Protein Engineering vol.12 no.5 pp.387–394, 1999

PSIC: profile extraction from sequence alignments with position-specific counts of independent observations

Shamil R.Sunyaev[1,2,3], Frank Eisenhaber[1,2,4], Igor V.Rodchenkov[3,5], Birgit Eisenhaber[1,2], Vladimir G.Tumanyan[3] and Eugene N.Kuznetsov[6]

[1]European Molecular Biology Laboratory, Meyerhofstrasse1, Postfach 10.2209, D-69012 Heidelberg, [2]Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13122 Berlin-Buch, Germany, [3]V.A. Engelhardt Institut of Molecular Biology, Russian Academy of Sciences, Vavilov Street 32, 117984 Moscow, [5]Moscow Institute of Physics and Technology, Institutsky per. 9, Dolgoprudny, Moscow Region and [6]Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznaya Street 65, 117806 Moscow, Russia

[4]To whom correspondence should be addressed.
E-mail frank.eisenhaber@embl-heidelberg.de

Sequence weighting techniques are aimed at balancing redundant observed information from subsets of similar sequences in multiple alignments. Traditional approaches apply the same weight to all positions of a given sequence, hence equal efficiency of phylogenetic changes is assumed along the whole sequence. This restrictive assumption is not required for the new method PSIC (position-specific independent counts) described in this paper. The number of independent observations (counts) of an amino acid type at a given alignment position is calculated from the overall similarity of the sequences that share the amino acid type at this position with the help of statistical concepts. This approach allows the fast computation of position-specific sequence weights even for alignments containing hundreds of sequences. The PSIC approach has been applied to profile extraction and to the fold family assignment of protein sequences with known structures. Our method was shown to be very productive in finding distantly related sequences and more powerful than Hidden Markov Models or the profile methods in WiseTools and PSI-BLAST in many cases. The profile extraction routine is available on the WWW (http://www.bork.embl-heidelberg.de/PSIC or http://www.imb.ac.ru/PSIC).

Keywords: fold recognition/ motif recognition/ profile extraction/position-specific independent counts/PSIC/sequence weighting

Introduction

Sets of similar sequences can be characterized by a multiple sequence alignment within common sequence domains (in the case of protein families) or just a small sequence region (motif). These alignments are used for the extraction of profiles or scoring matrices which subsequently find application in searches for other remotely homologous proteins (Bork and Gibson, 1996).

Correlation among observation data cannot be ignored if it is sufficiently extensive. This generally accepted rule is important for data treatment in multiple sequence alignments because alignments frequently contain many similar (and even duplicated) sequences. A typical protein family in sequence databases is a highly non-random sample of sequences where taxonomic units with a long-term research tradition, with medical or economic impact are heavily over-represented irrespective of their evolutionary role. Besides such technical issues, statistical correlation between similar sequences may arise from their common evolutionary origin or as a result of similar functional requirements. Closely related sequences are largely redundant, i.e. they provide less information than more distant family members.

To balance the representation of different classes of sequences in multiple alignments, various concepts of 'sequence weightings' have been developed. Numerical coefficients ('sequence weights') are associated with each sequence to denote the degree of independence of this sequence from the others in the multiple alignment. In its most drastic form, additional similar sequences are discarded from the set of sequences studied, i.e. their weight is assigned to zero and only highly different ('independent') sequences remain for subsequent analysis (Heringa *et al.*, 1992; Hobohm *et al.*, 1992; Neuwald *et al.*, 1995, 1997).

More elegant techniques use the full sequence information. The diverse techniques described in the literature are of two types, evolutionary tree-based or sequence distance-based. Tree-based approaches assume that the sequences in the multiple alignment have a common evolutionary origin and are a result of divergent evolution and that an evolutionary tree (or a set of alternative trees) can be constructed from sequence, taxonomic or additional information (Altschul *et al.*, 1989; Thompson *et al.*, 1994). The distance-based methods (Vingron and Sibbald, 1993; Gerstein *et al.*, 1994; Henikoff and Henikoff, 1994) avoid the problems of tree topology and root placement and even do not require that the sequences are related at all. Sequence weights are calculated from a matrix of pairwise sequence–sequence similarities (Vingron and Sibbald, 1993) or from the amino acid type diversity observed at each alignment position (Henikoff and Henikoff, 1994). Therefore, such methods are applicable also to alignments of short sequences ('motifs') or very distantly related proteins.

All variants of the sequence weighting approach described above have in common that a single sequence weight is applied to all sequence positions, i.e. the efficiency of phylogenetic changes is believed to be identical at every alignment position, an assumption which is obviously not true. Bruno (1996) has proposed a calculation scheme within a tree-based approach that assigns position-specific weights to each sequence position. However, even with simplifying approximations, the calculation costs of this technique are large and restrict its applicability to exemplary alignments, i.e. frequent usage in large-scale applications with genomic data is not practical.

As a main new aspect in this paper, we present a simple and fast computation approach (the PSIC approach) for the assignment of position-specific sequence weights within the sequence distance framework. For profile calculation, these weights do not have to be explicitly determined; instead,

© Oxford University Press

they are implicitly estimated in the form of position-specific probabilities to observe each of the amino acid types (position-specific independent counts, PSIC). In addition to an outline of possible further developments, we describe also a software prototype incorporating the main new ideas. As an application, profiles derived from multiple alignments with this method are shown to be highly productive in detecting members of protein fold families. They proved more powerful than HMM-, PairWise- or PSI-BLAST-derived profiles (Thompson *et al.*, 1994; Eddy *et al.*, 1995; Birney *et al.*, 1996; Altschul *et al.*, 1997) in many cases. As another application, the PSIC method has been successfully used for the characterization of the post-translational glycosylphosphatidylinositol-modification site (Eisenhaber *et al*., 1999).

