Pairwise iterative superposition of distantly related proteins and assessment of the significance of 3-D structural similarity

Alex C.W. May

Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

1Present address: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK

Introduction

Although most protein 3-D structures can be classified into families with common folds (Chothia, 1992; Blundell and Johnson, 1993; Murzin et al., 1995), such families can comprise members lacking any significant sequence similarity (Murzin and Chothia, 1992; Orengo et al., 1994). The classification of protein folds into families depends on methods for comparison and the clustering of protein tertiary structures (for reviews see Holm and Sander, 1994; Orengo, 1994).

Rigid-body superposition with the aim of minimizing the r.m.s. distance (r.m.s.d.) between corresponding Cα positions is often used to compare the 3-D structures of proteins. The identification of equivalences for superposition is a problem with such approaches. Alternatively, a structure alignment can be derived using the dynamic programming sequence alignment algorithm of Needleman and Wunsch (1970) to compare structural features and relationships (Taylor and Orengo, 1989; Šali and Blundell, 1990). Recently, Holm and Sander (1993c) used Monte Carlo optimization to align fragment inter-molecular Cα distance matrices, and Boutonnet et al. (1995) described a superposition method based on the clustering of fragments with similar conformations.

Methods that apply rigid-body superposition may fail because of differences in the orientation and position of secondary structure elements of proteins, despite the latter often retaining the same topology (Chothia and Lesk, 1986). However, most methods of comparative modelling are based on the assembly of rigid fragments from known 3-D structures. For instance, in the COMPOSER approach (Blundell et al., 1988), an ‘average’ (‘framework’) structure for the structurally conserved regions (SCRs) of the known structures, which usually comprises a set of helices or strands that are conserved in the family, is determined by direct multiple superposition (Sutcliffe et al., 1987) without bias to any one structure in the set.

To cluster protein folds, a measure of degree of structural similarity (analogous to a distance) is required. A matrix of ‘distances’ computed between all pairs of proteins can be used to construct trees that describe the relationships among them. Johnson et al. (1990a,b) calculated a structural distance measure from fractional topological equivalence and r.m.s.d. for a pairwise superposition that correlates well with the sequence metric (see below). By developing an algorithm to estimate the expected r.m.s.d. between a native structure and a random one of the same size, Cohen and Sternberg (1980) suggested a criterion for the significance of structural similarity. Recently, the statistical significance of local structural similarities in terms of r.m.s.d. has been estimated by Alexandrov and co-workers (Alexandrov et al., 1992; Alexandrov and Go, 1994). In contrast, Maiorov and Crippen (1994) proposed a nonstatistical definition of the significance of r.m.s.d. values in structure comparisons: two conformers are said to be intrinsically similar if their r.m.s.d. is smaller than that when one of them is mirror inverted. Recently, Maiorov and Crippen (1995) proposed a measure of similarity—based on r.m.s.d. and the radii of gyration of the two structures — which, unlike r.m.s.d. on its own, has the useful property of being size independent.

Similarity of structural features and relationships other than r.m.s.d. have also been used as measures of significance of structural correspondence. Johnson et al. (1990a,b) extended their distance metric from rigid-body superposition to incorporate sequence and structural features, including relationships such as hydrogen bonding patterns (Šali and Blundell, 1990). One improvement made by Orengo et al. (1992) to the original approach of Taylor and Orengo (1989) was the derivation of a logarithmic score to measure the average similarity of residue structural environments. This score correlates well with the r.m.s.d. after rigid-body superposition on the basis of structural equivalences (Orengo et al., 1992). Yee and Dill (1993) described a similarity measure based on weighted distance maps.

Keywords: protein structure comparison algorithm/r.m.s. distance/significance of structural similarity/structural similarity measure/topological equivalence
Of course, a more flexible definition of topological equivalence than that based solely on rigid-body superposition allows for the more robust identification of structural similarity where the best rigid group fitting is not necessarily the 'best' answer. By definition, such approaches have been held to be essential for the recognition of distant structural relationships (Holm and Sander, 1994; Orengo, 1994). Here I demonstrate that it is possible to extend the usefulness of the dynamic programming-based rigid-body superposition method so that it can detect remote similarity (e.g. the globin fold and colicin A), recognized previously only by slower, more flexible methods, and improve framework extraction for modelling by assembly of the rigid fragments. Of course, the use of dynamic programming means that the method cannot recognize structural similarity in cases of different topology or reversal of the chain direction. The point is to demonstrate how the dynamic programming-based superposition method can be made even more powerful so that distant structural similarity can be identified and quantified.

