Protein Engineering vol.8 no.10 pp.999–1009, 1994

A reliable sequence alignment method based on probabilities of residue correspondences

Sanzo Miyazawa

Faculty of Technology, Gunma University, Kiryu, Gunma 376, Japan

Probabilities of all possible correspondences of residues in aligning two proteins are evaluated by assuming that the statistical weight of each alignment is proportional to the exponent of its total similarity score. Based on such probabilities, a probability alignment that includes the most probable correspondences is proposed. In the cases of highly similar sequence pairs, the probability alignments agree with the maximum similarity alignments that correspond to the alignments with the maximum similarity score. Significant correspondences in the probability alignments are those whose probabilities are >0.5. The probability alignment method is applied to a few protein pairs, and results indicate that such highly probable correspondences in the probability alignments are probably correct correspondences that agree with the structural alignments and that incorrect correspondences in the maximum similarity alignments are usually insignificant correspondences in the probability alignments. The root mean square deviations in superimposition of corresponding residues tend to be smaller for significant correspondences in the probability alignments than for all correspondences in the maximum similarity alignments, indicating that incorrect correspondences in the maximum similarity alignments tend to be insignificant correspondences in probability alignments. This fact is also confirmed in 109 protein pairs that are similar to each other with sequence identities between 90 and 35%. In addition, the probability alignment method may better predict correct correspondences than the maximum similarity alignment method. Probability alignments do, of course, depend on a scoring scheme but are less sensitive to the value of parameters such as gap penalties. The present probability alignment method is useful for constructing reliable alignments based on the probabilities of correspondences and can be used with any scoring scheme.

Key words: DNA sequence/probability alignment/protein sequence/reliable alignment/sequence alignment

Introduction

A first step in assessing and analysing relationships between two or more proteins is the establishment of alignments of their amino acid sequences. The prediction of an unknown protein structure is often studied on the basis of known structures of homologous proteins. The correct correspondences of residues in an alignment will be critical in such a study.

Since the three-dimensional structure of a protein is more conservative than its amino acid sequence (Chothia and Lesk, 1986), structural alignments, which may be obtained by superimposing atomic positions (Rao and Rossmann, 1973) or by a dynamic programming method with a scoring system based on structural information involving long-range interactions among residues (Taylor and Orengo, 1989a,b; Subbarao and Haneef, 1991; Luo *et al.*, 1993), may be more appropriate for distantly related protein pairs than alignments based on sequence information alone. However, in cases where protein structures are unknown, we must depend on alignments based on sequence information alone.

There are two types of alignment methods based on dynamical programming: global alignment methods for aligning a whole sequence with maximum similarity or minimum distance, and local homology search methods for finding significantly similar segments between sequences. The global alignment algorithms were originally introduced by Needleman and Wunsch (1970) and Sellers (1974). The Smith and Waterman algorithm (Smith and Waterman, 1981) is known to be one of good local homology search algorithms. If two sequences are similar not over whole sequences but in limited regions, the local homology search methods will be better than the global alignment methods; usually this is the case when comparing DNA sequences. Even in proteins, local alignment methods may be better for finding some features, such as that a protein consists of domain structures or has evolved as a chimera via gene fusions of segments from different genes. However, the global alignment method is more appropriate for comparison between protein domains.

The same alignment program can frequently produce significantly different alignments under different parameter settings. The effects that parameter choice has on resulting alignments have been studied (Fitch and Smith, 1983; Vingron and Waterman, 1994). Gotoh (1990) also studied the effects of variation of gap penalties. Lesk *et al.* (1986) pointed out that in globin sequences deletions and insertions are infrequently observed to occur in the interiors of helical regions of proteins because of the importance of the stability of the structures of the packing of helix–helix interfaces, and tried to vary a gap penalty between helical regions and inter-helical and loop regions. Barton and Sternberg (1987) also showed the superiority of their secondary structure dependent alignment method under various gap penalties. Fischel-Ghodsian *et al.* (1990) modified a dynamic programming method to include predicted secondary structure information. On the other hand, Kanaoka *et al.* (1989) assigned larger gap penalties to the hydrophobic core.

Such improvements are important for obtaining better alignments. In any method, however, each residue correspondence in an alignment has essentially a different probability of being aligned. Such a probability, of course, depends on a scoring method including a gap penalty scheme and the values of parameters. Since the correct correspondences of residues between sequences are so critical in assessing and analyzing relationships between proteins, it will be useful to know the probability of each residue correspondence. Here we present a new method of alignment that is based on dynamic program-

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

ming and the probability of each residue correspondence, and which yields alignments consisting of highly probable correspondences of residues. This method can be employed with any scoring scheme, whether it is based on sequence information or structure information.

Materials and methods

Similarity for an alignment

Let us define $S(A_l)$ as a similarity score of a specific alignment A_l that is defined as

$$A_l \equiv \begin{bmatrix} \ldots & a_2 & a_3 & \phi & \phi & a_4 & \ldots \\ \ldots & \phi & b_3 & b_4 & b_5 & b_6 & \ldots \end{bmatrix} \quad (1)$$

ϕ means a deletion, and a_i and b_i are the ith amino acid of sequences a and b respectively in a protein alignment or the ith base in a DNA comparison. In this paper, we consider only the case where $S(A_l)$ can be defined as the simple sum of similarity scores of match/mismatch pairs and penalties for gaps as follows.

$$S(A_l) \equiv \sum_{\{(i,j)\in A_l\}} s(a_i,b_j) - (\text{penalty for gaps}) \quad (2)$$

$s(a_i,b_j)$ is a score for the correspondence of amino acids a_i and b_j, and is usually assumed to depend only on the types of these amino acids.

The maximum similarity alignment A of two sequences a and b is defined as an alignment with the maximum similarity, i.e.

$$\begin{aligned} &\text{maximum similarity alignment} \\ &\equiv A \text{ such that } S(A) = \max_l S(A_l) \end{aligned} \quad (3)$$

Formulations of the maximum similarity score $S(A)$ are given for some cases of Equation 2 in the Appendix.

Scoring matrix and log-odds

A scoring matrix $s(a,b)$ that is frequently used in protein homology search and alignment is one devised from observed data of amino acid replacements by Dayhoff *et al.* (1978). It is defined as a log-odds matrix multiplied by 10 as follows.

$$s(a,b) = 10 \log_{10}\left(\frac{M_{a,b}}{f_a}\right) \quad (4)$$

$M_{a,b}$ is an element of a mutation probability matrix for a given time interval and gives the probability that an amino acid of the type b in a sequence will change to a in the evolutionary process of this time interval. f_a is the composition of amino acid type a and gives the probability that an amino acid of type a will be found in a randomly shuffled sequence. This definition of the amino acid similarity score indicates the essential relationship between the similarity score for alignment and the transition matrix for substitution.

Karlin and Altschul (1990) proved in their study of local sequence similarity that as the length of a random sequence grows without bound, the frequency of amino acid a in any sufficiently high-scoring segment, especially in maximal segment, approaches $f_a \exp(s_a/T)$ with probability 1, where s_a is a score for the amino acid a to appear in a segment and T is a constant. In other words, the score s_a can be written in the form of log-likelihood ratio,

$$s_a = T \log\left(\frac{q_a}{f_a}\right) \quad (5)$$

where f_a is the frequency with which the amino acid a appears by chance and q_a is the amino acid's implicit target frequency. Therefore, any scoring matrix essentially corresponds to a log-odds matrix (Karlin and Altschul, 1990; Altschul, 1993), as Dayhoff *et al.* (1978) originally defined a scoring matrix as a log-odds matrix.

Thus, without losing any generality, it can be assumed that a scoring matrix can be represented in the form of a log-odds matrix, and then the total score $S(A_l)$ of alignment A_l corresponds to the log-likelihood ratio of the alignment. Therefore, the statistical weight of an alignment is proportional to $\exp(S(A_l)/T)$.

$$\text{Statistical weight of alignment } A_l = \exp\left(\frac{S(A_l)}{T}\right) \quad (6)$$

In the case of Dayhoff's PAM score of Equation 4, the constant T corresponds to

$$T = \frac{10}{\log_e 10} \quad (7)$$

In cases such as an identity scoring matrix and some other scoring methods, an appropriate value for T may not be defined *a priori* but may be determined empirically with some trials.

Partition function

Equation 6 indicates that the similarity score $S(A_l)$ may be treated as if it is negative energy. Thus a partition function Z can be defined as

$$Z \equiv \sum_l \exp\left(\frac{S(A_l)}{T}\right) \quad (8)$$

and the probability $P(A_l)$ of alignment A_l is calculated as

$$P(A_l) = \frac{1}{Z}\exp\left(\frac{S(A_l)}{T}\right) \quad (9)$$

The maximum similarity alignment that satisfies Equation 3 is, of course, identical to the most probable alignment:

$$\begin{aligned} \text{The most probable alignment} &\equiv A \text{ such that } P(A) = \max_l P(A_l) \\ &= A \text{ such that } S(A) = \max_l S(A_l) \end{aligned} \quad (10)$$

Probability alignment is devised on the basis of Equation 9. How to calculate the partition function Z is given in the Appendix.