For clarity, we want to emphasize that the sensitivity of profile techniques depends on several aspects, not just from the treatment of observed residue frequencies. Profile extraction methods face two difficulties: (1) the problem of interdependence between sequences and (2) the problem of a small number of sequences. This paper addresses topic (1). Problem (2) was extensively studied in the framework of Bayesian statistics and found a solution in the concept of pseudocounts (Lawrence *et al.*, 1993; Tatusov *et al.*, 1994; Bruno, 1996; Henikoff and Henikoff, 1996; Sjölander *et al*., 1996; Altschul *et al.*, 1997). Additionally, the application of profiles for database searches depends also on (3) the treatment of gaps and (4) the type of profile-sequence alignment method (Gribskov *et al.*, 1987; Birney *et al.*, 1996). Concerning aspects (2)–(4), we rely on published methodologies.

A preliminary version of our method has already been presented at the 2nd Annual International Conference on Computational Molecular Biology (RECOMB98; see Sunyaev *et al*., 1998). Here, we publish a more detailed and fully formalized description of the methodical approach with new technical improvements and more extensive comparisons with other available profile methods.

Theory and methodical details

From alignments of related sequences to counts of independent observations

We consider an alignment $A = \{S_k\}$ involving N amino acid sequences S_k ($k = 1, 2, \ldots, N$). The length of the alignment (i.e. the number of alignment positions) is denoted by L. Each sequence $S_k = (s_{1k}, s_{2k}, \ldots, s_{nk})$ is described as an L-tupel of amino acids (or gaps) s_{ik} at the alignment positions i ($i = 1, 2, \ldots, n$). We define also a Kronecker symbol $\delta(a,i,k)$, which is equal to 1 if amino acid type a is observed at alignment position i in sequence S_k and equal to zero otherwise:

$$\delta(a,i,k) = \begin{cases} 1, & \text{if } s_{ik} = a, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The profile matrix $W = \{W(a,i)\}$ represents the characteristic features of the sequence set in the alignment region in form of scores for the placement of amino acid types a at alignment positions i. In traditional approaches (Gribskov *et al.*, 1987; Tatusov *et al.*, 1994; Thompson *et al.*, 1994), the elements of the profile matrix are calculated by averaging scores from an amino acid substitution matrix $D(b,a)$:

$$W(a,i) = \sum_{k=1}^{N} w_k \sum_b \delta(b,i,k) D(b,a) \quad (2)$$

where sequence S_k contributes to the profile W with weight w_k. This equation is purely heuristic and does not rely on any statistical model of a protein family evolution. The notion of amino acid substitution matrices implicitly accepts that the mutation probabilities are identical at every position of protein family, an assumption which is somehow opposite to the basic idea of profile methods to derive position-specific scores.

In the PAM model, the process of amino acid substitutions during protein evolution is modelled as a Markov process (Dayhoff *et al.*, 1978). Theoretically, substitution matrices can be normalized to (raising a transition probability matrix to the power of) any evolutionary distance in PAMs. However, for very long distances the probabilities of occurrence of amino acid types converge to the invariant distribution of Markov process. Since substitution probabilities in the PAM model are independent of protein and position, the invariant distribution of the process is just the amino acid composition of the protein databank, i.e. the probability of observing a particular amino acid type at a given alignment position after infinitely long evolution coincides with the probability of finding this amino acid type in the sequence databank. Therefore, the PAM model of protein evolution is inefficient for detecting very remote relationships in the world of proteins.

A more realistic model of sequence family evolution can be constructed with the assumption that substitution probabilities are position-dependent. In this case, the invariant distribution of Markov process at the particular position i is the probability $p(a,i)$ of observing amino acid a at the position i after infinitely long evolution. This probability is determined by the environment of the residue in the 3D structure and by functional constraints. If the probabilities described above are known, the optimum equation (from a statistical viewpoint, see Appendix A) for the profile matrix element is given by the following log likelihood ratio (Kendall and Stuart, 1977; Karlin and Altschul, 1990; Lawrence *et al.*, 1993):

$$W(a,i) = \ln\left[\frac{p(a,i)}{q_a}\right] \quad (3)$$

where q_a is the default probability of observing amino acid a, for example, in a database of proteins.

The estimation of $p(a,i)$ is a very difficult task since, as a rule, the number of sequences in the alignment is small and the sequences themselves cannot be considered independent (the independent evolution has not been sufficiently long for every pair of sequences). Previous studies estimated $p(a,i)$ as raw or weighted frequencies (Tatusov *et al.*, 1994; Altschul *et al.*, 1997) or with a maximum likelihood approach (Bruno, 1996).

We suggest here a new, simple, empirical method for the estimation of the probabilities $p(a,i)$. If all sequences were independent, the best estimator for $p(a,i)$ is the raw frequency, i.e. the number $n(a,i)$ of observations of the amino acid a at position i of the multiple alignment normalized by the total number of sequences N. Since the sequences may be strongly dependent, we attempt to compute a normalized effective number $n(a,i)_{eff}$ of observations (i.e. the number of independent observations that carry the same amount of information as available dependent ones) and to determine $p(a,i)$ as

