Protein Science (1994), *3*:769–778. Cambridge University Press. Printed in the USA.

Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding

DANIEL FISCHER,[1,2] HAIM WOLFSON,[1] SHUO L. LIN,[3] AND RUTH NUSSINOV[2,3]

[1] Computer Science Department, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[2] Sackler Institute of Molecular Medicine, Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[3] Laboratory of Mathematical Biology PRI/Dynacorp, National Cancer Institute—Frederick Cancer Research Facility, Frederick, Maryland 21702

(RECEIVED December 16, 1993; ACCEPTED February 25, 1994)

Abstract

We have recently developed a fast approach to comparisons of 3-dimensional structures. Our method is unique, treating protein structures as collections of unconnected points (atoms) in space. It is completely independent of the amino acid sequence order. It is unconstrained by insertions, deletions, and chain directionality. It matches single, isolated amino acids between 2 different structures strictly by their spatial positioning regardless of their relative *sequential* position in the amino acid chain. It automatically detects a recurring 3D motif in protein molecules. No predefinition of the motif is required. The motif can be either in the interior of the proteins or on their surfaces. In this work, we describe an enhancement over our previously developed technique, which considerably reduces the complexity of the algorithm. This results in an extremely fast technique. A typical pairwise comparison of 2 protein molecules requires less than 3 s on a workstation. We have scanned the structural database with dozens of probes, successfully detecting structures that are similar to the probe. To illustrate the power of this method, we compare the structure of a trypsin-like serine protease against the structural database. Besides detecting homologous trypsin-like proteases, we automatically obtain 3D, sequence order-independent, active-site similarities with subtilisin-like and sulfhydryl proteases. These similarities equivalence isolated residues, not conserving the linear order of the amino acids in the chains. The active-site similarities are well known and have been detected by manually inspecting the structures in a time-consuming, laborious procedure. This is the first time such equivalences are obtained automatically from the comparison of full structures. The far-reaching advantages and the implications of our novel algorithm to studies of protein folding, to evolution, and to searches for pharmacophoric patterns are discussed.

Keywords: computer vision; protease active sites; protein database structural comparison; protein folding; 3D protein motifs

We have recently developed an extremely fast, template-free, sequence order-independent technique for the comparisons of the 3-dimensional structures of proteins. There are 3 novel aspects in our method. First, we compare the 3D structures of proteins (or any biological macromolecule) completely regardless of the order of the residues in the chain. This allows us to detect similarities between protein molecules, whether these are on their surfaces or in their interior. Our computer vision-based algorithm views atoms as collections of unconnected points in space. This truly 3D approach overcomes a major limitation inherent in other structural comparison techniques which require that the linear order of the amino acid sequences be conserved (e.g., Matthews & Rossmann, 1985). Although some techniques overcome the insertion/deletion difficulties, the constraint of chain order is still a problem. Other methods allow some degree of nonsequential matching by comparing fragments of consecutive amino acids (e.g., Alexandrov et al., 1992) or by matching secondary structure elements (e.g., Mitchel et al., 1989). Nevertheless, a strict linear order within the fragments is still required. Our approach is unique because it obtains a spatial similarity between isolated atoms (residues) belonging to protein mole-

Reprint requests to: Ruth Nussinov, Building 469, Room 151, NCI-FCRF, Frederick, Maryland 21702; e-mail: ruthn@fcrfv1.ncifcrf.gov.

cules, completely regardless of the order and the directionality of the residues in the chain. In particular, unlike some previous 3D approaches, it does not require matches of fragments of consecutive residues. The latter limits the generality and applicability of a methodology for 3D comparison. Second, we are able to detect recurring substructural, "real" 3D motifs in a set of structures without a predefinition of the motifs. Furthermore, all molecules in the database can be compared simultaneously. Third, the method is extremely fast, with a typical running time of less than 3 s for a comparison between 2 proteins on a Silicon Graphics Indigo workstation, or about 8 min for a comparison of 1 protein against a representative set of proteins from the crystallographic database, consisting of 170 protein structures.

Insight into evolution can be gained from 3D comparisons. In particular, the question of divergence versus convergence of proteins can be addressed. On the one hand, if the result of a 3D structural comparison is such that the linear order of the sequences is conserved, additional evidence of divergent evolution can be deduced because the sequential order is ignored during the comparison. Being completely blind to the order in the sequences, in these cases our method "rediscovers" the dual sequence-structure homology typical of divergent species. On the other hand, if "real" 3D matches are obtained (i.e., the sequential order is not conserved), either convergent evolution or genetic exchanges may be implicated (Fischer et al., 1993b).

Originally, interest in automated structural comparison methods arose from the need to superimpose the structures of divergently evolved proteins. In such comparisons, a strict linear order conservation (allowing insertions and deletions) has been enforced. Sequence order-dependent methodologies are adequate for the comparison of divergently evolved structures, although our method performs well also, even for remote similarities. Ideally, comparisons of such structures are carried out using both methodologies.

Recognition of common substructural features that do not conserve the linear order of the amino acid sequence entails application of sequence order-independent methodologies. Examples of such features may include similarities between active sites of convergently evolved structures, between different folding motifs, between the scaffolds of unrelated proteins, and between recurring stable configurations in the interior of proteins. As shown below, our method, without using the sequential order of the chains, succeeds in finding the similarities of divergently evolved proteins as well as those of convergently evolved ones.

Here we apply our method to the comparison of a trypsin-like serine protease against a representative data set of the crystallographic database (see Kinemage 1). Linear matches are obtained with other (divergently evolved) trypsin-like proteases such as trypsinogen, kallikrein A, elastase, and proteinases A and B. In particular, active-site similarities between the trypsin-like and (the convergently evolved) subtilisin-like and sulfhydryl proteases are automatically detected. These equivalences do not conserve the sequential order of the chains. Although these similarities have long been established (by human observation), it is the first time they are obtained (1) automatically, (2) without the predefinition of the active sites, and (3) efficiently (under 8 CPU min for the entire data set comparison). We are able to achieve this high level of performance owing to the unique approach we have adopted. In particular, recently we have been able to further improve our technique, reducing its complexity from $O(n^3)$ to $O(n^2)$, where n is the number of residues of the largest protein. This considerable improvement is described in some detail below.

