Protein Engineering vol.12 no.2 pp.85–94, 1999

Twilight zone of protein sequence alignments

Burkhard Rost[1,2,3]

[1]EMBL, 69 012 Heidelberg, [2]LION Bioscience AG, Im Neuenheimer Feld 517, 69 120 Heidelberg, Germany and [3]Columbia University, Department of Biochemistry and Molecular Biophysics, 650 West 168 Street, New York, NY 10032, USA

Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20–35% sequence identity. Here, more than a million sequence alignments were analysed between protein pairs of known structures to re-define a line distinguishing between true and false positives for low levels of similarity. Four results stood out. (i) The transition from the safe zone of sequence alignment into the twilight zone is described by an explosion of false negatives. More than 95% of all pairs detected in the twilight zone had different structures. More precisely, above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were. (ii) Whether or not sequence homology implied structural identity depended crucially on the alignment length. For example, if 10 residues were similar in an alignment of length 16 (>60%), structural similarity could not be inferred. (iii) The 'more similar than identical' rule (discarding all pairs for which percentage similarity was lower than percentage identity) reduced false positives significantly. (iv) Using intermediate sequences for finding links between more distant families was almost as successful: pairs were predicted to be homologous when the respective sequence families had proteins in common. All findings are applicable to automatic database searches.

Keywords: alignment quality analysis/evolutionary conservation/ genome analysis/protein sequence alignment/sequence space hopping

Introduction

Protein sequence alignments in twilight zone

Protein sequences fold into unique three-dimensional (3D) structures. However, proteins with similar sequences adopt similar structures (Zuckerkandl and Pauling, 1965; Doolittle, 1981; Doolittle, 1986; Chothia and Lesk, 1986). Indeed, most protein pairs with more than 30 out of 100 identical residues were found to be structurally similar (Sander and Schneider, 1991). This high robustness of structures with respect to residue exchanges explains partly the robustness of organisms with respect to gene-replication errors, and it allows for the variety in evolution (Zuckerkandl and Pauling, 1965; Zuckerkandl, 1976; Doolittle, 1979, 1986). Structure alignments have uncovered homologous protein pairs with less than 10% pairwise sequence identity (Valencia *et al.*, 1991; Holmes *et al.*, 1993; Holm and Sander, 1996; Brenner *et al.*, 1996; Hubbard *et al.*, 1997). Indeed, most similar protein structure pairs appear to have less than 12% pairwise sequence identity (Rost, 1997). Furthermore, the average sequence identity between all pairs of similar structures is supposedly 8–10%, and the observed distribution (Gaussian peaking around 8% identity) marks another region, the midnight zone (Rost, 1997). The midnight zone is populated by protein structure pairs that may have become similar by convergent or divergent evolution (Doolittle, 1994; Rost, 1997). Threading algorithms ultimately aim at revealing homologous pairs from the midnight zone (Wodak and Rooman, 1993; Bryant and Altschul, 1995; Sippl, 1995; Rost and Sander, 1996; Sippl and Floeckner, 1996; Fischer *et al.*, 1996; Rost and O'Donoghue, 1997). Conventional sequence alignment methods become problematic at much higher values of sequence identity. Methods often fail to correctly align protein pairs with 20–30% pairwise sequence identity. Hence, Doolittle (1986) coined the term twilight zone for sequence alignments in this region. Do the difficulties of alignment methods in this zone reflect merely technical difficulties (statistical significance of detection), or is the twilight zone defined by a particular feature of evolution?

Length-dependent cut-off for significant sequence identity

Pairwise sequence identity (percentage of residues identical between two proteins) is not sufficient to define the twilight zone. Instead, analysing the relatively small number of structure pairs available in 1990, Sander and Schneider (1991) defined a length-dependent threshold for significant sequence identity. The threshold curve defined (dubbed HSSP-curve) was roughly proportional to the inverse square-root of the length for alignments between 7 and 80 residues, and was clipped to saturate at 25% sequence identity over more than 80 residues. In 1990, no pair with more than 30 identical residues of 100 aligned had different structures (Sander and Schneider, 1991). Was this still true for the five times larger PDB (Bernstein *et al.*, 1977) of 1997?

Hopping in sequence space

If we could plot the space of protein sequences, would we observe the protein families as islands? Unfortunately, we cannot tell. Nevertheless, useful information has been extracted from sequence (Casari *et al.*, 1995) and structure (Maiorov and Crippen, 1995) space. In everyday database searches, protein families are widened by exploiting the transitivity of homology (Pearson, 1996): (i) a query sequence U is aligned to a database, say SWISS-PROT (Bairoch and Apweiler, 1997); (ii) all sequences aligned at levels of significant similarity are used as new seeds U_i, and for each U_i SWISS-PROT is searched again; (iii) this procedure is repeated until no new sequences are found. Sequence space hopping may be used in combination with knowledge from structures to widen families (Holm and Sander, 1997), or to increase the information contained in multiple sequence alignments input to prediction methods (Rost, 1996, 1997). Recently, the transitivity of protein families has been exploited successfully to automatically increase the yield in database searches [Ruben Abagyan presented the 'multi-link recognition' method 1996 at the CASP2 meeting (Abagyan and Batalov, 1997); Park *et al.* (1997) presented the 'intermediate sequence search' method and Neuwald *et al.* (1997) implemented the same concept (Neuwald, *et al.*, 1997)]. Here, I confirm the original findings based on a different data set, and analysed in detail how the gain depended on the number of intermediate sequence, and their similarity.

Here, I present results of aligning a set of 792 sequence-unique (no pair in set has more than 25% sequence identity) proteins of known structure against PDB. The following questions were investigated. Is the number of protein pairs of non-similar structures proportional to the distance from the HSSP-curve (eqn 1), or do false positives increase more rapidly in the twilight zone? Is the curve defined by Sander and Schneider (1991) still valid? Would using sequence similarity rather than identity improve accuracy (as speculated by Schneider and Sander)? Finally, can the accuracy be improved for pair alignments by expert rules? The results verify, partially, earlier work based on a 1000-fold larger data set (Sander and Schneider, 1991). The novel aspects were (i) a definition of a threshold for similarity (eqn 2), and a refinement of the threshold for identity; (ii) an introduction of various expert rules. Aspects largely complementing other analyses were (Abagyan and Batalov, 1997; Park *et al.*, 1997; Brenner *et al.*, 1998): (i) a large-scale evaluation of exploiting intermediate sequences (sequence-space-hopping); (ii) a detailed analysis of true and false positives providing estimates for accuracy and coverage of database searches; and (iii) a comparison with BLAST, one of the most popular methods for rapid databases searches (Altschul *et al.*, 1990; Altschul and Gish, 1996).

Methods

Data set: 792 sequence-unique protein structures

Protein databases are biased towards particular protein families. To reduce this bias, analyses are usually restricted to representative data sets (Hobohm *et al.*, 1992). Here, I chose the maximal set of sequence-unique proteins of known structure available in early 1997 (Holm and Sander, 1996). 'Sequence-unique' was defined as 'no pair in the set falls above the HSSP-curve (eqn 1; Sander and Schneider, 1991). As a rule-of-thumb, no pair had more than 25% pairwise sequence identity. Each of these proteins was aligned against the subset of PDB contained in the early 1997 release of the FSSP database of protein structure alignments (Holm and Sander, 1996). This subset amounted in total to about 5646 protein chains. Obviously the second step (792 versus 5646) re-introduced bias into the results. However, aligning the 792 sequence-unique pairs against themselves would not have yielded any result for most of the twilight zone analysed here. Thus, 792 versus 5646 was the best compromise in reducing bias *and* monitoring the biased region. The resulting test set was the largest possible set of proteins for which structural information was available (and thus false and correct hits could be automatically distinguished).