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff}} \quad (4)$$

The value $n(a,i)_{eff}$ is thought to depend on the overall similarity

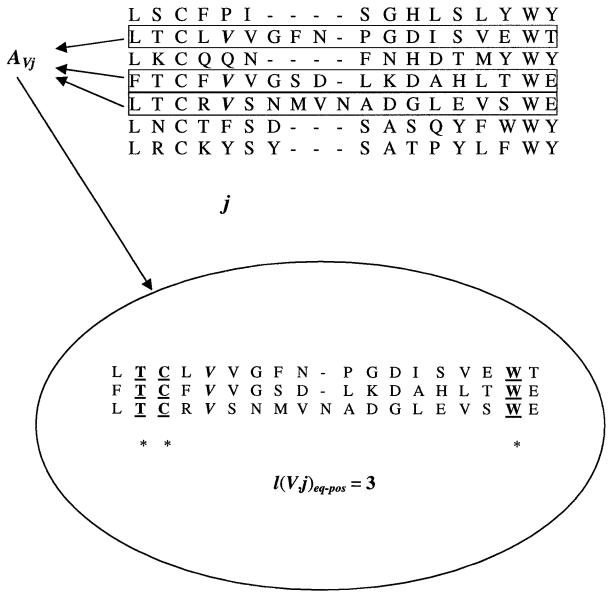

Fig. 1. Relationship between the effective number of amino acid type observations and the overall similarity of sequences. To compute the effective number of valines *V* at position *j* in accordance with Equations 8–10, we compute the number of conserved positions in the subalignment A_{Vj}.

of the sequences having the common amino acid type *a* at the alignment position considered. The idea is that the observation of amino acid type *a* at the given alignment position in a subset of sequences provides less new information compared with a single observation of *a* [implying a smaller $n(a,i)_{eff}$], the more similar the sequences in the subset are. The procedure outlined below complies with all intuitive requirements but does not pretend to be mathematically rigorous.

Algorithm for the determination of position-specific independent counts

Let us assume that a subset A_{aj} of sequences in the alignment *A* has the same amino acid type *a* at the alignment position *j* ($1 \leqslant j \leqslant L$). In this case (Figure 1), the number of observed occurrences of amino acid type *a* at position *j* is

$$n(a,j)_{obs} = \sum_{k=1}^{N} \delta(a,j,k) \tag{5}$$

In general, not all of these observations are independent. The more similar the sequences in the subset A_{aj} are to each other, the closer $n(a,j)_{eff}$ should be to unity (as in the case of identical sequences in A_{aj}). As a suitable similarity measure within the subset A_{aj}, we propose the number $l(a,j)_{eq\text{-}pos}$ of identical alignment positions (not including *j*) within the subset A_{aj}:

$$l(a,j)_{eq\text{-}pos} = \sum_{i=1,\, i\neq j}^{L} \left[\sum_{b} \prod_{k=1}^{N} \delta(a,j,k)\delta(b,i,k) \right] \tag{6}$$

where the two summations are over all alignment positions $i \neq j$ and all amino acid types *b*, respectively. The term

$$\prod_{k=1}^{N} \delta(a,j,k)\delta(b,i,k) \tag{7}$$

equals unity if every sequence with amino acid type *a* at position *j* (i.e. $s_{jk} = a$) has the same amino acid type at position *i* and to zero in all other cases.

The probability $P(A_{aj})$ of finding the same amino acid type at any alignment position $i \neq j$ for all sequences in the subset A_{aj} can be estimated with the value $l(a,j)_{eq\text{-}pos}$ via

$$P(A_{aj}) = \frac{l(a,j)_{eq\text{-}pos}}{m} \tag{8}$$

where $m \leqslant L-1$ is the number of alignment positions of the sequences in subset A_{aj} excluding position *j* and positions with gaps.

On the other hand, if the set A_{aj} were a set of $n(a,j)_{eff}$ randomly aligned independent Bernoulli sequences (given the amino acid composition $\{q_b\}$ in the sequence database, *b* is any of the amino acid types), the probability $P(A_{aj})$ would be equal to

$$P(A_{aj}) = \sum_{b} q_b^{n(a,j)_{eff}} \tag{9}$$

The central idea of the PSIC approach is to equate the right sides in Equations 8 and 9, i.e. the frequency of identical positions in a given alignment and the probability of identical alignment positions for random sequences in order to define the effective number $n(a,j)_{eff}$ of observations (= position-specific independent counts). Thus, the solution of the equation

$$\frac{l(a,j)_{eq\text{-}pos}}{m} = \sum_{b} q_b^{n(a,j)_{eff}} \tag{10}$$

for $n(a,j)_{eff}$ is an estimate for the number of independent observations of amino acid type *a* at position *j* in the alignment *A* and can be applied in Equations 3 and 4. The value $n(a,j)_{eff}$ defined in this way agrees well with intuitive requirements: For very similar (identical) sequences, it is close (equal) to 1, whereas for a subset A_{aj} of divergent sequences, $n(a,j)_{eff}$ is much larger than 1.

It may appear difficult to solve Equation 10 with respect to $n(a,j)_{eff}$, but a simple recursive, binary search procedure can easily help since the solution is enclosed between 1 and $n(a,j)_{obs}$ and the sum on the right-hand side decreases monotonously with $n(a,j)_{eff}$. The solution *x* is assumed to be the midpoint of a test interval (which is $[1, n(a,j)_{obs}]$ at the beginning of the recursion) and the right-hand sum Σ is calculated. If Σ is larger than the left-hand side of Equation 10, the lower half of the interval is used as test interval for the next recursion level. Otherwise, the upper half is taken. This procedure is continued until the length of the test interval is smaller than a user-defined epsilon (we used 10^{-4}); the value *x* is then considered the solution for $n(a,j)_{eff}$. The calculated values $n(a,j)_{eff}$ for all 20 amino acid types *a* enter Equation 4 for the computation of the probabilities $p(a,j)$ which in turn serve for the computation of the profile values in accordance with Equation 3. After having obtained $p(a,j)$, the explicit weight w_{jk} of sequence S_k at position *j* can be calculated (see Appendix B).

The profile matrix $W(a,j)$ having the dimensionality 20 times the number of alignment positions can be used in the traditional way for calculating scores of alignments of protein sequences with the given profile (Gribskov *et al.*, 1987). Any scheme for gap treatment (Birney *et al.*, 1996) or pseudocount heuristics (Lawrence *et al.*, 1993; Tatusov *et al.*, 1994; Bruno, 1996; Henikoff and Henikoff, 1996; Sjölander *et al.*, 1996; Altschul *et al.*, 1997) can be combined with the PSIC methodology.