Results

3D structural comparison

Our algorithm is based on the geometric hashing paradigm (Lamdan et al., 1988) adapted from computer vision to macromolecular comparison by Nussinov and Wolfson (1991). The problem of finding similar, though frequently partially occluded, structures between model objects and an observed scene is of central importance in computer vision. The geometric hashing paradigm for model-based object recognition is especially geared toward recognition of partially occluded objects belonging to large object databases, and its complexity is a low-degree polynomial on the size of the object. Rather than superimpose one protein on another in all rotations and translations, an explosive, time-consuming, gridlike search, our method uses a rotational and translational invariant representation of the molecules. This simple procedure allows fast detection of local, "good" matches first. The transformations of these "good" matches are subsequently calculated, achieving very high efficiency. Previously, Fischer et al. (1992) and Bachar et al. (1993) have reduced the complexity of the algorithm from $O(n^4)$ in Nussinov and Wolfson (1991) to $O(n^3)$, where n is the number of C_α atoms in the larger of the 2 structures being compared. In this work we describe an improvement of the method that reduces the complexity to $O(n^2)$, although, as described below, in practice its running time is almost linear. This considerably enhanced performance allows a rapid scan of the full crystallographic database in minutes.

Three steps are involved

We briefly summarize the 3 major steps of the method and refer the reader to our previous publications for a detailed description (Nussinov & Wolfson, 1991; Fischer et al., 1992; Bachar et al., 1993). The present improvement stems from the reduction of the number of invariants computed per molecule and is described below. When comparing two 3D structures, the transformation that best superimposes them is initially unknown. We solve the problem of finding a global match between the structures in 3 steps. In the first step, we search for a subset of atoms in one structure that matches a subset in the other structure. We divide each of the proteins being compared into spheres of a predefined radius (12.5 Å), each centered on 1 C_α atom of the protein, which we refer to as "balls." The balls around each C_α atom contain atoms close in space but not necessarily close in the sequence. A match is searched between the rotational and translational invariants of the atoms of every ball of one structure with those of every ball of the other. If a "good enough" match is obtained, the transformation that best superimposes the balls is computed. This is a local, seed match. Because more than 1 pair of balls can match, all the matching ball pairs are remembered. In the second step, pairs of balls (seed matches) having similar transformations are clustered together. In the last step, the (clustered) seed matches are extended by searching for additional matching pairs of atoms that are not contained in the matched balls. Pairs of atoms lying at a distance of 2.5 Å or less (after superimposition) are considered to match. At the end, only the largest global matches (largest number of matched pairs of C_α atoms) are produced.

Rotational and translational invariant representation

Previously, we built 1 local reference frame using pairs of C_α atoms. Here, only 1 reference frame is defined per C_α atom. The origin of the reference frame is placed on the C_α atom. Consider the vectors that connect a C_α atom to its previous and successor C_α atoms. These 2 vectors determine a plane. The x-axis is one of the vectors, and the y-axis is perpendicular to it in this plane. The z-axis is the normal to this plane. In this local reference frame, the coordinates of all C_α atoms contained within a ball of a given radius, centered on the same C_α, are computed. The coordinates of all the C_α atoms in each of the n balls of one of the proteins (the probe) are stored in a hash table. This enables efficient and simultaneous comparison of each ball from the second structure to all the balls of the probe (for details see Nussinov & Wolfson, 1991). Two balls are considered to match if at least 20 pairs of C_α atoms within the balls have "similar" coordinates (i.e., they differ by less than 1.0 Å in each of their x, y, z components). Twenty pairs of similar coordinates correspond to a relatively large percentage of matched atoms within the balls, as for example, in β-trypsin, the average number of C_α atoms contained in a ball of radius 12.5 Å is 34.

Performance

Because the number of C_α atoms contained within the balls is bound by a constant, the number of transformational invariants computed is linear on the size of the larger protein. The actual running time of the algorithm is thus almost linear. This results in extremely short execution (CPU) times, of 3 s on average, for a pairwise comparison of 2 protein structures. The only use we make of the sequence order of the proteins is in the definition of the reference frame. We are currently considering a different choice of reference frame, which is based on 3 atoms belonging to the same residue. (Requests for structural comparisons under our program can be sent by e-mail to fischer@fcafv1.ncifcrf.gov.)

Scanning the database

We have scanned the database using dozens of protein probes from major protein families: hemoglobins, immunoglobulins, dehydrogenases, lysozymes, cytochromes, and others. In all cases, the results of the scans can be classified as follows:

1. The highest scores (i.e., the best matches) correspond to proteins from the same family as the probe protein. In these cases, the match obtained is a linear one (conserving the sequential order of the sequences), even though no sequence information was used by the algorithm. Because the best geometric match corresponds to a linear alignment, these matches may imply evolutionary divergence between members of the family.
2. The next set of scores corresponds to proteins containing features similar to some substructures in the probe protein. These matches contain fewer matched pairs than above, usually have a larger RMS, and do not necessarily conserve the linear order of the sequences. Proteins demonstrating this type of substructural matching may have converged during evolution.
3. The lowest scores correspond to spurious matches between unrelated and dissimilar structures. These may contain equivalences of single α-helices, β-strands, or randomly matched isolated residues.

