Detecting Structural Similarity from Protein Sequence Comparison

Geoffrey J. Barton

Laboratory of Molecular Biophysics, South Parks Road, Oxford OX1 3QU, UK.

Introduction

The first step in solving the phase problem by molecular replacement is to identify a suitable structure to use as a search object. If *ab initio* structure prediction techniques were able to provide an accurate three dimensional model of a protein from the amino acid sequence, then this would be a straightforward task. Fortunately for the protein crystallographer, *ab initio* methods are still some way from providing this ultimate solution! However, there are a number of powerful techniques available to detect similarity between a protein sequence and a protein of known three dimensional structure. Some methods improve the sensitivity and selectivity of conventional sequence alignment methods by incorporating secondary or tertiary structural information from the protein of known structure. However, it can be difficult to decide when any given method is indicating genuine structural similarity, or merely a spurious match. Accordingly, in this article, I first briefly review the range of available methods and show their relative success in identifying the globin fold. I then describe an analysis that identifies the limits of detection for a standard pairwise sequence comparison method. For more detailed information on current sequence/structure comparison algorithms see volume 183 (1990) of *Methods in Enzymology*.

Overview of Methods

The available comparison techniques may be loosely divided into five categories of increasing complexity and sensitivity. In practice there is a lot of overlap between the methods in categories 2-5.

1. Pairwise sequence comparison.
2. Pairwise sequence comparison with secondary/tertiary structure information.
3. Multiple alignment comparison with/without secondary structure information.
4. Flexible patterns and templates.
5. Environment specific weighting and optimal threading.

(6. Three dimensional structure prediction!)

Pairwise sequence comparison may be applied to any two protein sequences. A pair score matrix is chosen that assigns a weight to the alignment of all possible pairs of amino acids, (e.g. AA might score 10 whilst AK scores -5), aligning any residue with a gap is assigned a negative value. A dynamic programming algorithm

(e.g. see Needleman & Wunsch 1970) is normally used to find the best alignment of the two sequences including a consideration of insertions and deletions. Although robust, the standard pairwise alignment methods take no account of secondary or tertiary structural constraints, e.g. insertions and deletions generally do not occur in the core of the protein. If we know the structure of one of the proteins, the position of the gaps can be encoded in the comparison to avoid core secondary structures. This approach yields alignments that are more consistent with structural features than straightforward sequence-only methods (Barton & Sternberg 1987a).

Often, the protein with which we are searching can be unambiguously aligned to other members of its family. The resulting **multiple alignment** may then be used as a more sensitive probe of other family members. The sensitivity is improved, since positions important to the protein fold (e.g. buried hydrophobic residues) are given a higher weight, and variable regions are given a lower weight when aligning to a further sequence. As with pairwise comparison, this method may be combined with secondary structural information. An extension of the multiple alignment method is to abstract only the most highly conserved regions into a **flexible pattern** (Barton, 1990, Barton & Sternberg 1990) or **template** (Bashford *et al.*, 1987; Taylor, 1986) that encodes the conserved secondary structures and other important features. As I will show in the next section, flexible patterns can yield great sensitivity and selectivity.

