J. Mol. Biol. (1996) **264**, 823–838

JMB

Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments

Osamu Gotoh

Department of Biochemistry Saitama Cancer Center Research Institute, 818 Komuro, Ina-machi, Saitama 362, Japan

The relative performances of four strategies for aligning a large number of protein sequences were assessed by referring to corresponding structural alignments of 54 independent families. Multiple sequence alignment of a family was constructed by a given method from the sequences of known structures and their homologues, and the subset consisting of the sequences of known structures was extracted from the whole alignment and compared with the structural counterpart in a residue-to-residue fashion. Gap-opening and -extension penalties were optimized for each family and method. Each of the four multiple alignment methods gave significantly more accurate alignments than the conventional pairwise method. In addition, a clear difference in performance was detected among three of the four multiple alignment methods examined. The currently most popular progressive method ranked worst among the four, and the randomized iterative strategy that optimizes the sum-of-pairs score ranked next worst. The two best-performing strategies, one of which was newly developed, both pursue an optimal weighted sum-of-pairs score, where the pair weights were introduced to correct for uneven representations of subgroups in a family. The new method uses doubly nested iterations to make alignment, phylogenetic tree and pair weights mutually consistent. Most importantly, the improvement in accuracy of alignments obtained by these iterative methods over pairwise or progressive method tends to increase with decreasing average sequence identity, implying that iterative refinement is more effective for the generally difficult alignment of remotely related sequences. Four well-known amino acid substitution matrices were also tested in combination with the various methods. However, the effects of substitution matrices were found to be minor in the framework of multiple alignment, and the same order of relative performance of the alignment methods was observed with any of the matrices.

Keywords: multiple sequence alignment; structural alignment database; iterative refinement; protein families; substitution matrices

Introduction

Multiple sequence alignment is now widely recognized as a valuable tool in studies of various aspects of molecular biology. It plays an essential role for predicting functionally and structurally important regions shared by a family of protein or nucleotide sequences. Multiple alignment is also a prerequisite for reliable reconstruction of the evolutionary history of a homologous set of biological sequences. Precision of secondary structure prediction is significantly improved by taking multiple alignment into consideration. Another important feature in which multiple alignment is superior to more conventional pairwise alignment is its reliability, because simultaneous comparison of many sequences can reduce the ''noise'' of various origins. In particular, accuracy of alignment

Abbreviations used: 3D, three-dimensional; CLW, CLUSTAL W; DNR, doubly nested randomized iteration; PAM, accepted point mutations; PWS, pairwise alignment; SP, sum-of-pairs; RIO, randomized iteration optimizing SP; RIW, randomized iteration optimizing WSP; WSP, weighted sum-of-pairs; CPU, central processing unit.

0022–2836/96/490823–16 $25.00/0

profoundly affects the outcome of molecular modeling studies. In the last 20 years, many multiple alignment algorithms, based on various principles, have been devised (Chan *et al*., 1992; Pevzner, 1992), and various programs have been developed, some of which are used widely. Continuous efforts have been devoted to solving two major problems: (1) how to evaluate the ''goodness'' of an alignment, and (2) how to get the alignment associated with the optimal score. Historically, the most parsimonious tree length was first assigned as the target for optimization in multiple alignment (Sankoff, 1975; Waterman *et al*., 1976). More recently, the sum-of-pairs (SP) score (Murata *et al*., 1984; Carrillo & Lipman, 1988) has been widely used as the target because of its simplicity and sensitivity. There seems to be no efficient algorithm for rigorously solving a multiple alignment problem in either scoring system (Wang & Jiang, 1994; Wareham, 1995), and all available methods ought to be approximate ones. To assess the reliability of these practical methods, particularly those adopting different scoring systems, we need objective criteria other than the target score itself.

Recently, the number of protein three-dimensional (3D) structures determined by X-ray crystallography and high-resolution NMR methods has been increasing rapidly. It has frequently been observed that proteins with only weakly related sequences share a highly similar 3D structure (Holm & Sander, 1994). Comparison of the 3D structures makes it possible to align distantly related protein sequences on the basis of their structural equivalence. A few collections of such structure-based alignments are now available (Pascarella & Argos, 1992; Šali & Overington, 1994; Holm & Sander, 1996). Thus it is possible to assess the quality of sequence alignments obtained by a given method by referring to the structural counterparts. Systematic investigations on various protein families will provide us with some idea of the general reliability of the sequence alignment method employed.

On the basis of similar considerations, Vogt *et al*. (1995) examined many published amino acid exchange matrices in aligning two protein sequences. The relative performances of the matrices were assessed by the degree of match between the resulting sequence alignment and the corresponding alignment obtained from structural superposition (Pascarella & Argos, 1992). Only two variations of pairwise alignment methods, global (Needleman & Wunsch, 1970) and local (Smith & Waterman, 1981) methods, were examined in combination with various exchange matrices and gap penalty values. McClure *et al*. (1994), on the other hand, examined 12 existing multiple alignment programs designed to generate either global or local alignment. Their criterion for assessing an alignment method was quite different from that of Vogt *et al*. (1995); i.e. they evaluated various alignment methods by their ability to correctly identify the ordered series of motifs characteristic of a protein family.

Of the 12 methods examined by McClure *et al*., AMULT (Barton & Sternberg, 1987) and TULLA (Subbiah & Harrison, 1989) adopt iterative refinement strategies, at least as an optional procedure. The idea of refining multiple sequence alignment by iteration has quite a long history (Sankoff *et al*., 1976; Hogeweg & Hesper, 1984; Waterman & Perlwitz, 1984). However, this old strategy has recently undergone significant renewal. Krogh *et al*. (1994) and Baldi *et al*. (1994) introduced hidden Markov model techniques into the multiple alignment problem. The model is described by a series of three kinds of states (match state, insert state, and delete state), the probabilities of transition between the states connected by given paths, and the position-specific amino acid (or nucleotide) distribution probabilities. The transition and distribution probabilities are adjusted so that the model best describes a given set of sequences. The underlying concept and the actual processes of computation are very similar to those used in iterative refinement methods with profile or consensus matching (Waterman & Perlwitz, 1984; Barton & Sternberg, 1987; Gribskov *et al*., 1987) with affine gap penalties (Gotoh, 1982; Searls, 1996). In the hidden Markov model approach, however, such parameters as amino acid substitution costs and gap penalty values are mostly learned from the data themselves, while these parameters are initially given in most conventional alignment procedures.

Another recent advance in the multiple alignment method stems from the idea of ''randomized'' iteration. Berger & Munson (1991) first proposed a strategy in which an initial unaligned family of sequences is randomly divided into two groups, which are then rejoined by the dynamic programming algorithm of Needleman & Wunsch (1970). The pairwise alignment process is repeated using different random divisions of the whole family into two groups. Gotoh (1993a) pointed out that the classical Needleman–Wunsch algorithm could fail to optimally align two groups of sequences when either group contains internal gaps, and proposed a rigorous group-to-group alignment algorithm that ensures monotonous improvement of the overall alignment score at each step of iteration. Gotoh (1994) later devised a generalized profile-matching algorithm that greatly reduces the computational cost for rigorous alignment between large groups. A large-scale experiment conducted by Hirosawa *et al*. (1995) showed that limited partitioning into two groups and/or recurrent application of randomized iteration to subgroups yielded much better cost-performance than the original strategy. Combination of the approach of Hirosawa *et al*. and the generalized profile-matching algorithm makes it possible to perform the total calculation effectively in proportion to the number of sequences to be aligned.

The alignment score most compatible with the iterative methods was the SP score, because partitioning and rejoining processes do not affect the intra-group contributions to the total score, and thus, the principle of divide and conquer is effective (Gotoh, 1993a). Although the simple SP scoring system might be inappropriate when some groups of sequences are over- or under-represented in a family, this drawback is correctable by introducing a proper weighting system (Altschul *et al*., 1989). A randomized iterative method optimizing such a weighted sum-of-pairs (WSP) score has recently been developed (Gotoh, 1995), and a preliminary experiment on the globin superfamily proteins indicated that the multiple alignment obtained by this method is closer to the structural counterpart than those obtained by any other methods examined, including the most popular progressive method and the former randomized iterative method targeting optimal SP.

In this paper, I report the results of more extensive examinations on 54 independent protein families (Table 1) that contain at least two distant members (amino acid sequence identity <40%) structurally aligned with one another (Šali & Overington, 1994). Four multiple sequence alignment strategies, a progressive method (CLUSTAL W = CLW: Thompson *et al*., 1994), a randomized iterative method that optimizes SP (RIO: Gotoh, 1994), another randomized iterative method that optimizes WSP (RIW: Gotoh, 1995), and a doubly nested randomized iterative method (DNR: see Methods) were applied to the same set of families. The classical pairwise alignment method (PWS: Gotoh, 1982) was also tested as a control. Close statistical tests detected clear differences in the accuracy of sequence alignment obtained by the four methods examined, although the difference between the results of RIW and DNR was insignificant. Similar tests were also done on the relative performance of four well-known amino acid exchange matrices (MDM250: Dayhoff *et al*., 1978; JTT250: Jones *et al*., 1992; BLOSUM62: Henikoff & Henikoff, 1992; GCB250: Gonnet *et al*., 1992) in combination with the four alignment methods. Several general factors that affect relative performance of the different alignment methods are discussed.