The PSIC approach has been implemented in a computer

program in the C-programming language. The computation time was found negligibly small on standard UNIX workstations.

PSIC implementation detail I: the case of very divergent sequences

Two types of difficulties arise in practical applications of this algorithm. The first problem is encountered in the case of very divergent sequences in the alignment A. The value $n(a,j)_{eff}$ approaches its maximum $n(a,j)_{eff}^{\max}$ for $l(a,j)_{eq\text{-}pos} = 1$ in accordance with Equation 6. It may also happen that $l(a,j)_{eq\text{-}pos} = 0$ (there may be no identical positions among the sequences in the alignment subset A_{aj} except for position j), i.e. $n(a,j)_{eff}$ cannot be estimated with Equation 6 owing to the lack of data (absence of similar and intermediate sequences). If this happens for only a few alignment positions j (regulated with a user-defined threshold t_0), we approximate $n(a,j)_{eff}$ with the smallest integer larger than $n(a,j)_{eff}^{\max}$ [we used the value 4; another possibility would be the value $n(a,j)_{obs}$]. Otherwise, it is necessary to subdivide the sequence family in alignment A into R subfamilies A_r and to compute $n(a,j)_{eff}$ as a sum over subfamilies:

$$n(a,j)_{eff} = \sum_{r=1}^{R} n_l(a,j)_{eff} \tag{11}$$

The PSIC software allows one to subdivide the set of sequences into subfamilies manually in accordance with additional information, for example, from a structural database such as CATH (Orengo *et al.*, 1997) or SCOP (Murzin *et al.*, 1995). In this case, the parameter t_0 is assumed constant and set equal to 1% of the alignment length.

PSIC implementation detail II: automatic subfamily division

We have also developed a simple clustering procedure based on pairwise sequence identity as a distance measure to group the sequences into subfamilies automatically. In the following, we describe in detail an iterative algorithm for the determination of both the number R of sequence sets and the threshold t_0 in dependence on the input sequence alignment.

If $R = 1$ (at the beginning of the iteration), all sequences form one set. In the case of $R > 1$, we determine as a first step R sequences (for a subdivision into R sequence sets) serving as cluster centers. If $R = 2$, the two most distant sequences are selected. For $R > 2$, the cluster centers are determined in an iterative manner: The Rth cluster center is the sequence having the largest sum of distances to the $R - 1$ previously selected clusters. The remaining sequences are assigned to the nearest cluster center.

The threshold t_0 can be determined with the following considerations. The *a priori* number n_0 of events $l(a,j)_{eq\text{-}pos} = 0$ can be calculated as

$$n_0 = (m+1)\sum_b q_b^{n(a,j)_{obs}}\left(1 - \sum_b q_b^{n(a,j)_{obs}}\right)^m \tag{12}$$

The value n_0 depends weakly on the number of alignment positions m and reaches its maximum $n_0^{\max}$ at $n(a,j)_{\text{obs}} = 3$ sequences for m up to about 100 and at $n(a,j)_{obs} = 4$ sequences for $m<1000$. In addition to the *a priori* expected events $l(a,j)_{eq\text{-}pos} = 0$ for each subfamily, we allow $l(a,j)_{eq\text{-}pos} = 0$ to happen for a user-defined fraction t_u of all alignment positions (for example, $t_u = 0.01$). Thus, we calculate the threshold t_0 as

$$t_0 = \text{int}(n_0^{\max} R + 1) + \text{int}(t_u L + 1) \tag{13}$$

where the function int(x) is the largest integer smaller than x. This equation takes into account the alignment length L (also via m) as well as the number of groups of highly different sequences in the alignment (via the number of subfamilies R). If the threshold t_0 is exceeded, the number of subfamilies R is incremented by one and the profile computation is restarted.

Subfamily creation in accordance with Equation 13 has been compared with the selection of a maximum subset of sequences (each representing a subfamily) with a pairwise sequence identity below a given threshold. Both approaches give comparable results (see Appendix C), hence Equation 13 can be considered a reasonable working assumption.

PSIC implementation detail III: the case of non-observed amino acid types

The second difficulty is connected with the problem of amino acid types x which have not been observed at a given alignment position j at all. In this case, the algorithm outlined above will result in $n(a,j)_{eff} = 0$ and it becomes impossible to take the log likelihood ratio in Equation 2. This is not only a formal difficulty but the amino acid type x may not be observed owing to the possibly small number of sequences in the alignment. This is a well known problem for profile methods based on the log likelihood ratio and for HMM techniques. Usually, it is solved with the so-called pseudocount approach (Lawrence *et al.*, 1993; Tatusov *et al.*, 1994; Bruno, 1996; Henikoff and Henikoff, 1996; Sjölander *et al.*, 1996; Altschul *et al.*, 1997). Generally, the PSIC approach for position-specific weighting of occurrences of observed amino acid types may be combined with any procedure of pseudocount evaluation. We assume a small amount of virtual effective observations n_x which are distributed among the non-observed amino acid types x in accordance with their database frequencies q_x. Equation 3 for observed amino acid types a is then changed to

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff} + n_x} \tag{14}$$

The probabilities of non-observed amino acid types x at alignment position j are estimated via

$$p(x,i) = \frac{n_x}{\sum_b n(b,i)_{eff} + n_x} \cdot \frac{q_x}{\sum_x q_x} \tag{15}$$

We tested values in the range 0.3–10 for n_x, the usual default value was 0.3. This PSIC version was used in the first part of the Results section.

