Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions

E. Krissinel* and K. Henrick

European Bioinformatics Institute, Genome Campus, Hinxton, Cambridge CB10 1SD, England

Correspondence e-mail: keb@ebi.ac.uk

Received 8 January 2004
Accepted 19 October 2004

The present paper describes the *SSM* algorithm of protein structure comparison in three dimensions, which includes an original procedure of matching graphs built on the protein's secondary-structure elements, followed by an iterative three-dimensional alignment of protein backbone C_α atoms. The *SSM* results are compared with those obtained from other protein comparison servers, and the advantages and disadvantages of different scores that are used for structure recognition are discussed. A new score, balancing the r.m.s.d. and alignment length $N_{\rm align}$, is proposed. It is found that different servers agree reasonably well on the new score, while showing considerable differences in r.m.s.d. and $N_{\rm align}$.

1. Introduction

Protein function is in significant part determined by spatial structure. It is commonly believed that the three-dimensional fold has a major impact on the ability of a protein to bind other proteins or ligands (drugs), stability and purely mechanical aspects of protein behaviour. The similarity analysis of protein structure is therefore a vital step in understanding that protein's role in the machinery of life. Comparison of protein structures is also essential for estimating the evolutionary distances between proteins and protein families.

Currently, there are more than 28 000 structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000), 10–20 structures being deposited on a daily basis. Analysis of the ever-growing number of known structures requires efficient tools for protein structure alignment in three dimensions. In contrast to commonly known sequence alignment (Smith & Waterman, 1981), three-dimensional alignment is based on the comparison of geometrical positions, rather than biochemical properties, of amino-acid residues. Two residues are considered as aligned if they satisfy certain distance and orientation criteria at the best mutual superposition of the structures. While high sequence similarity almost always implies structural similarity, the opposite is not true. It is therefore expected that three-dimensional alignment will provide more significant clues to protein function and properties than sequence alignment alone.

Several approaches to protein structure alignment have been explored over the past decade. The techniques used include comparison of distance matrices (*DALI*; Holm & Sander, 1993), analysis of differences in vector distance plots (Orengo & Taylor, 1996), minimization of the soap-bubble surface area between two protein backbones (Falicov & Cohen, 1996), dynamic programming on pairwise distances

between the proteins' residues (Subbiah *et al.*, 1993; Gerstein & Levitt, 1996, 1998) and secondary-structure elements (SSEs) (Singh & Brutlag, 1997), three-dimensional clustering (Vriend & Sander, 1991; Mizuguchi & Go, 1995), graph theory (Mitchell *et al.*, 1990; Alexandrov, 1996; Grindley *et al.*, 1993), combinatorial extension of alignment path (*CE*; Shindyalov & Bourne, 1998), vector alignment of SSEs (*VAST*; Gibrat *et al.*, 1996), depth-first recursive search on SSE (*DEJAVU*; Kleywegt & Jones, 1997) and many others (Zuker & Somorjai, 1989; Taylor & Orengo, 1989; Godzik & Skolnick, 1994; Russell & Barton, 1992; Sali & Blundell, 1990; Barakat & Dean, 1991; Leluk *et al.*, 2003; Jung & Lee, 2000; Kato & Takahashi, 2001).

None of the existing methods gives an exact solution to the problem. Typically, all the methods agree relatively well on the alignment of highly similar structures but often disagree over details if the similarity is low. Partially, this discrepancy is due to the absence of a commonly accepted measure for structural similarity. Most similarity measures are based on the evaluation of the size of common substructures, for example the length of alignment (the longer, the better), and a measure of the distance between them, such as r.m.s.d. (the lower, the better). However, except for the case of highly similar structures, it is always possible to enlarge the common substructures at the expense of the distance measure between them. Therefore, algorithms of three-dimensional alignment typically involve a number of heuristic elements and empirical parameters, which naturally causes differences in results. The effect of employed heuristics or choice of empirical parameters are rarely, if ever, published as systematic studies. There are few data on the comparison of three-dimensional alignments produced by different algorithms.

Structural alignment of proteins is known to be a computationally expensive procedure. Alignment of a new structure of a few hundred residues to the whole of the PDB with publicly available web servers (*DALI*, *VAST*, *CE*, *DEJAVU* and some others) may take several hours, with response time growing sharply as the size of the query structure increases. Our aim was to provide the community with an interactive web server, which would be capable of delivering protein structure alignments and database searches in less than a minute, with high quality of alignments. The goal was achieved with the help of an advanced graph-matching algorithm, recently developed for serving structural queries in the EBI-MSD database (Krissinel & Henrick, 2004). The new tool, *SSM*, has been available for public use from June 2002 at http://www.ebi.ac.uk/msd-srv/ssm. The efficiency of the structure-alignment algorithm, described below in this paper, was found sufficient for serving all queries in real time, and therefore, in contrast to most of the other similar servers, *SSM* does not maintain a database of pre-aligned structures. Furthermore in this paper, we compare *SSM* with some publicly available web servers in order to examine the quality of alignments and to demonstrate the range of difference between the servers. Finally, we discuss the problem of measuring the quality of three-dimensional alignments for more reliable identification of potentially significant matches.

2. Graph-theoretical approach to matching protein structures

Problems of structure comparison and recognition are conveniently addressed by the graph-theoretical approach (see *e.g.* Rouvray *et al.*, 1979, and references therein). The approach typically includes three major steps: (*a*) graph representation of the objects in question, (*b*) matching the graphs representing the objects and (*c*) evaluating the common subgraphs found in order to form conclusions about similarity.

Traditionally, three-dimensional graphs of chemical structures connect all atoms with distance-labelled edges (see *e.g.* Gardiner *et al.*, 2000) and should have special labels for atoms representing chiral centres in order to distinguish between mirror-reflected structures. The graphs are then matched with a tolerance to the difference in edge lengths using one of many algorithms available [see the review by Raymond & Willett (2002)]. This approach, however, is not applicable to protein structures because of the prohibitively high cost of graph matching. One of the most frequently used optimal graph-matching algorithms, based on maximal clique detection (Levi, 1972; Bron & Kerbosch, 1973), has time complexity of the order of $O[(mn)^n]$, where n and m ($n \leq m$) denote the size of input graphs. This limits the application of this algorithm to graphs having 20–30 unlabelled vertices. Non-optimal algorithms give an approximation to the optimal (exact) solution at a lower cost; however, the quality of approximation is not well controllable. The fastest optimal algorithm, based on the decision-tree approach, has been reported by Shearer *et al.* (2001). This algorithm shows time complexity of only $O(2^n n^3)$. However, the algorithm is not applicable to the matching of three-dimensional graphs because of its space complexity, depending exponentially on the number of vertex and/or edge labels involved (edge labels of three-dimensional graphs are derived from the edge length and thus form a nearly continuous label space). Recently we have described a new optimal bactracking algorithm for common subgraph isomorphism (Krissinel & Henrick, 2004), *CSIA*, which represents an advancement of the widely known algorithm of Ullman (1976) for exact subgraph isomorphism. The time complexity of *CSIA* is bounded by $O\left(m^{n+1}n\right)$, which makes it applicable to graphs having up to $n, m \simeq 70$ unlabelled vertices. It follows from the above that even in the case of a simplified representation of proteins by their backbone C_α atoms, the traditional approach can be applied only to the shortest protein chains.

Size limitations of the graph-theoretical approach may be overcome if less elementary objects are used as graph vertices (Bessonov, 1985; Raymond *et al.*, 2002). Thus, protein secondary-structure elements represent a natural and convenient set of objects for building three-dimensional graphs, partly because secondary structure provides most of the protein functionality and is often conserved through the evolution of the molecule. The idea of using SSEs as elementary motifs for the identification of protein folds was exploited in many studies (Mitchell *et al.*, 1990; Grindley *et al.*, 1993; Gibrat *et al.*, 1996; Singh & Brutlag, 1997; Kleywegt &

Jones, 1997). The largest proteins contain up to 100 SSEs per chain, which form a very big graph for optimal graph-matching algorithms. However, as we shall see, using SSEs as graph vertices results in a variety of vertex and edge labels, which considerably speeds up most graph-matching algorithms, including *CSIA*. We calculated SSEs with the help of the algorithm *PROMOTIF* (Hutchinson & Thornton, 1996).

Most details of protein fold may be expressed in terms of just two types of SSEs, namely helices and strands. SSEs may be used as graph vertices, v_i, with composite labels $\{T_i, L_i\}$, where T_i denotes the type of vertex (helix or strand, and if it is a helix then what type of helix) and L_i specifies the number of residues in the ith SSE. Any two vertices, v_i and v_j, in the graph are connected by an edge e_{ij}. Edge labels are composed so as to describe the geometry of mutual position and orientation of the connected SSEs, as shown in Fig. 1.