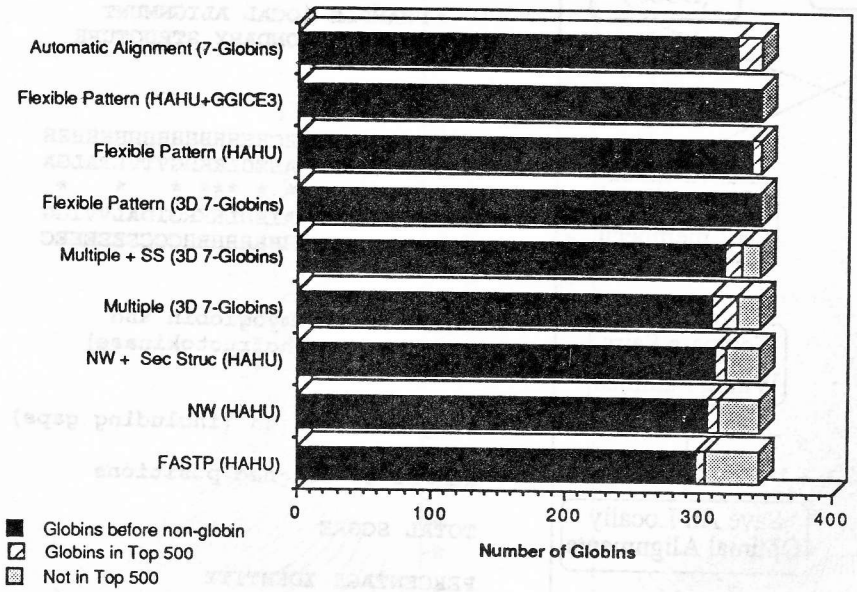
Environment specific weights are derived by studying the preferred environment, (i.e. exposed/buried, polar/non-polar contacts, secondary structure preference etc.) of each amino acid type, these weights are then applied to each residue in the protein of known three-dimensional structure to define a structural profile. Conventional dynamic programming is used to determine which proteins of unknown structure give the best alignment with the profile. Published accounts suggest, rather disappointingly, that environment weight methods give similar performance to conventional sequence comparison techniques (Bowie *et al.* 1991; Overington *et al.* 1992). **Pairwise potentials** take the idea of encoding the local environment a stage further. A residue-residue pair potential is derived from proteins of known structure, the sequence of unknown structure is then fitted to the core of the known structure and the lowest energy threading determined. This is a difficult optimization process that can not be solved by conventional dynamic programming techniques. However, preliminary results suggest this method has promise for detecting proteins that have similar folds, yet rather dissimilar sequences (e.g. two "Rossmann" beta/alpha/beta folds).

Evaluation of comparison methods

One approach to evaluating comparison methods is to select a well characterised protein family, then scan the entire sequence databank for members of that family. The globins provide a good test case with over 300 protein sequences known, and with representatives of diverse families with known crystal structures. Figure 1 illustrates the comparative success of different techniques for detecting the globin fold. In the databank scanned, there were 345 complete globin sequences, the query sequence or pattern was compared to all sequences (>6000) in the databank

Figure 1

Globin Scans (345 Whole Globins in Database)



and the resulting scores ranked. The results are presented as three values - number of globins before first non-globin, number of globins in top 500 scoring sequences, and number of globins not found in the top 500. These values give a measure of the selectivity and sensitivity of the different methods. Moving from the bottom of Figure 1, the methods get progressively better as more information is included in the scan. Starting with the simple, but fast FASTP algorithm (Lipman & Pearson, 1985), through Needleman & Wunsch (1970) (NW) , NW with secondary structural information, structurally derived multiple alignment and multiple alignment with secondary structure information, to flexible pattern, the methods improve. A flexible pattern derived from a single sequence and secondary structure does slightly worse than that derived from 7 structures, whilst adding a further sequence to the pattern recovers the sensitivity. Finally, deriving a pattern from an automatically determined alignment is slightly less specific than the structurally based alignment (see Barton & Sternberg, 1990 and Barton, 1990, for further details and discussion).

Guidelines for interpreting pairwise sequence alignments

Multiple alignment and flexible pattern comparisons provide good discrimination for the encoded protein fold. However, these techniques are not as frequently used as traditional methods. It is therefore important to have clear guidelines for interpreting conventional pairwise sequence alignments. For example, if I am given an alignment of two protein sequences that shows 32% identity, should I believe that the two proteins share similar folds? This question applies equally whether the proteins are of known or unknown structure. The following study was performed to establish general guidelines for the structural interpretation of sequence alignments.