Structural feature alignment using dynamic programming at the amino acid level identifies equivalenced residues for an initial superposition. This initial fit can then be improved by cycles of dynamic programming and Cα superposition. Topological equivalences are assigned on the basis of a cut-off distance for intermolecular Cα distances. A novel feature of the method shown here is that the final superposition obtained in this way can itself be refined further by using all the aligned positions, and not just the equivalenced ones, as initial equivalences to start another iteration with the same cut-off distance. The relationship between the number of equivalences, r.m.s.d. and cut-off distance is investigated. I also show how a previously published distance metric from rigid-body superposition (Johnson et al., 1990a,b) can be adapted to assess the quality of a given superposition not derived using a distance-based definition of topological equivalence. It is possible to discriminate between alternative superpositions for a given pairwise comparison.

Materials and methods

The algorithm

The use of dynamic programming (Needleman and Wunsch, 1970) to update topological equivalences following an initial superposition (Johnson et al., 1990a,b) means that rigid-body superposition requires only the specification of an initial set of equivalences. Unfortunately, such a designation is often not a trivial task.

The COMPOSER approach uses a multiple sequence alignment between all known structures to identify identical residues to be used as initial equivalences. This approach is fine when the members of a protein family have closely related sequences and are of similar length. Alternatively, a (slow) genetic algorithm-based procedure for rigid-structure superposition, in which the assignment of initial equivalences is not required, has been described (May and Johnson, 1994, 1995).

Subbiah et al. (1993) identified initial equivalences for rigid-body superposition from trials using five different schemes. In the approach described here (Figure 1), encoded in the completely automatic program YASCA (Yet Another Structure Comparison Algorithm), the local alignment method of Smith and Waterman (1981) is used first to compare inter-Cα torsion angles as represented in the four-letter classification of Ring et al. (1992). [The name of the program was inspired by the

Fig. 1. Flow chart for the program YASCA.

UNIX utility YACC (Yet Another Compiler-Compiler). Despite a multitude of automatic methods, structure comparison is still not resolved. This is not too surprising given that the concept of structural similarity and its significance are operationally defined (for a review see Mizuguchi and Go, 1995). Initial equivalences for superposition (Part 0 in Tables I and III) are selected as those identical positions in the structural sequence alignment where the difference in angles is less than a cut-off value (10° is used here because it seemed to select appropriate positions). Levine et al. (1984) compared the backbone dihedral angles of protein structures without using dynamic programming. One of the five schemes used by Subbiah et al. (1993) to define initial equivalences for updating is a dynamic programming-based comparison of inter-Cα torsion angles.

The initial equivalences are then iteratively updated by dynamic programming (Smith and Waterman, 1981) after each round of superposition. The definition of topological equivalence is a distance-based one: those intermolecular Cα distances less than a cut-off value are deemed equivalent (3.0 Å is used here). The cycles of superposition and dynamic programming terminate when the number of equivalences does not increase and the r.m.s.d. is stable. The novel development at this stage is the use of a two-step iterative updating of equivalences (Tables I and III). First, the initial equivalences from the structural sequence alignment are used to seed a superposition in a standard fashion. This is the point at which current dynamic programming-based methods finish (Part I in Tables I and III). A problem with using a cut-off distance to define topological equivalence is that those intermolecular Cα
Table I. Superposition of the 3-D structures of human deoxyhaemoglobin (1HBB, A chain) and E. coli colicin A (1COL, A chain) using YASCA before screening of the final equivalences

<table>
<thead>
<tr>
<th>Part</th>
<th>No. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>Pairwise structural distance metric (cut-off, Å)</th>
<th>No. of aligned positions</th>
<th>Time taken (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>73</td>
<td>9.78</td>
<td>130</td>
<td>104</td>
<td>1.00</td>
</tr>
<tr>
<td>1d</td>
<td>37</td>
<td>1.92</td>
<td>104</td>
<td>18.00</td>
<td></td>
</tr>
<tr>
<td>2°</td>
<td>47</td>
<td>1.92</td>
<td>103</td>
<td></td>
<td>61.00</td>
</tr>
<tr>
<td>Iteration 1</td>
<td>54</td>
<td>2.01</td>
<td>105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration 3</td>
<td>65</td>
<td>1.94</td>
<td>102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration 4</td>
<td>65</td>
<td>1.94</td>
<td>102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Refit N-terminal SCR</td>
<td>4 (0)</td>
<td>0.60</td>
<td>4</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Refit C-terminal SCR</td>
<td>22 (10)</td>
<td>0.75</td>
<td>22</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Final superposition</td>
<td>112</td>
<td>3.61</td>
<td>112</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*a*According to Johnson et al. (1990a,b). Cut-off is the distance threshold for the definition of topological equivalence (see text).

*b*Silicon Graphics INDY: 100 MHz IP22 processor, MIPS R4600 CPU, 32 MB RAM. Times are shown to two decimal spaces.