The SSEs are represented by the vectors $\mathbf{r}_{\rm SSE} = \mathbf{r}_{\rm b} - \mathbf{r}_{\rm e}$ where

$$\mathbf{r}_{\rm b} = \big(0.74\mathbf{r}_p + \mathbf{r}_{p+1} + \mathbf{r}_{p+2} + 0.74\mathbf{r}_{p+3}\big)/3.48,$$
$$\mathbf{r}_{\rm e} = \big(0.74\mathbf{r}_{q-3} + \mathbf{r}_{q-2} + \mathbf{r}_{q-1} + 0.74\mathbf{r}_q\big)/3.48, \qquad (1)$$

for helices and

$$\mathbf{r}_{\rm b} = \big(\mathbf{r}_p + \mathbf{r}_{p+1}\big)/2,$$
$$\mathbf{r}_{\rm e} = \big(\mathbf{r}_{q-1} + \mathbf{r}_q\big)/2, \qquad (2)$$

for strands (Singh & Brutlag, 1997). In (1) and (2), indices p and q denote the serial numbers of the first and last residues in the SSE, and we neglect strands shorter than three residues and helices shorter than five residues. Each edge, e_{ij}, is then labelled with a property vector $\{\rho_{ij}, \alpha_1^{ij}, \alpha_2^{ij}, \alpha_3^{ij}, \alpha_4^{ij}, N_i, N_j, C_i, C_j\}$, where ρ_{ij} is the edge length (in Å), $\alpha_{1/2}^{ij}$ is the angle between the edge and vertices v_i/v_j, and $\alpha_{3/4}^{ij}$ is the torsion angle between v_i and v_j. $N_{i/j}$ is the serial number of v_i/v_j in their protein chains (as counted from N to C termini), $C_{i/j}$ is the vertex chain identifier. Both $N_{i/j}$ and $C_{i/j}$ are used for controlling the SSE connectivity along the chains (see below).

The set of vertices, edges and their labels gives a full definition of a graph. In order to compare (match) the graphs, a graph-matching algorithm should also be provided with a set of rules for the comparison of individual vertices and edges. Obviously, these rules may be formulated in a number of different ways, each of which would involve a number of empirical parameters to be chosen in the course of multiple trials. In our definition, vertices v_i and v_j compare if

$$T_i = T_j \quad \text{and} \quad \big|L_i - L_j\big| < \varepsilon_L(L_i + L_j)/2 + \sigma_L. \qquad (3)$$

Two edges, e_{ij} and e_{kl}, are considered as comparable if all of the following hold true:

$$\big|\rho_{ij} - \rho_{kl}\big| < \varepsilon_\rho(\rho_{ij} + \rho_{kl})/2 + \sigma_\rho, \qquad (4)$$

$$\big|\alpha_{1,2}^{ij} - \alpha_{1,2}^{kl}\big| < \delta_1 \quad \text{and} \quad \big|\alpha_3^{ij} - \alpha_3^{kl}\big| < \delta_2, \qquad (5)$$

$$\operatorname{sign}\big(\alpha_4^{ij}\big) = \operatorname{sign}\big(\alpha_4^{kl}\big) \quad \text{at} \quad \big|\alpha_s^{ij,kl} + n\pi\big| > \delta_3,\ s = 1, 2, 4, \qquad (6)$$

$$\text{Connect}(e_{ij}, e_{kl}) \text{ (see below) returns true.} \qquad (7)$$

The tolerances ε_L, σ_L, ε_ρ, σ_ρ, δ_1, δ_2 and δ_3, used in (3)–(7), are chosen empirically. In our implementation, they are tabulated for five levels of match specificity, as listed in Table 1. Parameters for the 'Normal' level were defined so as to maximize the number of correct fold identifications at cross-matching the SCOP (Murzin *et al.*, 1995) domains. Levels of higher and lower specificity were chosen more arbitrarily, in order to provide a facility for widening or narrowing the search when necessary.

Table 1
Empirical parameters for the comparison of vertices and edges [equations (3)–(7)].

Specificity	ε_L	σ_L	ε_ρ	σ_ρ (Å)	δ_1 (°)	δ_2 (°)	δ_3 (°)
Highest	0.125	1	0.10	0.5	15	12	12
High	0.150	2	0.15	1.0	20	15	15
Normal	0.200	4	0.20	1.5	30	22	20
Low	0.300	4	0.30	2.0	36	30	30
Lowest	0.350	6	0.50	2.5	45	36	36

As seen from expressions (3) and (4), we use both relative and absolute differences in vertex and edge lengths for the comparison (thus allowing for larger absolute differences for longer vertices and edges), while only absolute differences in angles [*cf.* equation (5)] are analysed.

Comparison of torsion angles $\alpha_4^{i,j}$ allows one to distinguish between mirror-symmetry mates. Apparently, this differentiation was not made by Singh & Brutlag (1997) and Mitchell *et al.* (1990), although it was taken into account in other studies (Mizuguchi & Go, 1995; Grindley *et al.*, 1993). It should be realized, however, that if edge–vertex angles $\alpha_{1,2}^{i,j}$ are small even a slight difference between them in the compared structures may cause a significant disagreement in torsion angles. We therefore compare only signs of torsion angles, which is sufficient for distinguishing between the symmetry mates, and only if both vertex vectors and the edge are far from collinearity [*cf.* equation (6)].

Until now we considered three-dimensional arrangements of SSEs regardless of their ordering along the protein chain. Usually the connectivity of SSEs is significant; however, there

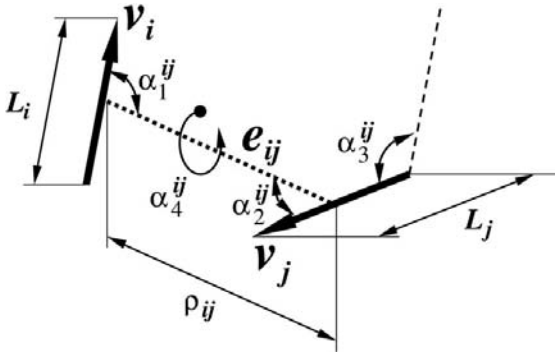

Figure 1
Properties of vertices and edges of the SSE graph. Vertices v_i and v_j are represented by vectors $\mathbf{r}_{\rm SSE}$ [*cf.* equations (1) and (2)]; edge e_{ij} connects their centres. Edge length ρ_{ij} and angles α_k^{ij}, $k = 1, \ldots, 4$, define mutual positions and orientations of all vertices in the graph. See text for more details.

are situations where it may or should be neglected (*e.g.* comparison of mutated or engineered proteins, or geometry of active sites). In previous studies, the SSE connectivity was either preserved (Singh & Brutlag, 1997) or, apparently, neglected (Mitchell *et al.*, 1990; Grindley *et al.*, 1993; Mizuguchi & Go, 1995). In order to handle the connectivity in a more flexible way, we have introduced a special function, Connect(e_{ij}, e_{kl}) [*cf.* equation (7)], providing for the following three options:

(i) Connectivity of SSEs is neglected. Connect(e_{ij}, e_{kl}) always returns true. Motifs *A* and *B*, shown in Fig. 2, would then match fully as $\{H, S_1, S_2, S_3|H, S_1, S_2, S_3\}$.

(ii) 'Soft' connectivity. The general order of matched SSEs along their protein chains is the same in both structures, but any number of missing or unmatched SSEs between the matched ones is allowed. In this case, Connect(e_{ij}, e_{kl}) returns false if

$$C_i = C_j, \quad C_k = C_l \quad \text{and} \quad \text{sign}\big(N_i - N_j\big) \neq \text{sign}(N_k - N_l).$$

Matching motifs *A* and *B* from Fig. 2 then yields five maximal common sub-motifs of size 2: $\{H, S_2|H, S_2\}$, $\{H, S_3|H, S_3\}$, $\{S_1, S_2|S_1, S_2\}$, $\{S_1, S_3|S_1, S_3\}$ and $\{S_2, S_3|S_3, S_2\}$.

(iii) 'Strict' connectivity. Matched SSEs follow the same order along their protein chains and may be separated only by an equal number of matched or unmatched SSEs in both structures. Connect(e_{ij}, e_{kl}) returns false if

$$C_i = C_j, \quad C_k = C_l \quad \text{and} \quad \big(N_i - N_j\big) \neq (N_k - N_l).$$

Matching motifs *A* and *B* from Fig. 2 then yields the only maximal common sub-motif of size 2: $\{S_2, S_3|S_3, S_2\}$.

Matching three-dimensional graphs built on secondary-structure elements gives a correspondence between *groups* of residues of the compared proteins, which allows for preliminary identification of protein folds and rough estimation of structural similarity. Fine comparative analysis requires information on the correspondence between *individual* residues, including those not found in SSEs. In order to obtain three-dimensional alignment of individual residues, we represent them by their C_α atoms and apply an additional procedure of aligning the latter in three dimensions, using the results of graph matching as a starting point. The alignment procedure is described in the next section.

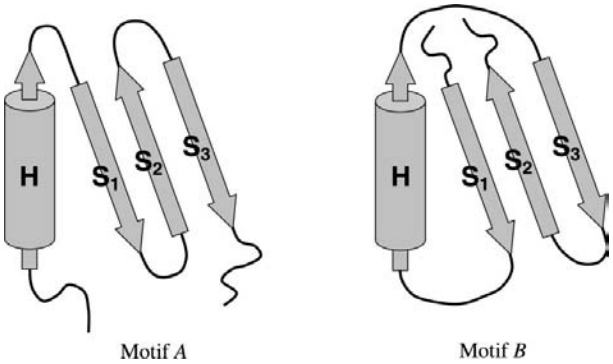

Figure 2
Example of SSE motifs (helix *H* and three strands, S_1, S_2 and S_3), each having different SSE connectivity. Motifs *A* and *B* form three-dimensional SSE graphs that are geometrically identical; however, the difference in connectivity may be also expressed in graphical terms (see text for details).