Results

Relative performance of amino acid substitution matrices

Vogt *et al*. (1995) examined some 80 amino acid substitution matrices for their performance in aligning a pair of protein sequences. My approach is analogous to theirs in the very basic principles, but differs considerably in detail. Most significantly, I regarded each family, rather than each structural pair, as an independent entity, because individual structural pairs within a family cannot be regarded as independent when multiply aligned. To test the sensitivity of this method, I first examined whether my approach could reproduce the major conclusions of Vogt *et al*. on the relative performance of various amino acid substitution matrices. The upper-left triangle of Table 2 shows the results of comparisons of the four matrices tested with the pairwise method. To obtain the data, I first selected the optimal gap-opening (v) and gap-extension (u) penalty values for each family and matrix so that the average consistency of sequence and structure alignments (or accuracy of sequence alignment) between all pairs of members in an entry of the Joy3.2 database (Šali & Overington, 1994) should be the highest. I then calculated the grand average of differences in average accuracy individually obtained for the 54 entries with the two matrices under comparison. The value shown in a cell is the mean difference in these average accuracies of all pairs in a family. Highly significant superiority of GCB250 over MDM250 or JTT250 ($P < 10^{-3}$), and marginal superiority over BLOSUM62 ($P < 10^{-2}$) was found by Wilcoxon matched-pairs signed rank test, although the grand average of difference in the consistency was only 1.5% or less. Although no other comparisons indicate statistical significance, the order of superiority of GCB250 > BLOSUM62 > JTT250 $\geqslant$ MDM250 appears definite. This order is in good agreement with that observed by Vogt *et al*. with the global alignment method and positively transformed matrices (Table 9, see Vogt *et al*., 1995), except that JTT250 was not included in their Table. Essentially the same order of performance of the four matrices was observed when average consistency for each family was calculated from a set of pairs of a fixed range of sequence identity, from 0% to 40%, from 0% to 20%, from 30% to 40%, or from 20% to 30% (data not shown). Note that all the sequence identity values referred to here and anywhere in this article were derived from the structural alignments in Joy3.2 database.

When I used the DNR multiple sequence alignment method, only a weak influence of the substitution matrices was recognized as summarized in the lower-right triangle of Table 2. Only the difference between MDM250 and GCB250 was statistically significant when either all pairs or pairs of sequences with less than 40% amino acid identity were considered, although the above order of the performance of the four matrices was basically retained. The effects of substitution matrices were even weaker with the RIO method, and no significant difference was detectable (data not shown).

Assessment of alignment methods

The accuracy of alignments was greatly affected by the methods tested. Although I examined all four substitution matrices in combination with three methods, PWS, RIO and DNR, exactly the same order of relative performance of these and the

CLW method was observed with each of these matrices. Hence, unless otherwise mentioned, the results obtained with GCB250 are presented below, since GCB250 was found to be the best among the four matrices (Table 2). I examined the RIW method only with GCB250, since RIW is time-consuming compared with other methods (see below).

Figure 1(a) shows the accuracy of sequence alignments averaged over all pairs of members in individual Joy3.2 entries calculated with the four methods, PWS, CLW, RIO and DNR. The results of RIW were almost indistinguishable from those of DNR and are not shown in this Figure. The differences in the performance of the pairwise and multiple alignment methods are emphasized in Figure 1(b), which plots the smallest accuracy value among all the pairs within a family. To evaluate the performance of the five methods more quantitat-

Table 1. Protein families used for alignment study

Family	Nstr[a]	Nseq[b]	MinID[c]	AvrID[d]	AvrLen[e]	Description
aa	3	23	20.75	36.94	483.0	Amylase
ace	2	47	25.93	25.93	534.5	α/β-hydrolase
adk	4	38	13.11	23.97	200.0	Nucleotide kinase
asp	10	62	22.82	34.18	331.3	Aspartic proteinase
az	8	63	16.98	32.73	104.3	Azurin/plastcyanin
bv	2	15	22.35	22.35	183.5	Plant virus coat protein
cbp	5	168	15.03	32.74	154.8	Ca binding protein
cox	2	11	15.64	15.64	535.0	Cholesterol oxidase
cryst	4	42	33.52	57.62	175.2	Crystallin
cyt3	2	7	32.38	32.38	112.5	Cytochrome *c*3
cyt5	3	26	21.25	37.35	80.7	Cytochrome *c*5
cytb	2	7	29.21	29.21	85.5	Cytochrome *b*
cytc	9	122	30.19	44.62	107.1	Cytochrome *c*
cytprime	2	17	20.61	20.61	129.0	Cytochrome *c′*
dhfr	4	37	23.67	34.69	173.0	Dihydrofolate reductase
egf	4	96	21.74	34.04	45.0	EGF-like domain
fer4	3	68	24.56	30.09	56.7	Ferredoxin (4Fe–4S)
flav	5	19	18.49	31.89	159.0	Flavodoxin
glob	18	666	11.51	27.88	146.1	Globin
gluts	4	36	20.83	36.76	212.5	Glutathione S-transferase
grs	5	34	15.73	29.07	466.8	Disulphide oxidoreductase
hemocyan	2	18	33.00	33.00	617.0	Haemocyanin
hip	4	11	16.92	24.61	72.2	High potential iron protein
hom	3	317	17.24	28.16	64.0	DNA-binding homeodomain
icd	2	29	26.93	26.93	367.5	Isocitrate dehydrogenase
igcon	11	71	13.68	32.82	91.5	Immunoglobulin constant region
igps	2	37	9.60	9.60	200.0	Tryptophan synthase
il8	2	54	18.18	18.18	66.0	IL8-like cytokine
ins	4	123	27.66	45.54	52.5	Insulin
intb	4	40	9.63	28.63	139.5	IL1β-like growth factor
ldh	9	63	17.18	35.87	324.4	Lactate/malate dehydrogenase
ltn	5	30	33.76	43.95	231.6	Lectin
neur	3	24	28.90	34.81	388.7	Neuraminidase
peroxi	3	25	13.90	24.47	295.3	Peroxidase
phoslip	6	51	28.93	44.74	122.0	Phospholipase A2
repr	3	16	19.12	31.52	67.7	Repressor
rhv	6	33	22.12	32.45	754.3	Picornavirus coat protein
ricin	2	22	27.97	27.97	262.0	Ricin-like protein
rnh	3	72	19.67	33.97	132.7	Ribonuclease H
rubisco	3	205	28.44	50.17	448.0	Ribulose-1,5-bisphosphate oxygenase
rvp	3	67	20.00	31.49	104.3	Retroviral proteinase
serbact	3	8	34.04	43.79	188.0	Bacterial serine proteinase
sermam	12	141	25.69	36.09	232.3	Mammalian serine proteinase
serpin	4	39	26.60	30.82	376.0	Serine proteinase inhibitor
sh2	2	69	34.34	34.34	99.5	SH2 domain
sh3	4	79	25.86	30.53	62.0	SH3 domain
sodfe	2	44	35.08	35.08	189.5	Fe/Mn superoxide dismutase
subt	7	45	32.35	50.82	273.4	Subtilase
sugbp	3	13	18.51	20.18	287.3	Periplasmic sugar binding protein
thiored	4	54	7.32	14.67	86.3	Thioredoxin
tim	4	17	38.40	44.99	248.0	Triose phosphate isomerase
tln	3	15	29.05	43.57	310.3	Zn metalloproteinase
tms	3	19	45.26	50.67	288.7	Thymidylate synthase
toxin	10	102	22.58	40.52	62.2	Snake toxin

[a] Number of aligned structures in the Joy3.2 entry.
[b] Number of sequences in the extended entry.
[c] Amino acid identity between most remotely related sequences in the Joy3.2 entry.
[d] Average sequence identity among all pairs of structures.
[e] Average length of sequences in the Joy3.2 entry.

Table 2. Comparison of performance of amino acid substitution matrices

Matrix	GCB250	BLOSUM62	JTT250	MDM250
MDM250	1.48 ± 0.31***	0.63 ± 0.40	0.03 ± 0.28	—
JTT250	1.45 ± 0.34***	0.60 ± 0.37	—	0.42 ± 0.41
BLOSUM62	0.85 ± 0.31**	—	− 0.11 ± 0.46	0.31 ± 0.32
GCB250	—	0.36 ± 0.43	0.25 ± 0.31	0.67 ± 0.38*

Difference in mean percentage of correctly matched residues obtained with the matrices indicated at the top and the left corners. The estimated standard error is shown after a ± symbol. Upper-left triangle: tested with the PWS alignment method (top matrix − left matrix; a positive value indicates that the top matrix performs better than the left matrix). Lower-right triangle: tested with the DNR method (left matrix − top matrix; a positive value indicates that the left matrix performs better than the top matrix). Significant differences are marked: *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

ively, I adopted a similar statistical test to that used for comparison of different substitution matrices. First, the average accuracy was calculated from all pairs of members in a Joy3.2 entry or from the pairs whose amino acid identity is less than 40%. Then, the accuracy values obtained with two methods were compared by the Wilcoxon matched-pairs signed rank test (Table 3). On the whole, the order of performance of the five methods was PWS < CLW < RIO < DNR ⩽ RIW. The differences in the results of PWS *versus* CLW and DNR *versus* RIW did not reach significant levels, but all other differences were highly significant. This order of the relative performance of the methods is exactly the same as that previously observed for the globin family (Gotoh, 1995), although only about half as many sequences (339 *versus* 666) and structures (7 *versus* 18) were used previously, and the new DNR method was not examined at that time.

