 621–632.

Mevissen,H.Th. and Vingron,M. (1996) Quantifying the local reliability of a sequence alignment. *Protein Eng.*, **9**, 127–132.

Miyazawa,S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, **8**, 999–1009.

Naor,D. and Brutlag,D.L. (1994) On near-optimal alignments of biological sequences. *J. Comp. Biol.*, **1**, 349–366.

Needleman,S.B. and Wunsch,Ch.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.

Pascarella,S. and Argos,P. (1992) A data bank merging related protein structures and sequences. *Protein Eng.*, **5**, 121–137.

Peterson,C. and Söderberg,B. (1989) A new method for mapping optimization problems onto neural networks. *Int. J. Neural. Syst.*, **1**, 3–22.

Peterson,C. and Söderberg,B. (1998) Neural optimization. In Arbib,M.A. (ed.), *The Handbook of Brain Research and Neural Networks*. Bradford Books/The MIT Press, Cambridge, MA.

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

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

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

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice: Review of concepts, case studies and implications. *J. Mol. Biol.*, **235**, 1–12.

Zuker,M. (1991) Suboptimal sequence alignment in molecular biology: Alignment with error analysis. *J. Mol. Biol.*, **221**, 403–420.