Probabilities of residue–residue correspondences

Let us consider the match/mismatch probability of a given site pair. The probability $p(a_i,b_j)$ that two sites a_i and b_j correspond to each other in all feasible alignments can be represented by

$$p(a_i,b_j) = \frac{1}{Z} Z_{i-1,j-1} \exp\left(\frac{s(a_i,b_j)}{T}\right) Z'_{i+1,j+1} \quad (11)$$

where $Z'_{i+1,j+1}$ is the partition function for partial sequences of a consisting of a_{i+1} to a_m and b from b_{j+1} to b_n. m and n are the lengths of sequences a and b. Then, the probabilities

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

that residue a_i in sequence a or b_j in sequence b is deleted in another sequence are represented as follows.

$$p(a_i,\phi) = 1 - \sum_{j=1}^{n} p(a_i,b_j)$$

$$p(\phi,b_j) = 1 - \sum_{i=1}^{m} p(a_i,b_j) \qquad (12)$$

The most probable correspondences for any site in sequences can be calculated easily from Equations 11 and 12. However, it should be noted here that a set of such site pairs does not always satisfy the conditions needed to constitute an alignment. When the most probable residue correspondence for a_i is b_j, a_i is not always the most probable residue correspondence for b_j. Also the sequence order among such site pairs is not always compatible with an alignment.

Probability alignment

Let us consider the construction of an alignment that consists of the most probable correspondences. Such an alignment can be made by iteratively choosing a site pair with the maximum probability as follows.

(i) Set $i_1 = 1$, $i_2 = m$, $j_1 = 1$, and $j_2 = n$.
(ii) Calculate a site pair (a_i,b_j) such that $p(a_i,b_j) = \max_{i_1 \leq k \leq i_2, j_1 \leq l \leq j_2} p(a_k,b_l)$, $p(a_i,b_j) \geq p(a_i,\phi)$, and $p(a_i,b_j) \geq p(\phi,b_j)$.
(iii) If there is no such site pair, align ϕ to all sites of $i_1 \leq i \leq i_2$ and of $j_1 \leq j \leq j_2$.
(iv) If (a_i,b_j) is such a site pair, choose it as one of the residue–residue correspondences in the alignment. Then, repeat steps (ii)–(iv) to align the remaining segments until all the sites are aligned.

This alignment may include residue correspondences that do not correspond to the most probable one, and whose probabilities are not significantly high. Probabilities of residue correspondences indicate the reliabilities of the correspondences. Is there any threshold of probability for reliable residue correspondences?

Let us consider site pairs with $p(a_i,b_j) > 0.5$. The number of correspondences with $p(a_i,b_j) > 0.5$ for a given site a_i and for b_j is limited to at most one, because the total sum of the probabilities must be equal to 1 (see Equation 12). The value 0.5 is the minimum value to ensure that the number of site pairs with $p(a_i,b_j) >$ (threshold) is either one or zero for a given site a_i and for b_j.

The number of b_j and ϕ such that

$$p(a_i, \{b_j,\phi\}) > 0.5 = 1 \text{ or } 0 \qquad (13)$$

Also, from Equation 13, the following condition is derived.

If $p(a_i,b_j) > 0.5$, then

$$p(a_i,b_j) = \max(\max_l p(a_i,b_l), p(a_i,\phi))$$

$$= \max(\max p(a_k,b_j), p(\phi,b_j)) \qquad (14)$$

That is, a site pair (a_i,b_j) with $p(a_i,b_j) > 0.5$ is the most probable correspondence for a given site a_i and for b_j.

However, the condition of Equation 13 is not sufficient to say that site pairs with $p(a_i,b_j) > 0.5$, with $p(a_i,\phi) > 0.5$, and with $p(\phi,b_j) > 0.5$ can constitute an alignment. In addition, the sequence order among residue correspondences must be compatible with an alignment, that is, the following condition must be satisfied for a set of site pairs to be able to constitute an alignment.

Lemma: let $p(a_i,b_j) > 0.5$ and $p(a_k,b_l) > 0.5$. If $i < k$, then $j < l$.

Proof: any alignment with the match/mismatch pair of a_i and b_j cannot have any match/mismatch pair of a_k and b_l with $i < k$ and $j \geq l$. Thus, if $p(a_i,b_j) > 0.5$, then $\sum_{l=1}^{j} p(a_k,b_l) < 0.5$ for $i < k$. Therefore, when $p(a_i,b_j) > 0.5$ and $p(a_k,b_l) > 0.5$, if $i < k$, then $j < l$.

Thus, all correspondence with $p(a_i,b_j) > 0.5$, with $p(a_i,\phi) > 0.5$, and with $p(\phi,b_j) > 0.5$ can constitute an alignment, and therefore are highly probable correspondences in the probability alignment that is constructed by the procedure already described in this section.

Since we are interested here in highly probable correspondences, probability alignments mean only correspondences with probabilities of >0.5, unless explicitly stated.

Results

The present method is applied to a few protein pairs to demonstrate the usefulness of the probability alignment method. The same scoring scheme is employed for both the maximum similarity alignment method and the probability alignment method, and is listed in Table I. Dayhoff's 250-PAM log-odd matrix (Dayhoff *et al.*, 1978), which is for the evolutionary distance of 250 accepted point mutations per 100 residues, is used as a scoring matrix (see Equation 4). T is a scaling factor for the Dayhoff's log-odd matrix (see Equation 7). A gap penalty scheme used here is the linear gap penalty scheme of Equation 31 that is described in Appendix, but the value of gap penalty is cut off at a certain value, satisfying Equation 21. A gap penalty is set to be smaller for terminal gaps than for gaps in the middle of a sequence. Because aligning termini together increases the number of possible alignments, a gap penalty scheme that is the same for all gaps makes large the probability of aligning termini together and often yields unrealistic correspondences between terminal residues of two sequences. Therefore, a smaller gap penalty as well as a smaller cut-off value is required for terminal gaps than for gaps in the middle of a sequence. Also, setting a gap penalty smaller for terminal gaps tends to yield better results even in maximum similarity alignments. However, it should be noted that if the gap penalty for terminal gaps becomes too small, a gap will often be inserted at termini in probability alignments. The algorithm used for maximum similarity alignments corresponds to Equations 22 and 23, and that for probability alignments corresponds to Equations 29 and 30; see the Appendix for these equations.

Table I. Scoring parameters used in both maximum similarity alignments and probability alignments

Parameter	Value
$s(a,b)$	Dayhoff's 250-PAM log-odd matrix[a]
$1/T$	0.23
$w(i,j-k,i,j) = w(i-k,j,i,j)$	$12 + 4(k-1)$ in the middle of a sequence; $6 + 2(k-1)$ at termini
Cut-off value of w	48 in the middle of a sequence; 24 at termini

[a]Dayhoff *et al.* (1978)

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

Table II. Protein pairs used in comparison of the probability alignment method with other alignment methods

Protein pair[a]	Length	Structural alignment[b]			Maximum similarity alignment						Probability alignment[c]					
		Identity[d]	Match/ mismatch[e]	r.m.s.d[f] (Å)	Identity[d]	Match/ mismatch[e]	Deletion/ addition[g]	Score	s.d.[h]	r.m.s.d[f] (Å)	Identity[d]	Match/ mismatch[e]	Deletion/ addition[g]	*T* log *Z*	s.d.[h]	r.m.s.d[f] (Å)
2HHB-A																
vs Lesk *et al.* (1986)	141 vs 153	23	138	3.2[i]	18/20	117/141	6/12	56	7.7		19/19	82/84	3/4	152.1	4.2	
vs GPYL2/JN0242[j]	141 vs 153				18/21	117/141	6/12	55	7.6		19/19	82/84	3/4	151.7	4.1	
vs 2LH4	141 vs 153				21/24	123/140	6/14	54	7.5	4.2	19/19	82/84	3/4	151.5	4.1	3.0
5CPV vs 3ICB	108 vs 75	19	68	4.2	18/20	51/69	38/45	43	6.3	4.6	18/20	41/43	38/40	98.5	6.9	2.1
3FAB VL vs CL	102 vs 102	14	89	4.2	13/22	62/99	3/6	13	3.9	9.0	11/15	41/49	3/7	90.9	2.6	4.3
1PCY vs 1AZU	99 vs 126	15	86	3.9	7/20	37/96	15/33	19	5.5	7.6	7/13	21/35	4/5	121.3	8.0	3.3
1RHD-A vs 1RHD-B	148 vs 145	16	124	4.1	8/25	39/139	1/15	8	4.0	14.1	6/18	24/71	1/3	128.3	2.6	12.5
3LZM vs 3LYZ	164 vs 129	6	88	6.7	0/3	0/23	93/270	−31	−0.0	5.3	0/5	0/10	1/2	84.6	0.8	10.3