Figure 2 illustrates the flow of the analysis. 477 chains from the Brookhaven PDB were grouped into two sets. Set 1 contained 89 unrelated protein chains selected from proteins known to have different folds. Set 2 comprised 57 distinct

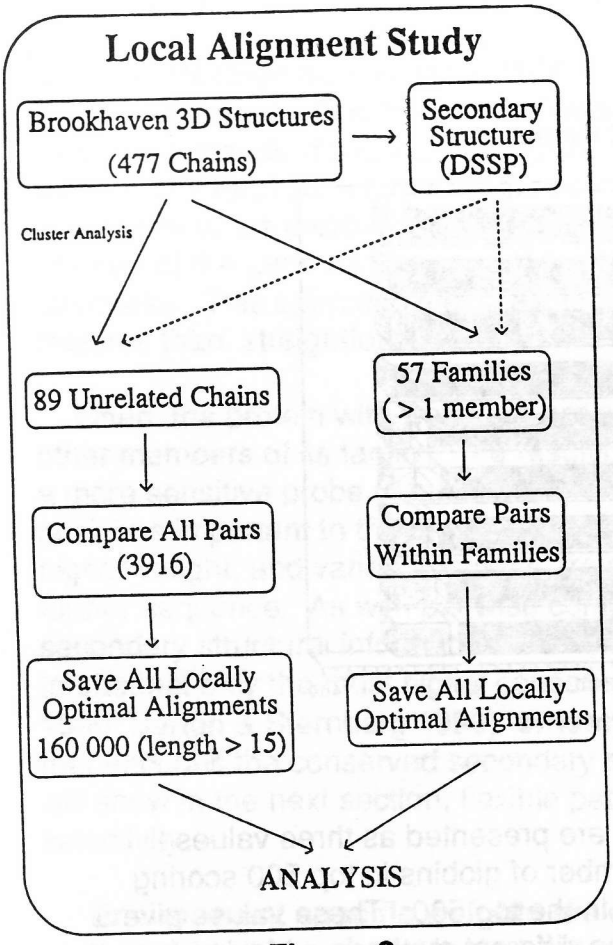


Figure 2

Figure 3

EXAMPLE LOCAL ALIGNMENT
SHOWING SECONDARY STRUCTURE

```
   CCCCCCCHHHHH HCHHHHHHHHHHHHHHHHH
45 RFKHLKTEAEMK ASEDLKKHGVTVLTALGA    74
   **     *     * * *** *   *   *
73 RFPEFRDENIRAVAIENLKKRGIDALVVIGG   103
   CCHHHHCHHHHHHHHHHHHHHCCCEEEEEEC

   (Sperm whale myoglobin and
    E.coli phosphofructokinase)

   Alignment length (including gaps) = 31

   Number of aligned positions       = 30

   TOTAL SCORE                       = 61

   PERCENTAGE IDENTITY               = 35.5

   PERCENTAGE ACCURACY               = 50.0
```

families with each family made up of at least two proteins. The secondary structure of all proteins was defined by the Kabsch & Sander (1983) program DSSP. The 89 unrelated protein sequences were then compared pairwise using a variant of the Smith-Waterman (1981) local similarity dynamic programming algorithm, scoring conservative substitutions with Dayhoff's matrix. This algorithm can locate *all* locally optimal alignments between two protein sequences and resulted in 160,000 alignments of length > 15 . For each alignment, the statistics shown in Figure 3 for two unrelated proteins were calculated . A similar process was also performed within each of the 57 families in Set 2.

The alignments obtained between the 89 unrelated proteins provide a set of scores and accuracies against which any sequence comparison may be measured. Figures 4a and 4b illustrate a plot of percentage accuracy against percentage identity for the unrelated and related proteins respectively. There is no correlation between the percentage identity and accuracy for proteins of unrelated structure. Indeed, local alignments of unrelated proteins can show up to 45% identity when optimally aligned, and many related pairs show less than this value! This would suggest that it is impossible to say whether or not our 32% identity alignment indicates structural similarity. Of course, the situation is not as bad as Figure 4

Figure 4a (Unrelated Proteins)