Generation of sequence alignments

Protein pairs were aligned by two different program types. (i) Full dynamic programming as implemented in the Smith–Waterman (Smith and Waterman, 1981) based method MaxHom (Schneider, 1994) (McLachlan metric, with minimum = –0.5, maximum = 1.00, and gap open = 3, gap elongation = 0.3); and (ii) quick database searches as implemented by the two versions of the BLAST series: BLASTP (Altschul *et al.*, 1990; Altschul and Gish, 1996), and PSI-BLAST (Altschul *et al.*, 1997). All 792 unique proteins were aligned against all 5646 proteins from the PDB subset. Alignments shorter than 10 residues were not considered, as identical polypeptides of up 10 residues are known to occur in different structure states (Kabsch and Sander, 1984; Cohen *et al.*, 1993). Technical limitations (CPU time) required the restriction of the dynamic-programming analysis to the best 2000 hits for each of the 792 unique proteins. (Note: this restriction applied only to the final displayed alignment. Of course, all possible combinations were explored initially by the alignment algorithm.) The resulting final data set comprised about 1.7 million pairwise alignments. For the comparison between the dynamic programming and the BLAST methods, the data set had to be reduced to all pairs that were aligned by all methods compared (the problem was that neither BLASTP, nor PSI-BLAST could be forced to report absolutely wrong, i.e. ALL pairwise alignments).

Definition of sequence identity and sequence similarity

(i) Pairwise sequence identity was defined by the percentage of residues identical between two aligned sequences (e.g. aspartic matching aspartic counts 1: D – D = 1; aspartic on glutamic was a non-match: D – E = 0). (ii) Pairwise sequence similarity was defined by the percentage of residues similar between two sequences (e.g. D – D $\leqslant$ 1; and aspartic on glutamic was now considered a match: D – E $>$ 0). Similarity scores depend on the particular metric used to capture physico-chemical properties of amino acids (note: most amino acids are not considered 100% similar to themselves by typical metrices, as such metrices are based on log-odds, e.g. for the McLachlan metric only F, W, Y and C yield 100% self-similarity). Consequently, levels of similarity are not directly comparable between different metrices. For comparability, I used the McLachlan metric (Gribskov *et al.*, 1987) also used in the HSSP database (Schneider *et al.*, 1997). In principle, there are two ways to convert similarity into percentage values: (i) by normalizing the similarity score by the maximal possible score observed in a given metric (percentage residue similarity); and (ii) by setting an arbitrary threshold of the similarity score to distinguish similar–not similar and counting the percentage of residues that are similar according to this threshold (percentage of similar residues). Again, I followed the practice of the HSSP database compiling the percentage residue similarity (normalized by maximal possible scores). When compiling percentages, the number of identical residues was normalized by the number of residues aligned, gaps were ignored.

Standard of truth for structural similarity

Similarity between two protein structures is not uniquely defined. Different structure alignment methods yield different scores (Alexandrov *et al.*, 1992; Holm *et al.*, 1993; Luo *et al.*, 1993; Orengo, 1994; Crippen and Maiorov, 1995; Gerstein and Levitt, 1996; Holm and Sander, 1996; Orengo and Taylor, 1996; Zu-Kang and Sippl, 1996). Such differences can be substantial, as illustrated by differences between the expert-based database of structural alignments SCOP (Murzin *et al.*, 1995; Brenner *et al.*, 1996; Hubbard *et al.*, 1997), and the automatically generated databases CATH (Orengo *et al.*, 1993, 1997) and FSSP (Holm and Sander, 1996). In general, FSSP tends to find more pairs of similar structure than do CATH and SCOP. However, this is only a trend. For many examples, SCOP finds structural similarity and FSSP does not. Here, I

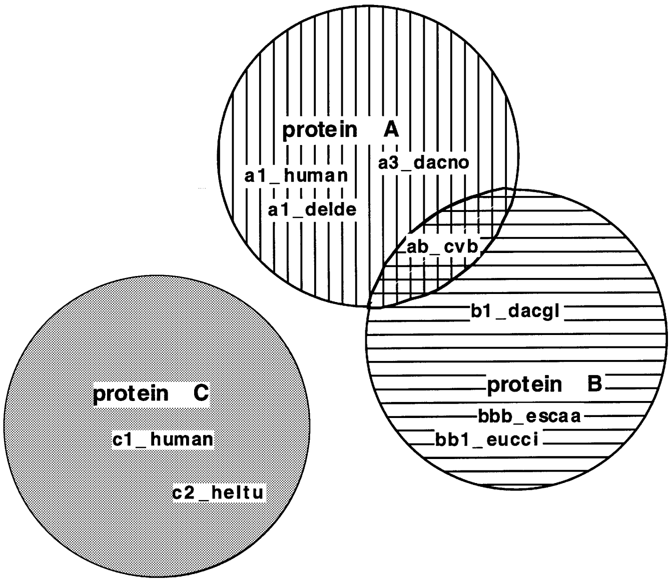

Fig. 1. Sketch of sequence-space-hopping. The triangle defines three search proteins (A, B and C) having mutually less than 25% sequence identity. The circles define the three families (all sequences inside the circle indicated by arbitrary names *aaa_species* have more than 25% sequence identity to the respective search proteins A, B and C). Sequence-space-hopping implies joining the circles representing the protein families (as shown for proteins A and B in the striped circles) if they contain identical proteins that are aligned in the same region (*ab_cvb* in the example given).

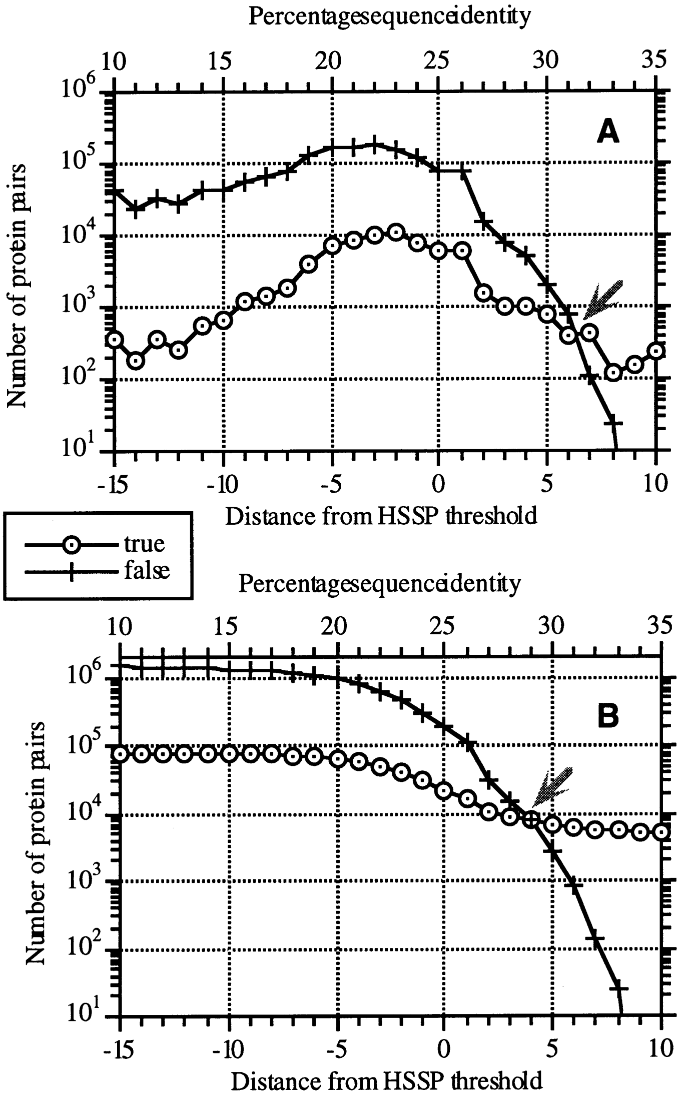

Fig. 2. Explosion of structurally dissimilar pairs in the twilight zone. Numbers of true (pairs with similar structure) and of false positives (pairs with no similar structure) plotted versus the distance to the HSSP-curve (Sander and Schneider, 1991), i.e. the horizontal axes give the distance from the threshold defined in eqn 1 (numbers refer to the parameter *n* in eqn 1). The levels of pairwise sequence identity corresponding to the distance were shown on top. (**A**) Number of pairs observed at any distance (logarithmic scale). (**B**) Cumulative number of pairs observed (logarithmic scale). For example, at a threshold corresponding to about 32% sequence identity for long alignments, the numbers of true and false positives were equal (arrow in A); at about 29% even the cumulative numbers of true and false positives were equal (arrow in B). Note: numbers of true negatives and false negatives result from the cumulative sums left of the threshold; percentages of true and false positives given in Figure 5.