[a]Unless stated, the protein codes are those in the Protein Data Bank (PDB).
[b]Taken from Taylor and Orengo (1989a) except for the alignment of Lesk *et al.* (1986).
[c]Only correspondences with probabilities of >0.5 are taken into account.
[d]The number of identical residues; for the maximum similarity and the probability alignments, the left number means correct residue correspondences in comparison with structural alignment.
[e]The number of residue–residue correspondences; for the maximum similarity and probability alignments, the left number means correct residue correspondences in comparison with the structural alignment.
[f]Root mean square deviation (Å) between matched/mismatched residues
[g]The number of deletions/additions; for the maximum similarity and probability alignments the left number means correct residue correspondences in comparison with the structural alignment
[h]Score in standard deviation units from the mean 100 Monte Carlo runs were employed to estimate the standard deviation and the mean of scores by chance.
[i]The coordinate set 2LH4 was used.
[j]Entry names in the Protein Information Resource (PIR).

A probability alignment approaches the maximum similarity alignment as the similarity between two sequences increases. Therefore, protein pairs that are dissimilar in sequence but similar in structure have been chosen as examples. These sequence pairs are listed in Table II. The percentages of identical residues in the alignments of these protein pairs are all < 21%; the proportion of identical residue correspondences is defined as (number of identical residues) · $2/(n + m)$, where n and m are sequence lengths. In Table II, sequence lengths, the numbers of identical residues, the numbers of residue–residue correspondences, the numbers of deletions and additions and total scores are also listed for these protein pairs.

The first example is the alignment of human α-haemoglobin and lupin leghaemoglobin. Lesk and Chothia (1980) compared the atomic structures of globins, and made alignments based on the three-dimensional structures of these molecules by superimpositions to determine which residues had similar relative spatial dispositions in the two structures. They pointed out that alignments based on sequence information alone tend to be different, especially in interhelical and turn regions, from those based on structural information in the case of distantly related globins. On the basis of such observations, Lesk *et al.* (1986) tried to improve an alignment method by varying gap penalties between helical regions and inter-helical and loop regions. The number of incorrect correspondences, in comparison with the alignment based on the three-dimensional structures, was decreased from 78 to 10 in the case of human α-haemoglobin and lupin leghaemoglobin by employing the variable gap penalty. Their results, namely, an alignment by structural superimposition and that with variable gap penalty, are shown in Figure 1A to allow comparison with a probability alignment. Figure 1A also shows maximum similarity alignments with uniform gap penalty and a probability alignment with the same uniform gap penalty scheme; only correspondences with probabilities of >0.5 are shown for the probability alignment. The uniform gap penalty scheme used here is a linear gap penalty scheme with a cut-off value rather than a constant gap penalty scheme in which a gap penalty does not depend on gap length; the values of the parameters in the linear gap penalty scheme are listed in Table I. In Figure 1A, alignments for three kinds of lupin leghaemoglobin sequences are shown; they are the amino acid sequence used by Lesk *et al.* (1986), the sequence of yellow lupin leghaemoglobin of entries GPYL2 and JN0242 found in the PIR database, and the leghaemoglobin sequence 2LH4 in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977), which are labelled as L1, L2 and L3, respectively. These sequences are different from each other by only one or two amino acids at the 85th or 86th position; the sequence used by Lesk *et al.* (1986) has A85 and S86; GPYL2 and JN0242 have V85 and S86; 2LH4 has V85 and T86.

As pointed out by Lesk *et al.* (1986), the alignment using the uniform gap penalty yields incorrect correspondences of residues in inter-helical regions between helices C and E and between helices E and F. In this parameter set, the maximum similarity method also yields incorrect correspondences in helix F for the leghaemoglobin lupin sequence used by Lesk *et al.* (1986), and for the yellow lupin leghaemoglobin II of entry GPYL2 and JN0242. In the case of the leghaemoglobin sequence 2LH4, helix F is correctly aligned but instead helix E is incorrectly aligned by this method. The different alignments of these regions for the sequences with single amino acid replacements indicate that this region of the alignments may be unreliable. In contrast, any of the residue correspondences in the interhelical region between E and F in the maximum similarity alignments are not assigned in the probability alignment, that is, the probability alignment indicates that these correspondences have probabilities of <0.5. Also it shows that the region between helices C and E cannot be aligned with probabilities of >0.5. Probabilities of all possible residue correspondences in alignments, which are calculated by Equations 11 and 12, are shown in Figure 1B. White plots represent probabilities of <0.1, while darker plots represent larger probabilities. Probabilities for deletions are shown at the zero residue position. These density plots clearly show the possibility of alternative alignments that are almost as probable as the maximum similarity alignment in some regions.

One of the interesting features of the probability alignment

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

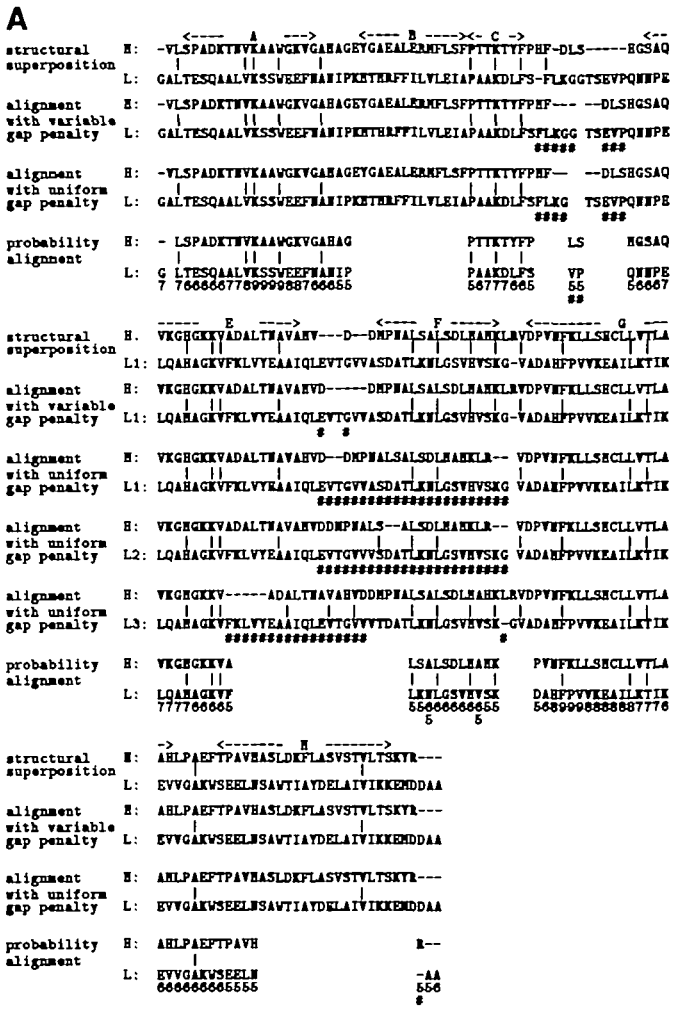

Fig. 1. (A) A comparison of the probability alignment method with other alignment methods; alignments of human α-haemoglobin (H) and lupin leghaemoglobin (L). The lupin leghaemoglobin sequence used by Lesk *et al.* (1986) is labelled as L1, the leghaemoglobin II sequence of entries GPYL2 and JN0242 in the PIR protein sequence database as L2, and the leghaemoglobin sequence of 2LH4 in the PDB protein structure database as L3. These sequences differ by only one amino acid at the 85th or 86th position; the sequence used by Lesk *et al.* (1986) has A85 and S86; GPYL2 or JN0242 has V85 and S86; 2LH4 has V85 and T86. 'Structural superposition' taken from Lesk *et al.* (1986) is based on superimpositions of the crystal structures. 'Alignment with variable gap penalty', also taken from Lesk *et al.* (1986), is based on sequence information alone, using variable gap penalties chosen to inhibit insertions or deletions within regions of secondary structures. α-helical regions A–H are taken from Lesk *et al.* (1986) and are shown at the top of the alignments. 'Alignment with uniform gap penalty' is the maximum similarity alignment based on sequence information alone with a linear gap penalty that is uniform over the entire sequences. 'Probability alignment' is based on the present method, and shows only residue correspondences that can make a unique alignment with probabilities of >0.5. The probability alignment is the same for all three sequences, L1, L2 and L3, except that the probabilities of residue correspondences vary slightly between them, especially in helix F. The numbers written under the sequence in the probability alignment represent probabilities with which the residue pairs are aligned; '5' means that the probability is >0.5 and <0.6. The gap penalty scheme used in both the maximum similarity alignments and the probability alignments is a linear gap penalty with a cut-off value; the values of these parameters are listed in Table I. Residue correspondences marked by # are those which do not agree with the alignment of structural superimposition. (B) Density plots of probabilities of residue correspondences in all feasible alignments between human α-haemoglobin (abscissa) and lupin leghaemoglobin 2LH4 (ordinate). Probabilities of all possible correspondences of residues were calculated using Equations 11 and 12 and are represented by different degrees of shading. Probabilities of deletions are shown at the zero residue position. White plots represent probabilities of <0.1, while darker plots represent larger probabilities.