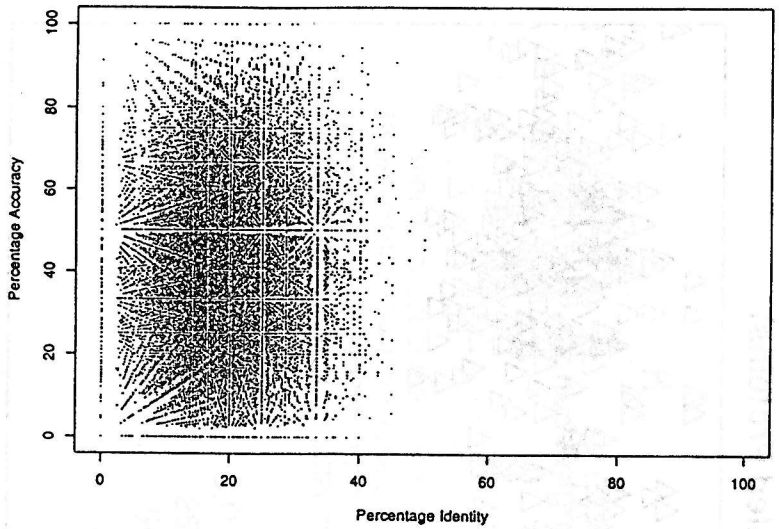


Figure 4b (Related Proteins)