3. C_α alignment in three dimensions

Alignment problems are traditionally approached by the technique of dynamic programming (Smith & Waterman, 1981), which may also be applied to structure alignment (Subbiah *et al.*, 1993; Gerstein & Levitt, 1996 1998; Singh & Brutlag, 1997). This technique, however, is not applicable if SSE connectivity is neglected and best alignment is achieved at misconnected SSEs (*cf.* the discussion above). We therefore employ a different procedure, which optimizes a quality function calculated at best superposition of aligned structures. The procedure is generally similar to those used in other studies (see *e.g.* Singh & Brutlag, 1997; Kleywegt & Jones, 1997); however, it involves a number of empirical elements, which are introduced and adjusted in the course of analysis of thousands of alignments. We therefore describe our algorithm 'as is', without discussion of its differences from similar techniques and exhaustive justification.

Our procedure is based on fast optimal superposition (FOS) of two sets of points in three-dimensional space (in our case, the positions of the C_α atoms of the two structures to be aligned), provided that correspondence between them is known. Several FOS techniques are available (McLachlan, 1972; Kabsch, 1976, 1978). We used a singular value decomposition of the correlation matrix, following the method described by Lesk (1986). The rotoinversion, if detected, is eliminated by changing the sign of the singular vector corresponding to the minimal singular value. The procedure is described in Appendix *A*.

Once the structures are superposed, their C_α atoms may be mapped onto each other using the procedure described below. The initial superposition of the structures is obtained by applying FOS to the representing vectors of matched SSEs [equations (1) and (2)]. This approach, however, does not work well if the SSE vectors $\mathbf{r}_{\rm SSE}$, forming a common SSE subgraph, are collinear. Applied to a set of collinear vectors, FOS yields a rotation matrix with arbitrary rotation about the vectors. We therefore add the edge-representing vectors to the sets of matched $\mathbf{r}_{\rm SSE}$ if the minimal absolute value of the cosine between any two of them exceeds 0.8. If addition of edge-representing vectors does not decrease the minimal cosine or if, in the case of structures with low similarity, the maximal common SSE subgraph includes only one SSE, we explore the whole rotation about the ambiguous axis in order to achieve a maximal overlap of other (type- and direction-compatible) SSEs or individual C_α atoms.

The mutual correspondence (mapping) between C_α atoms of superposed structures is found through analysis of the distances between them. While the simplest contact-seeking approach would suffice in many cases, we found that for the best quality of alignment one should employ a more sophis-

ticated technique. We suggest that the following steps are performed in order of their numbering.

(i) Mapping C_α atoms of matched SSEs. For each pair of matched SSEs, we find n_a ($n_a = 3$ for strands and $n_a = 4$ for helices) neighbouring pairs of C_α atoms with minimal separation, mark them as mapped and then expand the mapping to the ends of the SSEs, leaving no unmapped atoms between the mapped ones (see Fig. 3). The value of $n_a = 4$ for helices ensures that full helix turns are always mapped properly, even if there is only a partial overlap between the helices.

(ii) Mapping C_α atoms of non-matched SSEs. All pairs of non-matched SSEs. v_i^1 and v_j^2, which are of the same type and collinear with cosine greater than 0.7, are ranged in order of increasing r.m.s.d. of their closest n_a C_α atoms (dark atoms in Fig. 3), and only pairs with the lowest r.m.s.d. ($< R_c$) are left in the list. If the r.m.s.d. of the two pairs (v_i^1, v_j^2) and (v_i^1, v_k^2) is less than R_c, only one pair with the lowest r.m.s.d. is left (the superscripts stand for structure ID). Then the C_α atoms of all SSE pairs in the list are mapped as described above, starting from the pair with the lowest r.m.s.d. Before mapping an SSE pair, it is necessary to check that the mapping will not violate the connectivity of already mapped atoms (as explained in Fig. 4), if connectivity should be preserved (*cf.* §2). The preliminary ranging of SSE pairs on increasing r.m.s.d. ensures that only the best-overlapping SSEs will be mapped in the case of a connectivity conflict.

(iii) Expansion of contacts. If atom A of structure 1 and atom B of structure 2 form a contact, the distance between A and B is less than the distance between A and any atom of chain 2, except B, *and* less then distance between B and any atom of chain 1, except A. Finding contacts is an expensive procedure, unless a bricking algorithm is employed [see for example the program *CONTACT* by Tadeusz Skarzynski in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994)]. Contacts are calculated for all yet unmapped but mappable pairs of atoms and are ranged by increasing contact distance, and only contacts with contact distances shorter than R_c are left in the list. We consider a pair of atoms as unmappable if one atom belongs to a helix (unless closer than three residues to the helix ends) and another one belongs to a non-helical part of the protein chain. Starting with the shortest contact, contacting C_α atoms are mapped onto each other, provided that such mappings do not violate the chain connectivity (*cf.* Fig. 4). After consideration of all contacts, the procedure tries to map all remained mappable pairs of atoms, starting from pairs that adjoin the contacts, as shown in Fig. 5.

(iv) Quality filter. The previous steps result in the mapping of up to $\min(N_1, N_2)$ C_α atoms, where N_1 and N_2 are the number of residues in the aligned structures. In general, such mapping includes both similar and less similar substructures. Quite often, the quality of alignment may be improved by unmapping C_α atoms of less similar parts. Usually this is achieved by introducing a cut-off distance of about 2–4 Å. Such an approach, however, does not work well in many instances where one structure is a distorted (by a few Å) replica of another, and therefore the r.m.s.d. is not a good measure of the alignment quality. An intuitive understanding of structural similarity suggests contradictory requirements of achieving a lower r.m.s.d. and a higher number of mapped (aligned) residues $N_{\rm align}$. This contradiction may be eliminated, in the first approximation, by a score that represents a ratio of $N_{\rm align}$ and the r.m.s.d. We therefore suggest the function

$$Q = N_{\rm align}^2 / \left\{ \left[1 + ({\rm RMSD}/R_0)^2 \right] N_1 N_2 \right\} \qquad (8)$$

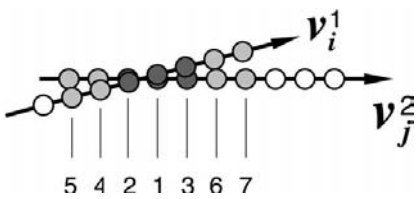

Figure 3
Mapping C_α atoms of superposed SSEs v_i^1 and v_j^2 (superscripts 1 and 2 stand for structure ID). Atoms are put into correspondence in order of their numbering in the figure. First, three dark atoms of v_i are mapped onto three dark atoms of v_j as having the least interatomic distances. Next, grey atoms are mapped in the direction from the dark atoms toward the SSE ends. White atoms remain unmapped.

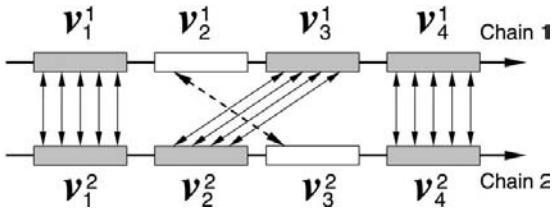

Figure 4
Schematic diagram of a connectivity conflict in the course of three-dimensional alignment. If SSE pairs (v_1^1, v_1^2) and (v_4^1, v_4^2) are matched by the three-dimensional graph-matching procedure then they are properly connected. If superposition of the structures shows that the C_α atoms of SSE pairs (v_2^1, v_3^2) and (v_3^1, v_2^2) may also be mapped, the algorithm first tries to map atoms of a pair with minimal RMSD of n_a closest atoms (*cf.* text), (v_3^1, v_2^2) in the figure. These atoms may be mapped without a conflict. However, then atoms of SSE pair (v_2^1, v_3^2) cannot be mapped without breaking the connectivity (dashed line in the figure), and therefore these atoms remain unmapped.

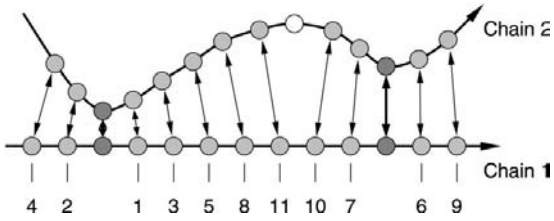

Figure 5
Expansion of C_α contacts. Found contacts (dark atoms, see text for details) are expanded in both directions gradually, starting from the shortest contact, such that the distance between newly mapped atoms (shown in grey) undergoes the minimal possible increase. For the example in the figure, pairs of C_α atoms are mapped in order of their numbering. If the procedure encounters unmappable pair of atoms (as defined in the text) it stops advancing in that direction from the contact. The procedure ensures that unmapped atoms (shown in white) are always the most distant ones in the region between two contacts.

as a measure of *quality* of alignment. In (8), R_0 is an empirical parameter (chosen at 3 Å) that measures the relative significance of RMSD and $N_{\rm align}$. Computer experiments have shown that a square dependence on $N_{\rm align}/{\rm RMSD}$ is as good as a cubic or linear one, and the second power was finally chosen only for technical convenience.