shown in Figure 1A is that significant correspondences in the probability alignment do not change in spite of the single amino acid replacements. An even more interesting fact is that in the region of helix F the probability alignment does not yield the same incorrect correspondences as the maximum similarity alignment, but rather yields correct ones. This indicates that in the case of distantly related sequence pairs, correct residue correspondences may be better predicted on the basis of probability than by the maximum similarity alignment. More improvements of the probability alignment can be achieved by employing the variable gap penalty scheme, as Lesk *et al.* (1986) demonstrated.

The second example is the EF-hand structure, which consists of two α-helices flanking a loop and which was first found in parvalbumin to be a calcium binding motif typically associated with the calmodulin binding superfamily (Kretsinger, 1980). Parvalbumin has two recurrent motifs, which consist of helices C and D and helices E and F, and also a redundant motif consisting of helices A and B that does not bind a calcium ion (Moews and Kretsinger, 1975). Two EF-hand motifs have also been identified in the intestinal calcium binding protein (Szebenyi and Moffat, 1985). Taylor and Orengo (1989a) demonstrated that their method of structural alignment could align the correct ion binding motifs in both protein structures, ignoring the redundant motif of A and B helices.

The similarity score of the maximum similarity alignment

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

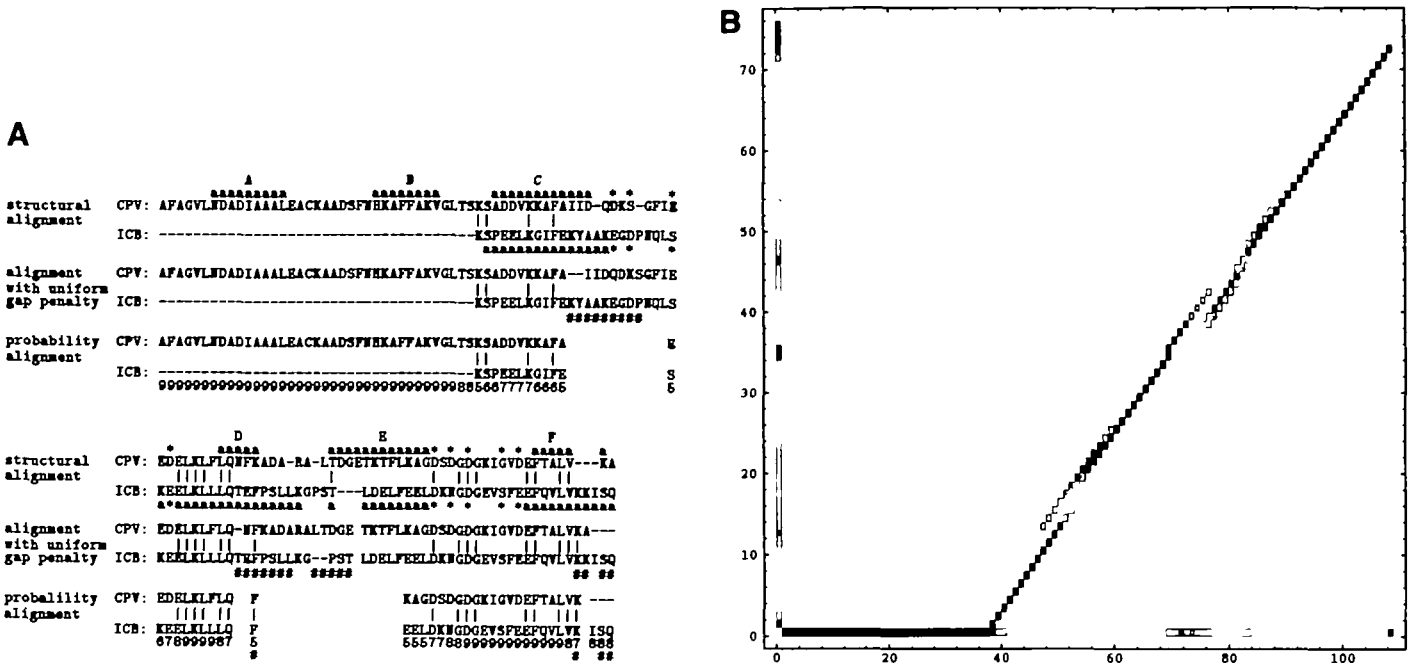

Fig. 2. (A) Alignments of calcium binding parvalbumin B (CPV) and intestinal calcium binding protein (ICB). 'Structural alignment' is taken from Taylor and Orengo (1989a) and is based on structural information alone, using their structural alignment method. α-helical regions are marked by 'a' and labelled from A to F; calcium binding residues are marked by asterisks; this information is also taken from Taylor and Orengo (1989a). Residue correspondences marked by # are those which do not agree with the structural alignment. See the legend of Figure 1A for other symbols. (**B**) Density plots of probabilities of residue correspondences in all feasible alignments between calcium binding parvalbumin B (abscissa) and intestinal calcium binding protein (ordinate). Probabilities of deletions are shown at the zero residue position. White plots represent probabilities of <0.1, while darker plots represent larger probabilities. See also the legend of Figure 1B.

with a uniform gap penalty for this protein pair is ~6 standard deviation units from the mean, which may be a lower limit of similarity score for homologous protein pairs, so that they belong to the 'twilight zone' in sequence comparison (Table II). The maximum similarity alignment, which is shown in Figure 2A, is correct except for the region of the C-terminal end of helix C through the N-terminal half of a loop region in the first motif and the inter-helical region between the first and second motifs. The probability alignment indicates that these regions may be incorrectly aligned in the maximum similarity alignment, although the deletion of the redundant motif consisting of helix A and B is highly probable, and helix C and the second motif consisting of helix E and F could be aligned with high probabilities as could the C-terminal half of the loop region in the first motif. In Figure 2B, the density plots of probabilities of residue correspondences indicate the possibility of alternative alignments which differ from each other with a few residues shifted.

The third example is the alignment of the variable region, V_λ, and the constant region, C_λ, of immunoglobulin light chain λ. The immunoglobulin light chain has two β-domains that are of two types, constant and variable regions. These domains are homologous to each other and also homologous to constant domains in immunoglobulin heavy chains. The core of each domain forms two β-sheets packed face-to-face consisting of 35 or 36 residues. The structural relationships of these domains have been well characterized by many researchers (Amzel and Poljak, 1979). Figure 3A shows the structural alignment of V_λ and C_λ domains done by Taylor and Orengo (1989a) and the assignment of residues to core β-sheets done by Lesk and Chothia (1982) for comparison with a probability alignment. Again, the probability alignment shown in Figure 3A indicates that most incorrect correspondences in the maximum similarity alignment have a probability of <0.5. The β-strands, B, C, F and G appear to be more conservative in their amino acid sequences than other β-strands, because correct correspondences are predicted with high probabilities for these regions in the probability alignment. This is consistent with the facts that β-strands, B, C, E, F, G and H are common to all structures of constant and variable regions (Taylor and Orengo, 1989a), and that only three residues common to all domains of V_κ, V_λ, V_γ, C_λ, $C_\gamma 1$, $C_\gamma 2$ and $C_\gamma 3$, (Lesk and Chothia, 1982) are two cysteines that form an intra-molecular disulfide bridge between strands B and G, and a tryptophan in strand C that packs against them. Hypervariable regions, which correspond to a loop between β-strands B and C, a region including a short β-strand D between C and E, and another loop between G and H, are all not aligned with probabilities of > 0.5.