*c*Superposition on the basis of initial equivalences identified by the alignment of inter-Cα torsion angles (see text).

*d*Superposition following iterative updating of the initial equivalences (Part 0) (see text).

*e*Superposition following iterative updating of the aligned positions of Part 1 as initial equivalences (see text).

*f*The number in parentheses is the number of new equivalences (those not included in the 65 from iteration 4 of Part 2) after refitting each terminal SCR.

Fig. 2. Structure-based local sequence alignment (Smith and Waterman, 1981) of human deoxyhaemoglobin (A chain) (Protein Data Bank entry 1HBB) and E. coli colicin A (A chain) (Protein Data Bank entry 1COL) derived from superposition with YASCA after iteration 4 of Part 2 (Table I). Distance-based topological equivalences (see text) are shown as uppercase matched residue positions. The cut-off distance for equivalences is 3.0 Å. The format for aligned structures is similar to that of the PIR sequence database and consistent with that used by Overington and co-workers in software for the derivation of rules for the relationships between sequence and structure from aligned 3-D structures (for a review see Blundell and Johnson, 1993; Sali and Overington, 1994).

Distances just greater than the cut-off value are discarded in the same way as those distances much greater than the cut-off value.

Although modified to incorporate dynamic programming (P. Thomas and M.S. Johnson, unpublished results), the COMPOSER approach (Sutcliffe et al., 1987) only produces an alignment for the SCRs. The alignment of the intervening regions (the structurally variable regions; SVRs), however, can be used to refine the superposition. The aligned positions from the SCRs and SVRs in the first fit (Part 1 in Tables I and III) are used to start another iterative superposition with the same cut-off distance. This process is repeated until the number of aligned positions stays constant. It usually leads to an improvement in the superposition (Part 2 in Tables I and III). Incidentally, a consideration ignored in the COMPOSER structure alignment is a check at each round of superposition that the current superposition is not the same as any previous one. At present, COMPOSER only checks whether the current superposition differs from the immediately previous one. This can lead to a situation where the program can oscillate between two superpositions until it terminates after a hard-coded number of cycles.

Use of the Smith and Waterman (1981) algorithm for best local alignment means that the YASCA structure alignment does not include terminal overhangs and the terminal aligned positions are topological equivalences (Figures 2 and 4). This suggests a strategy to determine whether the structure alignment can be extended at the termini. Of course, if one of the terminal topological equivalences comprises a terminal residue of a structure, then the alignment cannot be extended at this position. At least three initial equivalences are selected for refitting of the terminal SCRs. These initial equivalences are chosen by examining the equivalences from the terminal position for a break in residue continuity at least one of the structures. If the contiguous terminal SCR consists of at least three pairs of residues, then these are used as the seed for the rigid-body superposition of segments from the terminal topological equivalences to the termini of the structures (Figure 3). If, however, the contiguous terminal SCR consists of fewer than three pairs of residues, then further equivalences are recruited in the same way until there are at least three initial equivalences. Refitting of the terminal fragments on the basis of the end SCRs (Tables I and III), if possible, means that differences in the orientation and position of secondary structure elements outside the range of local structure alignment can still be accommodated using the same cut-off distance. Clearly, this approach could be repeated, so allowing the detection of domain movements.

A final superposition is performed on the basis of the topological equivalences and nonequivalent aligned positions from the Part 2 structure alignment, as well as any new equivalences obtained by refitting the terminal SCRs (Tables I and III). The rationale for including the nonequivalent aligned positions is as follows. Aligned residue positions within
equivalent segments of the local structure alignment may not be considered topologically equivalent at a given cut-off distance because the optimal rigid-body superposition of equivalent segments of two structures will not coincide perfectly with the fit of the entire structures (Figure 3). It is not necessary, however, to refit the equivalent segments as independent substructures because the nonequivalent aligned positions are bounded within the local structure alignment by topological equivalences. However, the terminal SCRs must be refitted because the terminal aligned positions cannot be nonequivalent due to the use of the Smith and Waterman (1981) local alignment algorithm. The problem with using the Needleman and Wunsch (1970) global alignment algorithm here would be that all those positions after the terminal topological equivalences would be accorded zero similarity by using a cut-off distance. This means that there would be no means of identifying those aligned positions at the termini (if any) that could be safely added to the main alignment. Selection of the initial equivalences from the terminal equivalences of the main alignment ensures that the new terminal structure alignments are consistent with the former.

Having calculated all equivalenced intermolecular Cx distances for the final fit, 0.5 Å increments to the starting cut-off distance are made until all the equivalenced distances are within a cut-off. Those equivalences of the final set within a given cut-off are then used as initial equivalences for the two-step iterative updating of equivalences, refitting of the terminal SCRs (if possible) and a new final superposition. The cut-off distance used is that which defined a subset of the above final equivalences (Tables I and III). The end result is a series of
superpositions for a given pair of structures (Table II), all of which differ from the final superposition obtained with the original cut-off distance (Tables I and III). The question is can we assess the quality of the individual superpositions so that a 'best' one might be identified?

