BIOINFORMATICS

Vol. 17 no. 8 2001
Pages 700–712

AL2CO: calculation of positional conservation in a protein sequence alignment

Jimin Pei[2] and Nick V. Grishin[1, 2,*]

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received on September 12, 2000; revised on February 23, 2001; accepted on February 28, 2001

ABSTRACT

Motivation: Amino acid sequence alignments are widely used in the analysis of protein structure, function and evolutionary relationships. Proteins within a superfamily usually share the same fold and possess related functions. These structural and functional constraints are reflected in the alignment conservation patterns. Positions of functional and/or structural importance tend to be more conserved. Conserved positions are usually clustered in distinct motifs surrounded by sequence segments of low conservation. Poorly conserved regions might also arise from the imperfections in multiple alignment algorithms and thus indicate possible alignment errors. Quantification of conservation by attributing a conservation index to each aligned position makes motif detection more convenient. Mapping these conservation indices onto a protein spatial structure helps to visualize spatial conservation features of the molecule and to predict functionally and/or structurally important sites. Analysis of conservation indices could be a useful tool in detection of potentially misaligned regions and will aid in improvement of multiple alignments.

Results: We developed a program to calculate a conservation index at each position in a multiple sequence alignment using several methods. Namely, amino acid frequencies at each position are estimated and the conservation index is calculated from these frequencies. We utilize both unweighted frequencies and frequencies weighted using two different strategies. Three conceptually different approaches (entropy-based, variance-based and matrix score-based) are implemented in the algorithm to define the conservation index. Calculating conservation indices for 35 522 positions in 284 alignments from SMART database we demonstrate that different methods result in highly correlated (correlation coefficient more than 0.85) conservation indices. Conservation indices show statistically significant correlation between sequentially adjacent positions i and $i + j$, where $j < 13$, and averaging of the indices over the window of three positions is optimal for motif detection. Positions with gaps display substantially lower conservation properties. We compare conservation properties of the SMART alignments or FSSP structural alignments to those of the ClustalW alignments. The results suggest that conservation indices should be a valuable tool of alignment quality assessment and might be used as an objective function for refinement of multiple alignments.

Availability: The C code of the AL2CO program and its pre-compiled versions for several platforms as well as the details of the analysis are freely available at ftp://iole.swmed.edu/pub/al2co/.

Contact: grishin@chop.swmed.edu

INTRODUCTION

Homologous proteins tend to form distinct families and superfamilies that are characterized by specific sequence motifs, common folds, and related functions. Multiple sequence alignments are routinely used for structure and function prediction and analysis, and for phylogenetic tree reconstruction of protein families. Analysis of positional conservation in an amino acid sequence alignment can aid in detection of motifs and functionally and/or structurally important residues, e.g. at the binding sites (Zuckerkandl and Pauling, 1965; Villar and Kauvar, 1994; Ouzounis *et al.*, 1998). Mapping the conservation information onto a protein 3D structure helps to visualize spatial conservation patterns and to deduce potential functional surfaces of a protein molecule (Sander and Schneider, 1991; Lichtarge *et al.*, 1996; Landgraf *et al.*, 1999; Makarova and Grishin, 1999; Zhang *et al.*, 2000). Several methods of conservation analysis have been used previously to extract functional information from sequence alignments. A vectorial method was proposed in predicting functionally important residues (Casari *et al.*, 1995). Another method, called evolutionary tracing, has been used for defining binding surfaces (Lichtarge *et al.*, 1996) and for identifying functional and structural features in protein families (Landgraf *et*

[*]To whom correspondence should be addressed.

al., 1999; Pritchard and Dufton, 1999). Entropy-based conservation analysis (Sander and Schneider, 1991; Shenkin *et al.*, 1991) has been utilized extensively for characterization of protein families (Atchley *et al.*, 1999; Lowry and Atchley, 2000) and for the analysis of protein folds (Mirny and Shakhnovich, 1999). Other methods of measuring conservation deal with amino acid substitution matrices (Levin and Satir, 1998; Landgraf *et al.*, 1999) or the deviance of amino acid frequencies from the mean values (Lockless and Ranganathan, 1999). However, comprehensive comparison between different estimators of positional conservation has not been done.

Vast numbers of algorithms have been developed to construct multiple sequence alignments (Barton and Sternberg, 1987; Feng and Doolittle, 1987; Taylor, 1988; Lipman *et al.*, 1989; Thompson *et al.*, 1994a; Eddy, 1995; Gotoh, 1996; Notredame *et al.*, 1998; Morgenstern, 1999; to cite a few). Despite significant progress in this direction, none of the available alignment algorithms is perfect (Thompson *et al.*, 1999), leaving the user to cope with the task of manual adjustment of automatically generated alignments. Several approaches have been developed for assessing the quality and reliability of sequence alignment (Vingron and Argos, 1990; Mevissen and Vingron, 1996; Notredame *et al.*, 1998; Thompson *et al.*, 1999; Domingues *et al.*, 2000). Since alignment construction is based on sequence conservation, it appears that a positional conservation estimator is suitable as a measure of alignment quality.

We developed a program (AL2CO) that performs conservation analysis in a comprehensive and systematic way. For a given protein multiple sequence alignment we calculate a conservation index for each position. Twelve different strategies of conservation index calculation have been implemented and their performance has been tested and compared on the alignments from the SMART database (Schultz *et al.*, 1998; Ponting *et al.*, 1999). We analyze the distribution of conservation indices, the correlation of conservation indices between different alignment positions, and the effects of gaps and the number of sequences on the conservation index. By comparing the SMART alignments and raw ClustalW alignments (Thompson *et al.*, 1994a), we test which method of conservation index calculation works best as a measure of a multiple alignment quality. For highly divergent sequences, where sequence-based alignment strategies are likely to fail and cannot be used as a reference, we make a similar comparison between representative structural alignments taken from FSSP database (Holm and Sander, 1996, 1998) and the corresponding raw ClustalW alignments.

ALGORITHM

The algorithm of AL2CO program performs calculations in two steps. First, amino acid frequencies at each position are estimated. The conservation index is then calculated from these frequencies. An optional third step allows the user to average the conservation indices over a window covering a selected number of positions.

Various methods to estimate position-specific amino acid frequencies have been developed. We divide them into three groups:

1.1. *Unweighted amino acid frequencies:* $f_a^u(i) = n_a(i)/n(i)$, where $n_a(i)$ is the number of sequences in which position i is occupied by amino acid a, and $n(i)$ is the total number of aligned sequences in which position i is present (no gap at this position): $n(i) = \sum_{a=1}^{20} n_a(i)$.