As seen from (8), Q reaches 1 only for identical structures ($N_{\rm align} = N_1 = N_2$ and ${\rm RMSD} = 0$), and decreases to zero with decreasing similarity (increasing RMSD or/and decreasing $N_{\rm align}$). Therefore, the higher Q, the 'better', in general, the alignment. Despite the fact that the Q score represents a very basic measure that does not take into account many factors related to the quality of alignment (the number of gaps and their size, sequence identity *etc.*), we found that maximization of the Q score produces good results.

In order to maximize the Q score of alignment, we first range all aligned pairs of C_α atoms by increasing interatomic distances: $R(i-1) \leq R(i)$, $i = 2, , N_{\rm align}$. Unmapping of the most separated C_α pair decreases both the alignment length $N_{\rm align}$ and RMSD. As may be found from the analysis of (8), such unmapping results in increasing Q at superlinear dependence $R(i)$ and in decreasing Q at sublinear $R(i)$. We therefore unmap C_α pairs one by one in order of decreasing interatomic distances until Q reaches a maximum (Q may change non-monotonically). Owing to empirical considerations, we do not unmap inner atoms of mapped SSEs without first unmapping their outmost atoms, and, in matched SSEs, we never unmap $n_{\rm a}$ atom pairs with minimal separation (dark atoms in Fig. 3).

(v) Unmapping short fragments. Pairs of C_α atoms, which form short (1 or 2 pairs) closures between gaps, most often correspond to purely incidental intersections of protein chains. However, they may effectively lock the structures in a particular orientation and thus prevent further optimization. We therefore unmap such pairs even if doing so decreases Q.

The mapping obtained may be used for the calculation of best structure superposition by applying FOS to the pairs of mapped C_α atoms. Since a change in orientation may affect the mapping, the cycle mapping FOS is repeated until the Q score of alignment ceases to increase over a sufficiently large number of successive iterations (ten by our choice). The contact distance R_c, used for mapping atoms of non-matched SSEs and in looking for contacts (*cf.* above) was found to be a very important parameter, which significantly affects the quality of results. In our implementation, R_c increases linearly from 3 to 5 Å during the first ten iterations.

The presented algorithm of C_α alignment converges to a local maximum of score function (8). Therefore, the results are highly dependent on the quality of the initial guess, which is provided by the identification of common subsets of SSE through the three-dimensional graph-matching procedure. In the course of analysis of many individual matches, we have found that a larger common subgraph is not an absolute indication of a better-quality match. Therefore, for each pair of structures, *SSM* performs C_α alignment starting from all common subgraphs that are larger than $N_{\rm SSE}^{\rm max} - 3$, where $N_{\rm SSE}^{\rm max}$ is the size of maximal common subgraph, and the alignment with the highest Q is accepted as a result. In our comparative study, presented below, we found that the overall procedure works very well if the structures show a reasonable degree of similarity. If structural similarity is very low (such that only one or two common SSEs may be identified), the procedure may result in a less accurate solution. In such cases, however, many imperfect alignments are usually possible, and choosing the best one is never self-evident.

4. Scoring the results

The score function Q [equation (8)] was found to be a good geometrical measure of structural similarity. As mentioned above, this function offers a compromise between contradicting requirements of achieving a lower r.m.s.d. and a higher number of aligned residues and, therefore, Q is expected to be a more objective indicator of quality of alignment than RMSD and $N_{\rm align}$ alone.

However, higher structural similarity does not necessarily imply higher significance of alignment. For example, a helix may be perfectly aligned with most of the PDB entries, but the significance of such alignments is very low because they are likely to be obtained simply by chance, by choosing the structures randomly from the database.

Our estimation of statistical significance is based on the same ideas as those employed by *VAST* (Gibrat *et al.*, 1996). The probability that matching two structures A and B is scored at value S or higher merely by chance may be estimated as the P value:

$$P_v(S) = 1 - \prod_k \left[1 - P_k(S)\right]^{M_k}. \qquad (9)$$

In this expression, $P_k(S)$ is the probability of achieving the score S in the event when matching two structures, picked randomly from the database, yields a common substructure containing k SSEs. M_k stands for the redundancy number, showing how many common substructures of size k may be formed from proteins A and B. We define score S as a sum of quality scores Q [*cf.* equation (8)] for the matched SSEs:

$$S = \sum_i Q_i = \sum_i N_i^2 \Big/ \left\{\left[1 + ({\rm RMSD}_i/R_0)^2\right] N_A^{(i)} N_B^{(i)}\right\}, \qquad (10)$$

where index i numbers the matched SSE pairs, $N_{A/B}^{(i)}$ is the number of residues in the ith matched SSE of protein A/B, N_i is the number of aligned residues in the ith SSE pair, and ${\rm RMSD}_i$ is the r.m.s.d. of the ith pair. Thus, for common substructures of size k, score S may vary from 0 (poorest alignment) to k (ideal alignment). Definition (10) allows one to calculate $P_k(S)$ as

$$P_k(S) = \int_S^k \rho_k(y)\,{\rm d}y = \int_S^k {\rm d}y \int_0^1 \rho_1(x)\rho_{k-1}(y-x)\,{\rm d}x, \qquad (11)$$

under a reasonable assumption that scores Q_i do not correlate. In (11), $\rho_k(x)$ is the density of the probability of finding a common substructure containing k SSEs with score x by randomly choosing the structures from the database. The functions $\rho_k(x)$ may be calculated for any k through their

recurrent relation given by (11). This recurrence starts from the function $\rho_1(x)$, which is calculated empirically by running *SSM* on all pairs of non-redundant protein structures [we used SCOP folds as found in SCOP Version 1.61 (Murzin *et al.*, 1995)].

The value of $P_k(S)$, corresponding to the actual alignment, may also be of interest. *SSM* reports it as a Z score, defined by the following equation,

$$P_k(S) = (2/\pi)\int_Z^\infty \exp(-t^2/2)\,\mathrm{d}t \qquad (12)$$

5. Implementation

The procedure described above for protein alignment in three dimensions has been implemented as a standalone application and as a web server, available for public use at http://www.ebi.ac.uk/msd-srv/ssm. The standalone application and the web server have identical functionality. The development is based on the new *CCP*4 Coordinate Library (Krissinel *et al.*, 2004) and runs on all Unix platforms. *SSM* allows for a number of different tasks, including three-dimensional alignment of protein pairs (uploaded or given as PDB/SCOP ID codes), alignment of a structure to all entries of the PDB/SCOP archives or any subset of SCOP, or alignment to an uploaded set of structures. For faster processing, *SSM* precompiles SSE graphs of all PDB and SCOP entries in fast-access files, which are updated automatically on a weekly basis. Serial alignments (a structure to a set of structures) are automatically scheduled on a number of CPUs depending on the anticipated task complexity. Unlike many other protein comparison services, *SSM* does not keep a database of precalculated three-dimensional alignments, because its performance was found to be sufficient for serving queries in real time.

A distinguishable feature of *SSM* is that its performance depends rather sharply on the minimal desired level of structural similarity set in advance. This feature follows from the properties of the original graph-matching algorithm that we employed (Krissinel & Henrick, 2004). The level of structural similarity is measured by the percentage of to-be-matched SSEs $p_{\rm SSE}$. The higher $p_{\rm SSE}$, the quicker is *SSM*. Typically, alignment of a few-hundred-residue protein to all PDB entries at $p_{\rm SSE} = 50-70\%$ takes much less than a minute.

6. Results and discussion

Fig. 6 shows a comparison of *SSM* with other publicly available web servers that deliver three-dimensional alignment of protein structures, namely *VAST* (Gibrat *et al.*, 1996), combinatorial extension (*CE*) (Shindyalov & Bourne, 1998) and *DALI* (Holm & Sander, 1993). The comparison is presented for the example of protein chain 1sar:A (ribonuclease SA; Sevcik *et al.*, 1991). As seen from Fig. 6, all the servers reveal fairly distinctive subsets of highly similar structural neighbours (from the whole PDB) and structures with intermediate similarity. *VAST* also returns some dissimilar structures (as identified by the alignment length $N_{\rm align}$). The number of returned hits differs from server to server, which probably reflects the different criteria used for the identification of insignificant matches. In *SSM*, we have chosen not to dispose of any hits found with similarity level higher than the requested $p_{\rm SSE}$. Instead, *SSM* provides a facility to sort the results by a variety of scores [Q score (default), P value, Z score, RMSD, $N_{\rm align}$, number of matched SSEs, number of gaps *etc.*] and tools for navigation through the matches. This approach was motivated by the consideration that none of the scores provides an absolutely reliable measure of structural similarity or statistical significance, and therefore the final decision of accepting a match should be reserved for the user.