**Results and discussion**

**Application of the method**

To demonstrate the broad applicability of the YASCA algorithm, three test cases in order of increasing difficulty will be described (the starting cut-off distance for equivalent Ca atoms is 3.0 Å). First, comparison between the mammalian serine proteinase bovine trypsin all-β 3-D structure (Protein Data Bank entry 3PTN) and an all-β bacterial one (Protein Data Bank entry 3SGB) is almost trivial when using a dynamic programming-based approach. This is because of the similarity in length of the proteins (223 and 185 amino acids), which means that alignment on the basis of the local environment in each structure provides robust initial equivalences for iterative updating. As expected, there is very little change between the Part 1 and 2 superpositions (113 equivalences, r.m.s.d. = 1.4 Å; 112 equivalences, r.m.s.d. = 1.4 Å), and only one extra equivalence is defined after refitting the terminal SCRs. With 54 nonequivalent Part 2-aligned positions there are 167 final equivalences that superpose with a r.m.s.d. of 4.9 Å. The superposition with a 4.0 Å cut-off defines an extra equivalence (168) with a lower r.m.s.d. between them (4.7 Å).

Second, identification of the common EF-hand fold for the two calcium binding proteins, *Drosophila* calmodulin (Protein Data Bank entry 4CLN) and leopard shark parvalbumin (Protein Data Bank entry 5PAL), is more challenging for a dynamic programming-based method. This is because of the difference in length of the proteins (148 and 109 amino acids) and because the EF-hand fold is an all-β one, meaning that the local environment in each structure will appear the same at the level of their residues. Iterative updating of the Part 0 initial equivalences defines only 14 Part 1 equivalences with an r.m.s.d. of 1.8 Å. Current dynamic programming-based methods terminate at this stage. However, consideration of the entire aligned positions from the derived alignment as seeds for further cycles of fitting affords the identification of 64 Part 2 equivalences with an r.m.s.d. of 1.3 Å. With no new equivalences obtained after refitting the terminal SCRs and four nonequivalent Part 2-aligned positions, there are 68 final

---

**Table II. Superposition of the 3-D structures of human deoxyhaemoglobin (1HBB, A chain) and *E.coli* colicin A (1COL, A chain) using YASCA: screening of the final equivalences (Table I) to select initial equivalences for new superpositions**

<table>
<thead>
<tr>
<th>Cut-off distance (Å)</th>
<th>Starting no. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>Final no. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>Pairwise structural distance metric (cut-off; Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>77</td>
<td>2.27</td>
<td>113</td>
<td>3.74</td>
<td>39.08 (10.0)</td>
</tr>
<tr>
<td>4.0</td>
<td>88</td>
<td>2.50</td>
<td>116</td>
<td>3.65</td>
<td>37.72 (10.0)</td>
</tr>
<tr>
<td>4.5</td>
<td>91</td>
<td>2.55</td>
<td>116</td>
<td>3.43</td>
<td>43.48 (8.0)</td>
</tr>
<tr>
<td>5.0</td>
<td>98</td>
<td>2.77</td>
<td>116</td>
<td>3.58</td>
<td>40.71 (9.0)</td>
</tr>
<tr>
<td>5.5</td>
<td>100</td>
<td>2.84</td>
<td>117</td>
<td>3.69</td>
<td>34.79 (11.0)</td>
</tr>
<tr>
<td>6.0</td>
<td>101</td>
<td>2.90</td>
<td>116</td>
<td>3.59</td>
<td>42.94 (8.5)</td>
</tr>
<tr>
<td>6.5</td>
<td>105</td>
<td>3.12</td>
<td>121</td>
<td>3.98</td>
<td>43.75 (9.0)</td>
</tr>
<tr>
<td>7.0</td>
<td>108</td>
<td>3.28</td>
<td>122</td>
<td>4.20</td>
<td>45.73 (9.0)</td>
</tr>
<tr>
<td>7.5</td>
<td>109</td>
<td>3.34</td>
<td>123</td>
<td>4.58</td>
<td>53.94 (8.0)</td>
</tr>
<tr>
<td>8.0</td>
<td>110</td>
<td>3.41</td>
<td>124</td>
<td>4.81</td>
<td>46.60 (10.0)</td>
</tr>
<tr>
<td>8.5</td>
<td>110</td>
<td>3.41</td>
<td>124</td>
<td>4.81</td>
<td>46.60 (10.0)</td>
</tr>
<tr>
<td>9.0</td>
<td>110</td>
<td>3.41</td>
<td>124</td>
<td>4.81</td>
<td>46.60 (10.0)</td>
</tr>
<tr>
<td>9.5</td>
<td>112</td>
<td>3.61</td>
<td>126</td>
<td>5.32</td>
<td>52.54 (9.5)</td>
</tr>
</tbody>
</table>

*aNumber of the 112 final equivalences (Table I) within a given cut-off distance.