1.2. *Weighted amino acid frequencies:* $f_a^w(i) = \sum_{k=1}^{n(i)} \delta(a,k,i)w_k / \sum_{k=1}^{n(i)} w_k$, where w_k is a given weight of a sequence k, and we put $\delta(a,k,i) = 1$ if amino acid a is in sequence k at position i, and $\delta(a,k,i) = 0$ otherwise. Setting equal weights $w_k = w_l$ for all sequences k and l results in unweighted frequencies. The idea behind the weights is to correct for unequal distances between different sequence pairs in the alignment. It appears logical that two close sequences with high similarity should influence amino acid frequencies less than a pair of divergent sequences. Thus, the weight attributed to each of a large family of similar sequences is less than the weight of a single divergent sequence. A wide variety of different methods have been proposed to calculate weights w_k (Altschul *et al.*, 1989; Sander and Schneider, 1991; Gerstein *et al.*, 1994; Henikoff and Henikoff, 1994; Thompson *et al.*, 1994b; Eddy *et al.*, 1995; Gotoh, 1995; Krogh and Mitchison, 1995). We used a modified method of Henikoff and Henikoff that is implemented in PSI-BLAST (Henikoff and Henikoff, 1994; Altschul *et al.*, 1997). In sequence weight calculation, we ignore positions with gaps present in more than 50% of the sequences and invariant positions.

1.3. *Estimated independent counts:* $f_a^{ic}(i) = n_a^{ic}(i)/n^{ic}(i)$ where $n_a^{ic}(i)$ is an estimate of the number of independent observations of amino acid a at position i and $n^{ic}(i) = \sum_{a=1}^{20} n_a^{ic}(i)$. The idea behind this approach is to correct for the correlation between aligned sequences. We use a modified method proposed by Sunyaev *et al.* (1999). The number of independent observations (= counts) of amino acid a at a position i is equal to the effective number of sequences that contain amino acid a at this position. The effective number of sequences in a sample is calculated in the following way. Given a sequence alignment, we define a function F whose value depends on the number of sequences in the alignment. For a given alignment, we can calculate the value of F_{real} (Sunyaev *et al.*, 1999). For a random alignment consisting of N random sequences we calculate $F(N_{\text{random}})$. The value of N_{random} for which $F_{\text{real}} = F(N_{\text{random}})$ corresponds to the effective number of sequences. Sunyaev *et al.* (1999) uses the number of invariant positions as the function F. This number can be easily calculated for a given alignment. Since the number of invariant positions is usually small in divergent sequence alignments, the estimate of F is often imprecise. We choose a more effective F, which is the average number of different amino acids per position. For a random alignment of N random sequences composed of equifrequent amino acids one gets $F = 20(1 - 0.95^N)$ (see appendix for the proof) which allows us to estimate the effective number of sequences as $N_{\text{eff}} = \ln(1 - F/20)/\ln 0.95$. For any F, if amino acid a is present in a single sequence at a position i, its count is $n_a^{ic}(i) = 1$. If amino acid a is present in $n_a(i)$ sequences at a position i, its count is $1 \leq n_a^{ic}(i) \leq n_a(i)$; $n_a^{ic}(i) = 1$ if all sequences with amino acid a at the position i are identical, and $n_a^{ic}(i) = n_a(i)$ if all sequences are independent.

Weighted frequencies have been used extensively in sequence analysis. However, we realize that sometimes researchers are particularly interested in a group of highly similar sequences that might be present in the alignment, and would like to see the conservation within that group not being influenced by divergent sequences. In this case, unweighted frequencies should be used. The presence of an amino acid a at a position i indicates that it is an admissible amino acid at this position, even if it is present in a single sequence only. If this sequence happens to be highly similar to other sequences, then the frequency of amino acid a will be reduced due to the weighting scheme. The strategy of independent counts can avoid this negative effect.

The conservation index is calculated in the next step from amino acid frequencies by one of the following strategies.

2.1. *Entropy-based measure:* $C^e(i) = \sum_{a=1}^{20} f_a(i) \ln f_a(i)$. Traditionally, the order of a system is measured by its entropy. Consequently, it can be used in particular for measuring sequence variability, as was proposed for example by Shenkin *et al.* (1991) and has been implemented in a number of studies (Sander and Schneider, 1991; Atchley *et al.*, 1999; Mirny and Shakhnovich, 1999; Lowry and Atchley, 2000). Entropy for a position i is maximal if all 20 amino acids at this position have equal frequencies. We use entropy with the reverse sign defined on position-specific frequencies $f_a(i)$ to estimate the conservation index. Entropy does not take into account possible bias in amino acid composition or similarities among amino acids. The latter defect can be partially corrected for by forming groups of amino acids with similar properties and calculating frequencies for these groups (Atchley *et al.*, 1999; Mirny and Shakhnovich, 1999).

2.2. *Variance-based measure:*
$C^v(i) = \sqrt{\sum_{a=1}^{20}(f_a(i) - f_a)^2}$, where f_a is the overall frequency for amino acid a in the alignment, i.e. $f_a = \sum_{i=1}^{l} n_a(i)/\sum_{i=1}^{l} n(i)$ if $f_a(i)$ were estimated using the methods ***1.1*** and ***1.3*** (see above), and $f_a^w = \sum_{i=1}^{l}\sum_{k=1}^{n(i)} \delta(a,k,i)w_k/\sum_{i=1}^{l}\sum_{k=1}^{n(i)} w_k$ for the method ***1.2***, and l is the total number of aligned positions. A similar method has been employed in the estimation of evolutionary conservation and coupling parameters (Lockless and Ranganathan, 1999). The position with amino acid frequencies $f_a(i)$ equal to the overall amino acid frequencies f_a in the aligned sequences will result in $C^v(i) = 0$. Alternatively, $C^v(i)$ reaches its maximum for the position occupied by an invariant amino acid whose frequency in the alignment is minimal. The advantage of this method is the use of overall amino acid frequencies, which differ for different protein families. This measure does not take into account similarities among amino acids. To utilize such information, usually presented as a scoring matrix, we opt for using

2.3. *Sum of pairs measure:*
$C^p(i) = \sum_{a=1}^{20}\sum_{b=1}^{20} f_a(i) f_b(i) S_{ab}$, where S_{ab} is an amino acid scoring matrix. This conservation index will be higher for the positions occupied by more similar amino acids. Since the diagonal scores might differ for different amino acids, conservation indices for invariant positions will depend on the amino acid type. For example, positions with invariant Trp will have the highest index if BLOSUM62 matrix is utilized. If the user wants to make conservation indices equal to each other for all invariant positions, the scoring matrix can be normalized: $S'_{ab} = S_{ab}/\sqrt{S_{aa}S_{bb}}$. We also allow the user to modify the scores according to the formula: $S''_{ab} = 2S_{ab} - (S_{aa} + S_{bb})/2$. This adjustment makes $C^p(i)$ equal to the original matrix score S_{ab} for the alignment of two sequences with amino acids a and b at a position i ($f_a(i) = f_b(i) = 0.5$, $C^p(i) = 0.5 * 0.5 * S''_{ab} + 0.5 * 0.5 * S''_{ba} + 0.5 * 0.5 * S''_{aa} + 0.5 * 0.5 * S''_{bb} = S_{ab}$).