In general, the average amounts of difference between the results of different methods, Δ(Method 1, Method 2), are greater as amino acid identity

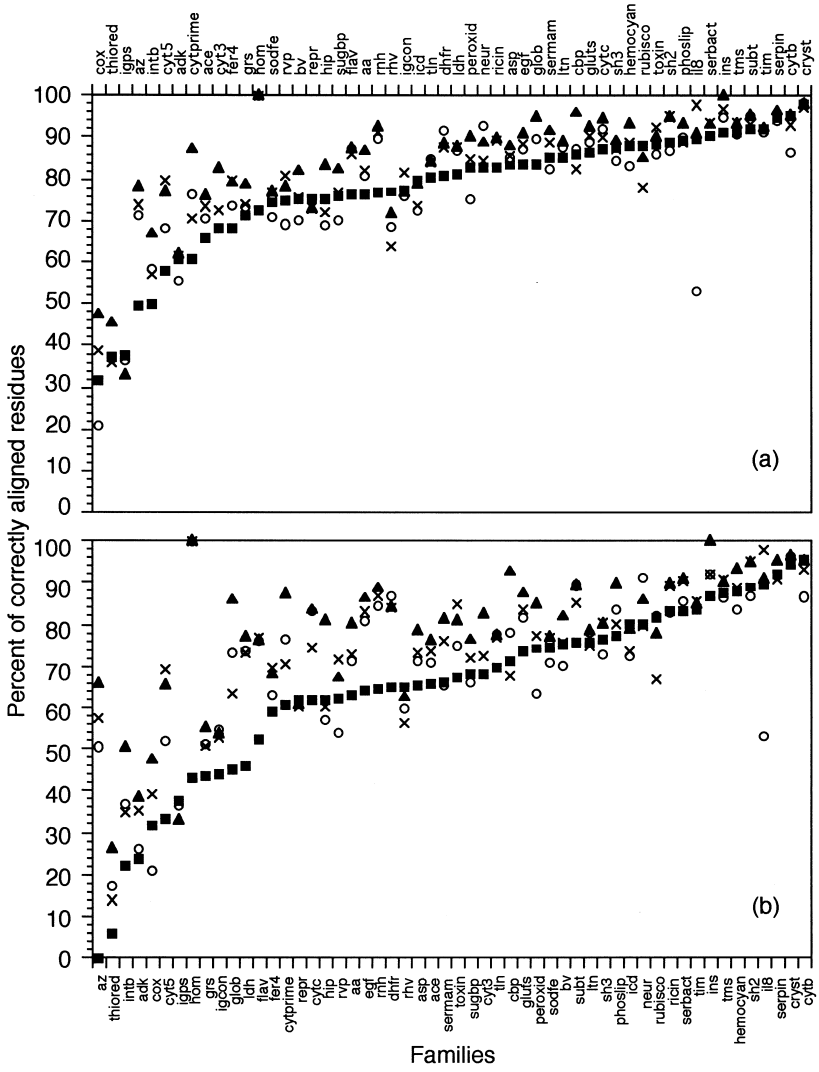

Figure 1. Comparison of the performance of various sequence alignment methods. Mean (a) and smallest (b) values for percentage of correctly aligned residues among all the pairs of sequences of known structures were obtained for each family. The methods used were PWS (■), CLW (○), RIO (x), and DNR (▲).

Table 3. Comparison of the performance of five sequence alignment methods

ID range[a]:	[0–100]	[Min]	[0–40]	[0–20]	[20–30]	[30–40]
No. of families:	54	54	53	19	36	31
Methods						
RIW–PWS	7.1 ± 1.0***	13.7 ± 1.9***	9.2 ± 1.4***	14.7 ± 2.7***	10.0 ± 1.7***	3.1 ± 1.0**
DNR–PWS	6.7 ± 1.0***	13.2 ± 1.8***	8.8 ± 1.3***	13.7 ± 2.8***	10.1 ± 1.7***	3.0 ± 1.1**
RIW–CLW	6.3 ± 1.0***	8.0 ± 1.2***	7.3 ± 1.1***	11.1 ± 2.8**	6.1 ± 1.2***	4.1 ± 0.9***
DNR–CLW	5.9 ± 1.0***	7.5 ± 1.1***	6.8 ± 1.0***	10.1 ± 2.6**	6.3 ± 1.0***	4.0 ± 1.0***
RIW–RIO	3.6 ± 0.7***	5.2 ± 1.0***	4.4 ± 0.8***	6.4 ± 2.3*	4.6 ± 1.1***	2.5 ± 0.7**
DNR–RIO	3.2 ± 0.6***	4.7 ± 0.9***	3.9 ± 0.7***	5.4 ± 2.0*	4.7 ± 1.0***	2.3 ± 0.8**
RIO–PWS	3.5 ± 0.9***	8.5 ± 1.8***	4.9 ± 1.3***	8.3 ± 2.9*	5.4 ± 1.7**	0.6 ± 1.0
RIO–CLW	2.7 ± 1.1**	2.8 ± 1.2*	2.9 ± 1.2**	4.7 ± 2.7*	1.5 ± 1.2	1.6 ± 0.9
CLW–PWS	0.8 ± 1.2	5.7 ± 2.0**	2.0 ± 1.5	3.6 ± 3.7	3.9 ± 1.6*	−1.0 ± 1.2
RIW–DNR	0.4 ± 0.4	0.5 ± 0.4	0.5 ± 0.4	1.0 ± 1.2	−0.2 ± 0.5	0.1 ± 0.3

Difference in mean correctly matched residues obtained with the methods indicated in the first column. GCB250 was used throughout the examinations. The estimated standard error is shown after a ± symbol. Significant differences are marked: *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

[a] Average accuracy for each family was calculated from the percentage of matched residues for the sequence pairs with the percentage amino acid identity (ID) within the range shown in square brackets. The data in the column labelled [Min] were obtained from the least accurate alignment among all pairs in each family.

decreases. For example, Δ(PWS, RIW) was only 3.1% for sequence pairs with amino acid identity between 30% and 40%, but about 10% for sequence pairs with amino acid identity between 20% and 30%, and as much as 14% for sequences with less than 20% of identity (Table 3). In this lowest range of sequence identity, Δ(CLW, DNR) or Δ(CLW, RIW) also exceeds 10%, implying that iterative refinement is most effective in augmenting the accuracy of alignment of distantly related sequences. The contrast between DNR (or RIW) and other methods in lower identity ranges was even more prominent when other substitution matrices were used; the greatest contrast was 15.8%, which was observed between the PWS and DNR methods with JTT250.

Computation time

Computation time is another important factor for evaluating different methods. Figure 2 presents the CPU time spent in obtaining an alignment averaged over various *u*–*v* pairs. The CPU time taken with CLW, but not with other methods, includes that

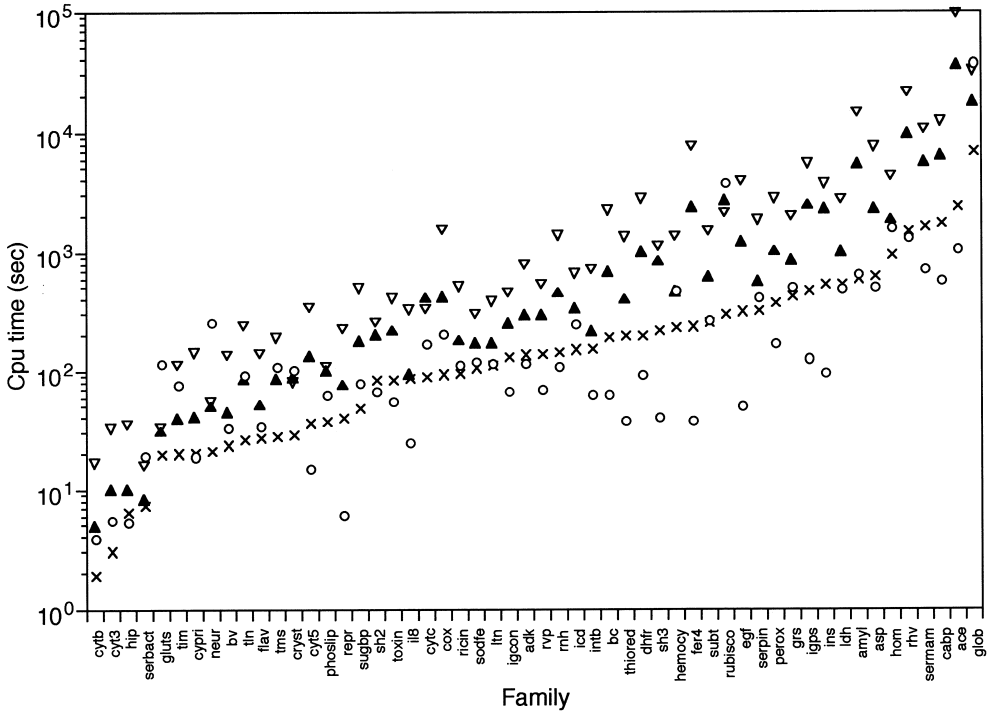

Figure 2. CPU time taken to obtain a multiple alignment for each family. The methods used were CLW (○), RIO (×), DNR (▲), and RIW (▽). The average CPU time of calculations with various gap penalties is presented. Calculations made on three workstations were calibrated to that on a SUN SPARCstation 2 according to the relative machine performances.