All these examples demonstrate that highly probable correspondences of residues will tend to be correct correspondences, if a scoring scheme is appropriate. This is also indicated by the values of root mean square deviation (r.m.s.d.) in superimposition of corresponding residues of the protein structures (see Table II). The r.m.s.d. values are always smaller for significant residue–residue correspondences with probabilities of >0.5 in the probability alignments of these protein pairs than for all correspondences in the maximum similarity alignments, indicating that incorrect correspondences in the maximum similarity alignments tend to be insignificant correspondences with probabilities of <0.5 in the probability alignments. To confirm this fact further, the same analysis was done for 109 protein pairs whose percent identities are between 90 and 35% in the PDB. The r.m.s.d. values are plotted in Figure 4. The abscissa represents the r.m.s.d. values for corresponding residues in the maximum similarity alignments, and the ordinate those for highly probable correspondences in the probability

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

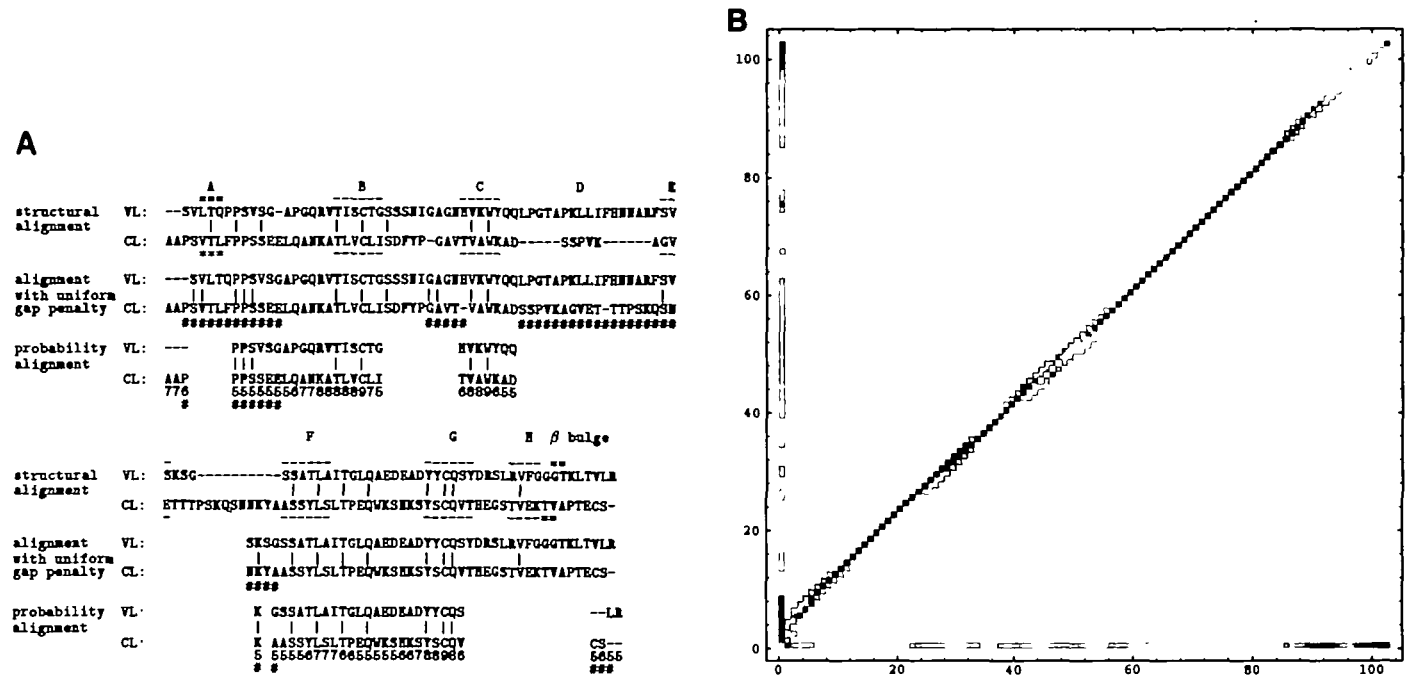

Fig. 3. (**A**) Alignments of immunoglobulin light chain λ [FAB(New)] variable (V_L) and constant (C_L) regions. 'Structural alignment' is taken from Taylor and Orengo (1989a). The labels of β-strands A–H, and β-bulge are the same designation as their topological equivalent in the C_H2 structure of the immunoglobulin heavy chain constant region, and are taken from Taylor and Orengo (1989a). Minus and equals signs represent the core regions aligned by Lesk and Chothia (1982). See legend of Figure 1A for other symbols. (**B**) Density plots of probabilities of residue correspondences in all feasible alignments between immunoglobulin light chain λ variable (abscissa) and constant (ordinate) regions. Probabilities of deletions are shown at the zero residue position. White plots represent probabilities of <0.1, while darker plots represent larger probabilities. See also the legend of Figure 1B.

alignments. This figure clearly shows that the r.m.s.d. values for the probability alignments tend to be lower than those for the maximum similarity alignments. Figure 4 also shows that probability alignments still yield relatively large r.m.s.d. values for some protein pairs that would result from incorrect residue–residue correspondences. Here it should be noted that even a small number of incorrect correspondences can cause such large r.m.s.d. values, ~6 Å.

The next example is the alignment of two copper binding proteins, plastocyanin and azurin, which consist of two β-sheets. Their structural alignment, done by Taylor and Orengo (1989a), is shown in Figure 5 as well as the probability alignment. The structural alignment includes regions that differ with minor insertions and deletions and displacements of a few residues from the structure alignment obtained from structural superimposition by Chothia and Lesk (1982). The probability alignment agrees with both the alignments only for two β-strands that are located at the C-terminus and constitute the second β-sheet (Chothia and Lesk, 1982), and a short β-strand that is located at the N-terminus and constitutes the first β-sheet. Figure 5 also shows the probabilities with which correct residue pairs are aligned. Only ~60% of the residue–residue correspondences agree with the Taylor and Orengo (1989a) alignment (see Table II). This proportion is significantly smaller than those for the three probability alignments already discussed, which are all >80%. However, the quality of the alignment is improved by the probability alignment method, because the value 60% is significantly larger than the proportion, 39%, of correct residue–residue correspondences in the maximum similarity alignment of this protein pair; in the case of immunoglobulin variable and constant domains, the proportion of correct residue–residue correspondences is 63% in the maximum similarity alignment and 84% in the probability alignment. The improvement by the probability alignment is also indicated by the fact that the large r.m.s.d. value (7.6 Å) for the maximum similarity alignment is improved to 3.3 Å for the probability alignment. This value, 3.3 Å, which is even better than the r.m.s.d. for the structural alignment (3.9 Å), may be too good to be expected from the fact that only 60% of residue–residue correspondences are correct in the probability alignment. Such a small r.m.s.d. value may be attained because the differences between the structural alignment and the probability alignment are mostly displacements of a few residues.

The probability alignment method will of course predict incorrect residue correspondences to be highly probable if the scoring scheme or the values of scoring parameters are inappropriate. Also, even if these are appropriate, the probable correspondences in a probability alignment can be incorrect if two sequences are distantly related. However, in the case of distantly related proteins, the number of significant residue correspondences can be limited; if so, it will not be a serious problem. The alignment of hen egg lysozyme (3LYZ) and T4 lysozyme (3LZM) is such an example. Only 10 residue–residue correspondences, 7% of the sequences, have probabilities of >0.5, and none of them are correct (see Table II).

On the other hand, the alignment of the N-terminal half (1RHD-A) and the C-terminal half (1RHD-B) of rhodanese indicates that the present scoring scheme or the employed values of the parameters such as the gap penalties and the scoring matrix may be inappropriate for aligning these sequences. Seventy-one residue–residue correspondences, ~48% of the sequences, are predicted to be highly probable in the probability alignment, but only 24 of 71 residue correspondences are correct, yielding a r.m.s.d. value of >10 Å (see Table II). The proportion of correct residue–residue