Despite the fact that conserved positions tend to cluster to form motifs, conservation indices for adjacent sequence positions usually show large variation. Averaging the indices over a window can smoothen the conservation profile along a sequence and facilitate sequence motif detection. For a given window of size w at position i, we average the indices from position $i-(w-1)/2$ to position $i+(w-1)/2$ if w is odd, and from position $i-(w-2)/2$ to position $i+w/2$ if w is even. The average value is assigned as a new index to position i. When averaging is applied to the positions near N-(C-) terminus (N terminus: $i < w/2$ if w is even and $i < (w+1)/2$ if w is odd; analogously for C terminus), the window sizes are reduced to completely cover the sequence at the N-(C-) terminus and the target position is placed in the middle of the window: e.g. for the N-terminus, the new window size is $w' = 2i - 1$ for positions i if $w' < w - 1$. To compensate for the increase of the index variance caused by the decrease of the window size, we also adjust the index as $C'(i) = \overline{C} + (C(i) - \overline{C})\sqrt{w'/w}$ where $C(i)$ is the conservation for position i near the alignment termini averaged over a smaller window of size w', w is the given window size, $\overline{C}$ is the mean conservation index with window size 1.

Calculated conservation indices can then be normalized to make comparisons possible among sets of indices calculated for different alignments or using different methods: $C_n(i) = (C(i) - \overline{C})/\sigma_C$, $\overline{C} = \sum_{i=1}^{l} C(i)/l$, $\sigma_C = \sqrt{\sum_{i=1}^{l} (C(i) - \overline{C})^2/(l-1)}$, where l is the number of positions in the alignment.

IMPLEMENTATION AND DISCUSSION

Multiple sequence alignments taken from SMART database (Schultz *et al.*, 1998; Ponting *et al.*, 1999) were used to compare different estimation methods of positional conservation. SMART database is well curated and in our opinion represents a large sample of alignments with high quality that are adjusted manually according to structural and/or functional considerations (Schultz *et al.*, 1998; Ponting *et al.*, 1999). Alignments that contain less than 20 sequences or less than 40 positions with gap fraction less than 0.5 were not considered in this analysis, resulting in 284 alignments that were used for conservation index calculation (for the information on the alignments see ftp://iole.swmed.edu/pub/al2co/SMART_list/). A total of $4 * 3 = 12$ methods that differ in weighting schemes and conservation–calculation strategies as discussed above were used. We designate the methods by two numbers with an underscore in between. The first number refers to the conservation–calculation strategy: 1, entropy-based measure; 2, variance-base measure; 3, sum-of-pairs measure using identity matrix; 4, sum-of-pairs measure using BLOSUM62 matrix. The second number refers to frequency estimation strategy: 1, unweighted frequencies; 2, Henikoff-weighted frequencies; 3, independent-count based frequencies.

Correlation between methods

For each of the 284 SMART alignments, positions with gaps in no less than 50% of sequences were discarded and conservation indices were calculated for the remaining positions (35 522 total) using all the $4 * 3 = 12$ proposed methods. The resulting conservation indices were then normalized to zero mean and unity variance for each individual alignment. ($C_k^m(i)$ is a conservation index calculated by the method $m = 1, \ldots, 12$ for position $i = 1, \ldots, l_k$ in alignment $k = 1, \ldots, 284$, $\overline{C_k^m} = \sum_{i=1}^{l_k} C_k^m(i)/l_k = 0$, $\sigma^2_{C_k^m} = \sum_{i=1}^{l_k} C_k^m(i)^2/(l_k - 1) = 1$). Correlation coefficients between the indices obtained by different methods m and s

$$\left(\rho(C^m, C^s) = \frac{\sum_{k=1}^{284} \sum_{i=1}^{l_k} C_k^m(i) C_k^s(i)}{\sqrt{\sum_{k=1}^{284} \sum_{i=1}^{l_k} C_k^m(i)^2 \sum_{k=1}^{284} \sum_{i=1}^{l_k} C_k^s(i)^2}}\right)$$

for all pairs of the 12 methods are presented in Table 1. All correlation coefficients are no less than 0.85, showing good correspondence among methods. For the four conservation–calculation strategies, strategy number 4, the one applying the BLOSUM62 matrix shows smallest correlation coefficients with the other three strategies. Calculations using the BLOSUM62 matrix take into account similarities among amino acids, while the remaining three methods do not. Figure 1 shows the plot of conservation indices of the two methods with the lowest correlation (Methods 2_1 and 4_3). These two methods differ in schemes for frequency calculation and in strategies for conservation–calculation: method 2_1 is a variance-based measure with no weighting and Method 4_3 is a sum-of-pairs BLOSUM62 measure with independent counts weighting scheme.

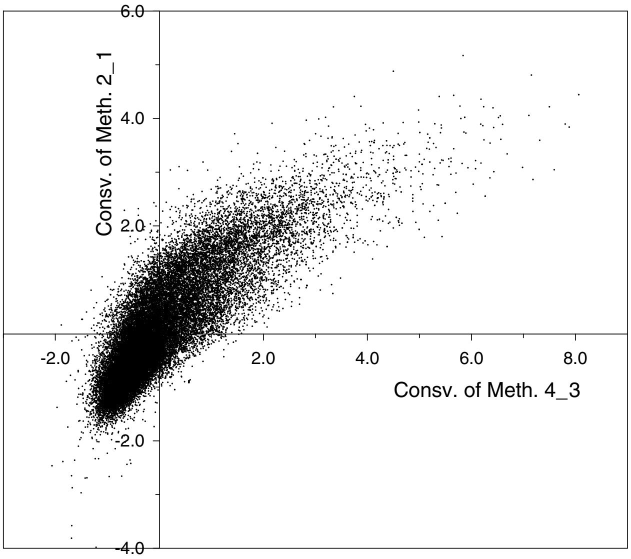

Fig. 1. Correlation plot between two methods that show the lowest correlation coefficient. The two methods are: 2_1 (variance based measure with no weighting) and 4_3 (sum-of-pairs measure with the BLOSUM62 matrix and independent count weights). 35 522 data points are shown. The correlation coefficient is 0.85.

Table 1. Correlation between conservation indices calculated by different methods[1]

1_1	100											
1_2	99	100										
1_3	97	98	100									
2_1	97	96	92	100								
2_2	97	98	94	99	100							
2_3	96	97	97	96	97	100						
3_1	97	96	93	97	97	96	100					
3_2	96	97	94	96	97	96	99	100				
3_3	92	94	96	90	92	97	95	97	100			
4_1	93	93	92	93	93	93	93	93	91	100		
4_2	92	93	92	91	93	94	92	93	92	99	100	
4_3	89	90	93	**85**	87	93	87	89	93	95	97	100
Method	1_1	1_2	1_3	2_1	2_2	2_3	3_1	3_2	3_3	4_1	4_2	4_3

[1] Correlation coefficients are shown in percent. See text for method abbreviations.