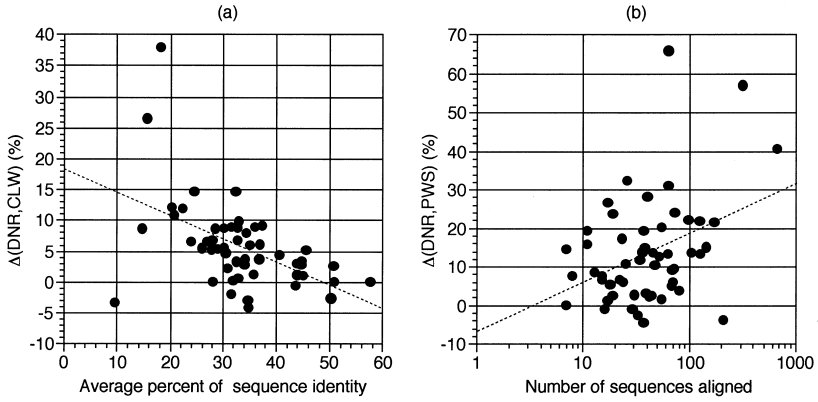

Figure 3. Correlation between improvement in accuracy of alignment by iterative refinement and some general factors. (a) Negative correlation (correlation coefficient $r = -0.5$, $P < 10^{-3}$) between sequence identity averaged over all the pairs of sequences of known structures in a family and difference in mean alignment accuracy obtained by the DNR and CLW methods. (b) Positive correlation ($r = 0.43$, $P \approx 10^{-3}$) between the number of sequences in an extended entry (family) and the difference in worst accuracy values observed with the DNR and CLW methods in that family. The abscissa is scaled logarithmically. A linear regression line is indicated by a broken line in each correlation diagram.

used to get the guide tree by a distance matrix method, where the distance matrix was calculated from pairwise alignment obtained by a FASTA-like rapid routine (Pearson & Lipman, 1988; Thompson *et al*., 1994). Since this routine takes time in proportion to the square of the number of sequences to be aligned, CLW uses more CPU time than other methods for large families such as globin. The quadratic routine was avoided by other methods, because a family-specific initial alignment was given to an iterative method. The CPU time used by RIO, DNR or RIW depends not only on the number and length of the sequences but also on other factors, in particular, the number and distribution of anchor points preset by the heuristic routine (see Methods). Hence, it is not easy to estimate the relative computational costs of different methods individually. The raw averages of CPU times relative to that used by the RIO method were 1.7, 3.3, and 7.9 for CLW, DNR and RIW, respectively. Even though the time spent for making the initial alignment was not included in the CPU time for the RIO, DNR and RIW methods, the results indicate that DNR and CLW take the same order of computation time.

Factors that affect improvement in alignment accuracy

Figure 1 indicates that the relative performance of different alignment methods, as well as the degree of alignment accuracy, varies greatly with families. To search for any general factors that affect the improvement in performance of one method over another, I analysed the correlation between the degree of improved accuracy and several general factors, such as average amino acid identity, sequence length, and number of sequences or structures in an entry. Correlation coefficients were calculated and significance was tested by either the standard Pearson method, or either the Kendall or the Spearman non-parametric method. The results for the average of all pairs or pairs with 0 to 40% sequence identity were essentially the same, and those for the worst pair in a family were also similar but somewhat more emphatic. The most significant and consistent finding was the negative correlation between the average sequence identity and the difference in accuracy between DNR and CLW, Δ(DNR, CLW), or to a lesser extent, Δ(RIO, CLW), implying again that the randomized iteration is effective for a set of more remotely related sequences (Figure 3(a)). On the other hand, a weak positive correlation was found between the average sequence identity and Δ(CLW, PWS). This observation may be accounted for by the fact that CLW dynamically adjusts gap penalties and substitution matrix depending on the degree of divergence of sequences to be aligned (Thompson *et al*., 1994).

Significant positive correlation was found between the number of sequences or structures that are highly correlated with each other and Δ(CLW, PWS) or Δ(DNR, PWS), particularly when the worst pairs in a family were examined (Figure 3(b)). This observation is reasonable because simultaneous alignment of a larger number of sequences would be more effective in improving accuracy, particularly for the most remotely related pairs in a family. The critical importance of the number effects was confirmed by another series of examinations, in which only the sequences of known structures were aligned without incorporating their homologues. Forty-one of the 54 families that contain more than two structures were used for the statistical tests with average consistency of all pairs in a family. Although the general tendency was the same as that observed above, the significance level and grand average of difference were much compressed; i.e. although the significant superiority of the three iterative multiple alignment methods over PWS was again confirmed, even the largest average difference between the results of DNR (or RIW) and PWS was only 2.9 ($\pm$0.9)% ($P < 10^{-4}$), less than a half of the corresponding value of 6.7 ($\pm$1.1)% ($P < 10^{-6}$) observed with the expanded data set containing homologous sequences.

A weak negative correlation was also found between the average sequence lengths and Δ(CLW,

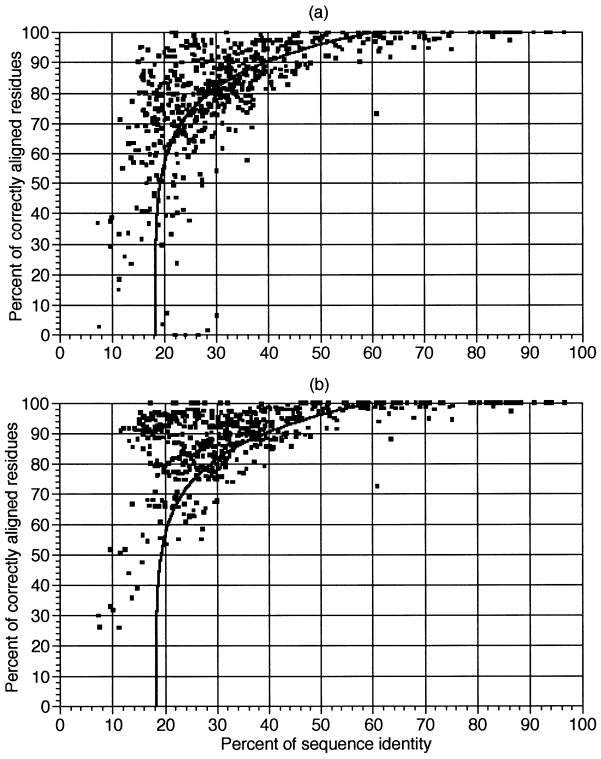

Figure 4. Plots of the percentage of correctly aligned residues against the percentage of identically conserved residues in the corresponding structural alignment. (a) All 682 pairs of sequences in the 54 Joy3.2 entries were aligned by the PWS method and the results were assessed by the degree of consistency with the corresponding structural alignments. (b) The sequences in an extended entry were multiply aligned with the DNR method, individual pairs were extracted therefrom, and the accuracy of an alignment was assessed as above. An identical curve, which was derived from an analytical function of the form of $x = A_1 \exp(B_1 y) + A_2 \exp(B_2 y)$ best fit to the points in (a), is shown in both (a) and (b) for easier comparison of the distributions of points.

PWS), Δ(RIO, PWS), or Δ(DNR, PWS), when the worst pairs in a family were examined (data not shown). This observation remains to be interpreted.

Accuracy of individual alignments of sequence pairs

So far, I have treated each Joy3.2 entry as an independent entity for the purpose of rigorous statistical tests. However, it is sometimes more interesting to know how much improvement in accuracy of sequence alignment between a given pair can be expected by using iterative multiple alignment methods, and how it depends on the degree of sequence divergence and other factors. Figure 4 plots the accuracy of sequence alignments obtained by PWS (a) and DNR (b) for all the 682 pairs in the 54 entries as a function of the percentage of amino acid identity. A histogram of the accuracy values for all the pairs with less than 20% sequence identity is shown in Figure 5. These

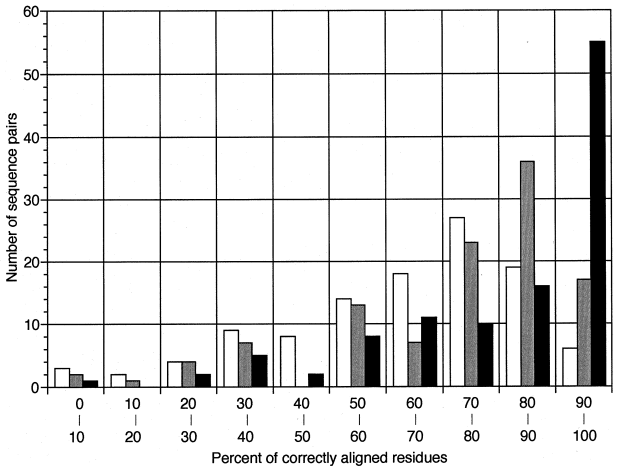

Figure 5. Histogram of the percentage of correctly aligned residues against the number of sequence pairs with less than 20% of identically conserved residues. Sequence alignment was calculated by the PWS (open bar), CLW (shaded bar) or DNR (filled bar) method.

Figures clearly indicate that DNR multiple alignment shifts the ''twilight zone'' towards lower identity ranges, where the twilight zone generally implies the sequence identity range in which the accuracies of calculated sequence alignments vary widely from one another (Doolittle, 1981).

Figure 6 shows the differences in alignment accuracy obtained for all the 682 pairs by the CLW and DNR methods. The positive region indicates that the result of DNR is superior to that of CLW, and the negative region indicates the opposite. The Figure shows that the difference in the accuracy of

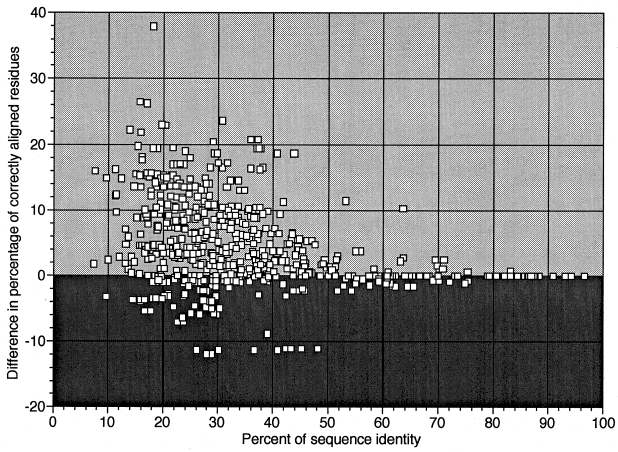

Figure 6. Difference in performance of progressive and randomized iterative multiple sequence alignment methods assessed with individual pairs of structurally aligned sequences. The percentage of correctly aligned residues observed with the DNR method minus the counterpart with the CLW method is plotted against the percentage of identically conserved residues in the corresponding structural alignment. Lightly shaded area indicates that the results with DNR are superior to those with CLW, and darkly shaded area indicates the opposite.