*bYASCA final superposition with a given cut-off distance to define topological equivalence (see text).

*cAccording to Johnson *et al.* (1990a,b). Cut-off is the distance threshold for the definition of topological equivalence (see text).

---

**Table III. Superposition of the 3-D structures of *T.dysxia* deoxyhemerythrin (1HMID, IA subunit) and *E.coli* apocytochrome *b*₅₆ (1APC) using YASCA before screening of the final equivalences**

<table>
<thead>
<tr>
<th>No. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>Pairwise structural distance metric (cut-off; Å)</th>
<th>No. of aligned positions</th>
<th>Time taken (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 0*</td>
<td>36</td>
<td>9.32</td>
<td>82</td>
<td>0.00</td>
</tr>
<tr>
<td>Part 1*</td>
<td>40</td>
<td>1.89</td>
<td>85</td>
<td>3.00</td>
</tr>
<tr>
<td>Part 2*</td>
<td></td>
<td></td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>Iteration 1</td>
<td>61</td>
<td>1.78</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Iteration 2</td>
<td>61</td>
<td>1.78</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Refit C-terminal SCR</td>
<td>19 (4)</td>
<td>1.27</td>
<td>19</td>
<td>0.00</td>
</tr>
<tr>
<td>Refit C-terminal SCR</td>
<td>8 (3)</td>
<td>1.35</td>
<td>10</td>
<td>0.00</td>
</tr>
<tr>
<td>Final superposition</td>
<td>79</td>
<td>3.41</td>
<td>79</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*aAccording to Johnson *et al.* (1990a,b). Cut-off is the distance threshold for the definition of topological equivalence (see text).

*bSilicon Graphics INDY: 100 MHz IP22 processor, MIPS R4600 CPU, 32 MB RAM. Times are shown to two decimal spaces.

*cSuperposition following iterative updating of the initial equivalences (Part 0) (see text).

*dSuperposition following iterative updating of the aligned positions of Part 1 as initial equivalences (see text).

*eThe number in parentheses is the number of new equivalences (those not included in the 65 from Iteration 4 of Part 2) after refitting each terminal SCR.
equivalences that superpose with an r.m.s.d. of 1.8 Å. This superposition corresponds to the equivalencing of the calcium ion ligands between sites III and IV of 4CLN and sites CD and EF of 5PAL. An alternative matching of sites I and II of 4CLN and sites CD and EF of 5PAL (May and Johnson, 1995) is not detected here because the initial equivalencing step restrains exploration of the alternative superpositions. The YASCA superposition with a 4.0 Å cut-off also defines 68 equivalences (of which 65 are common to the above 68) but with a lower r.m.s.d. of 1.6 Å.

Third, the globins, phycocyanins and colicin A make up a family described by a common fold (the three-on-three helical sandwich) but lack any significant sequence similarity—the globin fold is hence described as a 'superfold' (Orengo et al., 1994). This unexpected similarity was first identified independently by two automatic structure comparison methods: DALI (Holm and Sander, 1993a,b,c) and a version of the SSAP algorithm (Taylor and Orento, 1989) that incorporates a local alignment method (Orengo and Taylor, 1993). The latter assigns 111 equivalent residue structural environments between human haemoglobin (A chain) (Protein Data Bank entry 1HBB) and colicin A (A chain) (Protein Data Bank entry 1COL), which superpose giving an r.m.s.d. of 3.5 Å (Table V). The YASCA superposition with a 4.5 Å cut-off, defining five extra equivalences (116) with a lower r.m.s.d. between them (3.4 Å) (Table II), is a 'better' one than that of Orengo and Taylor (1993) with both a higher number of equivalences and a lower r.m.s.d. Of course, interpretation of the significance of the other YASCA alternatives is not as straightforward given the trade-off between maximizing the number of topological equivalences and minimizing the corresponding r.m.s.d. There is no a priori information that can be used to predict the nature of the relationship between the number of equivalences and the r.m.s.d. for a given pair of structures, so there is no choice but to perform iterative rigid-body superpositions with various cut-off distances. For instance, although superposition of C-phycocyanin (A chain) (Protein Data Bank entry 1CPC) and colicin A (A chain) (Protein Data Bank entry 1COL) with YASCA identifies the common fold, the relationship between r.m.s.d. and the number of equivalences against cut-off differs (data not shown) from that of a globin and colicin (Table II). The point here is that the number of equivalences for a given cut-off distance can be less than that obtained with a smaller cut-off distance. The advantage of the YASCA algorithm is that automatic selection of the most appropriate cut-off distance is still possible after superposition with the initial cut-off distance.