We paid special attention to the matrix-based sum-of-pairs method since it is the only one that can take into account similarities among amino acids. To evaluate the effects of different scoring matrices on the conservation index, we performed correlation analysis of conservation indices obtained using eight different scoring matrices: 30, 45, 62, and 80 from the BLOSUM series (Henikoff and Henikoff, 1992), PAM250 (Dayhoff *et al.*, 1978), GONNET (Gonnet *et al.*, 1992), and two recent structure-derived matrices, namely Structure-Derived Matrix (SDM) and Homologous Structure-Derived Matrix (HSDM) (Prlic *et al.*, 2000). Independent counts scheme was used to estimate amino acid frequencies. All correlation coefficients are no less than 0.93 (data not shown), suggesting that different matrices perform in a similar way. PAM250 matrix shows the smallest correlation coefficients to the others (data not shown). PAM250 is the oldest one and is derived by matrix multiplication from the mutation rates estimated using very similar sequences (Dayhoff *et al.*, 1978), while other matrices are obtained from direct statistical estimation of frequencies of aligned amino acid pairs in more divergent sequences. The highest correlation coefficient (99.7%) is found between the two structure-derived matrices SDM and HSDM. HSDM is obtained from a subset (77 presumably homologous proteins pairs out of the 122 structurally aligned pairs that might contain analogs) of a database that is used to derive SDM (Prlic *et al.*, 2000).

Correlation between positions

It is well known that alignments contain regions of high conservation (sequence signatures or motifs) with variable regions between them (Henikoff *et al.*, 1999). To clarify applicability of conservation indices calculated by different methods for motif detection, we calculated correlation of conservation indices at positions i and $i + j$ ($j = 1, \ldots, 20$) for all of the 284 SMART alignments using 12 methods. All 12 methods show rather similar traits in positional correlation (four of them are shown in Figure 2). For $j \leq 12$, the corresponding positions i and $i + j$ display significant positive correlation (P-value <0.05). This correlation pattern shows that the positions that are sequentially close to each other tend to have the same conservation properties (high conservation or low conservation), for stretches on average up to 12 residues in length. Directly adjacent positions (i and $i + 1$) have, on average, the highest correlation. The correlation drops when the positions get further apart in sequence. All correlation coefficients do not differ from 0 significantly for j larger than 12, indicating that long-range sequential coupling between positions are not the same in different protein families. Interestingly, the four peaks ($j = 1, 4, 7, 11$) show periodicity that is consistent with that of an α-helix where residues in positions $i, i+1$, $i+4, i+7$, and $i+11$ are spatially close. It appears that the medium-range coupling ($3 \leq j \leq 12$) is mainly caused by α-helices. β-strands, on the other hand, tend to contribute to short-range coupling ($1 \leq j \leq 4$) since they are usually short and adopt extended conformation.

Despite statistically significant correlation between conservation indices for positions close in sequence, the smoothness of the conservation index versus position number is low. Averaging of conservation indices over a window of w positions smoothens the indices. We calculated correlation coefficients between conservation indices at positions i and $i + j$ ($j = 1, \ldots, 21$) for different window sizes ($w = 1, \ldots, 20$) and different methods of conservation index estimation. Again all 12 methods show similar properties. One example is shown in Table 2 for method 1_3 of window size w up to 5 and position difference j up to 15. If window size w is larger than j, then correlation between i and $i + j$ is biased (and always significant) since the two windows for i and $i + j$ overlap (Table 2, shown in italic numbers).

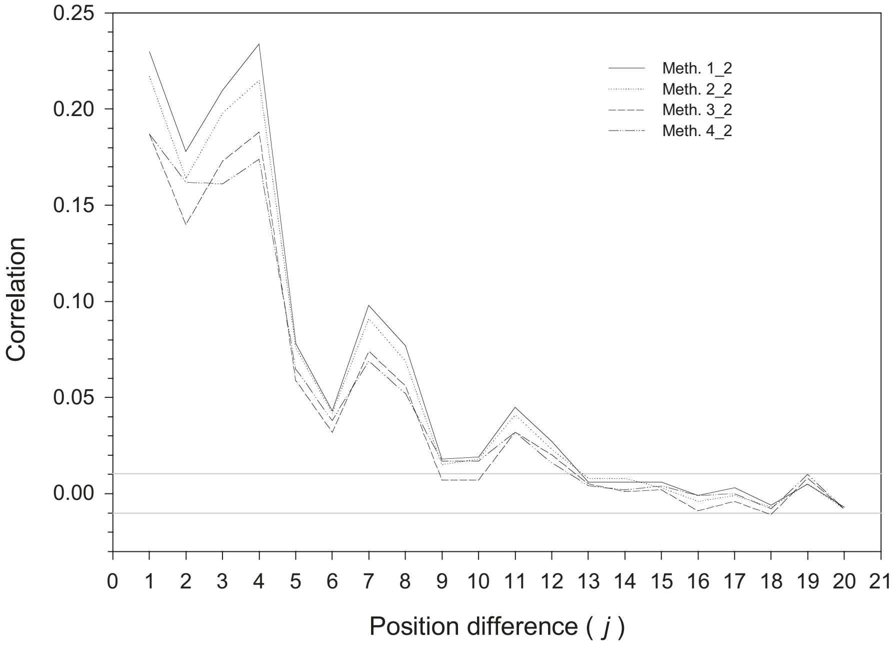

Fig. 2. Correlation between conservation indices at positions i and $i + j$ for different methods. The two gray lines parallel to the horizontal axis mark the area of insignificant difference from zero correlation ($P > 0.05$) in between. Methods are designated by two numbers with an underscore in between (x_2). The first number refers to the conservation–calculation strategy: 1, entropy-based measure; 2, variance-base measure; 3, sum-of-pairs measure using identity matrix; 4, sum-of-pairs measure using BLOSUM62 matrix. The second number (2) refers to Henikoff weighting scheme.

Diagonal elements ($w = j$, in bold) mark the start of an unbiased correlation along each row in Table 2. For all window sizes, statistically significant correlation lasts for about 11–12 residues beyond the window size. For each method, the first unbiased correlation (diagonal elements) increases to its maximum value at a window size of 3 (underlined, in bold), and then gradually drops with the increase of window size. The increase is due to the increase of smoothness, and the drop is caused by the averaging effect and the increased window size. Thus we propose that window size 3 should be ideal for smoothing of conservation indices and optimal for motif detection.

Distribution of conservation indices

The histogram of the normalized (to zero mean and unit variance) conservation indices for the 35 522 positions in 284 alignments is shown as discrete points in Figure 3 (bin size 0.2σ). It exhibits a single sharp mode at about -0.7σ, drops fast at its left side and has a shoulder at its right side. Such shape indicates a mixed distribution that is likely to have several distinct components. We use the sum of two Gaussian distributions to fit the data (Figure 3). The two Gaussians may serve as a rough approximation of low-conservation and high-conservation components respectively, although the overall fit is far from perfect. The low-conservation component (on the left) contributes mostly to the sharp peak. The high conservation component (on the right) gives rise to the shoulder. It has a larger variance than the low conservation component and may actually be further decomposed into several sub-components. The low and high conservation components cover almost equal areas, indicating that about 50% of all positions display significant conservation while the remaining positions are not conserved.