chose the FSSP database ‘a standard of truth’: any pair for which FSSP listed a significant score [zDALI $>$ 4 (Holm and Sander, 1996)] of structural similarity was considered to be structurally similar. In order to distinguish between true and false positives this decision implied that all pairs not listed at the given cut-off of the FSSP database were structurally not similar. However, this brought up the problem of different structure alignment methods. For example SCOP may consider a pair structurally similar, and FSSP may not. Thus, additionally all pairs were excluded from the analysis that were listed in FSSP but with lower z-scores. Even that still left pairs of proteins with clear levels of sequence identity (more than 40%) which were not found listed in FSSP. Thus, I had to refine this procedure by semi-automatically checking the structural similarity for about 2000 protein pairs all of which had levels of above 30% pairwise sequence identity [note this number was negligibly small, as only 1% of all pairs were found above this value (Fig. 2B)!]. The particular way in which the standard-of-truth was constructed implied that estimates for true positives might be slightly optimistic, estimates for false negatives slightly pessimistic.

Concept of true and false hits

When Chothia and Lesk (1986) first analysed the relation between sequence and structure similarity, they monitored the details of structural differences, and found that the differences are inversely proportional to the level of sequence identity. The binary notion of ‘similar structure’ (true or false) used in this analysis reflected a different focus: the goal was to estimate the accuracy in correctly detecting rather than in correctly aligning homologues. Did this imply that correct detection and correct alignment were not correlated (as often the case for threading: Bryant and Altschul, 1995; Lemer *et al.*, 1995; Sippl, 1995; Fischer *et al.*, 1996)? Not necessarily, but the fact is that two homologues can be detected although part—or even the entire—alignment is wrong. (However, this extremely irritating point was not pursued further in this analysis.) The following cases were distinguished: (i) true positives, alignments between proteins of similar structure that fall above a given threshold (defined by the sequence alignment method); (ii) false positives, alignments between proteins of dissimilar structure that fall above a given threshold of the sequence alignment; (iii) true negatives, alignments between proteins of dissimilar structure that fall below a given threshold; and (iv) false negatives, alignments between proteins of similar structure that fall below a given threshold. Note that ‘negatives’ and ‘positives’ represent two sides of the same coin: at

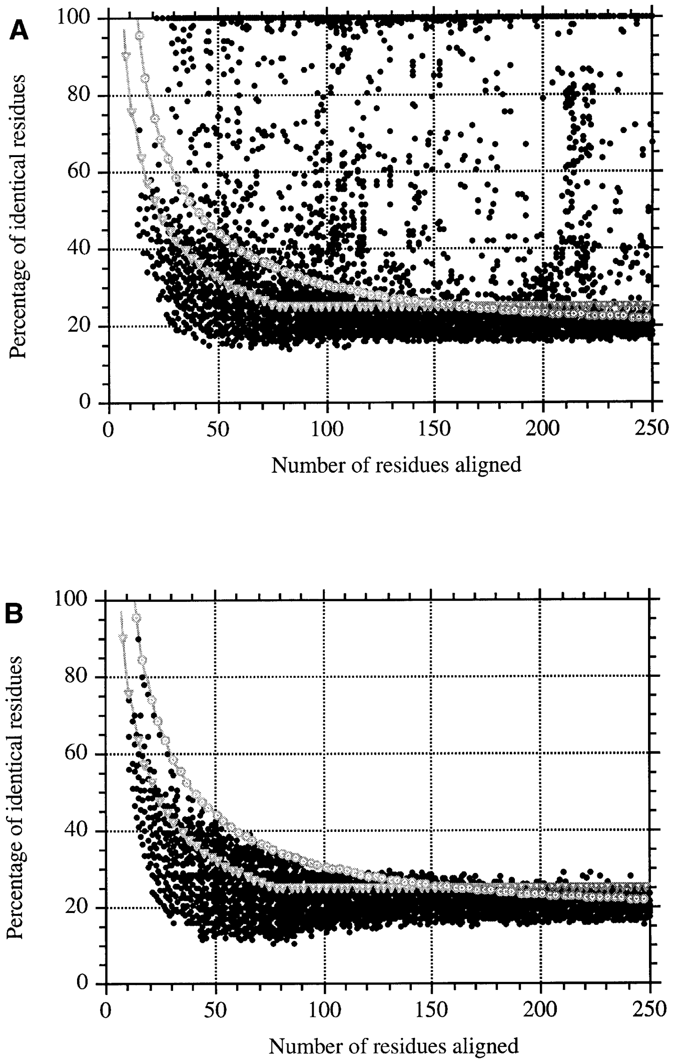

Fig. 3. Pairwise sequence identity versus alignment length. The original HSSP-curve (Sander and Schneider, 1991) (dotted circles, eqn 1) appeared to fit the true positives (homologues, **A**) better than the false positives (**B**). In contrast, the new curve proposed here (filled diamonds, eqn 2) was more conservative in excluding false positives. Note that due to the huge number of pairs the plots for true (A) and false (B) positives appeared almost equally densely populated (Figure 2 revealed the problem of such a scatter plot).

any threshold extracted from the sequence alignment *n*, the following equations hold (for cumulative numbers):

false negatives + true positives = all pairs of similar structure

true negatives + false positives = all pairs of dissimilar structure.

Distance to HSSP threshold

The HSSP-curve was originally defined by (Sander and Schneider, 1991):

$$p^I(n) = n + \begin{cases} 290.15 \cdot L^{-0.562}, \text{ for } L < 80 \\ 25 \quad , \text{doe } L \geqslant 80 \end{cases} \qquad (1)$$

where *L* gave the number of residues aligned between two proteins; p^I the cut-off percentage of identical residues over the *L* aligned residues; and *n* described the distance in percentage points from the curve ($n = 0$ corresponds to the original HSSP-curve; $n = 5$ to the official HSSP database releases; curve plotted in Figure 3). Once Schneider and Sander (1991) had discovered the basic functional dependence between sequence identity and alignment length, they merely had to fix two free parameters: the factor and the exponent. Both were chosen to fit the data observed in 1991, in particular to reach values of 25% around alignment length of 80, and values of 100% around alignment length of 10. The principle functional dependence described by eqn 1 also follows from statistics, as was recently shown in an elegant work (Alexandrov and Soloveyev, 1998). Let p_i ($i = 1,..., 20$) be the probability that amino acid *i* occurs in a protein, and m_{ij} the score for randomly aligning two amino acids *i* and *j*. The score *S* of an entire alignment can then be approximated by:

$$S = \langle m \rangle \cdot L$$

where <m> is the expectation value of m_{ij}, and *L* the alignment length. If the values of m_{ij} are independent, Gaussian distributed variables, it follows (after some elementary operations) that the relation between the standard deviation of the values of m_{ij} (σ_m), and the resulting score distribution (σ_S) is:

$$\sigma_m = L^{-0.5} \cdot \sigma_s$$

In their original article Alexandrov and Soloveyev work out an appropriate re-scaling of the dynamic programming alignment. However, this scheme cannot be applied after the alignment has been completed (as the threshold functions used in this work), rather it has to be implemented into the alignment method.