Table 4. Ternary consistency among sequence alignment and two structural alignments

Family	Nstr[a]	ID_j[b]	ID_3[c]	Joy_3D[d]	Joy_Seq[e]	3D_Seq[f]	Css3[g]	CLW[h]	RIO[h]	DNR[h]	RIW[h]
asp	5	36.6	36.9	94.2	87.4	88.2	85.3	87.2	88.6	90.6	91.4
az	5	28.7	28.9	73.0	75.0	67.9	61.2	76.2	80.8	81.9	81.9
cbp	2	50.3	50.3	98.3	97.3	96.0	96.0	96.6	99.7	97.6	97.6
cryst	2	80.5	80.5	99.1	100.0	99.1	99.1	100.0	100.0	100.0	100.0
cyt3	2	32.4	31.3	84.4	82.7	75.6	73.3	73.2	85.3	86.8	83.7
cytb	2	30.2	32.1	91.8	95.3	87.1	87.1	85.3	92.4	94.9	94.9
cytc	4	40.8	40.0	95.7	96.2	94.0	93.0	89.5	93.4	97.2	97.5
dhfr	4	34.7	34.7	91.5	88.6	86.8	84.2	91.8	90.0	91.8	92.3
glob	11	29.3	29.0	95.5	95.7	93.2	92.5	89.1	87.1	96.9	97.5
ldh	6	36.7	36.9	91.7	89.3	87.8	85.5	92.3	94.6	92.8	93.2
phoslip	3	56.5	56.2	98.4	92.9	94.3	92.9	90.5	89.7	94.5	94.2
repr	3	31.5	32.5	82.7	70.7	82.7	68.8	81.9	81.2	81.4	82.5
rvp	2	26.0	27.0	90.2	67.3	66.4	63.5	51.8	73.1	70.5	74.1
sermam	9	35.2	35.1	96.2	91.9	90.0	89.5	87.5	89.3	93.1	93.2
subt	4	43.4	44.2	83.3	92.8	83.0	80.6	96.4	95.2	96.7	97.4
sugbp	2	18.6	18.8	80.4	76.5	70.7	68.4	75.2	83.8	85.1	85.9
toxin	3	42.5	38.9	76.7	91.2	79.9	75.7	88.3	98.7	98.7	98.7
Average		38.5	38.4	89.6	87.7	84.8	82.2	85.5	89.6	91.2	91.5
SD[i]		14.1	13.9	8.0	9.8	9.9	11.7	11.5	7.3	7.8	7.3

[a] Number of structures used for the assessment.
[b] Mean amino acid identity between sequences in the subset of a Joy3.2 multiple structural alignment.
[c] Mean amino acid identity between sequences in the subset of a 3D_ALI multiple structural alignment.
[d] Mean percentage of consistently aligned residues between corresponding structural alignments in Joy3.2 and 3D_ALI.
[e] Mean percentage of consistently aligned residues between a Joy3.2 alignment and the sequence alignment obtained by the DNR method.
[f] Mean percentage of consistently aligned residues between a 3D_ALI alignment and the sequence alignment obtained by the DNR method.
[g] Degree of ternary consistency among Joy3.2, 3D_ALI, and the sequence alignment obtained by the DNR method.
[h] $100 \times$ Css3/Joy_3D, where Css3 was calculated from the sequence alignment obtained by the indicated method.
[i] Standard deviation.

the CLW and DNR methods is small for sequence pairs with more than 50% identical residues, and that it becomes gradually apparent in the identity range from 50% to 40% and widens further in lower identity ranges. These observations confirm those reported above on the set of independent families.

Effects of variation in structural alignments

The validity of the investigations described above obviously depends on the quality of structural alignments used as references. To what extent are published structural alignments mutually consistent? Does such variation in structural alignments affect the assessment of sequence alignment methods? Does mutual inconsistency in structural alignments correlate with inconsistency between structural and sequence alignments? To answer these questions partly, I simultaneously compared sequence alignments and two sets of structural alignments of the same proteins. One of the structural alignment databases was Joy3.2, as used above, and the other was 3D_ALI, constructed by Pascarella & Argos (1992). I selected 17 families so that each contains at least two structures in common in the corresponding entries of the two databases (see Methods). Table 4 summarizes typical results of the ternary comparison of alignments. The column labelled Joy_3D shows the degree of consistency of structural alignments averaged over all pairs of common structures in the corresponding database entries. The commonality of structures was identified by Protein Data Bank codes. The grand average for the 17 families is 89.6 ($\pm$8.0)%. This value is close to the grand average of accuracy of sequence alignments obtained by the DNR method for the same 17 families, as assessed by referring to either Joy3.2 (87.7 $\pm$ 9.8) or 3D_ALI (84.9 $\pm$ 9.9) (columns labelled Joy_Seq and 3D_Seq). This observation seems to imply that the quality of sequence alignments obtained by the DNR method has closely approached the range of variation among structural alignments. The column labelled Css3 shows the average degree of ternary consistency among the sequence alignment and the two structural alignments taken from Joy3.2 and 3D_ALI. This value divided by the degree of consistency between structural alignments gives a rough estimate of the attainment of sequence alignment relative to that of structural alignment. As shown in the columns labelled by the alignment methods, the attainment is about 90% for the RIO, DNR and RIW methods, which is 5% better than that of CLW.

The order of relative performance of the four multiple alignment methods was the same as that observed above with 54 Joy3.2 families (Table 3) when either 3D_ALI or Joy3.2 data of the 17 families were used as references, or when the ternary consistency was used as an indicator of accuracy. The difference in performance between any two methods, except between DNR and RIW, was significant to a similar level regardless of the reference data set (data not shown). Evidently, the

reference data sets have little affect on the results of assessment of relative performance of alignment methods.

All data fields in Table 4 including average sequence identity are correlated to each other. Particularly interesting is the weak but positive correlation between the relative attainment (columns indicated by alignment methods) and the degree of consistency between structural alignments, in spite of the fact that the latter is used as the denominator in calculation of the former; e.g. the correlation coefficient observed with the DNR method was 0.42 ($P < 0.1$). Figure 7 visualizes the degree of consistency between structural alignments (a), the degree of consistency between sequence alignment and structural alignment (b), and the attainment of sequence alignment relative to structural counterpart (c) as a function of sequence identity of individual pairs of proteins in the 17 families. The three plots show similar profiles to each other, while the relative attainment shows the least tendency to decrease in sequence identity. Thus, as would naturally be expected, both sequence alignment and structural alignment become harder with a decrease in sequence identity, although the rate of decline of accuracy in multiple sequence alignment is rather milder than that of uncertainty in structural alignments, at least within the range of sequence divergence examined here.

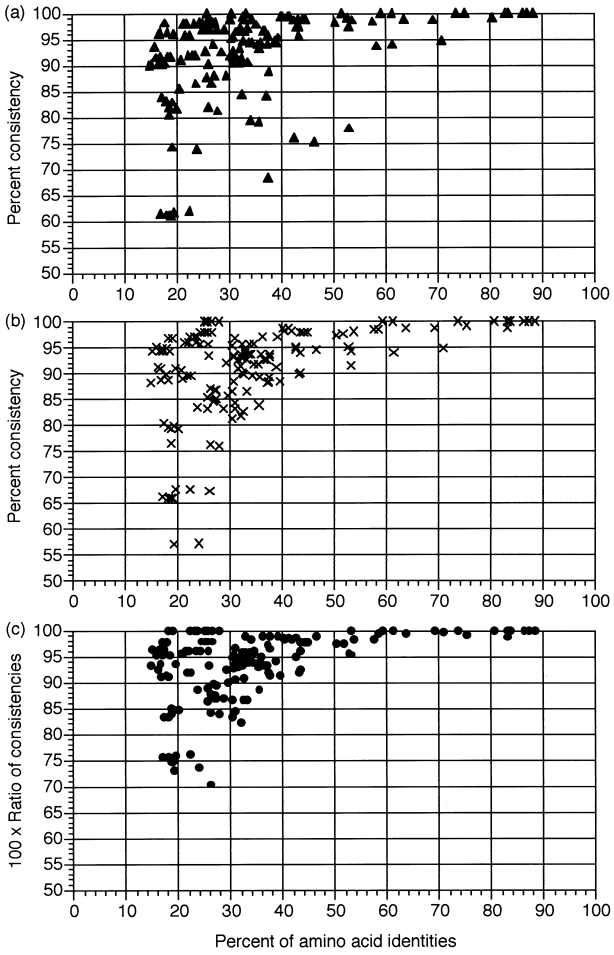

Figure 7. Ternary comparison of sequence and two structural alignments. (a) A pairwise alignment extracted from a multiple structural alignment in the Joy3.2 database (Šali & Overington, 1994) was compared with the corresponding alignment taken from the 3D_ALI database (Pascarella & Argos, 1992) by the same procedure as that used for assessment of accuracy of a sequence alignment. The percentage of consistently aligned residues is plotted against the percentage of identically conserved residues in the Joy3.2 alignment. (b) The plot is the same as that in Figure 4b but shows only a subset of sequence pairs with which ternary comparison is possible. (c) Attainment of sequence alignment is defined as the fraction of residues consistently aligned in all the three alignments (sequence alignment obtained by the DNR method and structural alignments extracted from Joy3.2 and 3D_ALI databases) divided by the fraction of residues consistently aligned in the corresponding Joy3.2 and 3D_ALI structural alignments.