It is useful to review the grounds for the cut-off distance approach to define topological equivalence. First, Chothia and Lesk (1986) defined 'common cores' of homologous proteins by superimposing individual secondary structure elements and then extending these superpositions until the terminal equivalenced distances exceeded 3.0 Å. Use of the same cut-off distance for the superposition of entire structures, however, means that the SCR assignment will only not differ from that of Chothia and Lesk (1986) if there are 'insignificant' (for this purpose) differences in the orientation and position of the corresponding secondary structure elements. This explains why the COMPOSER algorithm (Sutcliffe et al., 1987) is not as successful in detecting distant structural similarity as those methods that identify equivalences without superposition. Second, not having any idea of the nature of the relationship between the number of equivalences and the r.m.s.d. for a given pair of structures, the use of a cut-off distance with dynamic programming guarantees a low r.m.s.d. Again, this approach is most successful for the comparison of highly structurally similar proteins where the r.m.s.d. is a good measure of the significance of structural similarity (Johnson et al., 1990a,b). For more distantly related structures, however, the COMPOSER algorithm (Sutcliffe et al., 1987) will underestimate the size of the common fold (the number of topological equivalences) because of the cut-off constraint: in the above YASCA superposition with a 4.5 Å cut-off distance (116 equivalences, r.m.s.d. = 3.4 Å), for instance, the worst equivalenced distance is 7.8 Å.

It is useful to consider two further examples of common folds which occur in proteins with different functions and low sequence identities—the superfolds (Orengo et al., 1994).

a. Part 1 alignment

>Pl;1hmd

CA-CA distances in superposition

IldDhEktIqNqiLLsquadnadhIlnELrrCTGKhFlneqIqLqSSQ---------
yagYAEHKKADhD---FihkldtwDGDV--TyaknwlVnhiKTidfKyrK*  

>Pl;1apc

CA-CA distances in superposition

EtnDNdLkvIekaDNA-------aqvkDAltKMRAAlaakkatPPKLeDkpspe

mkdFRHGFIDtvGqIddalklEnEGKVkeAqaaeeqLK--TTrnAHqK*Y

b. Part 2 alignment

>Pl;1hmd

CA-CA distances in superposition

IDDEHKLFLNGIILLLsquadnAdhIlnELRRCTGKhFLEQQL---------mqss

QyAgYaeHKKADHDFIHKLDTW-------GGD-VTYakK*  

>Pl;1apc

CA-CA distances in superposition

NMETLNDNKLKEK------dNAA-QVkdDAlTkMRAAlDAKQAtppkledkpspe

MkDfRhgFDILGVGQIDDLALKAnegkvkeAQaaEeq-LK*
Despite structural similarity, these proteins may not be related by divergence from a common ancestor. Murzin et al. (1995) attempted to consider the evolutionary relationships in their structural classification.

The all-α four-helical up-and-down bundle fold (for a review see Harris et al., 1994) is currently found in six superfamilies (<http://scop.mrc-lmb.cam.ac.uk/scop/>; Murzin et al., 1995). The use of the current approach is shown in the superposition of a member of the hemerythrin superfamily [Themiste dyscrita deoxymerythrin (IA subunit); Protein Data Bank entry 1HMD] and one of the cytochrome superfamily (Escherichia coli apocytochrome b_{562}; Protein Data Bank entry 1APC). In view of the trypsin-like serine proteinase fold comparison, it might be expected that there would be very little change between the Part 1 and 2 superpositions because of the close similarity in length of the pair (113 and 106 amino acids). In fact, the number of equivalences increases by 53% and the r.m.s.d. between them decreases in the Part 2 superposition (Table III and Figure 4). The structural similarity in terms of the up-and-down packing of the four helices in a bundle is recognized.

Murzin et al. (1995) described four superfamilies (<http://scop.mrc-lmb.cam.ac.uk/scop/>), which adopt the all-β trefoil fold. Murzin et al. (1992) analysed three 3-D structures to characterize the β-trefoil fold: Erythrina cafra trypsin inhibitor (Protein Data Bank entry 1TIE), human interleukin-1β (Protein Data Bank entry 111B) and bovine acidic fibroblast growth factor (Protein Data Bank entry 1BAR). The Cambridge group identified 84 structural equivalences in the 12 common antiparallel β-strands. These equivalences are defined not just on the basis of a simple distance-based definition, as demonstrated by specifying the 84 structural equivalences as initial equivalences for a MNYFIT (Sutcliffe et al., 1987) multiple superposition with a cut-off distance of 3.0 Å.