For the calculations presented in Fig. 6, we used $p_{\rm SSE} = 0$; therefore, *SSM* produced tens of thousands of matches, from which we present only those also returned by *VAST*, *CE* and *DALI* in Fig. 6. It appeared that *SSM* has found all the structural neighbours found by the other servers, with the exception of a few structures represented by C_α atoms only. Such structures do not allow for the determination of hydrogen-bond patterns and, consequently, for a reliable calculation of secondary structure. Many (from tens to hundreds) hits in the similarity range of Fig. 6, returned by *SSM*, are not found in other servers' outputs. Although it is difficult to explain this result without knowing all details of the algorithms and their implementations, a higher fraction of newer structures in *SSM* results suggests that the difference is partially due to outdated databases. Other reasons may include narrowing of the search by different similarity criteria and use of representative structures instead of actual screening of the whole PDB.

Fig. 6 suggests that *SSM* fully agrees with the other servers in the identification of highly similar, less similar and dissimilar structures. The alignment length shows clear stepwise 'transitions' between the subsets of structures with different structural similarity to 1sar:A. In this particular comparison, *SSM* alignments are longer than those given by *VAST*, somewhat shorter than those from *CE* and of approximately the same length as those of *DALI*. A thorough examination of plots for RMSD and $N_{\rm align}$ reveals that longer alignments always come at the expense of higher r.m.s. deviations, and therefore the observed differences between the servers should be mostly due to the different criteria employed to balance these characteristics. This conclusion is corroborated by the observation that all servers agree on the r.m.s.d. for highly similar structures, when all residues of 1sar:A are aligned to targets.

The balance of RMSD and $N_{\rm align}$ is indicated by the Q score [equation (8)]. As seen from the Q score plots in Fig. 6, all servers deliver three-dimensional alignments with very close values of Q. Although the Q scores differ in value (with the largest difference found between *SSM* and *VAST*, while Q scores from *SSM* and *DALI* nearly coincide), they show a very good correlation, which is seen in that the Q curves are much

smoother than the corresponding curves of RMSD and $N_{\rm align}$. Fig. 6 also suggests that the Q score allows for a clearer identification of highly similar and less similar structural neighbours than do RMSD and $N_{\rm align}$ alone. A particular example can be found in the *SSM*–*CE* comparison (column *b* in Fig. 6). Judging by the alignment lengths obtained from *CE*, PDB entries numbered 120–220 may be qualified as closer structural neighbours to 1sar:A than entries numbered 30–119, which contradicts the *SSM* findings. Analysis of the corresponding RMSD plot suggests that the above conclusions may be incorrect. If, however, the Q score is taken as a measure of structural similarity, *SSM* and *CE* agree on rating the entries numbered 120–220 below those numbered 30–119. Since the Q score takes both RMSD and $N_{\rm align}$ into account, the last conclusion seems to be more justified. Another example of this kind may be found in the *VAST* results for structural neighbours numbered 220–265, showing a decrease in both $N_{\rm align}$ and RMSD, which therefore may be interpreted as either a decrease or an increase in structural similarity. At the same time, $N_{\rm align}$ and RMSD obtained from *SSM* unambiguously suggest decreasing similarity for these structures. A clear fall of Q scores obtained from both servers implies that the structures are, indeed, ordered by the decrease of their similarity to 1sar:A.

We performed a comparative study, similar to that described above, for a number of structures belonging to different protein folds. The results shown in Fig. 6 were found to be of a common nature. More results are published on the *SSM* web site.

The results obtained suggest that different similarity scores perform equally well if structure similarity is high. The situation, however, changes drastically as the similarity decreases. For example, PDB entry 1kn0:A (human RAD52 protein; Kagawa *et al.*, 2002) does not have an exact match in SCOP Version 1.61. All potential matches to 1kn0:A from SCOP 1.61 domains represent relatively remote structural neighbours, and therefore *SSM* should be run with a low similarity threshold. Choosing $p_{\rm SSE} = 15\%$ results in a total of 33 588 hits returned, none of which represents a perfect match. Fig. 7 shows superpositions of 1kn0:A with best-matching SCOP

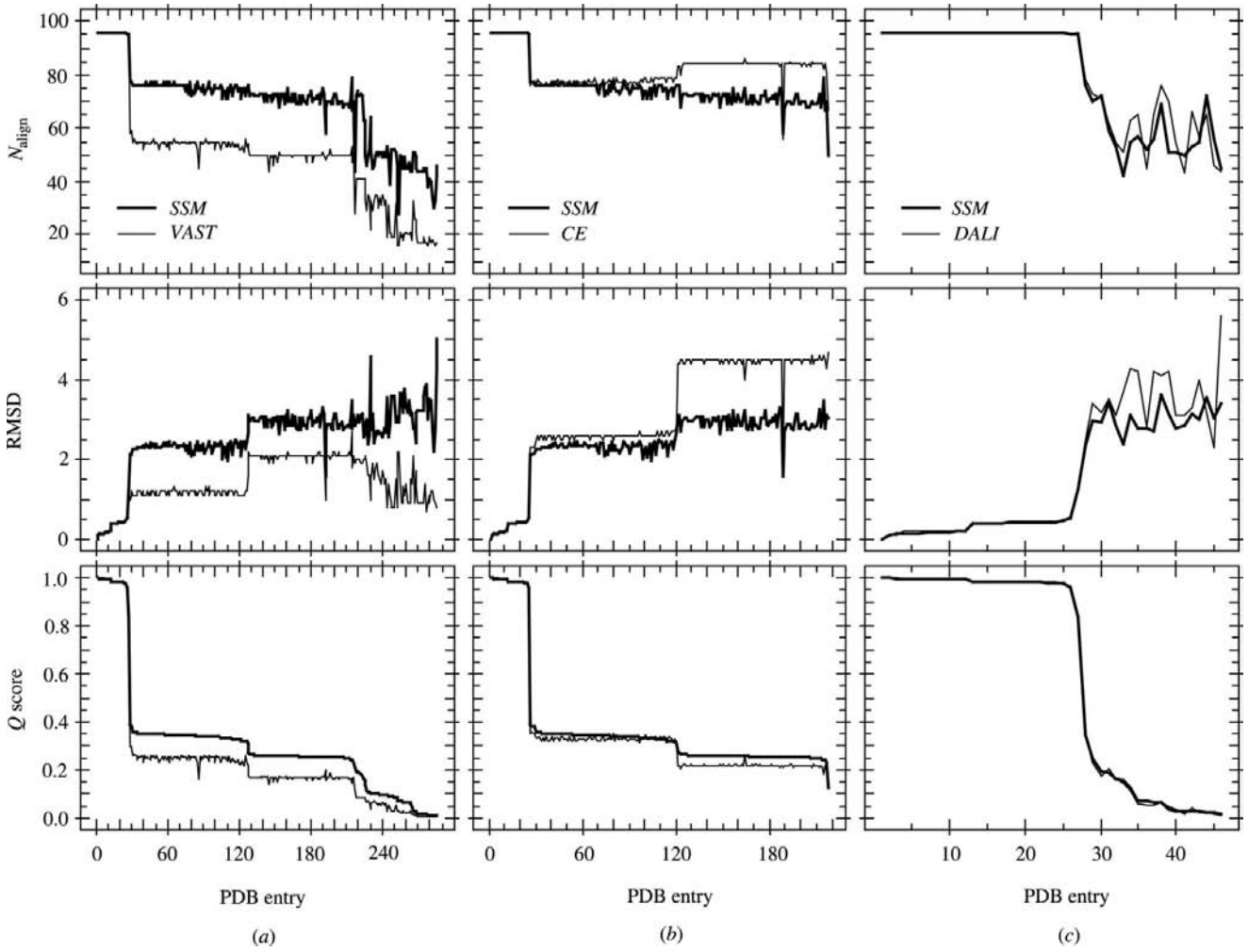

Figure 6
Comparison of *SSM* and *VAST* (*a*), combinatorial extension (*CE*) (*b*) and *DALI* (*c*). PDB chain 1sar:A (Sevcik *et al.*, 1991) was used as a query for screening the whole PDB. Results for all of the structural neighbours identified by *VAST*, *CE* and *DALI* were selected from *SSM*'s output and ordered by decreasing *SSM*'s Q score [equation (8)]. Thick lines: *SSM* results; thin lines: results obtained from *VAST*, *CE* and *DALI* as indicated in the figure.

Table 2
Scores of four matches to PDB entry 1kn0:A (184 residues) from SCOP 161, shown in Fig. 7, with best scores in bold (RMSD given in Å).

The last column shows the number and type of matched SSEs ('H' for helices, 'S' for strands). SI is the sequence identity [equation (7)], in %. See discussion in the text.

Domain	N_{res}	Q	RMSD	N_{align}	Z	SI	SSEs
*d1di2a_*	69	**0.213**	2.43	67	2.53	**16**	HS
d1emn_1	43	0.019	**0.90**	13	2.93	15	S
*d1elxb_*	449	0.020	5.82	**89**	0.01	7	**HHS**
*d1qmca_*	52	0.028	1.37	18	**5.09**	6	S

domains, as suggested by Q score, RMSD, N_{align} and Z score. The achieved scores are presented in Table 2.

As seen from Fig. 7, the highest Q score indicates a match (Fig. 7*a*; *d1di2a_*; double-stranded RNA binding protein A; Ryter & Schultz, 1998) that is (geometrically) best according to common intuition. Although the overlap is not perfect, the common substructures are compact and form most of the target structure. The matches with the lowest r.m.s.d. (Fig. 7*b*) and highest Z score (Fig. 7*d*) represent alignments that are too short to be rated high. The match with the maximal number of aligned residues (Fig. 7*c*) shows a poor superposition of common substructures with high r.m.s.d.; the alignment is fragmented and the overall overlap seems to be incidental.