Discussion

When new methods for multiple sequence alignment have been developed, it has become customary to evaluate the resulting alignment by referring to a model derived from structural equivalence (Barton & Sternberg, 1987; Feng & Doolittle, 1987; Smith & Smith, 1992; Lipman *et al*., 1989). In most cases, however, such assessment was done with only a few protein families, such as globins and mammalian serine proteinases. Moreover, the evaluation was often subjective in terms of quantifying accuracy of a sequence alignment. One reason for the difficulty in objective assessment has been an insufficient number of structural alignments usable as references. This situation has recently changed greatly, since several well-qualified databases of structural alignments have become publicly available (Pascarella & Argos, 1992; Šali & Overington, 1994; Holm & Sander, 1996). In addition, the variety of multiple sequence alignment strategies that can deal with a large number of sequences has recently become enriched (Thompson *et al*., 1994; Krogh *et al*., 1994; Baldi *et al*., 1994; Gotoh, 1994, 1995; Hirosawa *et al*., 1995). Hence, we are now able to objectively evaluate relative performances of various multiple alignment methods by referring to the structural models of many independent protein families. To my knowledge, this paper reports the first such examination.

The work of Vogt *et al*. (1995) is most similar in spirit to the present one. Of the 37 families they used, 26 are common to the 54 families examined here. Gap penalties were optimized similarly in both studies, and the accuracy of an alignment was quantified in a similar way. However, the major purpose of Vogt *et al*. was to test relative performance of various amino acid substitution matrices, and only pairwise alignment methods

were examined. McClure *et al*. (1994) tested 12 existing multiple sequence alignment programs. They evaluated an alignment method by its ability to identify correctly the ordered series of motifs characteristic of a protein family. Of the 4 families they examined, three families (globins, acidic proteases and ribonuclease H) are common to those tested here. Although McClure *et al*. did not rank the tested methods by a single criterion, they reported several general findings. First and unexpectedly, global methods were better than local methods in correctly aligning corresponding motifs. Second, progressive or iterative strategies consisting of repeated use of dynamic programming algorithms performed better than those based on consensus word-matching such as that developed by Waterman (1986). Third, space-saving approaches (Lipman *et al*., 1989) are not applicable in practice to more than ten sequences. Fourth, accuracy of alignment was improved with increasing number of sequences to be aligned. Because the four multiple alignment methods examined in this study are all global methods based on progressive or iterative use of dynamic programming algorithms, they are expected to perform better than most existing methods based on other principles.

The ranking of performances of the five sequence alignment methods elucidated in this study (PWS < CLW < RIO < DNR ≈ RIW) is reasonable. It has often been stated that accuracy of sequence alignment is improved by multiple alignment methods compared with conventional pairwise methods (Barton & Sternberg, 1987; Johnson & Overington, 1993), and the ranking of the pairwise method PWS below all multiple alignment methods examined accords with these observations. The iterative refinement strategy (Berger & Manson, 1991; Gotoh, 1993a, 1994) was designed to optimize the SP score on the expectation that an improved score would lead to improved accuracy of alignment. The expectation is borne out by the significant superiority of RIO over CLW (Table 3). A drawback of the RIO method was that the SP score, the target for optimization of RIO, ignores over- or under-representation of some subgroups in a family. The pair weights proposed by Altschul *et al*. (1989) were devised to correct for this problem. The results presented in Table 3 show that introduction of such a weighting system actually improves accuracy of sequence alignments, although the weights employed were approximate ones that conform to fast refinement procedures (Gotoh, 1995). The observed ranking is clearly not due to differences in range or density of examined parameters, since three times and twice as many sets of gap penalty values were tested with the PWS and CLW methods, respectively, than with the higher-ranked methods (see Methods). Most importantly, the performance of DNR (or RIW) relative to that of PWS or CLW increases with decreasing average sequence identity. This tendency is shown in several ways in Table 3 and Figures 3(a), 5, and 6. Considering that accurate alignment is more difficult for remote sequences, this feature of the DNR and RIW methods is very desirable when remote sequences must be aligned, e.g. for homology-based 3D structural modeling.

Although it was not the primary purpose of this study to evaluate various amino acid substitution matrices, the results shown in Table 2 indicate that GCB250 (Gonnet *et al*., 1992) performs best and BLOSUM62 (Henikoff & Henikoff, 1992) performs next best for pairwise sequence alignment. These findings are generally in good agreement with those of Vogt *et al*. (1995) and Johnson & Overington (1993). Since the test procedure used in this study is simple and easily automated, it might be applicable to more extensive studies of a larger set of matrices. However, the effects of substitution matrix were lessened under the framework of multiple alignment (Table 2). Probably, the simultaneous consideration of many homologous sequences masks the effect of subtle differences in relative performance of substitution matrices.

Compared with the relatively minor contributions of amino acid substitution matrices, gap penalties profoundly affected the quality of alignments. This is most prominent with the PWS method, as illustrated in Figure 8(a) by the wide distribution of optimal or nearly optimal gap penalties individually adjusted for the 54 families; the *Z*-axis of this contour map shows the normalized frequency of a u–v pair that generated an alignment as accurate as at least 99% of the optimal one in each family. There seem to be a few separate clusters of preferable u–v pairs along the line between $(u, v) = (1, 15)$ and $(3, 5)$. Similarly, two clusters near $(0, 11)$ and $(3, 10)$ are recognized with CLW, although less conspicuous (Figure 8(b)). On the other hand, the range of optimal gap penalty values is much narrower with RIO (Figure 8(c)), and sharply concentrated around $(3, 10)$ with DNR (Figure 8(d)). Essentially the same tendencies were observed when only the optimal u–v pairs were totalized (data not shown). This observation appears to indicate that the choice of gap penalty values is less critical when the DNR (or RIW) method is used. However, more comprehensive studies covering wider parameter ranges would be necessary to evaluate conclusively the effects of gap penalties on performance of different methods.

The definitions of SP and WSP are simple, and the same substitution matrix and gap penalties are used throughout an alignment process. Considering that gaps tend to occur outside secondary structures (Lesk *et al*., 1986), that the best log-odds matrix would vary with the overall sequence divergence (Dayhoff *et al*., 1978), and that higher-order structures and environmental conditions in protein molecules would affect preferable amino acid substitutions in a site-specific manner (Wako & Blundell, 1994a,b), it is reasonable to incorporate these factors into the alignment process. Position- and residue-specific gap penalties have been introduced in some published alignment methods (Zhu *et al*., 1992; Thompson

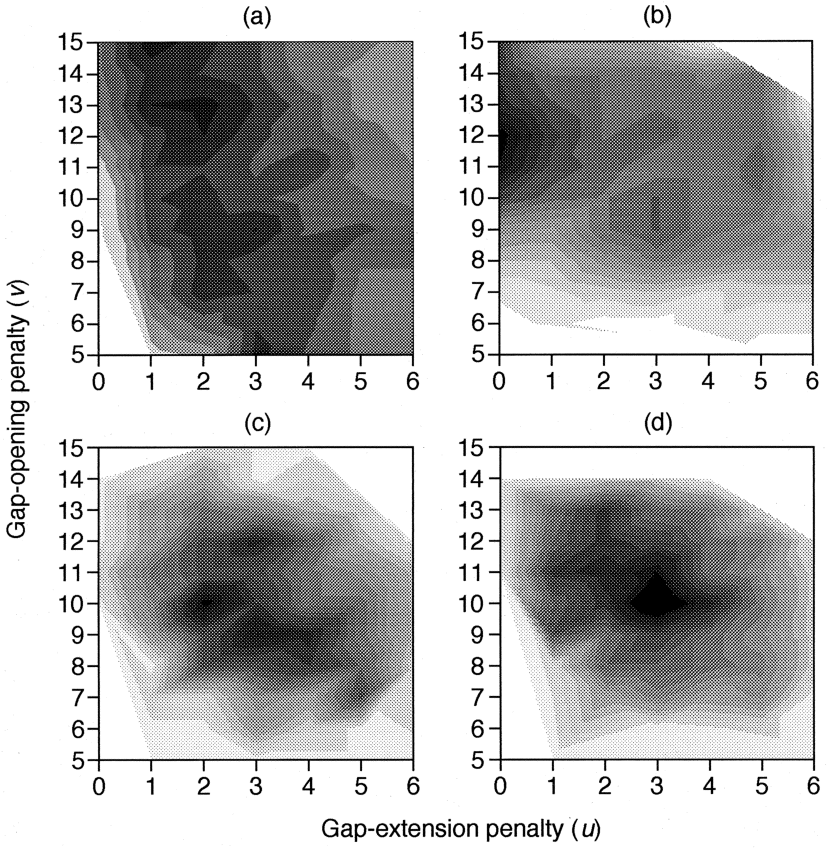

Figure 8. Contour maps of optimal or nearly optimal gap penalty values. The *Z*-axis of this contour map shows the normalized frequency of a *u*–*v* pair that generated an alignment as accurate as at least 99% of the optimal one in each family. Alignment methods used were (a) PWS, (b) CLW, (c) RIO, and (d) DNR. The total numbers of alignments satisfying the above conditions are (a) 1095, (b) 764, (c) 257, and (d) 331. The maxima of the *Z*-axis, which are most densely shaded, are (a) 2.5%, (b) 4.0%, (c) 5.0% and (d) 5.0%.

et al., 1994; Taylor, 1995). Furthermore, CLUSTAL W (Thompson *et al*., 1994) dynamically selects an appropriate substitution matrix, in addition to weighting individual sequences for correction of the phylogenetically biased distribution. The present results (Table 3 and Figure 6) indicate that these modulations as a whole are less effective than iterative optimization of SP or WSP for improving the quality of multiple alignment. This can be ascribed partly to the intrinsic characteristics of optimized SP or WSP, with which gaps in homologous sequences tend to match each other. However, it is likely that iterative refinement of a score incorporating such modulatory factors may lead to further improvement in the quality of alignment.