The common fold for the three structures is reduced to 75 equivalences. With this constraint of a 3.0 Å cut-off, we can assess the success of the YASCA Part 1 and 2 superpositions [those before refitting each terminal SCR (if possible) and the cycles with less strict cut-off distances (Figure 1)] with respect to the alignment of Murzin et al. (1992) (Table IV). Clearly, the 12 common antiparallel β-strands are identified by this procedure.

### Assessment of the significance of structural similarity

The distance metric from rigid-body superposition (Johnson et al., 1990a,b) was used to cluster multiple protein structures on the basis of all-against-all pairwise comparisons. The cut-off distance for the definition of topological equivalence in the pairwise comparisons was used in deriving the similarity measure. The metric can also be applied to assess the series of superpositions for a given pair of structures (Tables II and V).

The worst equivalenced distance in each final superposition was rounded up to the nearest 0.5 Å to provide an effective cut-off distance for the similarity measure. In fact, in light of theoretical concerns about using r.m.s.d. as a measure of structural similarity (see, for example, Yee and Dill, 1993), it may be useful anyway to report the worst, best and mean equivalenced distances for a given superposition as well as the r.m.s.d. According to the metric, the ‘best’ YASCA superposition is that with a 5.5 Å cut-off distance: 117 equivalences, r.m.s.d. = 3.7 Å, where the worst equivalenced distance is 10.7 Å (Table II). In fact, all of the worst equivalenced distances in Table II are greater than that obtained in Orengo and Taylor’s (1993) superposition (7.0 Å), demonstrating the success of the YASCA method.

This approach can be generalized to assess the quality of a given superposition not derived using a distance-based definition of topological equivalence by using the worst equivalenced distance to afford an effective cut-off value for the metric (Table V). Furthermore, alternative structure-based global sequence alignments can be evaluated in the same way on the basis of superimposing aligned positions (Table V). The two alternative superpositions (for an explanation, see Table V) for the Orengo and Taylor (1993) 117 aligned positions (r.m.s.d. = 3.81 and 3.84 Å) are estimated to be very slightly better than the best YASCA superposition (117 equivalences, r.m.s.d. = 3.7 Å; Table II) because the worst equivalenced distances in the former are 11.4 Å while that for the latter is 10.7 Å. The best YASCA superposition for porcine myoglobin (A chain) (Protein Data Bank entry 1PMB) and colicin A (A chain) (Protein Data Bank entry 1COL) is that of 116 equivalences and

---

**Table IV. Superposition of the 3-D structures of three proteins with the β-trefoil fold using YASCA (Parts 0, 1 and 2)**

<table>
<thead>
<tr>
<th>Protein pair*</th>
<th>No. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>YASCA stage</th>
<th>No. of common equivalences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1TIE:111B</td>
<td>96</td>
<td>1.58</td>
<td>Part 2</td>
<td>75 (76)</td>
</tr>
<tr>
<td>1TIE:1BAR</td>
<td>90</td>
<td>1.60</td>
<td>Part 1</td>
<td>72 (74)</td>
</tr>
<tr>
<td>111B:1BAR</td>
<td>99</td>
<td>1.42</td>
<td>Part 2</td>
<td>74 (79)</td>
</tr>
</tbody>
</table>

*Protein codes are those used in the Brookhaven Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987).

---

**Table V. Assessing the quality of published superpositions derived from equivalences not selected using a simple intermolecular Cα distance-based criterion and their associated structure-based global sequence alignments**

<table>
<thead>
<tr>
<th>Protein pair*</th>
<th>No. of equivalences</th>
<th>R.m.s.d. (Å)</th>
<th>Pairwise structural distance metric (cut-off; Å)*</th>
<th>No. of aligned positions</th>
<th>R.m.s.d. (Å)</th>
<th>Pairwise structural distance metric (cut-off; Å)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HBB:1COLA</td>
<td>111</td>
<td>3.54</td>
<td>48.37 (7.5)</td>
<td>117</td>
<td>3.81</td>
<td>34.35 (11.5)</td>
</tr>
<tr>
<td>1PMB:1COLA</td>
<td>112</td>
<td>3.21</td>
<td>46.62 (7.5)</td>
<td>125</td>
<td>5.62</td>
<td>29.22 (20.5)</td>
</tr>
</tbody>
</table>

*Protein codes are those used in the Brookhaven Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987). The last letter is the chain identifier.

---

*Orengo and Taylor (1993).*

The two answers correspond to the two possible matches for Lys16 of 1hbb: this residue was omitted in Figure 7 of Orengo and Taylor (1993).
an r.m.s.d. = 3.6 Å (worst equivalenced distance = 13.6 Å) obtained with a 4.0 Å cut-off value, which scores very slightly better than the Holm and Sander (1993a,b) 125 aligned positions superposition (r.m.s.d. = 5.6 Å, worst equivalenced distance = 20.5 Å) (Table V).