The effect of gaps

A question remains regarding the treatment of gaps in conservation-index calculation. In the former analysis, gaps were ignored and only positions with gaps present in no more than 50% of sequences were considered. It is clear that a gap should not be treated as a 21st letter for calculating frequencies since in that case the positions containing mostly (or entirely) gaps will be described as highly (or completely) conserved. However, the presence of a gap character at a position means the absence of the corresponding amino acid in the protein structure. This indicates a lack in the backbone chain and, thus, seems to be ‘less conserved’ than the mere change of an amino acid side chain corresponding to the substitution of one amino acid by another. In any case, deletion should represent

Table 2. Correlation coefficients between positions in alignments for different window sizes[1]

Window	Position difference (j)														
size (w)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.23**	0.19	0.21	0.23	0.08	0.04	0.10	0.08	0.02	0.02	0.04	0.03	0.00	0.00	0.00
	0	0	0	0	0	0	0	0	1.13	0.28	1e−8	2e−3	47.9	61.1	55.1
2	*0.67*	**0.34**	0.35	0.31	0.18	0.11	0.13	0.11	0.05	0.04	0.06	0.04	0.02	0.01	0.00
	0	0	0	0	0	0	0	0	0	2e−8	0	6e−8	0.93	24.8	59.2
3	*0.82*	*0.63*	***0.42***	0.36	0.26	0.19	0.15	0.13	0.09	0.07	0.06	0.05	0.03	0.02	0.01
	0	0	0	0	0	0	0	0	0	0	0	0	3e−4	1.78	15.4
4	*0.88*	*0.74*	*0.58*	**0.41**	0.32	0.25	0.20	0.15	0.12	0.10	0.08	0.06	0.04	0.03	0.01
	0	0	0	0	0	0	0	0	0	0	0	0	7e−7	7e−3	4.33
5	*0.90*	*0.78*	*0.66*	*0.52*	**0.36**	0.28	0.23	0.18	0.14	0.11	0.09	0.07	0.05	0.04	0.03
	0	0	0	0	0	0	0	0	0	0	0	0	0	2e−5	2e−2

Method 1_3 (entropy-based measure with independent counts frequencies) was used. For each element in the table, the upper number is the correlation coefficient between the sites i and $i + j$ for the window size w, the lower number is its significance (P-value, in percent) of difference from zero correlation. P-value in percent is shown as 0 if it is less than 10^{-8}.

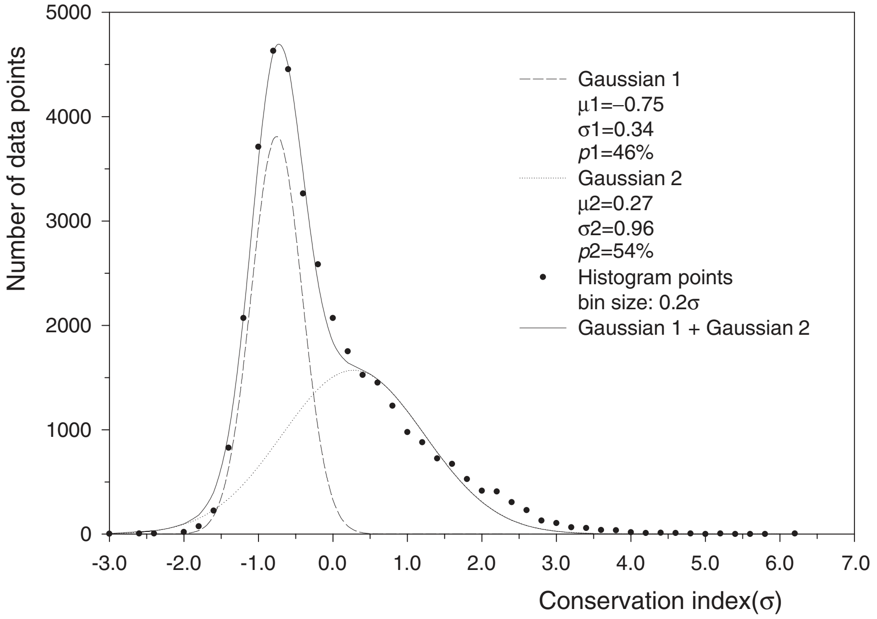

Fig. 3. Distribution of conservation indices for the method 1_3. 35 522 normalized conservation indices are used. Bin-size is 0.2σ. The number of data points in each bin is shown as circles. These data were fitted to the sum of two normal distributions using the SigmaPlot software: $0.2\sigma N\left(\frac{p1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right]+\frac{p2}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right]\right)$, where N is the total number of data points (35 522), $p1$ and $p2$ are the fractions of the two Gaussians. The best-fit parameters are shown in the upper right corner.

an event different from a substitution. It is also clear that ignoring gaps is not appropriate. For example, when at a given position only a few sequences contain amino acids and most have gaps, estimation of conservation that ignores sequences with gaps in that position will be on average higher, and the position would be completely conserved if only one sequence contains an amino acid.

To study the conservation properties at positions with varying fractions of gaps, we employed the following strategy. Conservation indices were calculated for every position in each of the 284 SMART alignments without considering the effects of gaps. However, the mean (μ) and standard deviation (σ) for normalization was calculated only using the positions that do not contain gaps at all. Then the positions containing amino acids in less than 20 sequences were discarded and the remaining positions with gaps were binned by the fraction of gaps with each bin containing the same number of data points. For each bin, the average of conservation indices was calculated and plotted as shown in Figure 4 for method 1_2. Positions with gaps indeed show lower average conservation than positions without gaps (mean value 0). The average conservation index drops sharply to about -0.8σ when the gap percentage increases from 0 to about 5% . The reasons for the decrease are two-fold. On the one hand, structurally or functionally important positions with high conservation tend to contain no gaps. On the other hand, the presence of gaps means that this position is to some extent unnecessary, so the conservation should be lower. Average conservation index stays about -1.0σ with gap percentage ranging from 10 to 70%, suggesting that there is little effect of gap percentage on positional conservation within this range. For gap percentages larger than 70%, the average conservation index gradually increases. Two reasons may account for this increase: the small number of the effective sequences for conservation calculation at positions with high gap fraction and, secondly, the fact that when a gap is present in most of the sequences, the remaining ungapped sequences tend to form subfamilies with distinct conservation features at that position. Since these subfamilies are not representative for the whole set of sequences, it seems reasonable to consider such positions as less conserved even if the apparent conservation indices calculated by ignoring gaps are high. Based on this analysis, we recommend to treat positions with gaps in the following way: (1) calculate conservation indices for all positions with gap percentage less than a given threshold (e.g. 50%) and estimate the mean (μ) and standard error (σ); (2) set the conservation indices for all positions with gap percentage higher than the threshold to $\mu - 1.0\sigma$.