New curve for length-dependent significance of pairwise sequence identity

I attempted to solve the problems of the original HSSP-curve (eqn 1; Results) by defining the following curve for the separation of true and false positives (Figure 3, grey line with dotted circles):

$$p^I(n = n + 480 \cdot L^{-0.32 \cdot (1 + e^{-L/1000})} \qquad (2)$$

where *L* gave the number of residues aligned between two proteins; p^I the cut-off percentage of identical residues over the *L* aligned residues; and *n* described the distance in percentage points from the curve ($n = 0$ plotted in Figure 3). The constraints in visually selecting the final function were (i) to maintain the functional form defined by eqn 1 (and suggested by the statistics of Alexandrov and Soloveyev, 1998); (ii) to hit the 100% mark at alignments that are too short to reveal anything about structural similarity (= 11 residues); (iii) to saturate at levels around 20% sequence identity (reached for length = 300); and (iv) to roughly reflect the observed gradient. Saturation for long alignments was realized by the functional form of the exponent (note: the term + $e^{-L/a}$ resulted in an exponential decay). This ‘saturation’ constraint also afflicted the particular value of the factor (0.32 rather than about 0.5 as suggested by the distribution of the data, Figure 4).

New curve for length-dependent significance of pairwise sequence similarity

The original HSSP-curve was derived for sequence identity, not for sequence similarity (Sander and Schneider, 1991). The functional dependence between similarity and length appeared comparable to the one between identity and length (Results). This prompted a similar definition for the separation between true and false positives based on similarity:

$$p^S(n = n + 420 \cdot L^{-0.335 \cdot (1 + e^{-L/2000})} \qquad (3)$$

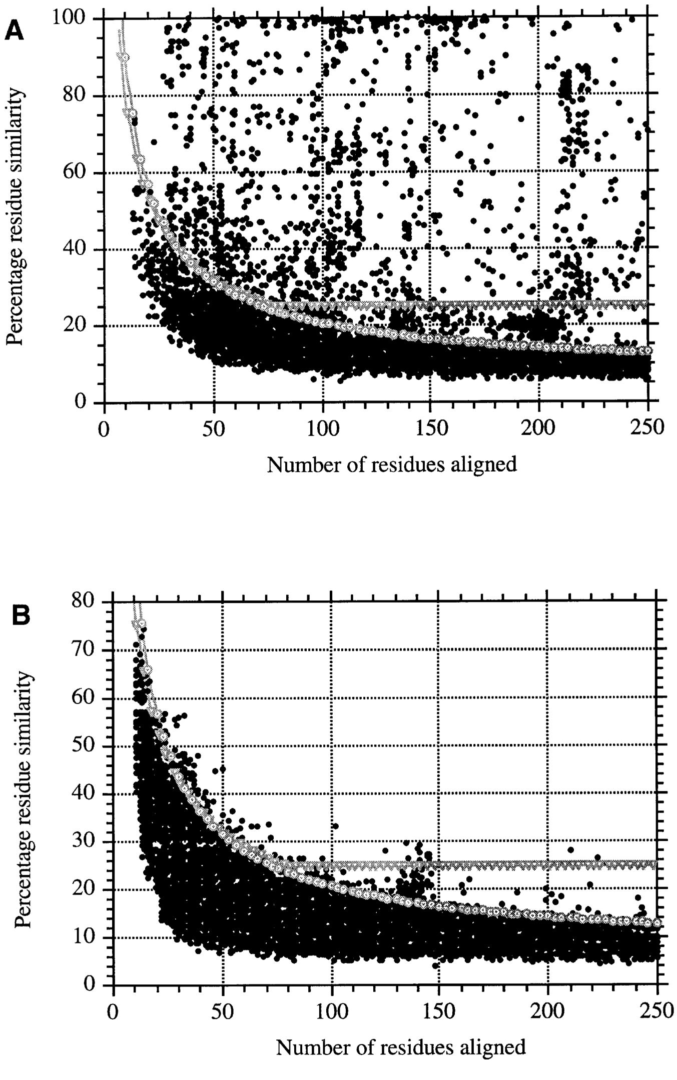

Fig. 4. Pairwise sequence similarity versus alignment length. (**A**) Correctly detected structural homologues; (**B**) false positives. Open circles, original HSSP-curve (Sander and Schneider, 1991) (eqn 1); filled triangles, new curve proposed here (eqn 3).

where L gave the number of residues aligned between two proteins; p^S defined cut-off for the percentage of residue similarity over the L aligned residues; and n described the distance in percentage points from the curve ($n = 0$ plotted in Figure 4).

Sequence-space-hopping

Suppose proteins A_0 and B_0 were less than 25% identical; family A is given by: $\{A_0, A_1,..., A_n\}$ (such that all proteins in the family A are more than 25% identical to A_0); analogously family B is given by: $\{B_0, B_1,..., B_m\}$. Although A_0 and B_0 differed by more than 75%, it may well be true that both were aligned to the same sequences, i.e. that for some i and j: $A_i = B_j$. If this is the case, 'sequence-space-hopping' refers to simply extending both families A and B to become: $\{A_0, A_1,..., \mathrm{A}_n, B_0, B_1,..., \mathrm{B}_m\}$ (Figure 1). Technically, I described this situation by compiling a simple matrix *H(A,B)* that contained the number of overlapping proteins (i.e. those contained both in family A and B) between all proteins in the test set (792 chains) and all proteins in the search set (5646 chains). For example, *H(A,B)* = 5 implied that test protein A and search protein B had five identical proteins in their family alignments. The family alignments were taken from the HSSP database (Schneider *et al.*, 1997) with a cut-off at: HSSP-curve + 10% ($n = 10$ in eqn 1), i.e. for alignments longer than 80 residues, 35% pairwise sequence identity was required. All protein pairs (A,B) in the twilight zone were investigated for which *H(A,B)* was larger than zero. Note, the concept of sequence-space-hopping explored here is being used in everyday sequence analysis. The novel idea introduced by others (Abagyan and Batalov, 1997; Neuwald *et al.*, 1997; Park *et al.*, 1997) was NOT to use sequence-space-hopping, but to use it for reducing false positives in large-scale sequence analysis. Here, I simply applied this concept was applied to the large data set explored, and investigated its usefulness in dependence on various parameters.

More-similar-than-identical rule

A simple rule-of-thumb was explored: accept hits only if the level of sequence similarity was higher than the level of sequence identity. This rule may appear to be non-selective in that similarity would always be larger than identity; however, for the given definition of similarity (using the McLachlan metric), this was not the case.

Results

Number of false positives exploded in twilight zone

In contrast to 1990, when Sander and Schneider (1991) compiled their data, now protein pairs of dissimilar structure were detected above the 30% cut-off (Figure 2A). And these were not exceptions: at a level of 32% (HSSP-curve + 7%, i.e. $n = 7$ in eqn 1), the number of false positives already equalled that of homologues. For the original HSSP-curve the number of false positives was 20-fold higher than the number of true pairs. The transition from 20 to 30% sequence identity was highly non-linear for true, and false positives (logarithmic scales in Figure 2): the number of true pairs rose by a factor of 5, that of false pairs by a factor of 200 (Figure 2B). Thus, below the region of significant pairwise sequence identity (>34%) the population of false positives exploded. However, also the vast majority of homologues had less than 30% sequence identity.

Functional shape of original HSSP-curve adequate

The functional shape of the original HSSP-curve proved to be basically correct (Figure 3, grey line with triangles). However, the larger data set analysed here revealed several problems in detail (Figure 3B). (i) A threshold of 25% was not reasonable for an alignment length below 150–200 residues. (ii) Above an alignment length of about 100 residues, the derivative of the curve separating true and false positives should be lower than at lengths below 80. I attempted to solve these problems by defining a new curve for separating true and false positives (eqn 2; Figure 3, grey line with dotted circles). The particular functional form guaranteed an approximate saturation for long alignments. For alignments shorter than 11 residues eqn 2 yielded values above 100%. However, this was acceptable as 100% identity for fragments of 10–11 residues does *not* imply structural similarity (Cerpa *et al.*, 1996; Minor and Kim, 1996; Muñoz and Serrano, 1996). The new curve saturated around 20% for alignments over more than 250 residues.