The results show that using an appropriate score is crucial for the similarity search. An idea of what it would take to find *d1di2a_* as the best match to 1kn0:A without using the Q score may be obtained from Table 3. The table lists the ten best matches, all of a comparable quality, rated by different scores. As may be seen from Table 3, *d1di2a_* is 1575th by RMSD, 3079th by Z score, 818th by P value and 872nd by alignment length (since N_{align} is an integer number, the last figure is subject to the sorting procedure). Thus, *d1di2a_* does not appear on top of result lists sorted by any of the traditionally used similarity scores, and it would take many hours, if not days, to find this match manually from the results.

It is commonly assumed that protein chains with similar sequences tend to fold into similar three-dimensional structures. This assumption is often used for narrowing the similarity search or for the selection of representative structures. Although using the assumption makes the search faster, a known side effect is that the results may be biased toward sequence similarity. Because our alignment procedure is completely indifferent to chain composition, we used *SSM* for studying the relationship between sequence and structure similarity. Fig. 8 shows correlations between sequence identity (SI), Q score, RMSD and the normalized alignment length N_m:

$$N_m = N_{align}/\min(N_1, N_2). \quad (13)$$

The sequence identity is defined as a fraction of identical residues in the total number of (structurally) aligned residues:

$$SI = N_{ident}/N_{align}. \quad (14)$$

The score correlations are represented by contour maps of the reduced density of the probability, $\rho_r(x, y)$, of obtaining three-dimensional alignment with particular values of scores x and y:

Table 3
Ten SCOP domains closest to 1kn0:A, as suggested by their Q scores.

Other scores (P value, Z score, RMSD and N_{align}) rate the hits as shown by the figures in brackets (see discussion in the text).

Domain	Q	$-\log(P_v)$	Z	RMSD	N_{align}
*d1di2a_*	0.2130	0.29(818)	2.53(3079)	2.431(1575)	67(872)
*d1di2b_*	0.1810	0.12(1135)	2.30(4384)	2.487(1801)	58(3720)
d1stu__	0.1490	0.00(2262)	1.44(12028)	3.086(6359)	62(2000)
*d1ekza_*	0.1310	0.00(3028)	0.84(20850)	3.143(7012)	62(2003)
*d1fx3d_*	0.1040	0.00(3304)	1.79(8121)	3.642(13866)	80(71)
d1hnwe2	0.0924	1.70(75)	3.88(245)	1.569(91)	57(4380)
*d1ibke_*	0.0898	0.00(4169)	3.61(380)	2.228(936)	62(2167)
d1hr0e2	0.0897	1.42(110)	3.61(379)	2.229(937)	62(2168)
*d1ible_*	0.0887	0.04(1473)	3.80(281)	1.950(403)	59(3129)
*d1avza_*	0.0875	0.00(14052)	0.77(21987)	3.823(16901)	65(1304)

$$\rho_r(x, y) = \rho(x, y)\left[\int_0^{x_{max}} \rho(x, y)\,dx \int_0^{y_{max}} \rho(x, y)\,dy\right]^{-1/2}, \quad (15)$$

where probability density $\rho(x, y)$ is calculated in the course of all-to-all alignment of all chains found in the PDB.

As seen from Figs. 8(*a*)–8(*c*), 100% sequence identity does not necessarily mean a perfect three-dimensional alignment in terms of either Q score, RMSD or alignment length. Values of $0.93 \leq N_m < 1$ at SI = 1 (Fig. 8*b*) indicate pairs of chains with sequence-identical common subchains. Despite the absolute sequence identity, these chains show structure differences with an r.m.s.d. of up to 1 Å (*cf.* Fig. 8*c*). Most of these differences are caused by the interaction between residues of matched and unmatched parts of the chains, and therefore 1 Å of

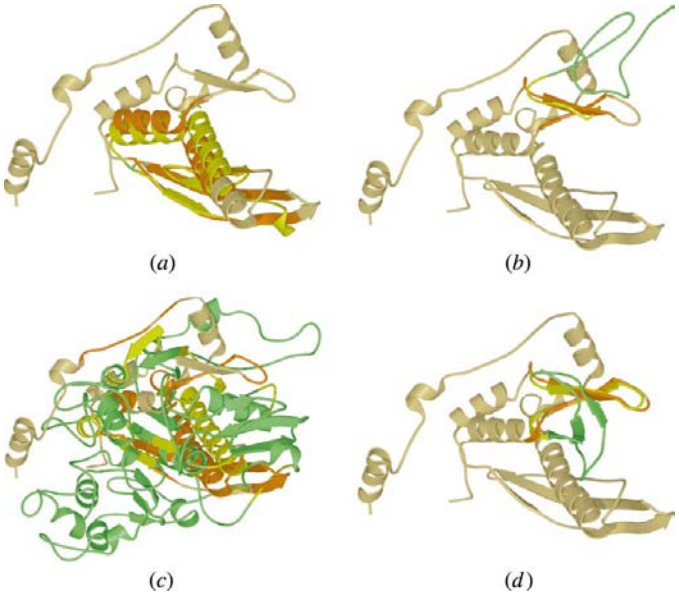

Figure 7
Superposition of PDB chain 1kn0:A (Kagawa *et al.*, 2002) with best-matching SCOP domains, as suggested by (*a*) Q score (*d1di2a_*) (*b*) RMSD (*d1emn_1*) (*c*) N_{align} (*d1elxb_*) (*d*) Z score (*d1qmca_*). Khaki/orange: unmatched/matched parts of 1kn0:A; dark green/green: unmatched/matched parts of the SCOP domains. The achieved scores are presented in Table 2. The hits are chosen from a total of 33 588 found by *SSM* in the course of matching with $p_{SSE} = 15\%$ SSE similarity threshold. The pictures were obtained using *MOLSCRIPT* (Kraulis, 1991) and *Raster*3d (Merritt & Bacon, 1997) software.

deviation per $1 - N_m \leq 7\%$ of difference in chain length may be considered as a measure of that interaction or as an effect of chain length. In order to estimate the effect of chain composition on its three-dimensional structure, consider matches with $N_m = 1$. The value of $N_m = 1$ corresponds to full-chain alignment and therefore indicates highly similar three-dimensional structures. As seen from Fig. 8(*b*), having as few as 20% of identical residues is already enough for chains to fold into highly similar structures. This conclusion generally agrees with previous findings (Chotia, 1992; Chotia & Lesk, 1986; Hubbard & Blundell, 1987). Comparison with Fig. 8(*c*) suggests that the difference in structure increases quite regularly with decrease in sequence identity, reaching 1–2.5 Å at SI $\simeq$ 20%. The decrease in structure similarity is seen as an exponential-like increase in RMSD, which has also been found in other studies (Chotia & Lesk, 1986; Hubbard & Blundell, 1987; Flores *et al.*, 1993; Russell & Barton, 1994; Russell *et al.*, 1997). Thus, the well defined ridge of the RMSD plot at $0.2 \leq \mathrm{SI} \leq 1$ in Fig. 8(*c*) represents the effect of chain composition on the three-dimensional structure of *similar* chains.

Structures with less than 20% sequence identity show a wide range of RMSDs and alignment lengths, while Q does not reflect this effect (with the exception of a few 'islands' at intermediate Q and SI). Fig. 8(*d*) demonstrates a clear reduction of the correlation between RMSD and N_m at RMSD > 2 Å and $N_m < 0.8$, which region, as may be derived from comparison with Figs. 8(*b*) and 8(*c*), corresponds to SI < 20%. These results lead to the conclusion that SI < 20% is a solid indication of low structural similarity, when reliable detection of common submotifs is not feasible. Usually, more than one common substructure with very close values of Q may be identified between remote structural neighbours. Then, alignment of structure A to its remote neighbours B and C is likely to lead to the result that the best common substructure for A and B is not the same as that for A and C, even if B and C are highly similar (but not identical). This uncertainty in the detection of common substructures arises due to small variations of Q at small variations of SI, and therefore the correlation between Q and SI should not be affected. However, close values of Q for different common substructures do not imply closeness of the corresponding values of RMSD and N_{align}. Simple considerations show that at lower structural similarity, the RMSD and N_{align} values of common substructures with close values of Q (and, consequently, SI), may show a wider range of variations. Therefore, with decreasing structural similarity, the correlation between RMSD, N_{align} and SI should vanish. This is exactly the picture seen in Fig. 8 at SI < 20%.

As shown by the obtained results, RMSD is a good score if the structure similarity is sufficiently high that more than 80–90% of residues are aligned. This situation corresponds to structures with *obvious* similarity, for which RMSD gives merely a measure of distortion. The alignment length does not perform well at any degree of similarity, and allows only for a rough indication that 80–90% of aligned residues correspond to highly similar structural neighbours. The Q score performs more or less uniformly in the whole similarity range, except for a few islands aside of the main ridge in Fig. 8(*a*). It is therefore expected that the Q score should be particularly useful if structural similarity is not obvious. This assumption is fully confirmed by the above example of 1kn0:A, which falls into the 'non-obvious' category, judging by the values of SI shown in Table 2. We have performed a series of experiments on the comparison of remote structural neighbours, which have convinced us of the above conclusion.