In summary, I have demonstrated that randomized iterative strategies significantly improve the quality of multiple sequence alignment compared with the currently most popular progressive method. Weighting sequence pairs was also proven to be useful for the improvement of alignment. Given an appropriate scoring system, these strategies are extended to more general alignment problems, such as multiple structural alignment or mixed alignment of sequences and structures.

Methods

Structural alignment database

The primary data set used as the reference of multiple structural alignments was Joy3.2 (see Šali & Overington, 1994). Structural alignments in this data set were obtained either by an automatic alignment program (Šali & Blundell, 1990) or by multiple structural superposition (Sutcliffe *et al*., 1987). The database consisted of 110 entries (families), of which 54 (Table 1) were selected according to the following criteria: (1) those entries consisting of only closely related members (amino acid identity >40%) or of short sequences (length <45 aa) are omitted, because I was interested in assessing the accuracy of sequence alignment of distantly related proteins; (2) the expected calculation time necessary for exhaustive tests under a variety of conditions should not be excessively long; and (3) no family is related to any other. The second criterion eliminated several families with many or long sequences (e.g. lipocalin and cytochrome P450) while the last criterion eliminated some immunoglobulin domains. Exceptionally, the thymidylate synthase family (tms), whose structural alignment contains a long gap, was retained in spite of the violation of the first criterion.

A small fraction of sites in some alignments in Joy3.2 was left blank because of the lack of well-resolved coordinates. Most of these blank sites were filled with the corresponding amino acids in the pertinent or most closely related sequence. Although the filling was done somewhat arbitrarily, the effects of potential errors should be negligible, because the fraction of blank sites was only 0.2% of the total residues.

Some of the structural alignments were further processed so that protruding amino- and carboxyl-terminal regions of few members were trimmed. This was desirable because all the methods used perform global alignment in which terminal and internal gaps are equally penalized. Although distinctive internal and terminal gap penalties might be assignable, further

complexity was avoided because quite a lot of calculations with various combinations of gap opening and extension penalty values, amino acid substitution matrices and alignment methods were already needed.

As a subsidiary set of structural alignments, I chose 17 of 38 familial entries in 3D_ALI (Pascarella & Argos, 1992), so that each 3D_ALI entry should share at least two structures in common with the corresponding Joy3.2 entry. The sequences were processed in the same way as described above.

Extended sequence-structure data set

SWISS-PROT Release 31 protein sequence database was searched for sequences similar to each member in a Joy3.2 entry by use of the BLASTP program (Altschul *et al*., 1989). Fragmental sequences partially homologous to the probe were abandoned, and only the embedded part of a sequence homologous to the probe was extracted when the sequence was longer than the structural domain. The union of the found sequences was multiply aligned together with the members in the structural alignment by a progressive method without referring to the structural alignment. By inspecting a phylogenetic tree obtained from the preliminary multiple alignment, I abandoned those sequences that were clearly not structural members. My intention was not to collect all known sequences belonging to the family, but to collect ''followers'' of each member in the structural alignment. The numbers of sequences thus picked up for the 54 families are listed in Table 1. The procedure ensures that the most remotely related pair of structural members is among the most remotely related sequence pairs in an extended entry.

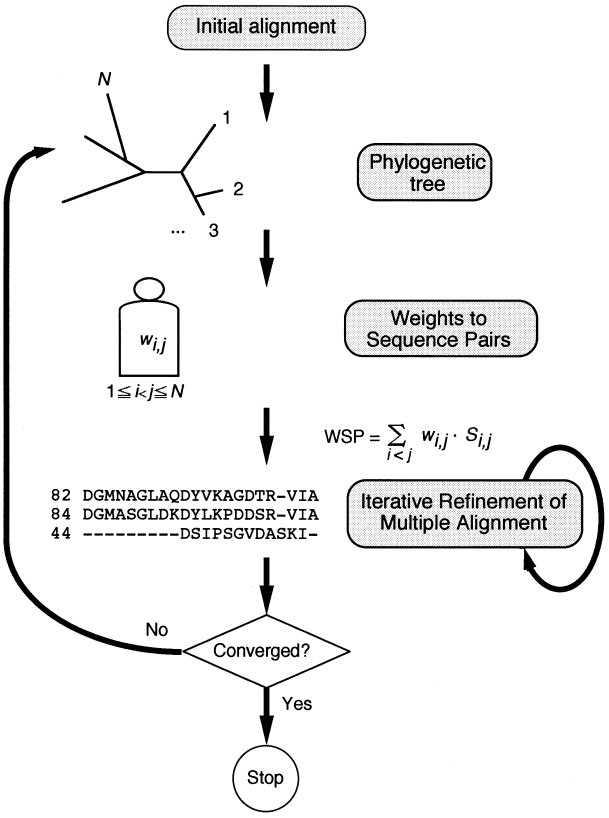

Figure 9. Schematic diagram of the procedures of the doubly nested randomized iterative (DNR) method for multiple sequence alignment. The details of the procedures are explained in the text.

Alignment methods

The four multiple sequence alignment strategies examined were: (1) a progressive method (CLW = CLUSTAL W, Thompson *et al*., 1994); (2) a randomized iterative (RIO) method that optimizes sum-of-pairs (SP) score (Gotoh, 1994); (3) a randomized iterative (RIW) method that optimizes weighted sum-of-pairs (WSP) score (Gotoh, 1995); and (4) a doubly nested randomized iterative (DNR) method optimizing WSP. In addition, the conventional pairwise sequence alignment (PWS) method based on dynamic programming (Gotoh, 1982) was used as a control. For the RIO and RIW methods, five independent series of iterations were carried out from the same initial alignment, and the alignment associated with the best score among them was reported.

The doubly nested randomized iterative method is a new strategy as illustrated in Figure 9. It starts with a preliminary multiple alignment, which may be obtained by any simpler method. In the present study, a single multiple alignment that had been obtained for each family as described in the previous subsection was used as the ''seed'' of all calculations. A set of weights assigned to all the pairs of sequences is calculated by the three-way algorithm (Gotoh, 1995) guided by the phylogenetic tree constructed from the distance values between members in the initial alignment. This set of pair weights is a good approximation of the rationale II weights proposed by Altschul *et al*. (1989). A fixed value for the ''equalization factor'', $F = 1.1$ (Gotoh, 1995) was used throughout all DNR and RIW calculations. The first cycle is the same as that of the RIW method but with only a single series of iteration. After the convergence, a new phylogenetic tree and pair weights are calculated, and the second cycle is started. In this way, the doubly nested iterations are continued until no change in the total WSP score is observed.

For tree reconstruction, either the UPGMA (Sneath & Sokal, 1973) or the Neighbor-Join (Saitou & Nei, 1987) method may be used. However, the present version of my program cannot properly deal with negative edge values, which sometimes appear with the Neighbor-Join method. Hence, I used the UPGMA method for all the calculations presented in this article.

A heuristic routine (Gotoh, 1993b) to locate well-conserved regions in a given multiple alignment is included in the RIO, RIW and DNR methods. These regions, if found, act as ''anchor points'', which are fixed during the following course of iterative refinement. Setting such anchor points can greatly reduce the overall calculation time at the expense of slight loss of rigor. Two factors are responsible for the reduction in computation time. First, the space scanned at each iterative group-to-group alignment step is narrowed depending on the number and the locations of the anchor points. Second, some closely related sequences are bundled to form a single cluster whose internal alignment is fixed. Since the effective number of independent sequences subjected to refinement can be diminished in this way, the total computational cost is also diminished accordingly.

Calculations were made on three network-connected workstations, SUN SPARCstation 2, IPX and 20 running under UNIX (SUN OS4.1.3 or Solaris 2.3). The source codes for DNR/RIW methods written in C programming language are available through anonymous FTP from

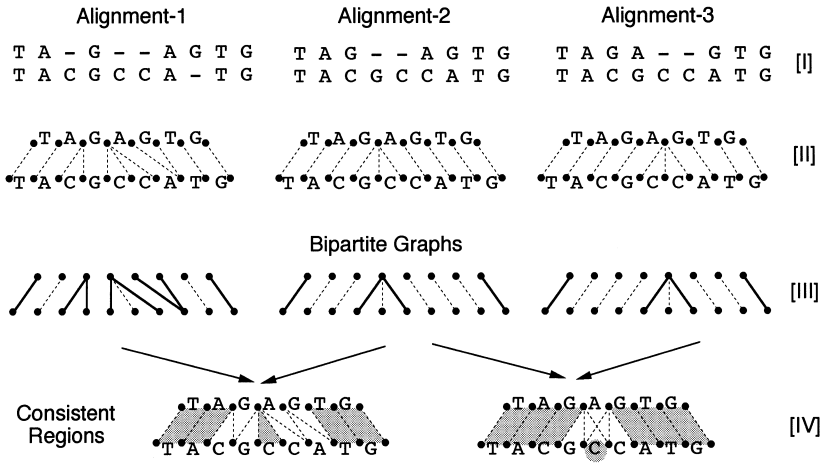

Figure 10. Schematic representation of the method for calculating the degree of consistency between two alignments. Three alignments of the same pair of sequences [I] are individually represented by bipartite graphs [II, III]. Alignment-2 is then superimposed on Alignment-1 ([IV] left) or Alignment-3 ([IV] right). The shaded areas indicate the consistently aligned regions. The C residue marked by a shaded circle ([IV] right) is regarded as consistently aligned under the relaxed conditions but not under the stringent conditions.

ftp.genome. ad.jp in the directory of ~/pub/genome/saitama-cc.