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

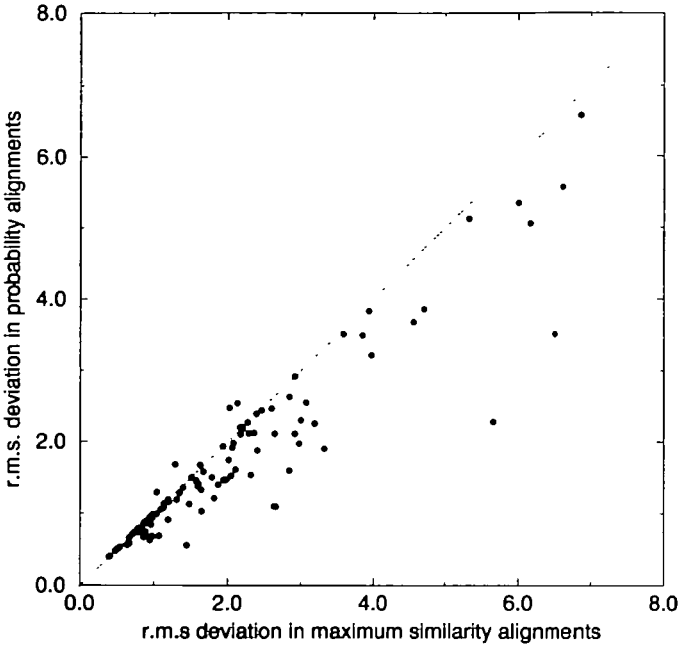

Fig. 4. Root mean square deviations (in Å) between corresponding residues in both maximum similarity alignments and probability alignments. Root mean square deviations for probability alignments were calculated by using only correspondences with probabilities of >0.5. The abscissa represents the values of r.m.s. deviation between corresponding residues in the maximum similarity alignments and the ordinate that in the probability alignments. A dotted line in this figure shows a line with equal values of r.m.s. deviation. The protein pairs used here were selected by the following procedure. Representative protein structures that differ from each other by at least 35% sequence identity in the PDB were selected by Orengo *et al.* (1993). They were used here to pick up protein structures from the PDB that were similar to one of the representatives with sequence identities between 35 and 90% but not more similar to each other than 95% identity. In this procedure, structures determined by NMR and structures whose resolution is <2.5 Å are removed from the set of proteins. Also, the pair of crystallin 1GCR and 2BB2 picked up by this procedure is not shown in this figure, because the relative orientation between domains is completely different between these structures. One hundred and nine protein pairs selected by this procedure were used here: 1ACX,1NOA; 1ACX,2MCM; 1ALD,1FBA-A; 1COB-A,1SDY-A; 1COB-A,1SOS-A; 1CSE-E,1MEE-A; 1CSE-E,1SO1; 1CSE-E,1TEC-E; 1CSE-E,2SIC-E; 1F3G,1GPR; 1FKF,1YAT; 1FXI-A,1FXA-A; 1FXI-A,3FXC; 1GCR,2GCR; 1GCR,3GCR-A; 1HBB-A,1FDH-G; 1HBB-A,1HDS-A; 1HBB-A,1HDS-B; 1HBB-A,1PBX-A; 1HBB-A,2HHB-B; 1HBB-A,2MHB-A; 1HBB-A,2MHB-B; 1LZ1,1ALC; 1LZ1,1FDL-Y; 1LZ1,2LZ2; 1MBC,1MBS; 1MBC,1PMB-A; 1PCY,7PCY; 1PI2,1TAB-I; 1R69,2CRO; 1TGS-I,1CGI-I; 1YCC,1CCR; 1YCC,1CYC; 1YCC,1YEA; 1YCC,2C2C; 2AZA-A,1AZR-A; 2CDV,1CTH-A; 2ER7-E,3APP; 2ER7-E,4APR-E; 2FB4-H,1FDL-H; 2FB4-H,1FVC-B; 2FB4-H,1FVD-B; 2FB4-H,1HIL-B; 2FB4-H,1IGJ-B; 2FB4-H,1IGM-H; 2FB4-H,1MAM-H; 2FB4-H,1NCB-H; 2FB;4-H,2FBJ-H; 2FB4-H,3FAB-H; 2FB4-H,6FAB-H; 2FB4-H,8FAB-B; 2FCR,1FLV; 2FCR,1OFV; 2HMZ-A,2MHR; 2LIV,2LBP; 2LTN-A,1LTE; 2OVO,1OVO-A; 2RHE,1FVC-A; 2RHE,1IGM-L; 2RHE,1REI-A; 2RHE,2FB4-L; 2RHE,2IMM; 2RHE,3FAB-L; 2RHE,3MCG-I; 2RHE,8FAB-A; 2SGA,3SGB-E; 2TS1,4TS1-A; 2TSC-A,4TMS; 3EBX,2ABX-A; 3EBX,2CTX; 3PSG,1RNE; 3PSG,1SMR-A; 3PSG,4APR-E; 3PSG,4CMS; 3SGB-I,1OVO-A; 4BP2,1BBC; 4BP2,1POB-A; 4BP2,1PP2-R; 4BP2,1PPA; 4BP2,4P2P; 4BP2,5P2P-A; 4CPV,1OMD; 4CPV,1PAL; 4CPV,1RTP-1; 4CPV,5PAL; 4PFK,2PFK-A; 4PTP,1CHO-E; 4PTP,1EST; 4PTP,1HGT-H; 4PTP,1TON; 4PTP,1TRM-A; 5HVP-A,1IVP-A; 5PTI,1AAP-A; 5PTI,1DTX; 5RXN,1CAA; 5RXN,1RDG; 5RXN,6RXN; 5RXN,7RXN; 5TIM-A,1TIM-A; 5TIM-A,1YPI-A; 6LDH,1LLD-A; 6LDH,9LDB-A; 6XIA,1XLA-A; 6XIA,4XIM-A; 7AAT-A,1SPA; 8DFR,2DHF-A; 8I1B,5I1B; 9RNT,1FUS; 9RNT,1RMS.

correspondences is hardly improved in the probability alignment: it is 28% in the maximum similarity alignment and 34% in the probability alignment. The structural alignment of these

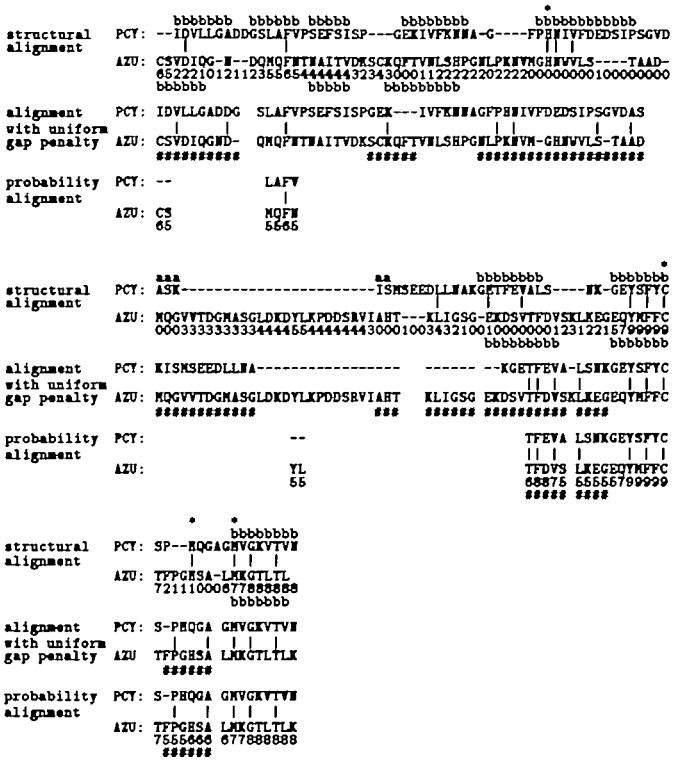

Fig. 5. Alignments of plastocyanin (PCY) and azurin (AZU). 'Structural alignment' is taken from Taylor and Orengo (1989a). β-strands from sheets I and II (Lesk and Chothia, 1982a) are labelled b (Taylor and Orengo, 1989a). Asterisks indicate the four copper ligands. The numbers written under the sequence in the structural alignment represent probabilities with which those correct residue pairs are aligned; '5' means that the probability is >0.5 and <0.6. See the legend of Figure 1A for other symbols.

sequences done by Taylor and Orengo (1989a) indicates that amino acid substitutions between these sequences are relatively less conservative in respect to the physico-chemical properties of amino acids. It is hard to predict the correct alignment of such proteins.

The score of these probability alignments may be measured by T log Z, which corresponds to 'negative free energy'. The significance of a maximum similarity alignment to other alignments can be measured by the difference between scores of a probability alignment, T log Z, and a maximum similarity alignment, $S(A)$. The more probable a maximum similarity alignment is, the smaller the difference between T log Z of a probability alignment and $S(A)$ of the maximum similarity alignment.

As the statistical significance of an alignment in the maximum similarity method is measured by representing its score from the mean in standard deviation units, the significance of a probability alignment may be measured by representing the value of T log Z in standard deviation units from the mean in the distribution of scores of randomly shuffled sequences. Table II lists such values; the mean and standard deviation in random comparison were calculated by 100 Monte Carlo runs of shuffling one sequence and aligning the sequences.

Discussion

It is usual that alignments are reported with a significance level but without mentioning which regions are more reliable or less reliable. The effects of amino acid replacements on protein structure are not uniform over a sequence: amino acid variability depends on residue position. It is well known that,

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

on average, residues are less conserved in open loops than in regular secondary structures, and more conserved in the interior of proteins than on the surface (Go and Miyazawa, 1980). The more that residues have been replaced between two sequences, the harder it is to align the two sequences. Thus, an alignment is less reliable in more variable regions. Since correct correspondences between residues are critical in assessing and analysing the relationship between proteins, it is useful to know how reliable each correspondence in an alignment is.