Defining a curve for pairwise sequence similarity

Compiling sequence identity neglects the physico-chemical nature of amino acids. Any multiple sequence alignment illustrates that, for example, the feature hydrophobicity is more

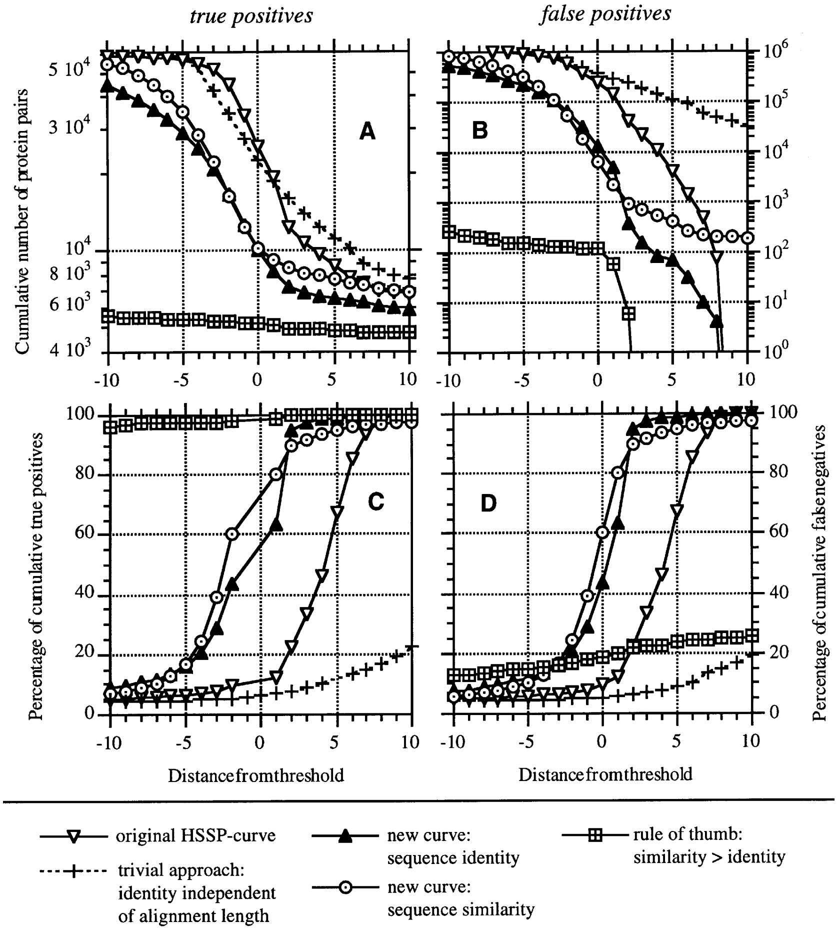

Fig. 5. Accuracy and sensitivity for detecting homologues in the twilight zone. How to choose the cut-off line for automatic database searches? The graphs A–D illustrate the pros and cons of particular choices. Given are the cumulative numbers of correctly detected homologues (true positives, **A**), and of false positives (**B**), as well as, the cumulative percentages of all correctly detected homologues (true positives, **C**), and of all homologues that were missed (false negatives, **D**) in dependence of the cut-off distance from the thresholds defined in eqn 1–3 (parameter *n*). Thresholds: (1) HSSP-curve (eqn 1), (2) new curve for sequence identity (eqn 2), (3) new curve for sequence similarity, (4) subset of proteins for which similarity is larger than identity (grey line in D: false negatives for this subset), (5) simple cut-off according to sequence identity disregarding alignment length (as often used in practice). Note: counts of true positives for the simple sequence identity cut-off (no alignment length) did not even fall into the interval displayed.

conserved than is the residue type. For the million protein pairs investigated here, this was reflected in a shift of the scatter plot towards lower percentages (Figure 4). In particular, for longer alignments false positives fall below 15% pairwise sequence similarity. This prompted the introduction of a threshold specifically for sequence similarity (eqn 3 in Methods; Figure 4, grey line with dotted circles). The curve surpassed 100% for alignments shorter than 12 residues and saturated at about 10% for alignments over more than 500 residues.

Better detection of homologues in twilight zone by new curves

The new curves for length-dependent cut-offs in sequence identity (eqn 2) and similarity (eqn 3) resulted in clearly lower false positive rates (higher accuracy) than the original HSSP-curve (Figure 5B and C). This was paid for by a lower number of true positives detected (lower coverage; Figure 5A). At the $n = 0$ (eqn 1–3), the old curve yielded about twofold more true positives, but more than 20-fold more false positives compared to the new curves for identity and similarity. Furthermore, at any level of true positives detected, the number of false positives was smaller for the new curves (eqn 2–3) than for the original HSSP-curve (eqn 1; Figure 7). When applying a cut-off according to mere sequence identity (ignoring alignment length), accuracy dropped below 10% at levels of 30% sequence identity (Figure 5C). Thus, detection accuracy rose almost 10-fold by the new curves.

Improving detection accuracy by expert rule

Experts often apply rules-of-thumb to visually distinguish true and false positives. However, many of such simple rules appeared not valid for automatic implementation. In particular, the distributions of the number and length of insertions did not, on average, differ between false and true positives (data not shown). Detection accuracy improved marginally by applying the following rules: (i) compile the distance for the similarity score n^S (eqn 3), and the identity score n^I (eqn 2), average over both ($[n^S + n^I]/2$), and accept pairs when this average is above some threshold *n*; (ii) take pairs whenever either identity or similarity surpassed the respective threshold (either n^S Ú $n^I > n$); (iii) take pairs if both values where above a given cut-off (n^S Ù $n^I > n$). In contrast, detection accuracy increased significantly by applying the ‘more-similar-than-identical’ rule: accept hits found in a database search only if percentage similarity is larger than percentage identity. This constraint resulted in >98% detection accuracy at $n = 0$ cut-off levels (eqn 2–3), while 2–4-fold less true positives were

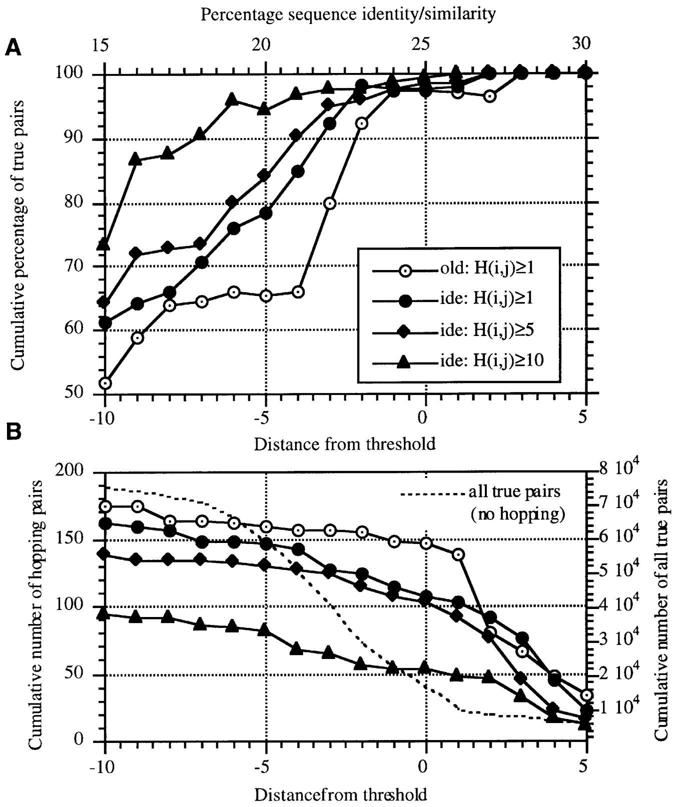

Fig. 6. Improving accuracy by sequence-space-hopping. Distances were compiled according to the old curve (eqn 1, 'old'), and to the new curve for identity (eqn 2, 'ide'). Corresponding levels of sequence identity shown on top. The cumulative percentages of true positives detected at a given cut-off distance were compiled for three different hopping strategies: hits were accepted if, at least, one (H(A,B) = 1), five (H(A,B) = 5) or 10 (H(A,B) = 10) proteins were common between two protein families (Methods). (**A**) Cumulative percentage of true positives (false positives = 100 – true); (**B**) cumulative number of true positives. The comparison of the true positives reached by intermediate sequences and all true positives (grey line in B, note: same as in Figure 2) showed that: (i) less than 1/1000 of the true positives were reached by intermediate sequences; (ii) the number of pairs reached by intermediate sequences did not explode in the twilight zone (scale on the left covers two orders of magnitude, that on the right only one). Numbers for true and false negatives would not make sense for this analysis: as we don't know all proteins, we cannot conclude that two families are unrelated only because we don't find a link between them.

found at this level (Figure 5A and C). Hence, applied as a conservative cut-off in automatic database searches, this rule proved rather powerful.