Recently, several methods performing relatively fast searches in the database have been developed (e.g., Taylor & Orengo, 1990; Vriend & Sander, 1991; Alexandrov et al., 1992; Grindley et al., 1993). In all of them, either the sequential order of the matches must be conserved, or the matching is carried out between fragments of contiguous residues in the chains. These methods are well suited to find similarities belonging to class 1 above, with various degrees of accuracy and speed. However, similarities belonging to classes 2 and 3 can be obtained by these methods only if sufficiently large fragments of consecutive residues in both proteins match. Our approach overcomes these limitations. In particular, in addition to obtaining matches that have been reported by the above methods (typically with a lower RMS and better performance), it is capable of obtaining matches of isolated residues not belonging to contiguous fragments or belonging to nonsecondary structure elements. To demonstrate these capabilities of the method, we scan the data set with different probes. First, the results of the scan using a trypsin-like serine protease as a probe are described in detail. Next we describe the results of scanning the database with subtilisin and actinidin as probes.

A serine protease scan

Here we show the results of scanning the data set with α-chymotrypsin from bovine pancreas, PDB code (Protein Data Bank; Bernstein et al., 1977) 1cho (Fujinaga et al., 1987). In addition to the expected matches with all the trypsin-like proteases in the database, substructural matches, not conserving the linear order of the chains, are obtained with other proteases.

Figure 1 shows the results of this scan. Each point in the figure corresponds to a comparison of α-chymotrypsin against 1 data set protein. The figure shows the relationship between the sizes of the data set proteins and the number of matched pairs obtained. A similarity score is defined (see legend of Fig. 1) to account for the difference in sizes between the probe protein and each of the data set proteins. The highest scoring proteins correspond to the 9 trypsin-like mammalian proteases in the data set (including 1cho). The next 3 highest scores correspond to the bacterial trypsin-like proteases. Just below these, the subtilisin-like and sulfhydryl proteases are found. Table 1 lists the names, sizes, scores, and RMS distances of the 20 top ranking proteins (circled in Fig. 1). A scan of the database using another trypsin-like serine protease, β-trypsin, has also been carried out, with very similar results.

The matches of the top 12 proteins (trypsin-like proteases) conserve the sequential order. They are equivalent to the reported comparisons between the serine proteases. The structures of the mammalian proteases are very similar to each other, as are the structures of the bacterial proteases. Mammalian and bacterial proteases share the same fold, but the similarity between them is considerably lower. All trypsin-like proteases seem to have evolved from a common ancestor and have a relatively high sequence and structural homology (see, e.g., Branden &