The effect of the number of sequences

All 284 SMART alignments used in this study contain at least 20 sequences to ensure appropriate sampling from sequence space. Here we address the effect of the number of sequences on the estimation of conservation properties. Of the 284 SMART alignments, those containing from 60 to 300 sequences were selected for this study. For each of these 107 alignments, sets of sub-alignments with various numbers of sequences (2, 4, 6, etc.) were generated by random sampling of sequences from the original complete alignment. Conservation indices were calculated for each sub-alignment and compared to those calculated for the complete alignment by calculating correlation coefficient between the two vectors of indices. The correlation coefficient increased rapidly with the increase of the number of sequences in sub-alignments. For most of alignments, less than 30 randomly selected sequences were enough to bring the correlation coefficient above 0.85. The number of sequences required to reach the correlation coefficient of 0.85 has a mean of 13.7 and a standard deviation of 5.6 for the 107 selected alignments. Based on this observation, we conclude that about 20 representative sequences are usually enough for estimating the conservation patterns in a protein family.

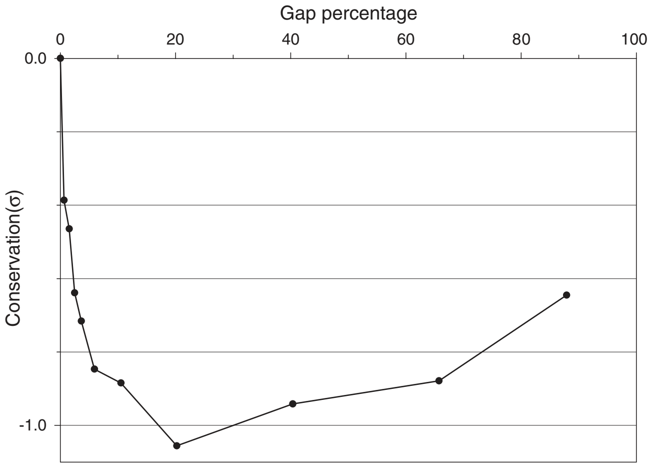

Fig. 4. Distribution of conservation indices at positions with gaps. The conservation indices generated by the method 1_2 were normalized using the mean and the variance of the conservation indices at positions without gaps. The conservation indices at positions with no less than 20 ungapped sequences and with at least one gap were binned into ten sets with equal number of data points in each set. The average conservation indices in each set are plotted against the average gap percentage in that set.

Conservation as a measure of alignment quality

Comparison between SMART and ClustalW alignments. To probe if conservation indices can aid in evaluation of alignment quality, we generated alignments from the sequences for each of the 284 SMART domains by ClustalW program (version 1.7) with default parameters (BLOSUM series matrices, gap open penalty 10, gap extension penalties for pairwise/multiple alignments 0.1/0.05) (Thompson *et al.*, 1994a; Jeanmougin *et al.*, 1998). Curated and manually adjusted SMART alignments are expected to be of higher quality than the raw ClustalW alignments. We compared the average values of conservation indices (without normalization) for positions with gap percentage less than 50% for SMART alignments and ClustalW alignments of SMART domains. In all 12 methods, the average conservation of SMART alignments is significantly higher than that of ClustalW alignments (P-value<0.05, Table 3). The difference is small, suggesting that ClustalW performs fairly well for sequences from SMART database. Higher conservation in SMART alignments versus ClustalW alignments illustrates well that it is possible to improve a multiple alignment algorithm, and a conservation index is

Table 3. Differences between SMART alignments and ClustalW alignments[1]

	Method											
	1_1	1_2	1_3	2_1	2_2	2_3	3_1	3_2	3_3	4_1	4_2	4_3
SMART average conservation	−1.73	−1.79	−1.97	0.433	0.412	0.344	0.297	0.275	0.214	1.18	1.02	0.609
ClustalW average conservation	−1.75	−1.80	−1.99	0.429	0.407	0.339	0.293	0.272	0.211	1.14	0.978	0.571
P-value of the difference	0.005	0.001	0.001	0.012	0.002	0.001	0.046	0.022	0.021	0.008	0.004	0.002
SMART correlation for $j = 3$ and $w = 3$	0.423	0.422	0.425	0.402	0.402	0.415	0.366	0.364	0.361	0.355	0.356	0.359
ClustalW correlation for $j = 3$ and $w = 3$	0.440	0.439	0.444	0.415	0.416	0.432	0.380	0.378	0.375	0.370	0.371	0.378
P-value of the difference	0.005	0.05	0.002	0.037	0.025	0.006	0.030	0.030	0.030	0.021	0.021	0.003

[1] Correlation coefficients are calculated between the sites i and $i + j$ for the window size w.

Table 4. Differences between FSSP structural alignments and ClustalW alignments[1]

	Method										
	1_2	2_2	3_2	4_2(B62)	BLOSUM30	BLOSUM45	BLOSUM80	PAM250	GONNET	SDM	HSDM
FSSP average conservation	−2.13	0.287	0.158	−0.040	0.436	0.213	−0.548	0.134	0.260	0.413	0.420
ClustalW average conservation	−2.18	0.269	0.146	−0.218	0.318	0.026	−0.843	−0.046	0.070	0.149	0.076
P-value of the difference	4e−4	5e−5	0.001	2e−8	0.001	2e−7	2e−8	2e−7	1e−8	2e−9	3e−9
FSSP correlation for $j = 3$ and $w = 3$	0.337	0.382	0.362	0.331	0.280	0.329	0.338	0.300	0.322	0.336	0.331
ClustalW correlation for $j = 3$ and $w = 3$	0.253	0.259	0.245	0.215	0.191	0.212	0.225	0.129	0.184	0.209	0.214
P-value of the difference	0.020	5e−4	0.001	0.002	0.017	0.001	0.002	6e−6	2e−4	5e−4	0.001

[1] Correlation coefficients are calculated between the sites i and $i + j$ for the window size w.

a reasonable indicator of the alignment quality. It is also apparent that given an index calculation method, using unweighted frequencies gives less significant differences between SMART and ClustalW alignments than weighted frequencies. The entropy-based scheme produces the lowest P-value of the difference between SMART and ClustalW alignments (Table 3).

Further, we compared the correlation between positions for normalized conservation indices in SMART and ClustalW alignments. The largest unbiased correlation coefficients (window size $w = 3$, positional difference $j = 3$) are shown in Table 3. For all 12 methods, The SMART correlation is slightly but significantly smaller than the ClustalW correlation. The lowest P-value results from the comparison of the entropy-based estimates. Additionally, the weighting scheme based on independent counts (Methods x_3) shows the smallest P-values for all strategies.

The results of comparisons between SMART and ClustalW alignments show that conservation index should be a valuable tool for alignment quality assessment. It appears that a weighting scheme is necessary and independent counts might be better than the Henikoff weights. Additionally, the usage of the simple entropy-based conservation measure does not seem to be inferior to that of the BLOSUM62-based measure.