Improving detection accuracy by sequence-space-hopping

Hopping in sequence space proved successful in discarding false positives. Already the minimal constraint to accept a pair if at least one protein was common between the two sequence families yielded levels of around 80% accuracy even down to cut-off levels corresponding to 20% sequence identity (Figure 6A, compared with <20% accuracy for the normal thresholds Figure 5C). Accuracy increased further when more proteins were required to be common to both families (Figure 6A). However, sequence space hopping was possible for only relatively few protein pairs (Figure 6B). Furthermore, the improvement in accuracy was less clear using sequence-space-hopping than by applying the 'more-similar-than-identical' rule (Figure 5).

Accuracy versus coverage for BLAST and full dynamic programming

The balance between accuracy (percentage of true pairs) and coverage (percentage of all true pairs) enables choosing automatic thresholds according to a particular purpose of a database search. It also permits comparing different methods (the higher the values, the better). (i) As expected, the commonly used simple level of sequence identity (disregarding alignment length) proved, again, an extremely bad choice. (ii) Surprisingly, the fast database searching method BLAST performed relatively well in comparison to the full dynamic programming (Figure 7A). (iii) Both BLASTP version 2 and PSI-BLAST were almost as good as the full dynamic programming with the previously defined HSSP-threshold (Sander and Schneider, 1991). (iv) Best performance was achieved by the new threshold for similarity (eqn 3). (v) However, the raw alignment score performed almost as well. (vi) BLASTP (Altschul *et al.*, 1990) performed rather similarly to the more elaborate and more recent PSI-BLAST (Altschul *et al.*, 1997) (and for 'high' accuracy even slightly better, Figure 7A inset; note: given that standard parameters were chosen, this was not surprising). The corresponding thresholds were given in Figure 5B for the dynamic programming, and in Figure 7B for the PSI-BLAST probabilities.

Many false negatives at reasonable cut-off values

The number of false negatives is often of interest, i.e. the number of proteins that belong to a structure family but were not detected above a given cut-off. For the data sets used here, the cumulative percentage of false negatives was extremely high for all reasonable cut-off levels (Figure 5D). The vast majority of all pairs of proteins with similar structure populate the midnight zone below 10% sequence identity (Rost, 1997). Thus, the extremely high false negative rates proved that methods aligning two proteins merely based on the pairwise levels of sequence homology clearly fail to find the gold mine of database searches (and that older analyses that failed to describe this effect were based on biased data sets).

Thresholds for practical use

For simplicity the functions (eqn 1–3) were explicitly provided in tables (Rost, 1998). At levels of $n = 0$ (eqn 1–3) the cumulative number of true positives were (Figure 5): HSSP-curve (eqn 1), 12%; new identity curve (eqn 2), 56%; new similarity curve (eqn 3), 73%. In order to achieve levels of 99% correct hits m percentage points have to be added to the curves, where m was HSSP-curve, $m = 8$; new identity curve, $m = 5$; new similarity curve, $m = 12$. For comparison, applying the 'more-similar-than-identical' rule yielded levels above 99% down to $m = -1$.

Conclusions

Rapid transition from trivial to needle-in-haystack problem

The twilight zone of sequence pair alignments (20–35% pairwise sequence identity) was characterized by two non-linear transitions. (i) The number of homologues (true positives) rose by a factor of about eight (Figure 2A). I obtained a similar result from analysing the first four entire genomes (Rost, 1997) which indicated that this result was general, rather than database dependent. (ii) The number of false positives rose by a factor of 5000 (Figure 2B). Hence, separating true and false positives switched from a trivial task (above 35%) to the problem of finding needles in a haystack (20–30%).

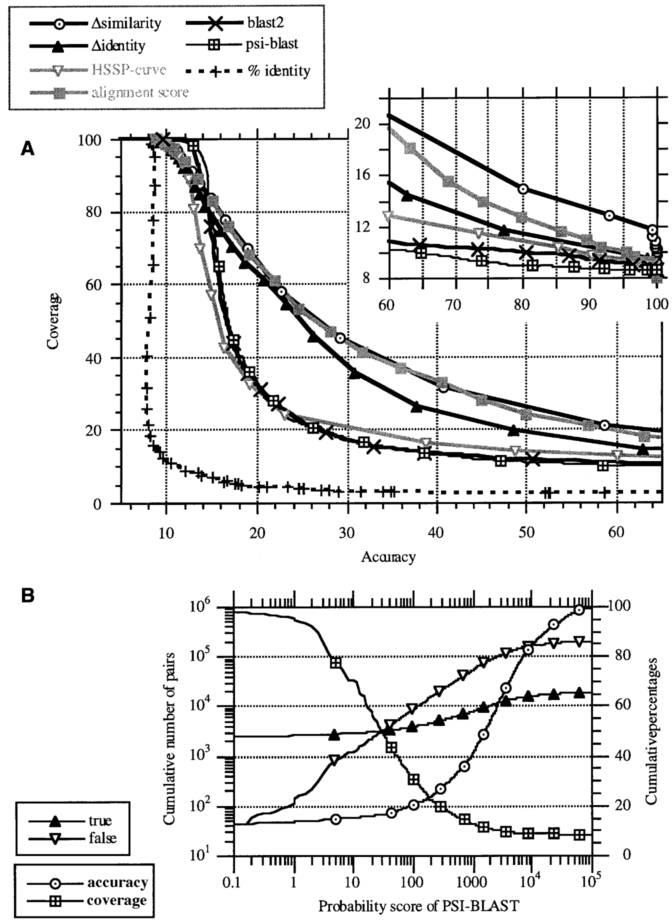

Fig. 7. Accuracy versus coverage for various methods and thresholds. Accuracy was defined as the cumulative percentage of true positives (actual true/all actual), coverage as the percentage of true positives that were detected at a given threshold (actual true/all true). (**A**) Thresholds and methods showed: Δ*identity,* new threshold for length-dependent sequence identity (eqn 2); Δ*similarity,* new threshold for length-dependent sequence similarity (eqn 3); *HSSP-curve,* curve proposed by Sander and Schneider (1991; eqn 1); *%identity,* threshold given by sequence identity alone, i.e., disregarding alignment length; *alignment score,* score used for the dynamic programming optimization MaxHom; *blast2,* BLASTP version 2 (Altschul and Gish, 1996); *psi-blast,* BLASTP version 3 (Altschul *et al.*, 1997), run with standard parameters. The values for the BLAST methods were based on the probability scores reported by these algorithms. The BLAST methods did not report all pairwise alignments, thus the data set had to be reduced to the subset for which aligned pairs were reported by all three methods (MaxHom, BLASTP2, BLASTP3). Note that whereas the curves for the BLAST methods, as well as for identity and similarity are likely to hold up, in general, the curve for the alignment score is valid for the particular implementation of the dynamic programming in MaxHom, and for the particular choice of parameters (Methods). (**B**) Detail of the relation between the BLAST probability (here for psi-blast), and the cumulative number of true/false hits, as well as percentage accuracy and coverage.