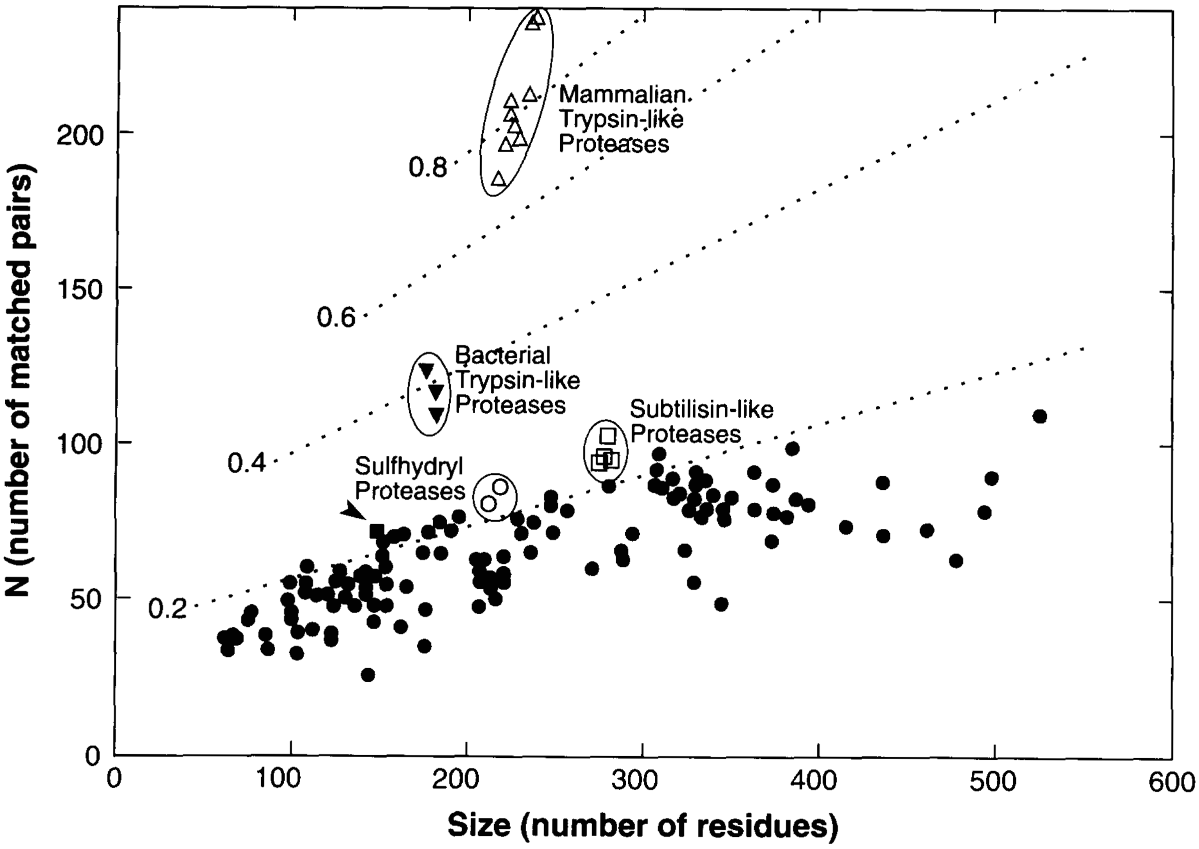

Fig. 1. Results obtained from the comparison of α-chymotrypsin (1cho) against a data set of 170 randomly selected proteins from the PDB (Bernstein et al., 1977) having a resolution of 3.0 Å or better. This data set covers the major protein families and includes several homologous entries for some families. It includes 12 trypsin-like serine proteases (9 mammalian, 3 bacterial), 5 subtilisin-like serine proteases, 2 sulfhydryl proteases, 15 globins, 11 immunoglobulins, 15 dehydrogenases, 6 DNA binding proteins, 4 lysozymes, 10 cytochromes, 9 calcium binding proteins, and others. The 170 proteins are listed below. Each dot in the figure represents 1 comparison between α-chymotrypsin and one of the data set proteins. The x-axis is the size (in number of residues) of the protein compared and the y-axis is the number of matched pairs found in the comparison with the α-chymotrypsin. A similarity score is computed for each comparison. This normalized score takes into account the number of matched C_α atoms and penalizes the difference in sizes between the probe (α-chymotrypsin) and each of the proteins. (What score should be used for structural similarity of proteins is controversial. Some commonly used scores do not equally penalize the difference in sizes of proteins larger or smaller than the probe.) The score is computed as: $score = pairs/(ProbeSize + TargetSize - pairs)$, where *pairs* is the number of matched pairs obtained in the comparison, *ProbeSize* is the number of amino acids of the probe (here, 238), and *TargetSize* is the size of the protein compared. In this score the number of matched pairs is divided by the sum of 3 components: (1) the number of unmatched residues from the probe, (2) the number of unmatched residues from the target, and (3) the number of matched pairs between the probe and the target. Explicitly, we divide *pairs* by $[(ProbeSize - pairs) + (TargetSize - pairs) + pairs]$. Similarity levels (dotted lines) are plotted at the values 0.2, 0.4, 0.6, and 0.8. Only 143 of the 170 proteins in the data set are shown in this figure. Lower similarity scores were obtained for the remaining 27 proteins. The 12 highest ranking scores correspond to the 9 mammalian (depicted by empty triangles) and the 3 bacterial (marked by inverted triangles) trypsin-like serine proteases in the data set. Ranked 13 is proteinase K, a subtilisin-like serine protease. The other subtilisin-like serine proteases are at ranks 15, 16, 18, and 19 (noted by empty squares). The 2 sulfhydryl proteases ranked 14 and 20 (depicted by empty circles). Flavodoxin is shown as a filled square pointed to by an arrow (see legend of Table 1). All other proteins are shown by filled circles. Table 1 lists the top 20 ranking proteins and the legend includes a list of the proteins having the next largest number of matched pairs. The scan of 170 proteins required 447 s of CPU time on a Silicon Graphics Indigo workstation, or roughly 3 s per pairwise comparison. The list of entries in our data set is detailed below (the 4-character PDB identifier is used, and if more than 1 chain exists in 1 entry, the fifth character denotes the chain used): 1abp 1ace 1acx 1alc 1ald 1atn 1azu 1bp2 1ca2 1cc5 1ccr 1cho 1choi 1cro 1cse 1csei 1ctf 1eca 1ecd 1etu 1f19h 1f19l 1fc1a 1fc2 1fcb 1fdh 1fdx 1fx1 1gcr 1gox 1gp1a 1gpd 1hip 1hkg 1hne 1hoe 1hsc 1i1b 1ldm 1lh1 1llc 1lrd 1lyz 1lz1 1mba 1mle 1nxb 1pcy 1pfc 1pfka 1pp2 1pyk 1pyp 1reia 1rhd 1sbt 2sc2 1srx 1tece 1tgse 1tgsi 1tima 1tmne 1tnc 1tnfa 1tpo 1ubq 1wsya 1wsyb 1ypi 2act 2alp 2app 2cdv 2cha 2cln 2cna 2cpp 2cro 2cts 2dhba 2dhbb 2est 2fbjh 2fbjl 2gd1o 2gn5 2hfll 2hmg 2i1b 2lbp 2lhb 2liv 2lzm 2ovo 2pab 2pka 2plv1 2plv3 2prk 2pt1 2pt2 2ptce 2ptci 2rspa 2sece 2sga 2sns 2sodo 2stv 2taa 2tbva 2ts1 2wrp 351c 3adk 3b5c 3bcl 3c2c 3cln 3cpv 3cyto 3dfr 3fabh 3fabl 3fxc 3gapa 3gapb 3grs 3hlaa 3hlab 3hvp 3icb 3ldh 3pcy 3pgk 3pgm 3rn3 3rp2a 3sgb 451c 4ape 4dfr 4enl 4fxn 4hhba 4hhbb 4ins 4mbn 4mdh 4rhv3 4tnc 5cpa5hmga 5hmgb 5ldh 5rub 5tnc 6cpa 6ldh 6lyz 6rsa 7adh 7xia 8adh 8atca 8cata 9apia 9apib 9pap.

Tooze, 1991). They are folded into 2 antiparallel β-barrel domains containing the Greek key motif. Each domain contains about 120 amino acids forming 6 β-strands folded into the same topology. The active site is in a crevice between the 2 domains and is formed by residues from 2 loop regions. Four structural features occur in an almost identical fashion in the serine protease family: (1) a catalytic triad consisting of the side chains of 3 residues: Asp, His, and Ser, which are close to each other in

Table 1. *The 20 top-ranking matches obtained in the comparison of α-chymotrypsin (1cho) against the 170 data set*[a]