Here a method that evaluates probabilities of all possible correspondences between residues and yields an alignment consisting of highly probable correspondences has been reported. The present method is based on the fact that any scoring matrix essentially corresponds to a log-odds matrix (Karlin and Altschul, 1990; Altschul, 1993) and, therefore, that in any scoring system the statistical weight of an alignment can be proportional to the exponent of its total score (see Equations 6 and 7). Then, the probabilities of all possible correspondences can easily be calculated. Similar methods, which are based on dynamic programming but in which the concept of the most probable state is employed rather than the minimum energy form, are found in many fields such as secondary structure prediction (Jernigan *et al.*, 1980), prediction of protein folding pathways (Miyazawa and Jernigan, 1982a,b) and a search for the most stable folds of protein chains (Finkelstein and Reva, 1991).

In the case of sufficiently similar sequences, a maximum similarity alignment corresponds to a sharp maximum on score surface, and can be a good prediction for alignment. In the case of such similar sequences, a probability alignment tends to agree with the maximum similarity alignment. However, when distantly related sequence pairs are aligned, the score surface becomes smooth at the maximum and the maximum similarity alignment becomes one of many possible alignments. Under such conditions, the probability alignment can be a better prediction than the maximum similarity alignment. Figure 1 shows such a case. The significance of the maximum similarity alignment to other alignments can be measured by the difference between the scores, $T \log Z$ of a probability alignment and $S(A)$ of a maximum similarity alignment. Large differences between those quantities indicate that alternative alignments cannot be ignored.

The present results of probability alignments clearly indicate that incorrect correspondences of residues in maximum similarity alignments often belong to unreliable regions in the alignments and therefore are likely to be predicted as insignificant correspondences in probability alignments. These characteristics of the probability alignment method are useful especially for analyses where correct residue correspondences are critical, such as protein structure predictions starting from the known structures of homologous proteins. Here it should be noted that the probabilities of residue correspondences depend on parameters such as gap penalties and the scoring matrix used. The present method cannot overcome the difficult problem of appropriately setting these parameters. However, in general, regions in an alignment that are sensitive to the values of gap penalties often belong to unreliable regions in the alignment. Therefore, highly probable correspondences in probability alignments would be relatively less sensitive to the values of gap penalties.

Local homology search methods (Smith and Waterman, 1981; Boswell and McLachlan, 1984) may be used to find significantly similar segments between sequences and to align them. However, similar segment pairs that are found by these methods may overlap each other and their sequence order may be incompatible with making an alignment. Therefore it is not usually possible to arrange these similar segments in an alignment. In the probability alignment method, only residue correspondences that are common in most of the possible global alignments are picked up. Residue correspondences that are more probable than 0.5 do not overlap each other and also satisfy the sequence order compatible with an alignment. Thus, the present probability alignment method would be better than local homology search methods for making a global sequence alignment consisting of highly probable residue correspondences.

In this paper we have shown and discussed only correspondences with probabilities of >0.5 in probability alignments. If necessary, entire alignments based on probabilities of correspondences could be constructed according to the procedure described in Materials and methods. The entire alignments include portions that consist of less probable correspondences and therefore may be unreliable. Such portions are likely to disagree with maximum similarity alignments.

The probability method has been applied here to a global alignment method with a scoring scheme based on sequence information alone, but it can be used for any dynamic programming method with any scoring scheme.

References

Altschul,S.F. (1993) *J. Mol. Evol.*, **36**, 290–300.
Amzel,L.M. and Poljak,R.J. (1979) *Annu. Rev. Biochem.*, **48**, 961–997.
Barton,G.J. and Sternberg,M.J.E. (1987) *Protein Engng*, **1**, 89–94.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Boswell,D.R. and McLachlan,A.D. (1984) *Nucleic Acids Res.*, **12**, 457–464.
Chothia,C. and Lesk,A.M. (1982) *J. Mol. Biol.*, **160**, 309–323.
Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C., pp. 345–352.
Finkelstein,A.V. and Reva,B.A. (1991) *Nature*, **351**, 497–499.
Fischel-Ghodsian,F., Mathiowitz,G. and Smith,T.F. (1990) *Protein Engng*, **3**, 577–581.
Fitch,W.M. and Smith,T.F. (1983) *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.
Go,M. and Miyazawa,S. (1980) *Int. J. Peptide Protein Res.*, **15**, 211–224.
Gotoh,O. (1990) *Bull. Math. Biol.*, **52**, 359–373.
Jernigan,R.L., Miyazawa,S. and Szu,S.C. (1980) *Macromolecules*, **13**, 518–525.
Kanaoka,M., Kishimoto,F., Ueki,Y. and Umeyama,H. (1989) *Protein Engng*, **2**, 347–351.
Karlin,S. and Altschul,S.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
Kretsinger,R.H. (1980) *CRC Crit. Rev. Biochem.*, **8**, 119–174.
Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–342.
Lesk,A.M., Levitt,M. and Chothia,C. (1986) *Protein Engng*, **1**, 77–78.
Luo,Y., Lai,L., Xu,X. and Tang,Y. (1993) *Protein Engng*, **6**, 373–376.
Miyazawa,S. and Jernigan,R.L. (1982a) *Biopolymers*, **21**, 1333–1363.
Miyazawa,S. and Jernigan,R.L. (1982b) *Biochemistry*, **21**, 5203–5213.
Moews,P.C. and Kretsinger,R.H. (1975) *J. Mol. Biol.*, **91**, 201–228.
Needleman,S.B. and Wunsch,C.B. (1970) *J. Mol. Biol.*, **48**, 443–453.
Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.
Rao,S.T. and Rossmann,M.G. (1973) *J. Mol. Biol.*, **76**, 241–256.
Sellers,P.H. (1974) *SIAM J. Appl. Math.*, **26**, 787–793.
Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
Subbarao,N. and Haneef,I. (1991) *Protein Engng*, **4**, 877–884.
Szebenyi,D.M.E. and Moffat,K. (1985) *Nature*, **294**, 327–332.
Taylor,W.R. and Orengo,C.A. (1989a) *J. Mol. Biol.*, **208**, 1–22.
Taylor,W.R. and Orengo,C.A. (1989b) *Protein Engng*, **2**, 505–519
Vingron,M. and Waterman,M.S. (1994) *J. Mol. Biol.*, **235**, 1–12.

Received October 17, 1994; revised May 8, 1995; accepted June 27, 1995

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

Appendix

Maximum similarity alignment

Let us consider a case that a similarity score of alignment A_l can be defined as a simple sum of similarity scores of match/mismatch pairs and penalties for gaps as follows.

$$S(A_l) = \sum_{\{(i,j)\in A_l\}} s(a_i,b_j) - \sum_{\{(i,j)\in A_l\}} w(i,j-k,i,j) - \sum_{\{(i,j)\in A_l\}} w(i-k,j,i,j) \quad (15)$$

The first term corresponds to the total sum of match/mismatch scores, and the sums in the second and the third terms are taken over all gaps in the sequence a and b, respectively, $s(a_i,b_j)$ is a score for the correspondence of amino acids a_i and b_j, and $w(i,j-k,i,j)$ is a penalty for an addition of a segment of b_{j-k+1} to b_j or a deletion of the segment at the position next to a_i. In this case, the maximum similarity score $S_{i,j}$ for the partial sequences of a consisting of a_1 to a_i and b from b_1 to b_j can be calculated as follows.

$$A_{i,j} = \max_{1\le k\le j} \{\max\{B_{i,j-k},C_{i,j-k}\} - w(i,j-k,i,j)\}$$

$$B_{i,j} = \max_{1\le k\le i} \{\max\{A_{i-k,j},C_{i-k,j}\} - w(i-k,j,i,j)\}$$

$$C_{i,j} = S_{i-1,j-1} + s(a_i,b_j)$$

$$S_{i,j} = \max\{A_{i,j}, B_{i,j}, C_{i,j}\} \quad (16)$$

with the following boundary conditions:

$$\begin{aligned} S_{0,0} &= C_{0,0} = 0 \\ S_{0,j} &= A_{0,j} = -w(0,0,0,j) \\ S_{i,0} &= B_{i,0} = -w(0,0,i,0) \end{aligned} \quad (17)$$

where $A_{i,j}$, $B_{i,j}$, and $C_{i,j}$ correspond to the maximum similarity score for each of the alignments $\begin{smallmatrix} a_i \\ \cdots \\ \Phi \\ b_j \end{smallmatrix}$, $\begin{smallmatrix} a_i \\ \cdots \\ \Phi \\ \Phi \end{smallmatrix}$, and $\begin{smallmatrix} \Phi \\ \cdots \\ b_j \end{smallmatrix}$.