To exclude the computation of pseudocounts as a possible source of performance differences in the comparison with the PSI-BLAST routine (Altschul *et al.*, 1997), we developed also a version with a pseudocount function as described by Henikoff and Henikoff (1996). This PSIC routine was used in test calculations presented in the second part of the Results section.

Results

Performance comparison with PairWise and HMMs

Our profile extraction method was applied to the problem of protein fold recognition. Good test cases are such sequence families which contain a sufficient number of PDB structures as well as many examples with non-trivial sequence variations, especially with small sequence identity. We used multiple alignments for 10 large and divergent families. Eight alignments were

Table I. Comparison of profile and HMM methods in a fold family assignment experiment

	Learning sequences		PairWise	HMM	PSIC	
	With known 3D structure	Without known 3D structure			Not optimized	Optimized
Acid proteases	16	109	51	50	52	55
Sh2 domain	5	125	17	17	17	17
Sh3 domain	8	98	23	23	23	23
Lysozymes	6	63	62	62	270	270
Globin	17	453	82	74	85	86
Cysteine knots	1	50	1	1	6	6
Four helix bundle	7	20	11	11	11	12
Lim domain	2	127	3	3	2	3
TIM barrel[a]	15	387	47	36	87	90
Immunoglobulin[a]	43	457	225	221	228	228

The first two columns characterize the learning set of sequences used for deriving the profile [taken from 3d_ali (Pascarella *et al.*, 1996) except for cysteine knots and the lim domain]. Each set is composed of (1) sequences with known 3D structure in the Brookhaven Protein Data Bank (PDB) and (2) other protein sequences taken from sequence databases (mostly SWISS-PROT) which are highly similar to the proteins with known structure. The remaining columns contain numbers of PDB sequences ranked before the first false-positive recognized with the respective method. This number includes also sequences of the training set if their score was sufficiently high. As a rule, all training structures have been found in addition to other proteins. The only exceptions with non-recognized training sequences have been observed for PairWise and HMM in some cases of non-trivial similarities (see text).

[a]Search results for manual division of the set of sequences in the multiple alignment into subfamilies. Data for the automatic clustering are described in the text.

taken from the 3d_ali (Pascarella *et al.*, 1996) database (TIM barrel fold, immunoglobulin type family, globins, acid proteases, four helix bundles, lysozymes, sh2 domains and sh3 domains). We studied also the lim domains and the cysteine knot family. Alignment positions with gaps for the majority of sequences were discarded.

Family members were searched in the PDB database (as of August 1997) with the commercial Bioaccelerator software Profilesearch using the Smith–Waterman algorithm. The gap initialization–gap extension model was used (gap opening score 4.00, gap extension score 0.05). Scores were normalized with the empirical model of Pearson (1995). The correctness of family assignments was checked manually by comparison with the CATH (Orengo *et al.*, 1997) or SCOP (Murzin *et al.*, 1995) databases.

The same test exercises were carried out with the WiseTools [the program PairWise was reported to be the best profile extraction method (Thompson *et al.*, 1994; Birney *et al.*, 1996)] and with the Hidden Markov Model (HMM) method [programs hmmb and hmmsw (Eddy *et al.*, 1995)]. For clarity, it should be emphasized that each method was tested with the same alignments of protein sequence families. Since the capacity of PairWise is limited to 500 sequences, we used always only the first batch of 500 sequences in each alignment for profile extraction. Since the results for PairWise profiles searches depend greatly on parameters such as gap opening and gap extension penalties and the number of gap positions not taken into account, we optimized these parameters. In the case of HMM, we used the default parameters of the program.

For every method, we counted the number of correctly recognized PDB sequences with a score higher than that of the first false positive. The results are shown in Table I. For the method presented in this paper, we report both non-optimized (default value of 0.3 for n_x) and optimized [best value for n_x in a few ($\leqslant$5) trials] data. Generally, none of the four techniques identifies all the family members. Hence there is still a lot of work to be done to improve the sensitivity of sequence search methods. At the same time, various methods sometimes pick different sequences, i.e. they appear trained for different features of the protein family (examples are given below with the notes for sequence families). It should be noted that, as a rule, all training protein structures have been recognized by the respective methods. Exceptions with non-recognition of training sequences have been observed for PairWise and HMM in some cases of non-trivial similarities with globins and TIM-barrels (see below).

The method described in this paper performs well compared with PairWise or HMMs. In particular, PSIC recognized all training structures in all tests. Even the non-optimized version yields best results in six out of 10 tested families. This is especially remarkable since parameter optimization was carried out in all applications of the PairWise algorithm.

In the case of sh2 and sh3 domains, all methods tested can detect only close homologues in PDB.

There is a considerable difference in predictive power for lysozymes between PairWise and HMMs on the one hand and PSIC on the other. Obviously, PairWise and HMM profiles do not recognize a whole subfamily that is highly represented in the PDB. This subfamily consists of many close homologues, mainly mutants of T4 lysozyme (an object of extensive mutational studies in structural biology). Thus, the large difference in numbers of recognized proteins reflects this database bias.

For globins, our method is the only one which can recognise all globins in the PDB based on the 3d_ali alignment. Two other details are of interest: none of the methods except ours can recognize the sequence of colicin in spite of the fact that it was included in 3d_ali alignment (1col was one of the training structures). Phycocyanin has a high score in our method and appears in the list after only six false positives (phycocyanin is not included in 3d_ali alignment of globin fold family). Phycocyanin was not listed at all in the search protocols obtained with the HMM- or PairWise-generated profiles.

In the case of the cysteine knot family, our method is the only one that can detect the similarity between chorionic gonadotropin and transforming growth factor-beta and also between the two chains of chorionic gonadotropin.

All methods found only a set of close homologous sequences of four-helix bundles included in the learning multiple alignment. Other more remote relatives were not detected, except

for the optimized PSIC which recognized also the sequence of the protein 1buc.