Consider now the relationship between the structure/sequence similarity and the statistical significance of the matches (Fig. 9). Since statistical significance depends on both the similarity of matched structures and the composition of the database, a perfect match does not necessarily correspond to the lowest P_v and highest Z. As may be seen from the figures, this is, indeed, the case, and at $Q \simeq 1$, SI $\simeq 1$, a wide range of P_v and Z values are attained. Although, on average, statistical significance increases with increasing structure/sequence similarity, the correlation decreases significantly at higher Q and SI [note that the effect of Z should be estimated through integral (12), and the significance of a hit changes in inverse proportion to P_v]. Therefore, statistical significance scores are very sensitive to small structural variations between close structural neighbours, being nearly indiscriminative if structural similarity is low. These findings agree with intuition. Indeed, one expects to find no more than one structure, *identical* to the query ($Q = 1$), in the whole PDB, which

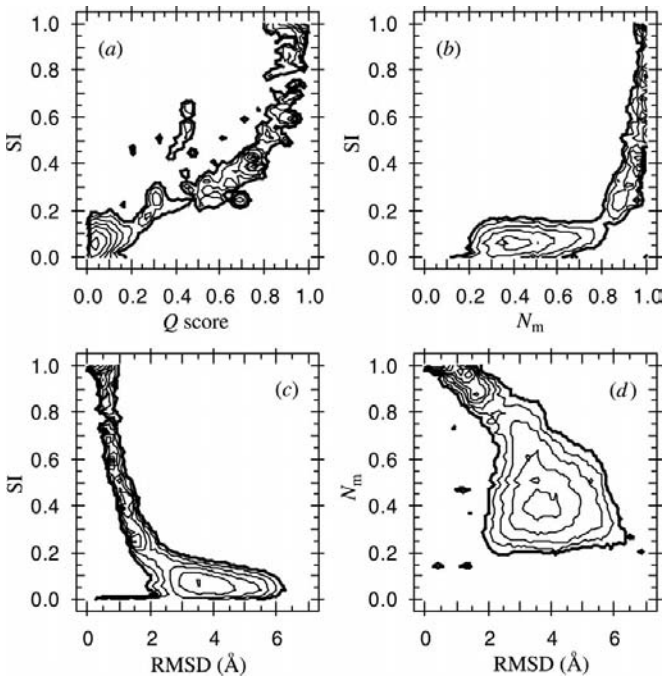

Figure 8
Correlations between (*a*) Q score [equation (8)] and sequence identity [SI; equation (14)], (*b*) SI and normalized alignment length [N_m; equation (13)], (*c*) RMSD and SI, and (*d*) RMSD and N_m, represented as contour maps of the reduced density of probability [equation (15)] of obtaining three-dimensional alignments with the corresponding scores in 'all-to-all' alignment of all chains found in PDB. The outermost contours correspond to the level of 0.05 of the maximum.

finding is then a highly significant event. However, that structure's fold or family will normally have a considerable number of highly similar structural neighbours, even with Q just slightly lower than 1. These matches will not be very surprising in statistical terms. Hence the difference in statistical significance of hits to similar structures with $Q \simeq 1$ should be high. Conversely, detection of low similarity is statistically insignificant, no matter how exactly dissimilar, in one of many million ways, the structures are. Therefore, small differences in $Q \ll 1$ correspond to relatively small differences in $\log(P_v)$ and Z.

Values of $P_v \simeq 1$ and $Z \simeq 0$ indicate hits that are completely expectable, for example, finding a structure containing a helix or a strand. The Q score of such hits does not exceed 0.3 at SI $\leq$ 0.26, which corresponds to low structural similarity. As seen from Fig. 9, the region of low similarity is bounded by $P_v > 10^{-3}$. This fact has a simple explanation as the non-redundant database, which we used for the calibration of P values [that is, the calculation of $\rho_1(x)$, *cf.* equation (11)], was composed of $765 \simeq 10^{2.8}$ folds of SCOP 1.61. Therefore, non-trivial matches are expected to emerge with probability lower than $10^{-2.8}$.

Comparison of Figs. 9(*a*) and 9(*b*) with Figs. 9(*c*) and 9(*d*) shows that the Q score correlates with statistical significance better than with sequence identity. The overal difference in the landscapes is explained by the relationship between Q and SI in Fig. 8(*a*), which shows that Q is not sensitive to SI at SI > 0.5. At the same time, it is curious enough to see that, with the exception of a few islands in Fig. 9(*c*), the P value does not show any evident dependence on chain composition at $0.5 < \mathrm{SI} < 0.95$.

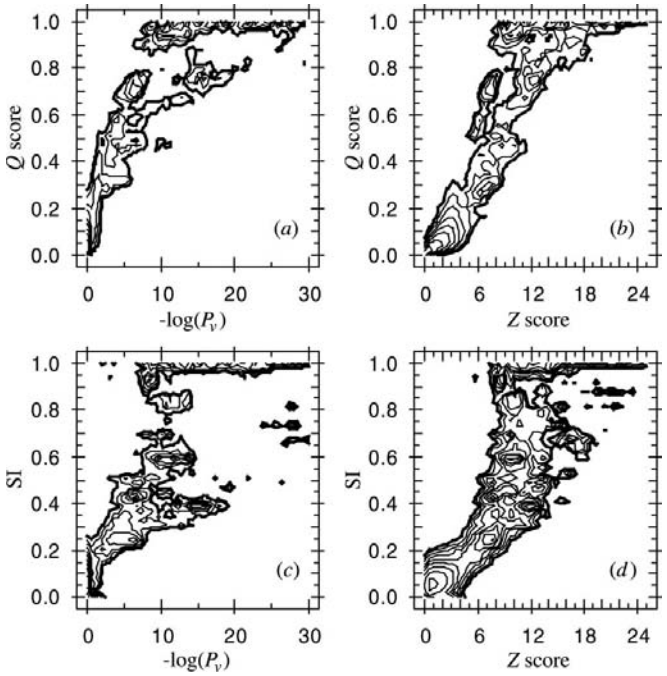

Figure 9
The same data as in Fig. 8, but for the correlations between the structure and sequence similarity, as measured by (*a*) and (*b*) Q score [equation (8)] and (*c*) and (*d*) sequence identity [SI; equation (14)], and statistical significance of matches represented by P value [equation (9)] and Z score [equation (12)]. The outermost contours correspond to the level of 0.05 of the maximum.

7. Conclusion

More than two years of working with *SSM* and studying the feedback from its users worldwide has convinced us that *SSM* represents a powerful, flexible and accurate tool for protein structure comparison in three dimensions. It is particularly efficient, as compared with other similar resources available, when applied to large protein structures (more than a few hundred amino-acid residues) and for matching a structure to a precompiled database of structures (PDB, SCOP or user-defined).

The competitive performance of *SSM* is mostly a result of the original graph-matching algorithm employed (Krissinel & Henrick, 2004). In the present study, we did not compare the efficiency of *SSM* with that of similar algorithms, although, in our experience, *SSM* is at least an order of magnitude faster. However, a direct and objective comparison is hardly obtainable. Many other services are not interactive, which prevents direct time measurements. Most of the existing services maintain a database of precalculated alignments or use sets of representative structures, so that the number of actual alignments is never the same. Finally, *SSM* runs on a CPU cluster, employing different numbers of CPUs depending of the task complexity, while little is known about the implementation and hardware basis of other developments.

The iterative procedure of C_α alignment as described in this paper includes a number of empirical elements and parameters. These elements were introduced and the corresponding parameters tuned in the course of analysing of thousands of alignments. As a result, comparison of *SSM* with other similar servers shows a good overall agreement, to the degree of difference between all of them.

Because of the ever-growing number of solved protein structures, automatic recognition of their structural motifs becomes an increasingly important task. The very definition of structural similarity remains, however, a vague issue in general. Unless the similarity is self-evident, there is no perfect quantitative measure for drawing a line between similar and dissimilar structures, and even for ranging structure pairs in order of their similarity. Because of this circumstance, any test on true/false positives/negatives is never fully convincing, and therefore such a test was omitted in the present study. In the numerical study presented in this paper, we considered a few scores applicable to measuring the structural similarity. As shown, the most obvious scores of RMSD and alignment length do not provide a sufficient level of confidence in structure recognition. The best quality of structure recognition is achieved by using the introduced Q score [equation (8)], which combines both RMSD and the alignment length. The Q score represents a measure of quality of three-dimensional alignment and is maximized by the *SSM*'s C_α alignment algorithm. Although the Q score should be viewed only as a model simplification of an intuitive understanding of the alignment quality, we found that in practice it works very well.

It should be noted that there are other scores combining the alignment length and relative remoteness of aligned residues (see *e.g.* Russell & Barton, 1992; Kleywegt & Jones, 1994), which we did not investigate in this study.