The explosion of false positives shed light on the shape of sequence space. From 100–35% sequence identity, any residue exchange resulting in a stable structure maintains structure. From 28–35% sequence identity, most residue exchanges maintain structure. From 20–28% sequence identity, the absolute majority of residue exchanges forming stable structures populate different protein families. Is the explosion caused by features of structure space? If one generates protein sequences at random (or randomly superposes non-related proteins), the counts for most of the region above 10% sequence identity are negligible (Rost, 1997). Thus, although it is obvious that we expect to find more pairs for lower levels of sequence identity based on mere statistics, the particular transition in the twilight zone seems not to be evident. However, this analysis did not provide answers to whether or not the observed explosion may reflect structural (Chung and Subbiah, 1996) and/or functional constraints.

Poor distinction between true and false positives by sequence identity, alone

Even journals such as Cell, or EMBO provide an ample source for the following fallacy: ‘these two fragments of 16 residues adopt similar structures as they have more than 10 similar residues’. Thus, one of the most important messages of this analysis might be the repetition of a point made by others (Sander and Schneider, 1991): high levels of sequence similarity or identity do *not* ascertain structural similarity (Figure 5). Instead, the levels of significant sequence identity and similarity depend on the alignment length (Figures 3 and 4), or the respective raw score of the alignment methods.

Better distinction by new curves for sequence identity and similarity

The length-dependent cut-off for significant sequence identity pioneered by Sander and Schneider (1991) needed refinement in several ways to account for the findings from a 1000-fold larger data set: (i) shift towards higher values for shorter alignments; (ii) saturation for alignments longer than 150 residues; (iii) definition of new curve for levels of sequence similarity. These tasks were solved by introducing threshold curves for significant sequence identity (eqn 2), and for significant sequence similarity (eqn 3). The precise definition of the two thresholds was entirely empirical. However, the essential functional dependency of the curves was kept similar to what would be expected from pure statistical considerations. Although not true for all problems (Nielsen *et al.*, 1996), on average, sequence similarity was marginally more successful than identity in distinguishing true and false positives. The new curves improved accuracy at a given coverage (Figure 5 and 7). Additionally, this analysis supplied detailed levels for expected accuracy and coverage for the curves defined, as well as for standard BLAST searches (Figures 5 and 7). Such estimates may have implications for automatic database searches. They also shed light on the comparison between sequence alignments and threading techniques that both only make use of pair comparisons (rather than using family specific profiles): already at levels of 25% sequence identity, pair alignments detect only 10–30% true positives. This is below the level of what threading techniques achieve in the interval 0–25% sequence identity (Sippl, 1995; Fischer and Eisenberg, 1996; Russell *et al.*, 1996; Rost *et al.*, 1997).

Improved accuracy by ‘more-similar-than-identical’ rule and sequence space hopping

The number of false positives was significantly reduced by two techniques (only the first of which was novel to this work). (i) The ‘more-similar-than-identical’ rule: 95% of all pairs for which percentage similarity was larger than percentage identity had similar structures. Thus, this constraint clearly improved detection accuracy. The cost was low coverage: for only 10% of the structurally similar pairs the percentage similarity was larger than percentage identity. This might be explained by the fact that half of the protein, on average, embedded in loop regions, may tolerate residue exchanges that do not conserve physico-chemical properties (and thus decrease the overall average more than the few to-be-conserved-regions increase it). (ii) The usage of 'multi-links' (Abagyan and Batalov, 1997), 'intermediate sequences' (Park *et al.*, 1997), 'transitivity' (Neuwald *et al.*, 1997), or 'sequence space hopping': most protein pairs that contained a similar subset of identical proteins in their respective sequence families were found to have similar structures even at low levels of sequence homology. Obviously, the validity of transitivity (detection accuracy) between protein families (Figure 1) depended on the distance between the families (Figure 6). Interestingly, the improvement of accuracy hardly depended on the number of proteins required to be common to two families. This suggested that although the vast majority of protein pairs with 25% sequence identity had dissimilar structures, the 'islands' populated by structure families were well separated. Unfortunately, for the data set explored here, the yield of this analysis was found to be very low: on average only one in 1000 pairs was reached via intermediate sequences (Figure 6). Furthermore, sequence-space-hopping resulted in clearly lower coverage/accuracy ratios than did the application of the 'more-similar-than-identical' rule (Figures 5 and 6).

Beginning of the 90's: over-estimation of sequence alignment methods

Until 1996, very few people had taken up the laborious task of objective large-scale analyses of protein sequence comparisons. Partially, because automatic structure comparison methods are fairly recent. The few earlier workers (Sander and Schneider, 1991; Vogt *et al.*, 1995; Gotoh, 1996) based their work on data sets of about 1000 pairs of protein structure alignments. Gotoh (1996) and Vogt *et al.* (1995) used the same set (Pascarella and Argos, 1992) for testing different alignment methods, and a variety of substitution metrices. They focused on monitoring the detailed accuracy in terms of number of residues correctly aligned. Due to the small data set Vogt *et al.* (1995) found about 98% true positives at 30% sequence identity (ignoring alignment length), and 50% true positives at 20% sequence identity. For the 1000-fold larger data set used here the corresponding values were quite different (ignoring alignment length): 11% true positives at 30% sequence identity, and 5% true positives at 20% identity. However, even the more conservative analysis introducing the importance of alignment length for levels of significant sequence identity (Sander and Schneider, 1991) still over-estimated the possible levels of sequence identity between proteins of dissimilar structure.

End of the 90's: database searches do not reach the gold mine, yet

The thresholds for sequence identity and similarity defined here, as well as those established by others (Abagyan and Batalov, 1997; Brenner *et al.*, 1998) complemented the levels for 'significance' provided by BLAST (Altschul and Gish, 1996), FASTA (Pearson, 1996) or other statistical analyses (Bryant and Altschul, 1995) by addressing the question 'how significant is the significance of the respective alignment method?'. Based on quite different data sets the principal messages were similar: (i) most proteins of similar structure were not found by pairwise sequence comparisons at reasonable cut-off thresholds; (ii) raw scores from dynamic programming methods were comparable to the original length-dependent cut-off thresholds for sequence identity (Sander and Schneider, 1991); (iii) dynamic programming was only slightly superior to BLAST searches (Altschul and Gish, 1996; Altschul *et al.*, 1997). However, in detail the numbers differed between the recent analyses. Obviously, the absolute values depended crucially on the particular choice of the data set. Abagyan and Batalov (1997) analysed various substitution metrices on a data set comparable to the one used in this analysis. They concluded that raw alignment scores provide better separations between true and false positives than do length-dependent cut-offs for sequence identity and similarity. The difference between their result, and the one shown here may result from the fact that Abagyan and Batalov (1997) used the optimal choice of all parameters for comparing the raw alignment score to sequence identity and similarity. Brenner and co-workers have analysed the accuracy and coverage for various statistical scores (Brenner *et al.*, 1998). They used a completely different data set than I did. An approximate comparison of the two analyses was possible by the reference point of simple identity (ignoring alignment length). It seems that the performance for the best separation method they find (new FASTA) was comparable to the improved, simple thresholds defined here (eqn 2–3). Here, the BLAST probability was found to be a relatively good way to separate true and false positives (Figure 7A): it was only slightly inferior to the raw dynamic programming alignment score, results for which hold up exclusively for the particular choice of parameters and the particular alignment algorithm used.

Thresholds in practice

The advantages of the length-dependent levels of identity and similarity (eqn 2–3) over other thresholds (Abagyan and Batalov, 1997; Alexandrov and Soloveyev, 1998) was that these thresholds, in principle, are applicable to any alignment, and may relate more explicitly to structure. Identity and similarity can be compiled easily without having to re-do the entire database search. In practice, this does not always hold up: (i) different parameters (e.g. the way in which gaps are treated) may result in different alignments; and (ii) the similarity values compiled hold for the choice of a particular metric (here McLachlan). Additionally, the thresholds introduced here provide independent evidence for the separation, and permitted the application of the successful 'more-similar-than-identical' rule.

Will the analysis hold up for the next 500 structures?