For the lim domain family, our non-optimized method (but not the optimized version) is slightly inferior to PairWise and HMMs. It does not recognize the lim domain fragment with the PDB code 1zfo. β-Ribbon cysteine-rich proteins which are as cysteine-rich as lim domains appear as false positives in the search output.

The TIM barrel family is one of the most divergent as well as most abundant amongst protein families. There is a significant difference between results of the method presented in this paper and other two methods tested. This can probably be explained by the fact that the learning multiple alignment consists of a set of divergent subfamilies. PairWise and HMM programs are unable to find some sequence subfamilies contained in the multiple alignment (PairWise did not find five learning structures, 1fcb, 1ald, 1wsy-A, 1did and 1xim; the HMM approach did not recognize seven training structures, 1fcb, 6taa, 1wsy-A, 1did, 1pii, 1gox and 1xim). It should be emphasized that the low performance of PairWise is not the result of a single false positive in the upper part of the output list; all positions 48 and 50–66 in the search protocol are wrong predictions. In contrast, our method with position-specific weightings is capable of finding some correct hits with no obvious similarity to any sequence of the learning alignment. The large and extremely divergent TIM barrel multiple alignment requires division into sequence subfamilies for application of the PSIC technique. We are pleased to note that the profile's predictive power does not change significantly whether the subdivison is made manually in accordance with CATH (Orengo *et al.*, 1997) or automatically as described in the PSIC implementation detail I section above (87 versus 84 recognized structures for the non-optimized version of the PSIC program).

All three methods recognize many immunoglobulin-like proteins such as HLA (chain M) proteins, CD4 proteins and some other related subfamilies. In the case of several distantly related subfamilies, only one or few of the methods can detect them. For example, PairWise finds the vascular cell adhesion molecules (structures 1vsc-A and 1vca) but not HMM or PSIC (neither non-optimized nor optimized versions). The human class I histocompatibility antigen structures 1hla-A and 1hhg-A have been registered by PSIC and HMM but not by PairWise. The human class II histocompatibility antigen structure 1dlh-A was found by all methods except for PSIC. However, a huge amount of immunoglobulin fold proteins that are more divergent from immunoglobulins was not detected at all. It should be noted that, resembling the case of TIM barrels, the alignment of the immunoglobulins consists of many subgroups of highly similar sequences forcing the PSIC algorithm to divide into subfamilies. Here, the automatic scheme is less powerful than the manual sequence family determination. After 198 correct hits, two false-positives appear in the output list (two structures 1ahh and 1ahi, of the 7α-hydroxysteroid dehydrogenase) followed by another 24 correct structures.

Performance comparison with PSI-BLAST

During this work, the PSI-BLAST program with an automatic and fast profile generation procedure (Altschul *et al.*, 1997) became generally available. We want to emphasize that PSI-BLAST calculates profile matrix elements in a two-step procedure: first, uniform (non-position-specific) sequence weights are calculated from the sequence alignment with the algorithm of Henikoff and Henikoff (1994). The observed frequencies of amino acid types at each alignment position are weighted by these sequence-specific (but not position-specific) weights. Second, these frequencies are complemented with position-specific pseudocounts based on the amino acid type variability at this alignment position. Thus, PSIC differs in a major way from PSI-BLAST by its position-specific *and* sequence-specific weighting of observed frequencies of amino acid types.

Table II. Comparison of PSI-BLAST and PSIC methods in a fold family assignment experiment

Profile Alignment	PSI-BLAST PSI-BLAST	PSI-BLAST Smith–Waterman	PSIC Smith–Waterman
2ohx	41	32	40
3sdh	300	318	318
1aoz	24	26	26
2mta	14	20	21
1hur	68	63	63
1etp	17	41	50
1ten	7	12	17
1rec	97	98	98
2hvp	107	111	107
1bbt	18	63	68
2trx	28	27	27
1raa	25	25	25
2cmd	27	27	27
4gcr	14	14	12
1hsq	47	62	62
3chy	32	34	36
1cid	5	5	5
1hpm	18	20	20
1tin	28	28	28
1xer	33	33	33

Representative proteins from the SCOP database were selected from families with many subfamilies and for which PSI-BLAST finds non-trivial homologues in the first iteration. We list data for searches with (1) the PSI-BLAST profile and the PSI-BLAST alignment routine, (2) the PSI-BLAST profile and the Smith–Waterman alignment procedure [as implemented in SearchWise on the Bioaccelarator (Birney *et al.*, 1996)] and (3) the PSIC-generated profile and the Smith–Waterman routine. Since the dispersion of the PSIC profile matrix values is about 30% of that from the PSI-BLAST profile matrix, the gap penalty parameters were changed to 30% of the standard values in the former case.

Results of 20 comparable fold recognition tests of PSI-BLAST and PSIC generated profiles are presented in Table II (status of PDB as of July 1998). We selected representative proteins from the SCOP database from families with many subfamilies and for which PSI-BLAST finds non-trivial homologues in the first iteration. In this step, we extracted both the computed multiple alignment and the PSI-BLAST-generated profile from the program. Since the accuracy of the alignment procedure influences the search results, we list data for database searches under the following conditions:

(1) the PSI-BLAST profile and the PSI-BLAST alignment routine;
(2) the PSI-BLAST profile and the Smith–Waterman alignment procedure (as implemented in 'profilesearch' on the massively parallel Bioaccelarator machine);
(3) the PSIC-generated profile and the Smith–Waterman routine.

Since the dispersion of the PSIC profile matrix values is about 30% of that from the PSI-BLAST profile matrix, the gap penalty parameters have just been changed to 30% of the standard values in the former case.