Comparison of alignments

The alignment obtained by the pairwise method was compared directly with its structural counterpart. In other cases, all sequences in an entry of the extended sequence-structure data set were aligned, and each pair of sequences of known structures was extracted from the large alignment and compared with the corresponding structural alignment. The degree of consistency of two alignments was evaluated as described (Gotoh, 1995). In brief, we represent each alignment of two sequences or structures by a bipartite graph and superimpose them (Figure 10). By examining the overlap of edges in the two graphs, we can easily count the number of residues that are consistently aligned. This number divided by the total number of residues in both sequences is defined as the degree of consistency between the two alignments. However, this measure somewhat underestimates the degree of consistency, because the displacement of a single residue may make the entire insertion region inconsistent. To relax the rather stringent definition of consistency, we regard a residue as consistent if it is in insertion regions in both alignments regardless of the precise locations of the insertion points (Figure 10). Ternary consistency among three alignments of the same pair of sequences was defined as the fraction of residues that are consistently aligned in all the tree alignments. Structural features, such as secondary structure and exposed or buried location in a protein molecule, were not taken into consideration.

Amino acid substitution matrices and gap penalties

The four amino acid substitution matrices examined are all log-odds matrices derived from different sets of protein sequence alignments used as the reference data. The classical mutation data matrix (MDM250) at 250 PAM (accepted point mutations) level was obtained by the pioneering work of Dayhoff *et al*. (1978). Jones *et al*. (1992) applied the procedure of Dayhoff *et al*. to a larger set of reference data. The matrix extrapolated to the 250 PAM level (JTT250) was used here. BLOSUM62 (Henikoff & Henikoff, 1992) is also one of a series of matrices obtained on the basis of the BLOCK database, which collects conserved cores in various protein families. The BLOSUM series of matrices are used as the default in the current versions of CLUSTAL W (Thompson *et al*., 1994) and BLASTP (Altschul *et al*., 1990). The fourth matrix examined was that proposed by Gonnet *et al*. (1992). This matrix (GCB250) was derived from exhaustive comparison of sequences in the SWISS-PROT database and normalized to the 250 PAM level. These matrices were chosen because they were rated relatively highly in the examination of Vogt *et al*. (1995). Since my algorithm is designed to minimize distance rather than maximize similarity, the sign of the matrix elements was reversed, but nothing was added to or subtracted from each element.

A total of 77 combinations of gap-opening ($5 \leqslant v \leqslant 15$) and gap-extending ($0 \leqslant u \leqslant 6$) penalty values was examined with the PWS method. The degree of consistency between the sequence alignment and the corresponding structural alignment was calculated as described above. About 25 of these u–v pairs centering around the one that yielded the best consistency value were selected and used in all the other methods except for CLW. Fractional numbers of u or v were not examined, since the optimal range of u–v pairs was found to be rather broad (see Results). Among those alignments obtained with the set of u–v pairs, the one showing the greatest average consistency with the structural one for a given family was subjected to further analysis, unless otherwise mentioned.

For CLUSTAL W, I examined only the default set of amino acid substitution matrices, which are dynamically selected from the BLOSUM series depending on the degree of divergence between sequences to be aligned. In addition, the GAPEXT parameter was varied from 0 to 6, whereas GAPOPEN values were chosen in a similar way to v values in the other multiple alignment methods. On average, about 50 sets of GAPEXT and GAPOPEN were examined for each family. Note that the roles of gap penalties in CLUSTAL W and in other methods are different in detail. To save computation time, /QUICKTREE switch was set in all calculations.

Statistical analysis

Significance in the difference in the performance between two methods or between two amino acid substitution matrices was studied by the Wilcoxon matched-pairs signed rank test in the SPSS program package (Version 6.1, SPSS Inc.). The test takes into account the sign and magnitude of the differences in the degrees of consistency between the structural alignment

and the sequence alignment obtained by the two methods or matrices under examination. Unlike more conventional tests of difference in averages between matched pairs of variables, the non-parametric test does not implicitly assume the probability distribution of the variables. In all tests, I regarded each family, rather than each structural pair, as an independent entity, because individual structural pairs within a family cannot be regarded as independent when multiply aligned. Correlation coefficients were also calculated with the SPSS program package. In addition to the ordinary Pearson method, the Kendall and the Spearman non-parametric methods were used.

Acknowledgements

I thank Dr J. P. Overington for sending me the Joy3.2 software and database, and Drs H. Fujiki and G. L. DeGeorge for reading the manuscript. This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas, ''Genome Informatics'', from the Ministry of Education, Science and Culture of Japan.

References

Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **207**, 647–653.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059–1063.

Barton, G. J. & Sternberg, M. J. E. (1987). A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327–337.

Berger, M. P. & Munson, P. J. (1991). A novel randomized iterative strategy for aligning multiple protein sequences. *Comput. Applic. Biosci.* **7**, 479–484.

Carrillo, H. & Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* **48**, 1073–1082.

Chan, S. C., Wong, A. K. C. & Chiu, D. K. Y. (1992). A survey of multiple sequence comparison methods. *Bull. Math. Biol.* **54**, 563–598.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation, Washington, DC.

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science,* **214**, 149–159.

Feng, D.-F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.

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

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

Gotoh, O. (1993a). Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Applic. Biosci.* **9**, 361–370.

Gotoh, O. (1993b). Extraction of conserved or variable regions from a multiple sequence alignment. In *Proceedings of Genome Informatics Workshop IV*, pp. 109–113, Universal Academy Press, Tokyo.

Gotoh, O. (1994). Further improvement in methods of group-to-group sequence alignment with generalized profile operations. *Comput. Applic. Biosci.* **10**, 379–387.

Gotoh, O. (1995). A weighting system and algorithm for aligning many phylogenetically related sequences. *Comput. Applic. Biosci.* **11**, 543–551.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355–4358.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Hirosawa, M., Totoki, Y., Hoshida, M. & Ishikawa, M. (1995). Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput. Applic. Biosci.* **11**, 13–18.

Hogeweg, P. & Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: An integrated method. *J. Mol. Evol.* **20**, 175–186.

Holm, L. & Sander, C. (1994). Searching protein structure databases has come of age. *Proteins,* **19**, 165–173.

Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure–structure alignment of proteins. *Nucl. Acids Res.* **24**, 206–209.

Johnson, M. S. & Overington, J. P. (1993). A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.* **233**, 716–738.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci.* **8**, 275–282.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. *J. Mol. Biol.* **235**, 1501–1531.

Lesk, A. M., Levitt, M. & Chothia, C. (1986). Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng.* **1**, 77–78.

Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA,* **86**, 4412–4415.

McClure, M. A., Vasi, T. K. & Fitch, W. M. (1994). Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11**, 571–592.

Murata, M., Richardson, J. S. & Sussman, J. L. (1985). Simultaneous comparison of three protein sequences. *Proc. Natl Acad. Sci. USA,* **82**, 3073–3077.

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

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

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.

Pevzner, P. A. (1992). Multiple alignment, communication cost and graph matchings. *SIAM J. Appl. Math.* **52**, 1763–1779.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Šali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A

procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403–428.
Šali, A. & Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3**, 1582–1596.
Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **78**, 35–42.
Sankoff, D., Cedergren, R. J. & Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5 S ribosomal RNA. *J. Mol. Evol.* **7**, 133–149.
Searls, D. B. (1996). Sequence alignment through pictures. *Trends Genet.* **12**, 35–37.
Smith, R. F. & Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5**, 35–41.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy*, Freeman, San Francisco.
Subbiah, S. & Harrison, S. C. (1989). A method for multiple sequence alignment with gaps. *J. Mol. Biol.* **209**, 539–548.
Sutcliffe, M. J., Haneef, I., Carney, D., Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, Part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.
Taylor, W. R. (1995). An investigation of conservation-biased gap-penalties for multiple protein sequence alignment. *Gene,* **165**, GC27–GC35.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **249**, 816–831.
Wako, H. & Blundell, T. L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. Mol. Biol.* **238**, 682–692.
Wako, H. & Blundell, T. L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol.* **238**, 693–708.
Wang, L. & Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1**, 337–348.
Wareham, H. T. (1995). A simplified proof of the NP- and MAX SNP-hardness of multiple sequence tree alignment. *J. Comput. Biol.* **2**, 509–514.
Waterman, M. S. (1986). Multiple sequence alignment by consensus. *Nucl. Acids Res.* **14**, 9095–9102.
Waterman, M. S. & Perlwitz, M. D. (1984). Line geometries for sequence comparisons. *Bull. Math. Biol.* **46**, 567–577.
Waterman, M. S., Smith, T. F. & Beyer, W. A. (1976). Some biological sequence metrics. *Advan. Math.* **20**, 367–387.
Zhu, Z.-Y., Šali, A. & Blundell, T. L. (1992). A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng.* **5**, 43–51.

Edited by F. E. Cohen

(Received 21 May 1996; received in revised form 5 September 1996; accepted 16 September 1996)