R	PDB	Title and source	Size	N	RMS	Score	Comments
1	1cho	α-Chymotrypsin (bovine pancreas)	238	238	0.00	1.00	Mam. tryp. lk
2	2cha	α-Chymotrypsin A (cow)	236	236	0.55	0.99	Mam. tryp. lk
3	2ptce	β-Trypsin (bovine pancreas)	223	211	0.85	0.89	Mam. tryp. lk
4	2est	Elastase (porcine pancreas)	234	213	1.00	0.82	Mam. tryp. lk
5	1tpo	β-Trypsin (bovine pancreas)	223	207	0.81	0.81	Mam. tryp. lk
6	1tgse	Trypsinogen (bovine pancreas)	225	203	0.89	0.78	Mam. tryp. lk
7	3rp2a	Rat mast cell proteinase II (rat intestine)	220	197	0.94	0.76	Mam. tryp. lk
8	2pka	Kallikrein A (porcine pancreas)	228	199	0.98	0.75	Mam. tryp. lk
9	1hne	Elastase (human neutrophils)	216	186	0.90	0.70	Mam. tryp. lk
10	2alp	α Lytic proteinase (*Lysobacter enzymogenes*)	175	124	1.40	0.43	Bact. tryp. lk
11	2sga	Proteinase A (*Streptomyces griseus*)	181	117	1.27	0.39	Bact. tryp. lk
12	3sgb	Proteinase B (*S. griseus*)	181	110	1.24	0.36	Bact. tryp. lk
	1snv	Sindbis capsid protein (Sindbis virus)	151	84	1.66	0.27	Tryp. lk[b]
13	2prk	Proteinase K (fungus)	279	103	1.67	0.25	Subtilisin lk
14	2act	Actinidin (kiwi fruit)	218	86	1.69	0.24	Sulfhydryl P
15	1sbt	Subtilisin (*Bacillus amyloliquefaciens*)	275	96	1.76	0.23	Subtilisin lk
16	2sece	Subtilisin Carlsberg (*Bacillus subtilis*)	274	95	1.78	0.23	Subtilisin lk
17	1fxl	Flavodoxin (*Desulfovibrio vulgaris*)	147	71	1.69	0.23	α/β[c]
18	1cse	Subtilisin Carlsberg (*B. subtilis*)	274	94	1.72	0.23	Subtilisin lk
19	1tece	Thermitase (*Thermoactinomyces vulgaris*)	279	94	1.79	0.22	Subtilisin lk
20	9pap	Papain (papaya)	212	81	1.74	0.22	Sulfhydryl P

[a] Notation is as follows: R, rank; Size, number of residues; N, number of matched pairs; RMS, root mean square distance; Score, see legend of Figure 1. Fourteen additional proteins with scores between 0.20 and 0.22 are listed below. In parentheses, for each protein we note R, Size, N, and Score, in this order. Most of these are α/β proteins involving equivalences between β-strands. The 14 proteins are: 4dfr (21, 159, 71, 0.21); 1gp1a (22, 183, 75, 0.21); 3adk (23, 194, 77, 0.21); 6cpa (24, 308, 97, 0.21); 3dfr (25, 162, 71, 0.21); 1srx (26, 108, 61, 0.21); 1tnfa (27, 152, 68, 0.21); 1etu (28, 177, 72, 0.21); 1ypi (29, 247, 83, 0.20); 5cpa (30, 307, 92, 0.20); 1pyp (31, 280, 87, 0.20); 1hkg (32, 190, 72, 0.20); 1tima (33, 247, 81, 0.20); 3fxc (34, 98, 56, 0.20).

[b] 1snv was not originally in the 170 data set (see text).

[c] Flavodoxin, of α/β type, obtained a relatively large score, although the number of matched residues was relatively low (71). This may be due to the scoring function used, which may not adequately penalize proteins smaller than the probe. As is the case for some of the 14 matches listed above, the match of trypsin with flavodoxin consists mainly of equivalences between β-strands.

the active site but far apart in the sequence of the chain; (2) an oxyanion hole, which is a pocket that stabilizes and tightly binds the tetrahedral transition state intermediate; (3) a nonspecific binding loop region that hydrogen bonds (through the main chains) to the substrate; and (4) a specificity pocket that fits a preferred substrate amino acid. The catalytic triad residues (in trypsin) are H57, D102, and S195. Residues 193–195 form the oxyanion hole, residues 214–216 form the nonspecific binding, and the specificity pocket is formed by residues 189, 216, and 226.

Strikingly, proteinase K (2prk; Betzel et al., 1988), a subtilisin-like serine protease, ranks very high (13th) in Table 1. It also ranks highly (15th) in the comparison of the β-trypsin against the database (not shown here). Subtilisin-like proteases are α/β open parallel sheets of about 275 residues formed by 7 strands and 4 helices, 2 on each side of the β-sheet. The active site is at the C-termini of the central β strands. This fold differs from those of the double antiparallel β-barrel structure of the trypsins (see, e.g., Kraut et al., 1972). However, it is remarkable that although the structures of subtilisin-like and trypsin-like proteases are globally dissimilar, their active sites are structurally similar. The 4 structural features described above also occur in the subtilisin-like proteases. The C_α atoms of proteinase K participating in the catalytic triad (H69, D39, and S224), in the oxyanion hole (N161), and in the substrate binding (100–102, 134–135) are in very similar positions relative to those of the trypsin-like proteins. This appears to be an example of convergent evolution where different ancestors converged to the same structural solution for a catalytic mechanism. We did not expect to find a subtilisin-like protease ranking so high in either of the lists due to the large structural differences between them and the trypsin-like proteases. The match is completely 3D, without any sequential order conservation except for small segments of contiguous residues. At first sight, it seems that it is a random match of small segments plus isolated residues. However, close inspection reveals that the match paired equivalent atoms in the active sites. The match between β-trypsin and proteinase K is very similar. Figure 2 shows the match between β-trypsin (PDB code 1tpo) and proteinase K. In what follows we refer to this match only. Residues 32–39 of 2prk were equivalenced to residues 109–102 of 1tpo (in reverse order). This corresponds to the matching of 2 β-strands with opposite directions. Particularly interesting is that D102 of 1tpo was matched to D39 of 2prk. Both Asp residues are part of the catalytic triads and lie on the edges of the matched β-strands. In addition, the serines of the catalytic triads are also equivalenced (S195 of 1tpo with S224 of 2prk) and the histidines of the catalytic triads are similarly spatially superimposed, only displaced by 1 residue (H57 of 2prk matched to G70 of 1tpo instead of H69). The cysteines at positions 58 of 1tpo and 73 of 2prk were also matched. Close to the oxyanion hole we find a match between residues G197 (1tpo) and V157