As in the case of Table II, the number of correctly recognized

sequences before the first false positive is listed. Surprisingly, PSI-BLAST profiles with the PSI-BLAST alignment routine find more correct sequences than the same profiles with the Smith–Waterman technique in the cases of 2ohx, 1hur and 2trx. The implementations of the PSI-BLAST and of the Smith–Waterman alignment procedures do not allow the introduction of a completely comparable set of parameters. In all other cases, the Smith–Waterman routine was at least as or often even clearly more powerful than the PSI-BLAST alignment routine with the same PSI-BLAST-generated profile, as would be expected.

Our results in Table II indicate that, as a tendency, PSIC profiles have a greater recognition potential than PSI-BLAST profiles (larger values for 2ohx, 2mta, 1etp, 1ten, 1bbt and 3chy). This is especially remarkable since the speed of profile generation is comparable for both techniques. Only in the cases of 2hvp and 4gcr were PSIC profiles less successful. A detailed analysis showed that a slight increase in the gap-open parameter changed the PSIC profile's performance to that of the PSI-BLAST level. Similarly to the tests in the previous section, most hits where found both with PSI-BLAST and PSIC. Sometimes PSI-BLAST and PSIC complement each other in their predictive power; for example, the calpain structures 1aj5 and 1alv have been detected with PSI-BLAST only or the sarcoplasmic protein 2sas is recognized by PSIC alone (for searching with the recoverin 1rec profile).

Discussion

The method described in this paper is a statistical and heuristic approach giving weights as a function of both sequence and alignment position to amino acid type occurrences. It amounts mostly to down-weighting columns of identities in a set of sequences when many positions are totally conserved in this set. In many cases, this should not lead to an outcome that is dramatically different from that obtained with more traditional methods, but our approach has the advantage of being independent of any phylogenetic assumption in addition to having small computational costs with an improved predictive power. These advantages are offset by a somewhat higher sensitivity to misalignments. Clearly, the profile values at a given alignment position are influenced by the rest of the alignment via Equation 8 and large errors there will influence the outcome of Equation 10. However, this effect may not be dramatic, as exemplified by the data in the second part of the Results section since the generally poor multiple alignments produced with the PSI-BLAST routine could be successfully used for predictions with the PSIC method.

We found that, in most cases, less than 10% of the hit lists from various profile methods are mutually exclusive but the potential of different techniques to recognize additional hits might sometimes be of practical importance.

It should also be noted that, instead of producing a single profile for several sequence subfamilies by the combination rule (11), it might be more efficient to create one profile for each subfamily and to integrate the search results only. In this case, the PSIC approach may serve as a quantitative measure to find the point where subfamily division can be useful. We observed also that our automatic sequence clustering procedures build profiles with generally reduced predictive power compared with those from manual subfamily divisions since the latter incorporate more biological sense although the effect is dependent on the sequence family studied.

We want to add a few thoughts on possible future developments. In spite of many years of research effort, there is not yet a good statistical model of protein family evolution even with the assumption of independent positions. We tried to incorporate statistical concepts into the problem of sequence weighting and profile extraction. However, this attempt is far from a complete solution of the problem and our method is more empirical than statistical at some points.

The notion of 'independent counts' is fundamental for our approach. Sequences are independent if one sequence does not carry information about the others in the alignment and vice versa. In such a case, the probability of observing amino acid a at any alignment position is just q_a, i.e. the conditional probability $p(a,i)$ is equal to the *a priori* probability q_a. The probability of observing amino acid a at a given alignment position totally N times (in N sequences) is just q_a^N. For instance, this is true for randomly chosen and randomly aligned sequences. Thus, $n(a,i)_{eff}$ calculated with Equation 10 is the real number of observations in the case of independent sequences.

The approach as outlined in the Theory section can be easily generalized. Independent sequences might be even more precisely defined as belonging to the same family but having undergone very long independent (divergent or convergent) evolution. In this case, we can define position-dependent amino acid probabilities $p(a,i)$ characterizing the sequence family. Then the probability of observing the amino acid type a at a given position in N independent sequences is $p(a,i)^N$. Equation 9 might be rewritten as

$$P(A_{aj}) = \frac{1}{L}\sum_{i=1}^{L}\sum_{b} P(a,i)^{n(a,j)_{eff}} \qquad (16)$$

where i runs over all alignment positions. This equation can be solved by an iterative procedure with starting values from Equation 10. Although derived with more general assumptions, the new profiles computed with Equation 16 have about the same predictive power in the fold family collection experiment as the simpler version Equation 10 as we observed in test calculations (data not shown). A more detailed analysis shows that the value $n(a,j)_{eff}$ does not change much for most alignment positions since it is linearly related to $\{\log[p(a,j)]\}^{-1}$ in the iteration but not to $p(a,j)$ itself.

It should be noted that the general idea of the PSIC algorithm is independent of the specific similarity measure as introduced with Equations 6–9 or 16 and that other similarity measures might be introduced.

To conclude, the PSIC approach with position-specific sequence weights is an important step forward towards a statistically sound method of sequence weighting and profile extraction from multiple alignments.

Availability of the program

The profile extraction routine can be accessed via the Internet at two identical mirror sites: http://www.bork.embl-heidelberg.de/PSIC and http://www.imb.ac.ru/PSIC (the latter URL may be difficult to access from outside Russia). Interested readers may also contact the authors by E-mail: sunyaev@embl-heidelberg.de or eisenhab@embl-heidelberg.de.

Acknowledgements

The authors are grateful to P.Bork for continuous support during this work and to H.Hegyi, C.Ponting and J.Schultz for technical advice, for an introduction to the HMM programs and for example multiple alignments.

Support from T.Gibson in using the WiseTools and the bioaccelerator is gratefully acknowledged.