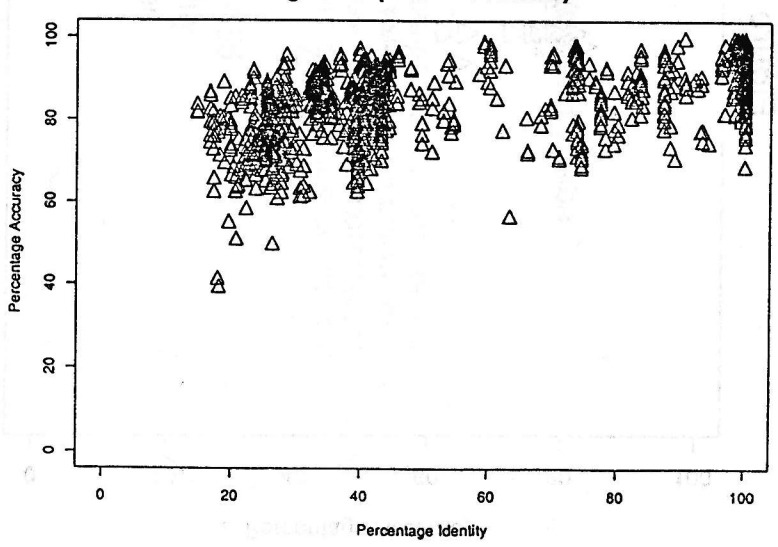


Figure 5a

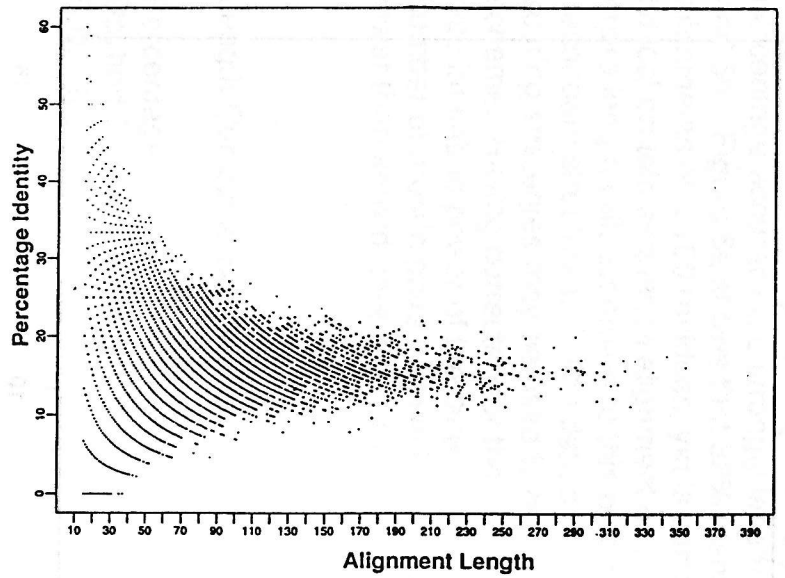


Figure 5b

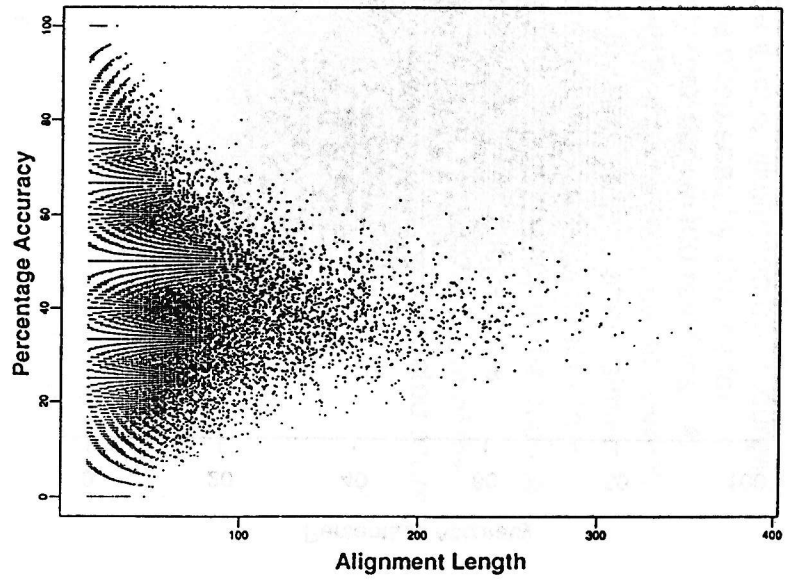


Figure 6a (Unrelated Proteins)