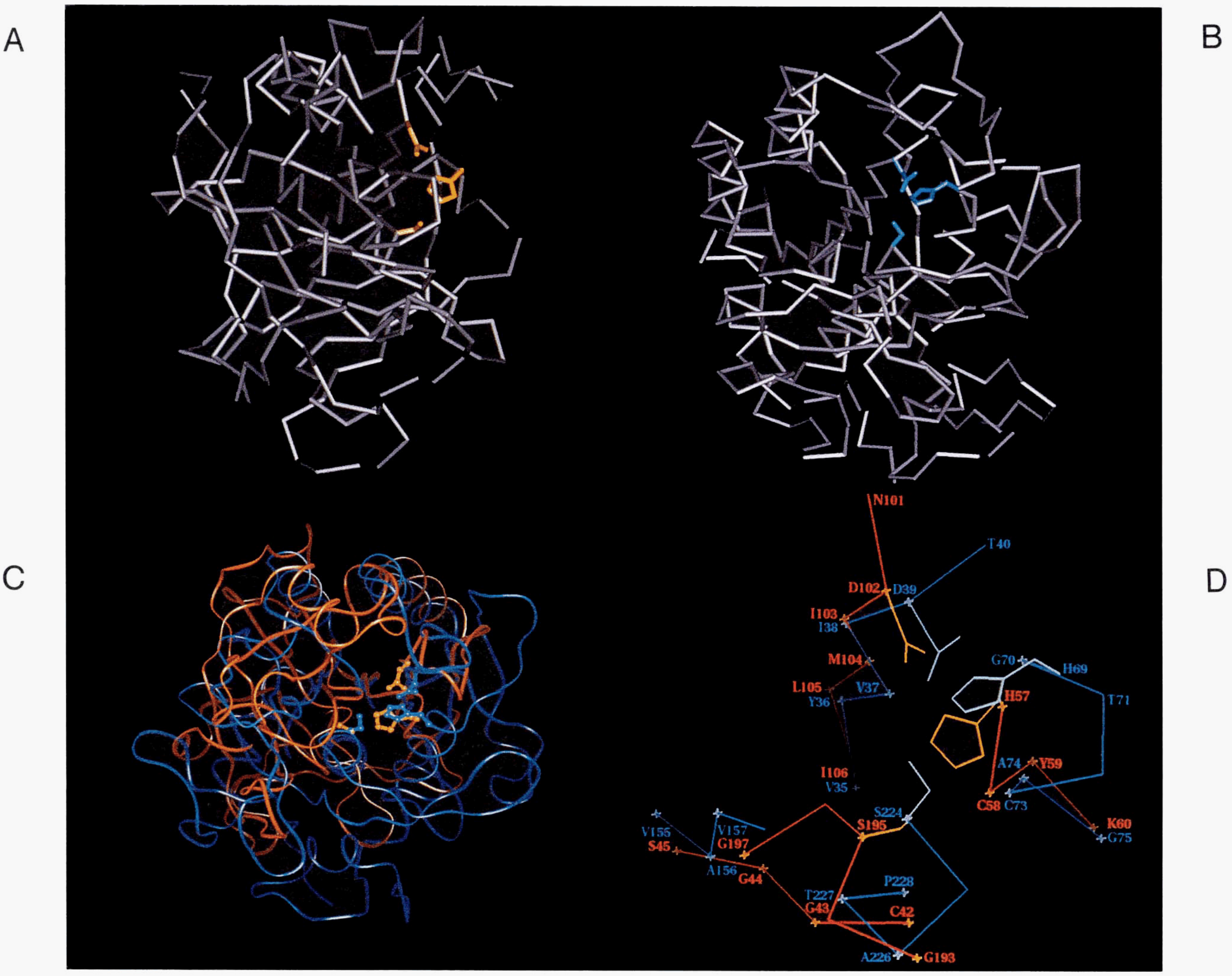

Fig. 2. Structural match between β-trypsin (1tpo) and proteinase K (2prk). **A, B:** Backbones of β-trypsin and proteinase K, respectively. The side chains of the catalytic triads are shown in yellow and blue, respectively. **C:** Ribbon diagram of the superposition of β-trypsin (brown) on proteinase K (blue). The catalytic triads are shown in yellow and light blue, respectively. This figure is viewed from a slightly different angle than A, B, and D. **D:** Details of the match of β-trypsin (orange) and proteinase K (blue). The catalytic triad side chains are shown in yellow and light blue, respectively. A cross represents a matched C_α atom. Only the C_α atoms of each protein were used to obtain the superposition and no information on the location of the catalytic triads was given to the program. As can be seen, the superposition of the C_α's found by the program brings the catalytic triads close together. A least-squares fit of the catalytic triads only produces a slightly better match (results not shown). Some segments of contiguous residues were also matched. These include portions of β-strands and some residues in the loops connecting them. Roughly, the equivalences include portions of the central 4 strands of 2prk and portions of 4 of the strands of the first domain of 1tpo. The actual matching of these segments is listed below:

2prk:	G32	S33	C34	V35	Y36	V37	I38	D39	
1tpo:	K109	L108	K107	I106	L105	M104	I103	D102	
2prk:	H69	G70			C73	A74	G75		
1tpo:		H57			C58	Y59	K60		
2prk:	G152	V153	M154	V155	A156				
1tpo:	N48	I47	L46	S45	G44				
2prk:	T88	Q89	L90	F91	G92	V93			
1tpo:	K86	K87	S88	I89	V90	H92			
2prk:	F113	V114	A115	S116	D117	K118	N119	N120	R121
1tpo:	W237	I238	K239	Q240	T241	I242	A243	S244	N245
2prk:	G126	V127	V128	A129	S130				
1tpo:	Q50	W51	V52	V53	S54				

The first 3 pairs of segments are partially shown in Figure 2D. In addition, the match between β-trypsin and proteinase K involves the equivalence of several isolated residues (not shown).