Comparison between FSSP and ClustalW alignments. The protein families from the SMART database usually consist of close homologues characterized by relatively high sequence similarity. To compare the performance of conservation indices in highly divergent but structurally similar proteins, where sequence-based alignment strategies are likely to fail (Vogt *et al.*, 1995; Jaroszewski *et al.*, 2000), we compared the structural-based alignments taken from the FSSP database (Holm and Sander, 1996, 1998)

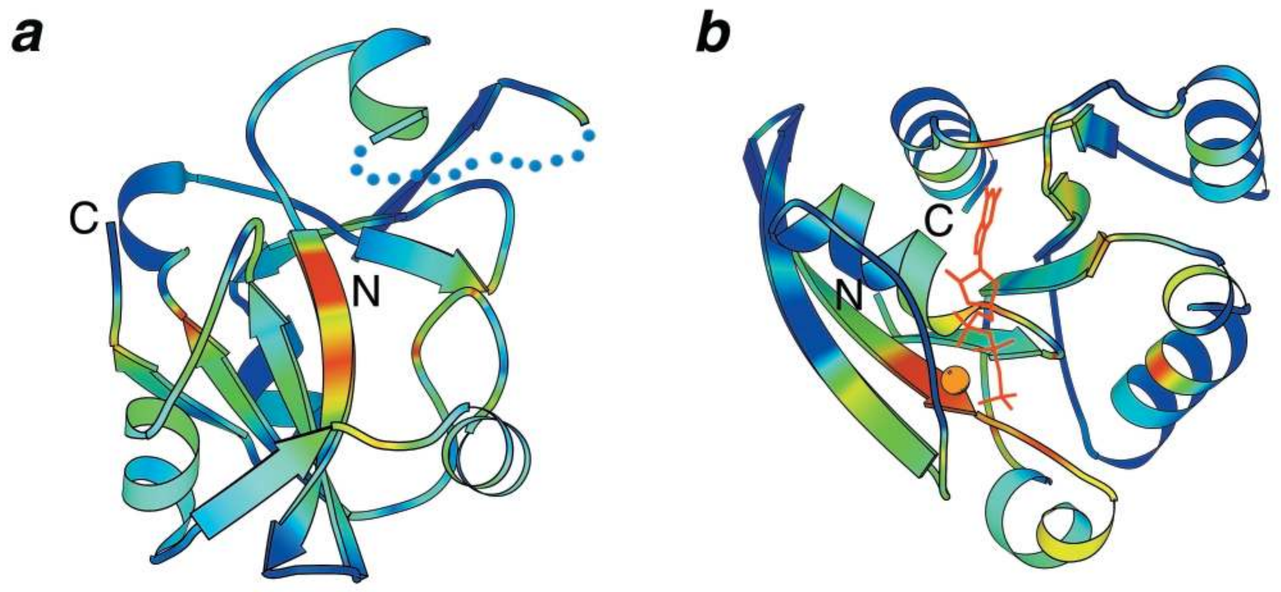

Fig. 5. Conservation mapping onto 3D-structures. The figures were drawn by BOBSCRIPT (Esnouf, 1997). Red and blue correspond to the highest and the lowest conservation respectively. Two examples are shown. (a) The YBAK family protein of unkown function. The alignment for conservation calculation and the structure (PDB entry 1DBU) are according to Zhang *et al.* (2000). (b) The Rab family of small G proteins. Conservation indices are calculated using the alignment from the SMART database and the structure template (PDB entry 3RAB) is according to Dumas *et al.* (1999). GppNHp is displayed in red lines; the Mg^{2+} is shown as an orange ball.

and the corresponding ClustalW alignments. Despite the rapidly increasing number of determined structures, 3D-structure database is still small compared to the protein sequence database. Conservation index calculation usually requires no less than 15 aligned sequences. Due to these restrictions, we selected the eight largest representative FSSP structural alignments, in which most of pairwise identities are in the 'twilight zone' (less than 25%) and the number of sequences is no less than 18 (the list is available at ftp://iole.swmed.edu/pub/al2co/FSSP_list/). Five of these alignments cover the proteins with the most widely spread folds, such as immunoglobin, OB, Rossmann, ferrodoxin-like, and TIM barrel, and have been studied previously by Mirny and Shakhnovich (1999). Three additional alignments were of globin-like superfamily, trypsin-like serine proteases, and P-loop nucleotide triphosphate hydolases (Murzin *et al.*, 1995). Structurally non-equivalent extensions at N-(C-) terminus were removed from the original FSSP alignments.

The procedures described above for comparing SMART and ClustalW alignments were carried out to compare the FSSP structural alignments and the ClustalW alignments (Table 4). Since the structure-based alignments were selected to contain very divergent sequences, results obtained with different weighting schemes do not differ significantly and are not shown. Implementation of the Henikoff weights is illustrated in Table 4. Both the average conservation values and the correlation coefficients are lower in Table 4 compared to Table 3 (methods 1_2, 2_2, 3_2, and 4_2), due to the lower sequence conservation in FSSP alignments than in SMART alignment. However, despite a much smaller dataset (about 1000 positions) from FSSP than that from SMART (about 35 000 positions), the P-values of the differences are more significant in FSSP-ClustalW comparison than those in SMART-ClustalW comparison. This observation suggests that conservation properties are still strong in structure-based alignments of divergent sequences, and ClustalW performs poorly when the sequences display low similarity with each other (Thompson *et al.*, 1999).

It is clear that for divergent sequences from FSSP, BLOSUM62 matrix sum-of-pairs measure (Table 4, method 4_2) shows the most significant differences between FSSP and ClustalW alignments (among methods 1_2, 2_2, 3_2, and 4_2), and sum-of-pairs measure based on identity matrix (Table 4, methods 3_2) offers the poorest discrimination. This is consistent with the notion that identity between divergent sequences is very low, but similarity still exists. Thus we also compared the conservation indices calculated with eight different amino acid scoring matrices: 30, 45, 62, and 80 from the BLOSUM series (Henikoff and Henikoff, 1992), PAM250 (Dayhoff *et al.*, 1978), GONNET (Gonnet *et al.*, 1992), and two structure-derived matrices, namely Structure-Derived Matrix (SDM) and Homologous Structure-Derived Matrix (HSDM) (Prlic *et al.*, 2000) (Table 4). Among these matrices, SDM and HSDM show the most significant differences in average conservation indices. This can be explained by the facts that SDM and HSDM are derived from 3D-structure comparisons similarly to the alignments in FSSP database, and all the protein pairs used to obtain these matrices had sequence identity below 30% (Prlic *et al.*, 2000). On the contrary, the BLOSUM30 matrix, which is derived from the sequence-based alignments of lower identity between sequences, offers the poorest discrimination between FSSP and ClustalW alignments.

The differences in positional correlation show different tendencies in FSSP–ClustalW and SMART–ClustalW comparisons (Tables 3 and 4). The FSSP correlation is significantly higher than the ClustalW correlation (Table 4), while the SMART correlation is lower than the ClustalW correlation (Table 3). This apparent contradiction can be easily explained. The observed behavior of correlation coefficients appears to be similar to the first-rise-then-fall phenomenon for the correlation coefficients depending on different window sizes (Table 2, bold numbers, discussed above). For the SMART–ClustalW case, the differences between SMART alignments and ClustalW alignments are small. However, ClustalW could misalign a few positions with relatively small shifts, resulting in a smoothing effect of the conservation indices. This may account for the slightly higher correlation between positions in ClustalW alignments than in SMART alignments. For more divergent sequences, such as the ones taken from FSSP, ClustalW alignments are significantly inferior, which results in much lower correlation between positions.