The results given here based on the largest possible data set for which structural alignments provided a well-defined distinction between true and false. One conclusion was that seven years ago (Sander and Schneider, 1991) the database was too small to capture the details. Will this also be true in 2005? Answers have to remain speculative. (i) Although the database used in 1990 was 1000-fold smaller than the one used here, some principle findings were verified. (ii) Assuming that there are only 1000 folds in nature (Chothia, 1992), and that these correspond to about 10 000 families, then even the full catalogue of all protein sequences would yield a data set essentially only 30 times larger than the one used here (note: the data set used corresponded to about 300 different folds aligned against about 1000 families).

Rather more accurate, or more sensitive?

An accurate and sensitive distinction between true and false positives is important for automatic database searches. The new curves introduced here (eqn 2–3) proved slightly more sensitive (higher coverage) and more accurate than the previously proposed curve (Sander and Schneider, 1991). The accuracy increased significantly by applying the 'more-similar-than-identical' rule, and by sequence space hopping. However, accuracy was gained at the expense of coverage. Which is more important? Clearly, the evolutionary information contained in multiple alignments is the single most important contribution to improving protein structure prediction in the 90's (Rost and Sander, 1996; Rost and O'Donoghue, 1997). Is the gain by increased diversity more important than the loss of accuracy when using alignments for structure prediction? The answer depends on the particular prediction goal. For example, for secondary structure prediction diversity is more important than accuracy (cut-off at 25% versus that at 30%), whereas for the prediction of solvent accessibility the opposite is true (unpublished). Furthermore, as databases grow coverage may be less important than accuracy. Irrespective of individual preferences, the sharper the knife cutting between true and false positives, the better. This analysis has sharpened the knife a little, and added new optional tools to it.

Acknowledgements

Particular thanks to Richard Friedman (Columbia, New York) for his extremely helpful comments on the manuscript. Furthermore, thanks to Reinhard Schneider (LION, Heidelberg) for discussions; to Jong Park (MRC, Cambridge) and Tim Hubbard (Sanger, Hinxton, UK) for providing their manuscript prior to publication, and for the two anonymous referees for their helpful comments. Financial support was provided by Matti Sarraste (EMBL, Heidelberg), Gerrit Vriend (EMBL, Heidelberg), Chris Sander (Millenium, Boston) and Friedrich von Bohlen (LION, Heidelberg). Last, but not least, thanks to all those who deposit experimental information about protein structures and sequences in public databases, and to those maintaining these sources of knowledge.

References

Abagyan,R.A. and Batalov,S. (1997) *J. Mol. Biol.*, **273**, 355–368.

Alexandrov,N.N. and Soloveyev,V.V. (1998) In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds) *HICCS' 98: Pacific Symposium on Biocomputing' 98.* Maui, Hawaii, USA, pp. 463–472.

Alexandrov,N.N., Takahashi,K. and Go,N. (1992) *J. Mol. Biol.*, **225**, 5–9.

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

Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A.G. (1996) *Methods Enzymol.*, **266**, 635–643.

Bryant,S.H. and Altschul,S.F. (1995) *Curr. Opin. Struct. Biol.*, **5**, 236–244.

Casari,G., Sander,C. and Valencia,A. (1995) *Nature Struct. Biol.*, **2**, 171–178.

Cerpa,R., Cohen,F.E. and Kuntz,I.D. (1996) *Folding Des.*, **1**, 91–101.

Chothia,C. (1992) *Nature*, **357**, 543–544.

Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.

Chung,S.Y. and Subbiah,S. (1996) *Structure*, **4**, 1123–1127.

Cohen,B.I., Presnell,S.R. and Cohen,F.E. (1993) *Protein Sci.*, **2**, 2134–2145.

Crippen,G.M. and Maiorov,V.N. (1995) *J. Mol. Biol.*, **252**, 144–151.

Doolittle,R.F. (1979) In Neurath,H. and Hill,R.L. (eds) *Protein Evolution*. Academic Press, New York, pp. 1–118.

Doolittle,R.F. (1981) *Science*, **214**, 149–159.

Doolittle,R.F. (1986) *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA, USA.

Doolittle,R.F. (1994) *TIBS*, **19**, 15–18.

Fischer,D. and Eisenberg,D. (1996) *Protein Sci.*, **5**, 947–955.

Fischer,D., Rice,D.W., Bowie,J.U. and Eisenberg,D. (1996) *FASEB J.*, **10**, 126–136.

Gerstein,M. and Levitt,M. (1996) In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R.F. (eds) *Fourth International Conference on Intelligent Systems for Molecular Biology*, AAAI, St Louis, MO, USA, pp. 59–67.

Gotoh,O. (1996) *J. Mol. Biol.*, **264**, 823–838.

Gribskov,M., McLachlan,M. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–5358.

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

Holm,L., Ouzounis,C., Sander,C., Tuparev,G. and Vriend,G. (1993) *Protein Sci.*, **1**, 1691–1698.

Holm,L. and Sander,C. (1996) *Nucleic Acids Res.*, **25**, 231–234.

Holm,L. and Sander,C. (1997) *Proteins*, **28**, 72–82.

Holmes,K.C., Sander,C. and Valencia,A. (1993) *TICB*, **3**, 53–59.

Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.

Kabsch,W. and Sander,C. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.

Lemer,C.M.-R., Rooman,M.J. and Wodak,S.J. (1995) *Proteins*, **23**, 337–355.

Luo,Y., Lai,L., Xu,X. and Tang,Y. (1993) *Protein Engng*, **6**, 373–376.

Maiorov,V.N. and Crippen,G.M. (1995) *Proteins*, **22**, 273–283.

Minor,D.L.J. and Kim,P.S. (1996) *Nature*, **380**, 730–734.

Muñoz,V. and Serrano,L. (1996) *Folding Des.*, **1**, R71–R77.

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

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

Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) *Proteins*, **24**, 165–177.

Orengo,C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 429–440.

Orengo,C.A. and Taylor,W.R. (1996) *Meth. Enzymol.*, **266**, 617–635.

Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.

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

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) *J. Mol. Biol.*, **273**, 349–354.

Pascarella,S. and Argos,P. (1992) *Protein Engng*, **5**, 121–137.

Pearson,W.R. (1996) *Methods Enzymol.*, **266**, 227–258.

Rost,B. (1996) *Methods Enzymol.*, **266**, 525–539.

Rost,B. (1997) *Folding Des.*, **2**, S19–S24.

Rost,B. and O'Donoghue,S.I. (1997) *CABIOS*, **13**, 345–356.

Rost,B. and Sander,C. (1996) *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 113–136.

Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.

Rost,B. (1997) WWW document (http://www.embl-heidelberg.de/predictprotein). EMBL.

Rost,B. (1998) WWW document (http://www.embl-heidelberg.de/~rost/Papers/Dfig/98twilight/app.html). EMBL.

Russell,R.B., Copley,R.R. and Barton,G.J. (1996) *J. Mol. Biol.*, **259**, 349–365.

Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.

Schneider,R. (1994) PhD, University of Heidelberg.

Schneider,R., de Daruvar,A. and Sander,C. (1997) *Nucleic Acids Res.*, **25**, 226–230.

Sippl,M.J. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.

Sippl,M.J. and Floeckner,H. (1996) *Structure*, **4**, 15–19.

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

Valencia,A., Kjeldgaard,M., Pai,E.F. and Sander,C. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 5443–5447.

Vogt,G., Etzold,T. and Argos,P. (1995) *J. Mol. Biol.*, **249**, 816–831.

Wodak,S.J. and Rooman,M.J. (1993) *Curr. Opin. Struct. Biol.*, **3**, 247–259.

Zu-Kang,F. and Sippl,M.J. (1996) *Folding Des.*, **1**, 123–132.

Zuckerkandl,E. (1976) *J. Mol. Evol.*, **7**, 269–311.

Zuckerkandl,E. and Pauling,L. (1965) In Bryson,V. and Vogel,H.J. (eds) *Evolutionary Divergence and Convergence in Proteins*. Academic Press, New York; London, pp. 97–166.

Received March 23, 1998; revised October 23, 1998; accepted October 27, 1998