(2prk) and near the main-chain substrate binding, a match between residues L133 of 2prk and S214 of 1tpo is observed. Besides matching the active sites, the match between 2prk and 1tpo equivalences several isolated residues and short segments of contiguous residues (see legend of Fig. 2). The other subtilisin-like proteases in the database ranked 15, 16, 18, and 19.

Strikingly, actinidin (2act; Baker & Dodson, 1980) and papain (9pap; Kamphuis et al., 1984), 2 sulfhydryl proteases, also rank very high in the scan with chymotrypsin. In Table 1 (and Fig. 1) they rank 14th and 20th, respectively. Similarly, scanning the database with β-trypsin, they are found at the top of the list as well (positions 13th and 17th, respectively). Sulfhydryl proteases are also formed by 2 domains with the active site between them. Below we refer to the match of β-trypsin and actinidin only (the matches between α-chymotrypsin and actinidin, as well as the match between both trypsins and papain are very similar). Remarkably, the active-site residue Ser 195 from 1tpo matches the active-site residue Cys 25 of actinidin (see Fig. 3). Asp 102 from 1tpo is matched to Ser 18 of actinidin, only 1 residue away from the active site residue Gln 19. Gln 19 is actually matched to Ser 214 of 1tpo. Ala 56, 1 residue away from His 57 of β-trypsin, is matched to Trp 184 in actinidin, an active-site residue. The fourth residue in the actinidin catalytic site is His 162. It is matched to Cys 42 of β-trypsin.

The structural similarity between the active sites of these proteases has previously been recognized by visual inspection. Nevertheless, the results presented here are unique. Our method succeeded in finding the rough similarity around the active sites of these proteases automatically, without any prior knowledge of their existence. Except for the match of Cys 25 with Ser 195, the matches obtained between actinidin and trypsin differ from the ones suggested by Garavito et al. (1977). We note, however, that only C_α atoms were compared, whereas the active sites contain mostly side chains. It is striking that although the number of matched pairs is not large (about 100), the matches of trypsin with these non-trypsin-like proteases appear so high in Table 1 (and in Fig. 1). The 12 trypsin-like serine proteases, the 2 sulfhydryl proteases, and the 5 subtilisin-like proteases present in our data set are within the 20 top-ranking scores. The matches of the subtilisin-like and sulfhydryl proteases are completely sequence order-independent, equivalencing single, isolated, though functionally similar residues, lying in loops. They are neither contiguous, and thus do not belong to fragments, nor do they conserve the linear order of the sequences. As such, they could not have been obtained using previously published methods. However, as can be seen from Figures 2 and 3, the matches also contain some segments of contiguous residues. These segments may provide a similar scaffold in which the active sites reside. Note also that some of the matches discussed above rank only marginally above the background ranks of random matches. This is a problem frequently encountered in sequence comparison applications, namely, how to discriminate meaningful matches from background noise, and how deeply into the "twilight zone" should one go.

Scanning the data set of 170 proteins required between 5 CPU min (for the smaller probes) to 8 CPU min (for the larger probes) on a contemporary Silicon Graphics workstation (less than 3 s per comparison on average). The running time of our method grows sublinearly with the probe size. This is due to the efficient organization of the probe's rotational and translational invariants in the hash table.

A recently refined structure with a fold and a catalytic triad similar to those of the trypsin-like serine proteases is the core protein of Sindbis virus (PDB code 1snv; Tong et al., 1993). This entry was not originally included in our data set. 1snv has no significant sequence homology to the trypsins. Comparison between 1tpo and 1snv resulted in 84 matched pairs at an RMS of 1.66 Å and a similarity score of 0.27 (1snv has 151 residues). This score would be ranked 13th in the scan with both α-chymotrypsin and β-trypsin, just below the 12 trypsin-like serine proteases of our data set (see Table 1). Analysis of the match showed it is a linear match (conserving the sequential order of the chains) and that the catalytic triads were properly equivalenced. This match is similar to the one reported by Tong et al. (1993). The latter was obtained by refining a manually determined initial transformation based on the superposition of 9 residues, with 3 residues around each catalytic triad residue. Because (1) all other trypsin-like sequences and structures are relatively similar and (2) the equivalence of the catalytic triads of trypsin with those of other nonhomologous proteases (e.g., subtilisins and sulfhydryl proteases) results in a nonlinear structural superposition, it is remarkable that although the Sindbis virus core protein has no significant sequence homology to the trypsins, the structural match fully conserves the linear order of the sequences. Because our method carries out the comparison without an initial equivalence and ignores the sequential order of the chains, this provides stronger evidence in favor of divergent evolution than a result obtained from a comparison which requires that the linear order of the sequences be conserved.

Other protease scans

To check the consistency of our results and scoring function, we have scanned the data set using 2 proteases that ranked highly in the comparisons with both the chymotrypsin and the trypsin as probes. Scanning with subtilisin as a probe, the first 5 matches correspond to the 5 subtilisin-like entries in the data set (3 subtilisins, proteinase K, and thermitase). Several matches belonging to α/β structures in the data set, along with both trypsin-like serine proteases and sulfhydryl proteases, rank next.

As a second experiment, and in order to demonstrate the speed of our method, we have scanned the complete Protein Data Bank. All chains of all entries in the data bank were generated. Only those chains that are almost identical were excluded. The resulting database contains 1,191 entries. Here the structure of actinidin, a sulfhydryl protease, is deployed as a probe of the database. As expected, the largest matches correspond to other sulfhydryl proteases in the database. These are followed by serine proteases, including trypsin-like and subtilisin-like entries. This complete PDB database scan required less than 40 CPU min and is shown in Figure 4. A supplementary table listing the names, number of matched residues, and scores of all the hits of this scan with a similarity score above 0.2 appears on the Diskette Appendix.