APPENDIX *A*
Fast optimal superposition in three dimensions

A number of methods have been reported for the calculation of the rotation matrix $\hat{\mathbf{R}}$, which optimally superposes two sets of points in three-dimensional space, $\mathbf{x}_i$ and $\mathbf{y}_i$, $i = 1, \ldots, N$, such that (both sets are brought into their centres of mass)

$$D = \sum_{i=1}^{N} w_i \left(\mathbf{x}_i - \hat{\mathbf{R}}\mathbf{y}_i\right)^2 \tag{16}$$

(w_i are weights) is minimal (see *e.g.* McLachlan, 1972; Kabsch, 1976, 1978; Lesk, 1986). The methods involve converging iterations, diagonalization or orthogonal decomposition of the correlation matrix $\hat{\mathbf{A}}$ (Lesk, 1986),

$$A_{jk} = \sum_{i=1}^{N} w_i x_{ij} y_{ik}, \qquad j, k = 1, 2, 3. \tag{17}$$

We found that the best results are obtained using singular value decomposition (SVD), which is a very stable numerical procedure applicable even to singular correlation matrices. According to Lesk (1986), $\hat{\mathbf{A}} = \hat{\mathbf{R}}^{\mathrm{T}}\hat{\mathbf{H}}$, where $\hat{\mathbf{H}}$ is a (unique) Hermitian positive definite matrix. Applying SVD to matrix $\hat{\mathbf{A}}$, we obtain

$$\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\lambda}\hat{\mathbf{V}}^{\mathrm{T}} = (\hat{\mathbf{U}}\hat{\mathbf{V}}^{\mathrm{T}})(\hat{\mathbf{V}}\hat{\lambda}\hat{\mathbf{V}}^{\mathrm{T}}), \tag{18}$$

where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are orthonormal matrices and $\hat{\lambda}$ is a diagonal matrix of (always non-negative) singular values. Considering that $\hat{\mathbf{V}}\hat{\lambda}\hat{\mathbf{V}}^{\mathrm{T}}$ represents a Hermitian positive definite matrix, we obtain

$$\hat{\mathbf{R}}^{\mathrm{T}} = \hat{\mathbf{U}}\hat{\mathbf{V}}^{\mathrm{T}}. \tag{19}$$

This procedure, however, does not guarantee that $\hat{\mathbf{R}}$ will represent a proper rotation. If $\det(\hat{\mathbf{R}}) < 0$ then the superposed set $\{\mathbf{y}_i\}$ is inverted (rotoinversion) (Kabsch, 1978). There is no way out of this problem other than to make an appropriate correction to the correlation matrix $\hat{\mathbf{A}}$. As follows from equation (19), changing the sign of any of the vectors $\mathbf{U}_i$ or $\mathbf{V}_i$ will change the sign of $\det(\hat{\mathbf{R}})$ and thus make $\hat{\mathbf{R}}$ the matrix of proper rotation. Such a change of sign is equivalent to a distortion of $\hat{\mathbf{A}}$. Since (Lesk, 1986)

$$D = \sum_{i=1}^{N} \left(|\mathbf{x}_i|^2 + |\mathbf{y}_i|^2\right) - \mathrm{trace}(\hat{\mathbf{R}}\hat{\mathbf{A}}), \tag{20}$$

such a distortion may result in increasing D. As may be derived from equations (18) and (20), this increase is least (and therefore the resulting proper rotation is the best possible one) if changing the sign is applied to the vector $\mathbf{U}_i$ or $\mathbf{V}_i$ that corresponds to the minimal singular value λ_i.

It is important to note that the calculation of the rotation matrix using SVD gives a meaningful result even if the correlation matrix $\hat{\mathbf{A}}$ is degenerate, which fact was taken into account in our choice of method. The optimal superposition is achieved by applying the rotation matrix $\hat{\mathbf{R}}$ to structures $\{x_i\}$, $\{y_i\}$ brought into their centres of mass.

The authors are thankful to Dr Stephen H. Bryant for a detailed explanation of the *P* value calculations in *VAST* (Gibrat *et al.*, 1996). EK is grateful for support from the BBSRC Collaborative Computational Project No. 4 in Protein Crystallography (Collaborative Computational Project, Number 4, 1994).

References

Alexandrov, N. N. (1996). *Protein Eng.* pp. 727–732.
Barakat, D. W. & Dean, P. M. J. (1991). *Comput. Aided Mol. Des.* **5**, 107–117.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* pp. 235–242.
Bessonov, Y. E. (1985). *Vychisl. Sistemy*, **112**, 3–22.
Bron, C. & Kerbosch, J. (1973). *Commun. ACM*, **16**, 575–577.
Chotia, C. (1992). *Nature* (*London*), **357**, 543–544.
Chotia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Falicov, A. & Cohen, F. E. (1996). *J. Mol. Biol.* **258**, 871–892.
Flores, T. P., Orengo, C. A., Moss, D. C. & Thornton, J. M. (1993). *Protein Sci.* **2** 1811–1826.
Gardiner, E. J., Willett, P. & Artymiuk, P. J. (2000). *J. Chem. Inf. Comput. Sci.* **40**, 273–279.
Gerstein, M. & Levitt, M. (1996). *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pp. 59–67. Menlo Park, California: AAAI Press.
Gerstein, M. & Levitt, M. (1998). *Protein Sci.* **7**, 445–456.
Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.
Godzik, A. & Skolnick, J. (1994). *Comput. Appl. Biosci.* **10**, 587–596.
Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willett, P. J. (1993). *Mol. Biol.* **229**, 707–721.
Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.
Hubbard, T. J. P. & Blundell, T. L. (1987). *Protein Eng.* **1**, 159–171.
Hutchinson, E. G. & Thornton, J. M. (1996). *Protein Sci.* **5**, 212–220.
Jung, J. & Lee, B. (2000). *Protein Eng.* **13**, 535–543.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.
Kagawa, W., Kurumizaka, H., Ishitani, R., Fukai, S., Nureki, O., Shibata, S. & Yokoyama, S. (2002). *Mol. Cells*, **10**, 359.
Kato, H. & Takahashi, Y. J. (2001). *Chem. Softw.* **7**, 161–170.
Kleywegt, G. J. & Jones, T. A. (1994). *CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 9–14.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 525–545.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Krissinel, E. & Henrick, K. (2004). *Softw. Pract. Exp.* **34.** 591–607.
Krissinel, E. B., Winn, M. D., Ballard, C. C., Ashton, A. W., Patel, P., Potterton, E. A., McNicholas, S. J., Cowtan, K. D. & Emsley, P. (2004). *Acta Cryst.* D**60**, 2250–2255.
Leluk, J., Konieczny, L. & Roterman, I. (2003). *Bioinformatics*, **19**, 117–124.
Lesk, A. M. (1986). *Acta Cryst.* A**42**, 110–113.
Levi, G. (1972). *Calcolo*, **9**, 341–354.
McLachlan, A. D. (1972). *Acta Cryst.* A**28**, 656–657.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Mitchell, E. M., Artymiuk, P. J., Rice, D. W. & Willett, P. J. (1990). *Mol. Biol.* **212**, 151–166.
Mizuguchi, K. & Go, N. (1995). *Protein Eng.* **8**, 353–362.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. J. (1995). *Mol. Biol.* **247** 536–540.
Orengo, C. A. & Taylor, W. R. (1996). *Methods Enzymol.* **266**, 617–635.
Raymond, J. & Willett, P. J. (2002). *Comput. Aided Mol. Des.* **16**, 521–533.
Raymond, J. W., Gardiner, E. J. & Willett, P. J. (2002). *Chem. Inf. Comput. Sci.* **42**, 305–316.
Rouvray, D. H., Balaban, A. T., Wilson, R. J. & Beineke, L. W. (1979). Editors. *Applications of Graph Theory*, pp. 177–221. NewYork: Academic Press.
Russell, R. B. & Barton, G. J. (1992). *Proteins*, **14**, 309–323.
Russell, R. B. & Barton, G. J. (1994). *J. Mol. Biol.* **244**, 332–350.
Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A. & Sternberg, M. J. E. (1997). *J. Mol. Biol.* **269**, 423–439.
Ryter, J. M. & Schultz, S. C. (1998). *EMBO J.* **17**, 7505–7513.
Sali, A. & Blundell, T. J. (1990). *Mol. Biol.* **212**, 403–428.
Sevcik, J., Dodson, E. J. & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240–253.
Shearer, K., Bunke, H. & Venkatesh, S. (2001). *Pattern Recognit.* **34**, 1075–1091.
Shindyalov, I. N. & Bourne, P. E. (1998). *Protein Eng.* **11**, 739–747.
Singh, A. P. & Brutlag, D. L. (1997). *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ISMB-97*, pp. 284–293. Halkidiki, Greece: AAAI Press.
Smith, T. F. & Waterman, M. S. (1981). *J. Mol. Biol.* **147**, 195–197.
Subbiah, S., Laurents, D. V. & Levitt, M. (1993). *Curr. Biol.* **3**, 141–148.
Taylor, W. & Orengo, C. J. (1989). *Mol. Biol.* **208**, 1–22.
Ullman, J. R. (1976). *J. Assoc. Comput. Mach.* **23**, 31–42.
Vriend, G. & Sander, C. (1991). *Proteins*, **11**, 52–58.
Zuker, M. & Somorjai, R. L. (1989). *Bull. Math. Biol.* **51**, 55–78.