If the following conditions are satisfied for any positive integers of k and l,

$$\begin{aligned} w(i,j-k-l,i,j) &\le w(i,j-k-l,i,j-k) + w(i,j-k,i,j) \\ w(i-k-l,j,i,j) &\le w(i-k-l,j,i-k,j) + w(i-k,j,i,j) \end{aligned} \quad (18)$$

then Equations 16 and 17 can be simplified as follows

$$S_{i,j} = \max\{\max_{1\le k\le j}\{S_{i,j-k} - w(i,j-k,i,j)\}, \max_{1\le k\le i}\{S_{i-k,j} - w(i-k,j,i,j)\}, S_{i-1,j-1} + s(a_i,b_j)\} \quad (19)$$

$$S_{0,0} = 0 \quad (20)$$

Equation 18 corresponds to the assumption that a gap penalty is a convex function of gap length. This assumption is biologically reasonable. However, it should be noted here that Equations 19 and 20 can be justified without the assumption of Equation 18, if it is assumed that a gap with length $k + l$ can occur with a single mutational event with a gap penalty $w(i,j-k-l,i,j)$ and also can occur with double mutations such as deletions or additions of length l with a penalty $w(i,j-k-l,i,j-k)$ and of length k with a penalty $w(i,j-k,i,j)$. However, if this is assumed but Equation 18 is not satisfied, that is, if a gap penalty $w(i,j-k,i,j)$ is defined to be a penalty for a single mutation and gap penalties for multiple mutations do not satisfy Equation 18, Equations 16 and 17 will no longer be correct.

Equation 16 and 19 are both $O(nm^2) + O(n^2m)$ algorithms; n and m are the sequence lengths of a and b. An algorithm of $O(nm)$ is well known for the case of a linear gap penalty scheme in which a gap penalty is assumed to be proportional to the gap length. In the case where there is an upper bound for gap penalty, an $O(nm)$ algorithm is also available. That is, if

$$\begin{aligned} &\exists k_b \text{ such that } w(i,j-k,i,j) = w(i,j-k_b,i,j) \text{ for } k \ge k_b \\ &\exists k_a \text{ such that } w(i-k,j,i,j) = w(i-k_a,j,i,j) \text{ for } k \ge k_a \end{aligned} \quad (21)$$

the maximum score can be calculated as follows.

$$\begin{aligned} E_{i,j} &= \max\{E_{i,j-1}, S_{i,j}\} \\ F_{i,j} &= \max\{F_{i-1,j}, S_{i,j}\} \\ S_{i,j} &= \max\{\max_{1\le k<k_b}\{S_{i,j-k} - w(i,j,-k,i,j)\}, \\ &\quad E_{i,j-k_b} - w(i,j-k_b,i,j), \max_{1\le k<k_a}\{S_{i-k,j} - w(i-k,j,i,j)\}, \\ &\quad F_{i-k_a,j} - w(i-k_a,j,i,j), S_{i-1,j-1} + s(a_i,b_j)\} \end{aligned} \quad (22)$$

$$S_{0,0} = E_{0,0} = F_{0,0} = 0 \quad (23)$$

Partition function

A partition function for the scoring scheme of Equation 15, which corresponds to Equation 16 of the maximum similarity, can be derived as follows by changing the maximum operations in Equation 16 to summations.

$$Z^A_{i,j} = \sum_{k=1}^{j} (Z^B_{i,j-k} + Z^C_{i,j-k}) \exp\left(-\frac{w(i,j-k,i,j)}{T}\right)$$

$$Z^B_{i,j} = \sum_{k=1}^{i} (Z^A_{i-k,j} + Z^C_{i-k,j}) \exp\left(-\frac{w(i-k,j,i,j)}{T}\right)$$

$$Z^C_{i,j} = Z_{i-1,j-1} \exp\left(\frac{s(a_i,b_j)}{T}\right)$$

$$Z_{i,j} = Z^A_{i,j} + Z^B_{i,j} + Z^C_{i,j} \quad (24)$$

with the following boundary conditions

$$Z_{0,0} = Z^C_{0,0} = 1$$

$$Z_{0,j} = Z^A_{0,j} = \exp\left(-\frac{w(0,0,0,j)}{T}\right)$$

$$Z_{i,0} = Z^B_{i,0} = \exp\left(-\frac{w(0,0,i,0)}{T}\right) \quad (25)$$

This is an $O(nm^2) + O(n^2m)$ algorithm like Equation 16. In contrast to the case of Equation 16, in which the additional assumption of Equation 18 yields the simpler expression of Equation 19 and the further assumption of Equation 21 yields the $O(nm)$ algorithm of Equation 22, these assumptions do not help to get a simpler expression and a faster algorithm, corresponding to Equations 19 and 22, for calculating a partition function. However, an alternative interpretation of gap formation, which can justify Equation 19 without the assumptions of Equation 18, can yield equations corresponding

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013

to Equations 19 and 22. In the scheme of Equation 24, we assumed that the probability of a gap with length k occurring is proportional to the Boltzmann factor of a gap penalty $w(i,j-k,i,j)$. Now let us assume instead that a gap with length $k+l$ can occur with a single mutational event whose occurrence probability is proportional to the Boltzmann factor of the gap penalty $w(i,j-k-l,i,j)$, and also occur with double mutations such as deletions or additions of length k and of length l. In this case, however, Equations 24 and 25 become no longer correct, but the following, simpler equation, which corresponds to Equation 19, is instead satisfied. If

$$w(i,j-k,i,j) \equiv \text{gap penalty for a single mutational event} \quad (26)$$

then

$$Z_{i,j} = \sum_{k=1}^{j} Z_{i,j-k} \exp\left(-\frac{w(i,j-k,i,j)}{T}\right) + \sum_{k=1}^{i} Z_{i-k,j} \exp\left(-\frac{w(i-k,j,i,j)}{T}\right) + Z_{i-1,j-1} \exp\left(\frac{s(a_i,b_j)}{T}\right) \quad (27)$$

with the boundary condition

$$Z_{0,0} = 1 \quad (28)$$

Here it should be noticed that the possibility of multiple mutations is explicitly taken into account to derive Equations 27 and 28.

If Equation 21 is satisfied, we can get the following equations corresponding to Equations 22 and 23.

$$Z^E_{i,j} = Z^E_{i,j-1} + Z_{i,j}$$
$$Z^F_{i,j} = Z^F_{i-1,j} + Z_{i,j}$$
$$Z_{i,j} = \sum_{k=1}^{k_b-1} Z_{i,j-k} \exp\left(-\frac{w(i,j-k,i,j)}{T}\right) + Z^E_{i,j-k_b} \exp\left(-\frac{w(i,j-k_b,i,j)}{T}\right) + \sum_{k=1}^{k_a-1} Z_{i-k,j} \exp\left(-\frac{w(i-k,j,i,j)}{T}\right) + Z^F_{i-k_a,j} \exp\left(-\frac{w(i-k_a,j,i,j)}{T}\right) + Z_{i-1,j-1} \exp\left(\frac{s(a_i,b_j)}{T}\right) \quad (29)$$

$$Z_{0,0} = Z^E_{0,0} = Z^F_{0,0} = 1 \quad (30)$$

In the case of a linear gap penalty scheme, that is,

$$w(i,j-k,i,j) = w(i-k,j,i,j) = \alpha + \beta(k-1) \quad (31)$$

where α and β are constants, Equations 27 and 28 can be simplified as follows.

$$Z^A_{i,j} = \begin{cases} Z_{i,j-1} \exp(-\frac{\alpha}{T}) & \text{for } j = 1 \\ Z_{i,j-1} \exp(-\frac{\alpha}{T}) + Z^A_{i,j-1} \exp(-\frac{\beta}{T}) & \text{for } j > 1 \end{cases}$$

$$Z^B_{i,j} = \begin{cases} Z_{i-1,j} \exp(-\frac{\alpha}{T}) & \text{for } i = 1 \\ Z_{i-1,j} \exp(-\frac{\alpha}{T}) + Z^B_{i-1,j} \exp(-\frac{\beta}{T}) & \text{for } i > 1 \end{cases}$$

$$Z_{i,j} = Z_{i-1,j-1} \exp\left(\frac{s(a_i,b_j)}{T}\right) + Z^A_{i,j} + Z^B_{i,j} \quad (32)$$

$$Z_{0,0} = 1$$
$$Z_{0,j} = Z^A_{0,j}$$
$$Z_{i,0} = Z^B_{i,0} \quad (33)$$

Note added in proof

The program used here for probability alignments is available through our e-mail server; send an empty mail to flat_netsev@smlab.eg.gunma_u.ac.jp for a brief manual.

Downloaded from http://peds.oxfordjournals.org/ at Harvard University on March 12, 2013