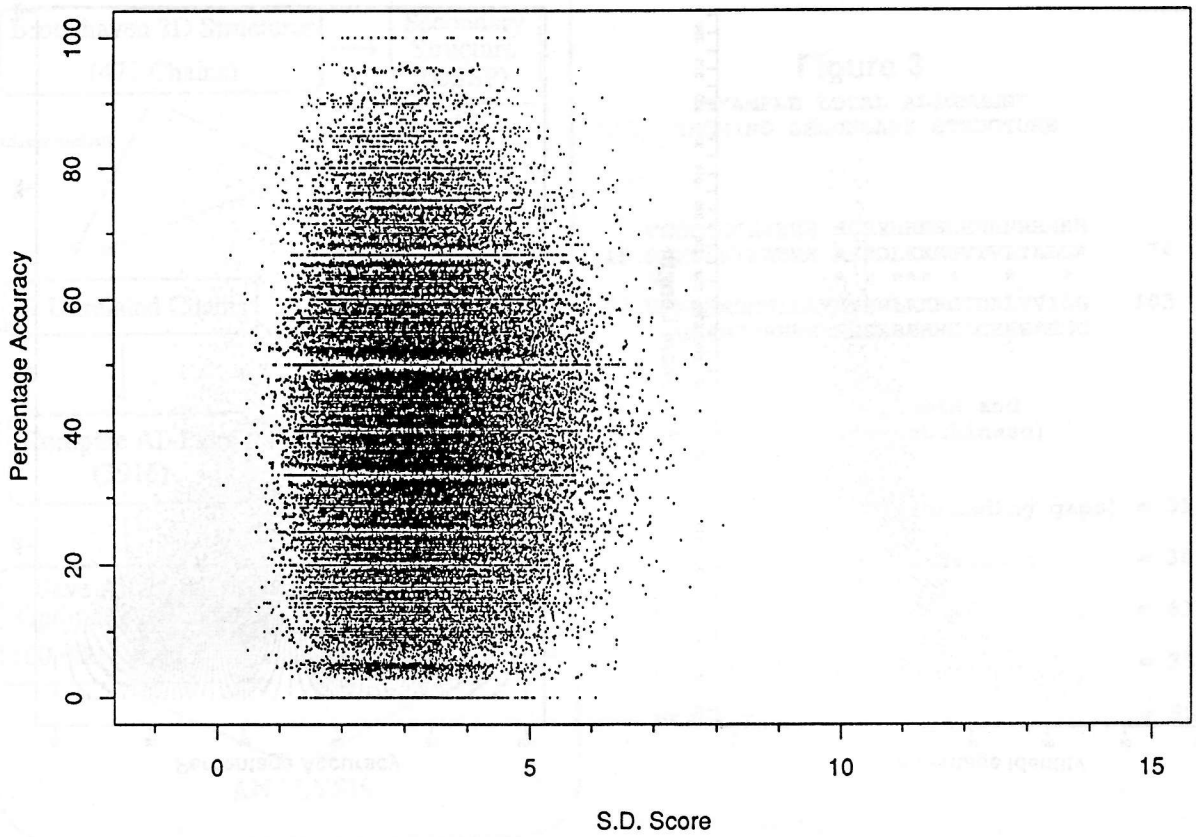
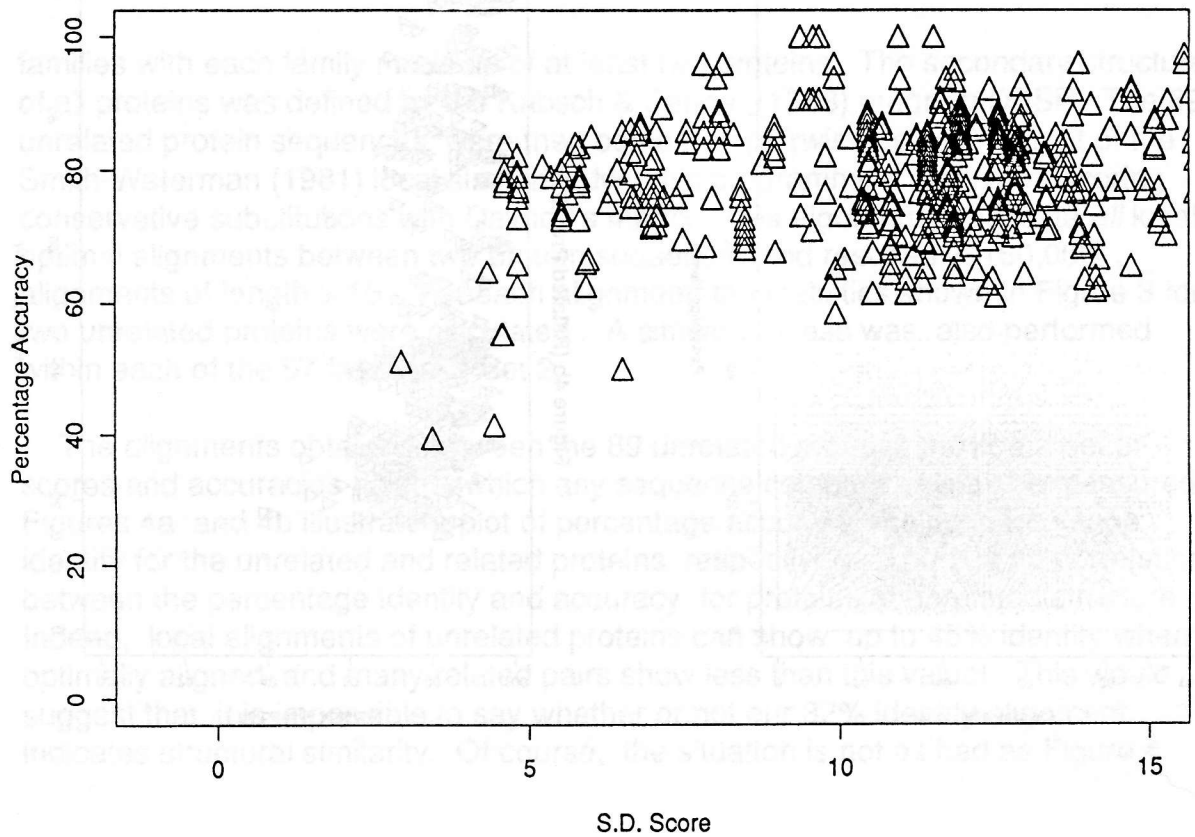