Comparing the curated (SMART) or structural (FSSP) alignments with those that are automatically generated by ClustalW (Version 1.7), we conclude that conservation properties should be a valuable tool for alignment quality assessment and might be used as an objective function for alignment refinement. The independent-count weighting scheme combined with the entropy-based indices seem to be a more sensitive measure to judge the quality of the alignments constructed from rather similar sequences (SMART), while the sum-of-pairs index using structurally derived amino acid scoring matrices appears to be superior for very divergent sequences.

Mapping conservation onto protein spatial structure

Mapping the conservation information onto the 3D-structure helps visualizing the conservation in three-dimensional space and facilitates prediction of structurally and/or functionally important sites (Sander and Schneider, 1991). Such an approach has already been applied in a number of cases (Lichtarge *et al.*, 1996; Landgraf *et al.*, 1999; Makarova and Grishin, 1999; Zhang *et al.*, 2000). The AL2CO program can be used to assist in mapping conservation indices onto a spatial structure. If the structure of a protein from a multiple alignment is available, the user has an option to specify the coordinate file in PDB format. B-factors in that file will be substituted by conservation indices. Bobscript or Molscript (Kraulis, 1991; Esnouf, 1997) can be used to draw a structure diagram colored by the B-factor (line in the Molscript/Bobscript input file: *colour ss from blue via green to red by b-factor from X to Y*, values $X = -1.0$ and $Y = 2.0$ are usually good for normalized indices), which in our case corresponds to the conservation index. We illustrate an application of this procedure to a domain of unknown function (YBAK, Figure 5a) (Zhang *et al.*, 2000). The conservation is maximal around a cavity on this structure showing the potential location of the ligand or substrate binding or catalytic site. Another example (Figure 5b) is for the alignment of the Rab family of small G proteins (Figure 3) (Dumas *et al.*, 1999). The mapping clearly shows that the regions of high conservation are clustered around the catalytic site and hydrophobic core of the molecule.

APPENDIX

The average number of different amino acids per position

For N random sequences of equifrequent amino acids (20 amino acids with frequency $1/20$ each), average number of different amino acids in a position is given by the formula $F = 20(1 - 0.95^N)$. The formula can be proven by induction. If $N = 1$, then $F = 1$ and the formula is true. Assume that the formula is true for $N = n$. Let $f(n, i)$ be the probability of i different amino acids to occur at the position: $F = \sum_{i=1}^{20} f(n, i)i = 20(1 - 0.95^n)$. When the number of sequences increases by 1 ($N = n + 1$), the number of different amino acids at the position either remains to be i with probability $i/20$ (by adding an amino acid of the same type to one of the existing i amino acids), or becomes $i + 1$ with probability $(20 - i)/20$ (by adding an amino acid different from any of the i amino acids). Thus the average number of different amino acids in the position for $n + 1$ random sequences is:

$$
\begin{aligned}
F &= \sum_{i=1}^{20} [i(i/20) + (i+1)(20-i)/20] f(n,i) \\
&= \sum_{i=1}^{20} (1 + 0.95i) f(n,i) \\
&= 1 + 0.95 * 20(1 - 0.95^n) = 20(1 - 0.95^{n+1})
\end{aligned}
$$

which shows that the formula holds for $n + 1$.

ACKNOWLEDGEMENTS

We are grateful to Hong Zhang for critical reading of the manuscript and to two anonymous reviewers for constructive suggestions and helpful comments.

REFERENCES

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

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Atchley,W.R., Terhalle,W. and Dress,A. (1999) Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J. Mol. Evol.*, **48**, 501–516.

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

Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

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

Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.

Dumas,J.J., Zhu,Z., Connolly,J.L. and Lambright,D.G. (1999) Structural basis of activation and GTP hydrolysis in Rab proteins. *Structure Fold Des.*, **7**, 413–423.

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *Ismb*, **3**, 114–120.

Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.

Esnouf,R.M. (1997) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.*, **15**, 132–134.

Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–60.

Gerstein,M., Sonnhammer,E.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.

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

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

Gotoh,O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.

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

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Jaroszewski,L., Rychlewski,L. and Godzi,A. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.*, **9**, 1487–1496.

Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.*, **23**, 403–405.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.

Krogh,A. and Mitchison,G. (1995) Maximum entropy weighting of aligned sequences of proteins or DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 215–221.

Landgraf,R., Fischer,D. and Eisenberg,D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.*, **12**, 943–951.

Levin,S. and Satir,B.H. (1998) POLINA: detection and evaluation of single amino acid substitutions in protein superfamilies. *Bioinformatics*, **14**, 374–375.

Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

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

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.

Lowry,J.A. and Atchley,W.R. (2000) Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. *J. Mol. Evol.*, **50**, 103–115.

Makarova,K.S. and Grishin,N.V. (1999) The Zn-peptidase superfamily: functional convergence after evolutionary divergence. *J. Mol. Biol.*, **292**, 11–17.

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

Mirny,L.A. and Shakhnovich,E.I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, foldingkinetics and function. *J. Mol. Biol.*, **291**, 177–196.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.

Ouzounis,C., Perez-Irratxeta,C., Sander,C. and Valencia,A. (1998) Are binding residues conserved? *Pac. Symp. Biocomput.*, 401–412.

Ponting,C.P., Schultz,J., Milpetz,F. and Bork,P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.

Pritchard,L. and Dufton,M.J. (1999) Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment. *J. Mol. Biol.*, **285**, 1589–1607.

Prlic,A., Domingues,F.S. and Sippl,M.J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences [in process citation]. *Protein Eng.*, **13**, 545–50.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Shenkin,P.S., Erman,B. and Mastrandrea,L.D. (1991) Information-theoretical entropy as a measure of sequence variability. *Proteins*, **11**, 297–313.

Sunyaev,S.R., Eisenhaber,F., Rodchenkov,I.V., Eisenhaber,B., Tumanyan,V.G. and Kuznetsov,E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.

Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, **28**, 161–169.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994a) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994b) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.

Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Villar,H.O. and Kauvar,L.M. (1994) Amino acid preferences at protein binding sites. *FEBS Lett.*, **349**, 125–130.

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

Vogt,G., Etzold,T. and Argos,P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.*, **249**, 816–831.

Zhang,H., Huang,K., Li,Z., Banerjei,L., Fisher,K.E., Grishin,N.V., Eisenstein,E. and Herzberg,O. (2000) Crystal structure of YbaK protein from Haemophilus influenzae (HI1434) at 1.8 A resolution: functional implications. *Proteins*, **40**, 86–97.

Zuckerkandl,E. and Pauling,L. (1965) In Bryson,V. and Vogel,H.J. (eds), *Evolving Genes and Proteins*. Academic Press, New York, pp. 97–166.