May and Johnson (1995) have described an alternative method to estimate the quality of structure-based sequence alignments founded on an alignment score calculated using a structure-based amino acid scoring matrix derived from observed amino acid substitutions in 97 families of aligned 3-D structures [updated from the 65 homologous sets used by Johnson and Overington (1993)]. This alignment score is then divided by the length of the alignment, ignoring terminal overhangs, to give the score per alignment position.

Conclusion
Of course, virtual torsion angle alignment using dynamic programming will not always provide the best initial equivalences for rigid-body superposition. For instance, Subbiah et al. (1993) used five different initial equivalences for a given comparison. The point of this paper is not to suggest that this structural feature alignment is the most successful way to identify initial equivalences, but rather to demonstrate that their refinement by dynamic programming-based rigid-body superposition can often be improved. Furthermore, such an approach extends the usefulness of the distance-based definition of equivalence so that a common fold can still be recognized for proteins with the same topological connections, despite shifts in the orientation and position of similar secondary structure elements. The use of dynamic programming to define a structure alignment means that, provided there are similar topological connections, the aligned residue positions not deemed to be equivalent within corresponding secondary structure elements can be still used to refine the final alignment. Matching secondary structure elements other than those at the termini of the alignment do not have to be superimposed in isolation. Clearly, the recognition of distant structural similarity will often require a cut-off distance for topological equivalences greater than the mean distance between neighbouring Cα atoms in a peptide chain (3.8 Å). The problem is that with a cut-off distance >3.8 Å there is a possibility of misequivalencing which, by definition, becomes more likely in cases of remote similarity. Like COMPOSER, YASCA uses a restrictive definition of equivalence at first which fixes the structure alignment in regions of close similarity but ignores those distances that exceed the initial cut-off distance. In the COMPOSER approach, however, the segments connecting the SCRs are discarded, regardless of the extent to which distances between aligned positions in these regions exceed the cut-off. This constraint on the definition of topological equivalence means that such an algorithm can only recognize close similarity. Following superposition with an initial restrictive cut-off distance, YASCA affords automatic selection of the most appropriate cut-off distance for a comparison.

While reliance on dynamic programming is a limitation, because its use imposes a linear constraint on any alignment, this method of comparison is the basis for almost all knowledge-based protein modelling. Derivation of the rules for the relationships between sequence and structure has been possible through dynamic programming-based alignment of topological features of protein family members (for reviews see Blundell and Johnson, 1993; Sali and Overington, 1994). These rules can then be applied to predicting sequences that will adopt a common fold, usually by constructing a tertiary template for each family fold (Sali et al., 1990; for a review see Bowie and Eisenberg, 1993). The key step in this application stage is alignment of the protein of known sequence but unknown structure with the appropriate template. Alignment allows extrapolation of the same set of rules derived from the known structures so that a 3-D model of the protein might be built and often more about its function learnt (for reviews see Bork et al., 1994; May et al., 1994; for a general review of protein structure prediction see Eisenhaber et al., 1995).

One referee stated that the present method is too similar to that of Taylor and Orengo (1989). In fact, this is not the case. The latter method does not use rigid-body superposition. They use double dynamic programming to compare intramolecular Cα vectors. The approach used here employs dynamic programming to compare intermolecular Cα distances following superposition. The methods described in May and Johnson (1994, 1995) are also superposition based. However, the current method is more flexible (it can accommodate rigid-body shifts not only in corresponding secondary structures but also at the termini of an alignment) and much faster, making it suitable for data bank searching and/or fold classification. Further work lies in the application of the method to this end.

Of course, it is well known that an optimum must be found between the number of residues equivalenced and r.m.s.d. What is perhaps not so obvious is the effect of the cut-off distance on this relationship for dynamic programming-based rigid-body superposition methods: an increase in the cut-off distance does not always lead to an increase in the number of equivalences and/or a higher r.m.s.d.

Despite the limitations of r.m.s.d. as a measure of structural similarity, the results of structure comparison methods, whether or not derived by superposition, are often evaluated in terms of this measure. I suggest a method to assess the quality of a given superposition not derived using a distance-based definition of topological equivalence.

Since submission of the first version of this paper, I have become aware of the procedure of Diederichs (1995) that also identifies the common fold between the globins and colicin A by direct superposition.

Acknowledgements
I thank Tom Blundell, Mark Johnson and N. Srinivasan for helpful discussions. I acknowledge financial support from the UK Biotechnology and Biological Sciences Research Council and Zeneca Agrochemicals.

References
Defining structural similarity


Received December 22, 1995; revised July 15, 1996; accepted July 25, 1996