Figure 6b (Related Proteins)



suggests, since both percentage identity (Sander & Schneider, 1991) and percentage accuracy are strongly length dependent as is illustrated by Figures 5a and 5b. Figure 5a shows that 32% identity will be often seen by chance for alignments of < 100 residues, yet is comparatively rare above 100 residues. A typical protein sequence alignment of 150 residues that gives 32% identity over its entire length will, according to this plot, show that the two proteins share similar secondary structures. The data shown in Figure 5a and similar plots for other scoring strategies may be used to correct for the length dependency in the scoring scheme. Having corrected for the length effects, we can then compare the discriminating power of the different approaches. Table 1 shows the count of the number of protein pairs that we know share similar folds, yet give a corrected score lower than known unrelated pairs of proteins.

Table 1.

Length Corrected Scoring Scheme	Number
Percentage Identity	1280
Dayhoff Score	865
S.D. Score	846

Figure 7
Aligned Fragments of 2cts and 2paba

```
     *         *         *    *  *    **        * ****
     HHHHHCCCCCCCCHHHHHHHHHHHCCCCCHHHHHHHHHHHHHCCCCCCCHHHHHHH
230  LYLTIHSDHEGGNVSAHTSHLVGSALSDPYLSFAAAMNGLAGPLHGLANQEVLV 283
  3  LMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTTEEQFV  56
     EEEEEEECCCCEECCCCEEEEEEECCCCCEEEEEEEECCCCCEECCCCCCCCCC
```

S.D. Score = 7.55
25.9% Identity
54 Residues

As one might expect, a scoring scheme that takes into account conservative substitutions (Dayhoff Score) performs somewhat better than the percentage identity scheme. The S.D. score is often calculated to estimate the similarity between protein sequences. This is done by first optimally aligning the two sequences, then shuffling the sequence orders, and re-aligning 100 or more times. The mean and standard deviation of the shuffled sequence alignment scores is then used to normalise the alignment score of the native sequences. S.D. scores also show a length dependency which is corrected for in Table 1. However, it is interesting to plot the raw S.D. scores against accuracy of alignment as shown in Figure 6a/b. It has long been known that S.D. scores higher than 5-6 are required to illustrate a genuine relatedness, or structural similarity (Barton & Sternberg, 1987b, Dayhoff, 1978), and Figure 6 graphically illustrates this phenomena. The mean S.D. score for unrelated proteins is close to 3.0 and not 0.0 as it would be if protein sequence alignments were random. Some scores for the unrelated proteins are as high as 7.5 as for the example in Figure 7. Clearly, although these sequences give an S.D. score of 7.55, the value of 26% identity in 54 residues is insignificant according to Figure 5a.

Summary

Protein sequence comparison methods can be sensitive and selective tools for detecting proteins of known structure that share a similar fold to a protein undergoing crystallographic analysis. The analysis presented here, in particular Figures 5 and 6 should be useful when assessing the suitability of a protein as a search object for molecular replacement.

References

Barton, G. J., Meth. Enzymol. *183* (1990) 403-428.
Barton, G. J. & Sternberg, M. J. E., Prot. Eng. *1*, (1987a) 89-94.
Barton, G. J. & Sternberg, M. J. E., J. Mol. Biol. *198,* (1987b) 327-337.
Barton, G. J. & Sternberg, M. J. E., J. Mol. Biol. *212*, (1990) 389-402.
Bashford, D. Chothia, C. & Lesk, A. M., J. Mol. Biol. *196,* (1987) 199-216.
Kabsch, W. & Sander, C., Biopolymers *22*, (1983) 2577-2637.
Lipman, D. J. & Pearson, W. R., Science *227* (1985) 1435-1441.
Needleman, S. B. & Wunsch, J. Mol. Biol. *48*, (1970) 443-453.
Overington, J. *et al.* Prot. Sci. *1* (1992)
Sander, C. & Schneider, R., PROTEINS, *9*, (1991) 56-68.
Smith, T. F. & Waterman, M. S., J. Mol. Biol. (1981) 195-197.
Taylor, W. R., J. Mol. Biol. *188,* (1986) 233-258.