References

Altschul,S.F., Carroll,R.J. and Lipman,D.J. (1989) *J. Mol. Biol.*, **207**, 647–653.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Birney,E., Thompson,J.D. and Gibson,T.J. (1996) *Nucleic Acids Res.*, **24**, 2730–2739.

Bork,P. and Gibson,T.J. (1996) *Methods Enzymol.*, **266**, 162–184.

Bruno,W.J. (1996) *Mol. Biol. Evolut.*, **13**, 1368–1374.

Dayhoff,M.O., Shwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequences and Structures*. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Eddy,E.R., Mitchinson,G. and Durbin,R. (1995) *J. Comp. Biol.*, **2**, 9–23.

Eisenhaber,B., Bork,P. and Eisenhaber,F. (1998) *Protein Engng*, **11**, 1155-1161.

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

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Henikoff,S. and Henikoff,J.G. (1994) *J. Mol. Biol.*, **243**, 574–578.

Henikoff,S. and Henikoff,J.G. (1996) *Comput. Appl. Biosci.*, **12**, 135–143.

Heringa,J., Sommerfeldt,H., Higgins,D. and Argos,P. (1992) *Comput. Appl. Biosci.*, **8**, 599–600.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.

Karlin,S. and Altschul,S.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Kendall,M. and Stuart,A. (1977) *The Advanced Theory of Statistics*. Griffin, London.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) *Science*, **262**, 208–214.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, **4**, 1618–1632.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) *Nucleic Acids Res.*, **25**, 1665–1677.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.

Pascarella,S., Milpetz,F. and Argos,P. (1996) *Protein Engng*, **9**, 249–251.

Pearson,W.R. (1995) *Protein Sci.*, **4**, 1145–1160.

Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I.S. and Haussler,D. (1996) *Comput. Appl. Biosci.*, **12**, 327–345.

Sunyaev, S.R., Rodchenkov, I.V., Eisenhaber, F. and Kuznetsov,E.N. (1998) in *Proceedings of the 2nd Annual International Conference on Computers in Molecular Biolology* (*RECOMB98)*, pp. 258–264.

Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Comput. Appl. Biosci.*, **10**, 19–29.

Udenfriend,S. and Kodukula,K. (1995) *Annu. Rev. Biochem.*, **64**, 563–591.

Vingron,M. and Sibbald,P.R. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 8777–8781.

Received October 19, 1998; revised February 5, 1999; accepted February 19, 1999

Appendix A

The problem of sequence-profile comparison can be considered in the framework of statistical hypothesis tests. For every query sequence *S*, the hypothesis H_0 that this sequence belongs to the family studied is tested against the alternative H_1 that this sequence is just taken from a databank by chance. With the assumption of independent sequence positions, the likelihood L_0 of the sequence *S* being a family member is given by

$$L_0 \propto \prod_{i=1}^{L} p(s_i|i) \tag{17}$$

The term $p(s_i|i)$ is the conditional probability for the amino acid type s_i at the alignment position i of the protein family considered. The product is taken over all alignment positions. The likelihood L_1 that the sequence *S* is assigned to the family by chance is given by

$$L_1 \propto \prod_{i=1}^{L} q(s_i) \tag{18}$$

where $q(s_i)$ is the general frequency of the amino acid s_i in the database.

In accordance with the Neyman–Person lemma (Kendall and Stuart, 1977), the log likelihood ratio test *T* is the optimum decision criterion between hypotheses H_0 and H_1 with the lowest possible type II error while the type I error is fixed:

$$T = \sum_{i=1}^{L} \log \left[\frac{p(s_i|s)}{q(s_i)} \right] \tag{19}$$

Therefore, the optimum profile element must be chosen as (3).

Appendix B

Finally, we want to derive an explicit equation for the sequence weighting that has implicitly found application in this profile extraction approach. The weight for amino acid *a* observed at position *j* in a subset A_{aj} of sequences S_k ($k = 1, \ldots, N$) is given by $p(a,j)$ in Equation 4. Since this amino acid has been observed $n(a,j)_{obs}$ times and the total weight can be symmetrically distributed among all $n(a,j)_{obs}$ sequences, the weight w_{jk} of sequence S_k at position *j* can be computed as

$$w_{jk} = \frac{p(a,j)}{n(a,j)_{obs}} = \frac{\sum_a \delta(a,j,k) \dfrac{n(a,j)_{eff}}{n(a,j)_{obs}}}{\sum_b n(b,j)_{eff}} \tag{20}$$

It should be emphasized that, for the purpose of profile computation, the weights do not actually have to be calculated. Equation 20 is given here for convenience and comparison with other sequence weighting methods.

Appendix C

It is interesting to compare subfamily creation in accordance with Equation 14 with the selection of a maximum subset of sequences (each representing a subfamily) with a pairwise sequence identity below a given threshold. As an independent test example, we used the set of glycosylphosphatidylinositol (GPI)-attachment site sequence segments, an extremely divergent set of protein sequence pieces with many subsets of high sequence identity. GPI anchoring to membranes is a common post-translational modification for extracellular eukaryotic proteins with diverse structure and functionality (Udenfriend and Kodukula, 1995). From SWISS-PROT (rel. 35), we extracted sequences of 38 protozoan and 99 metazoan proteins with known GPI-attachment and propeptide cleavage site (ω-site) and with a C-terminal propeptide length of 17–31 amino acids. The sequence segments consisting of positions between ω – 15 and ω + 25 were aligned at the ω position without any gaps. Using the threshold t_0 as in Equation 13, we obtained five protozoan and 44 metazoan families. If the maximum subset of non-homologous sequence segments with less than 30% pairwise sequence identity is searched (Heringa *et al.*, 1992; Hobohm *et al.*, 1992), 14 protozoan and 44 metazoan sequences remain from the original set. Hence the number of subfamilies has the same order of magnitude in both approaches.