Discussion

Advantages of 3D sequence order-independent structural comparison

We have demonstrated the advantages inherent in our method as outlined above, namely, its ability to detect spatial similar-

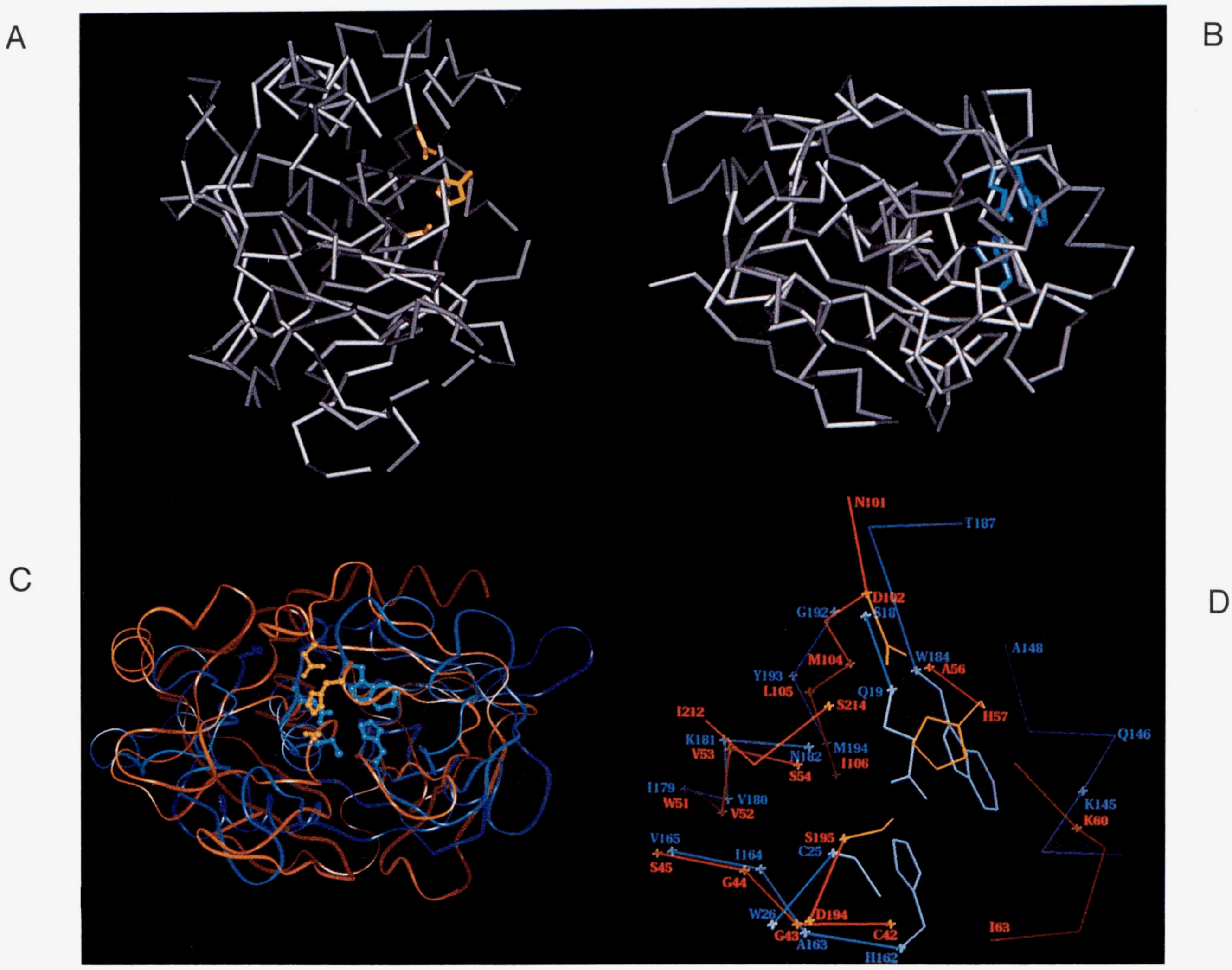

Fig. 3. Three-dimensional match between β-trypsin (1tpo) and actinidin (2act). **A, B:** Backbones of β-trypsin and actinidin, respectively. The side chains of the catalytic triads are shown in yellow and blue, respectively. **C:** Ribbon diagram of the superposition of β-trypsin (brown) on actinidin (blue). The catalytic triads are shown in yellow and light blue, respectively. This figure is viewed from a slightly different angle than that of A, B, and D. **D:** Detail of the match of β-trypsin (orange) and actinidin (blue). The catalytic triad side chains are shown in yellow and light blue, respectively. A cross represents a matched C_α atom. Only the C_α atoms of each protein were used to obtain the superposition, and no information on the location of the catalytic triads was given to the program. As can be seen, the superposition of the C_α's found by the program brings the catalytic triads close together. A least-squares fit of the catalytic triads only produces a slightly better match (results not shown). Some segments of contiguous residues were also matched. These include portions of β-strands and some residues in the loops connecting them. The actual matching of these segments is listed below:

2act:	H162	A163	I164	V165	I166	V167	G168	Y169
1tpo:	C42	G43	G44	S45	L46	I47	N48	S49
2act:	W178	I179	V180	K181	N182	S183	W184	
1tpo:	Q50	W51	V52	V53	S54	A55	A56	
2act:	G192	Y193	M194	R195	I196			
1tpo:	M104	L105	I106	K107	L108			
2act:	C25	W26						
1tpo:	S195	D194						
2act:	T153	F152	I151					
1tpo:	S84	A85	S86					
2act:	V133	S134	V135	A136	L137	D138		
1tpo:	Y29	Q30	V31	S32	L33	N34		

The first 4 pairs of segments are partially shown in Figure 3D. In addition, the match between β-trypsin and actinidin involves the equivalence of several isolated residues (not shown).

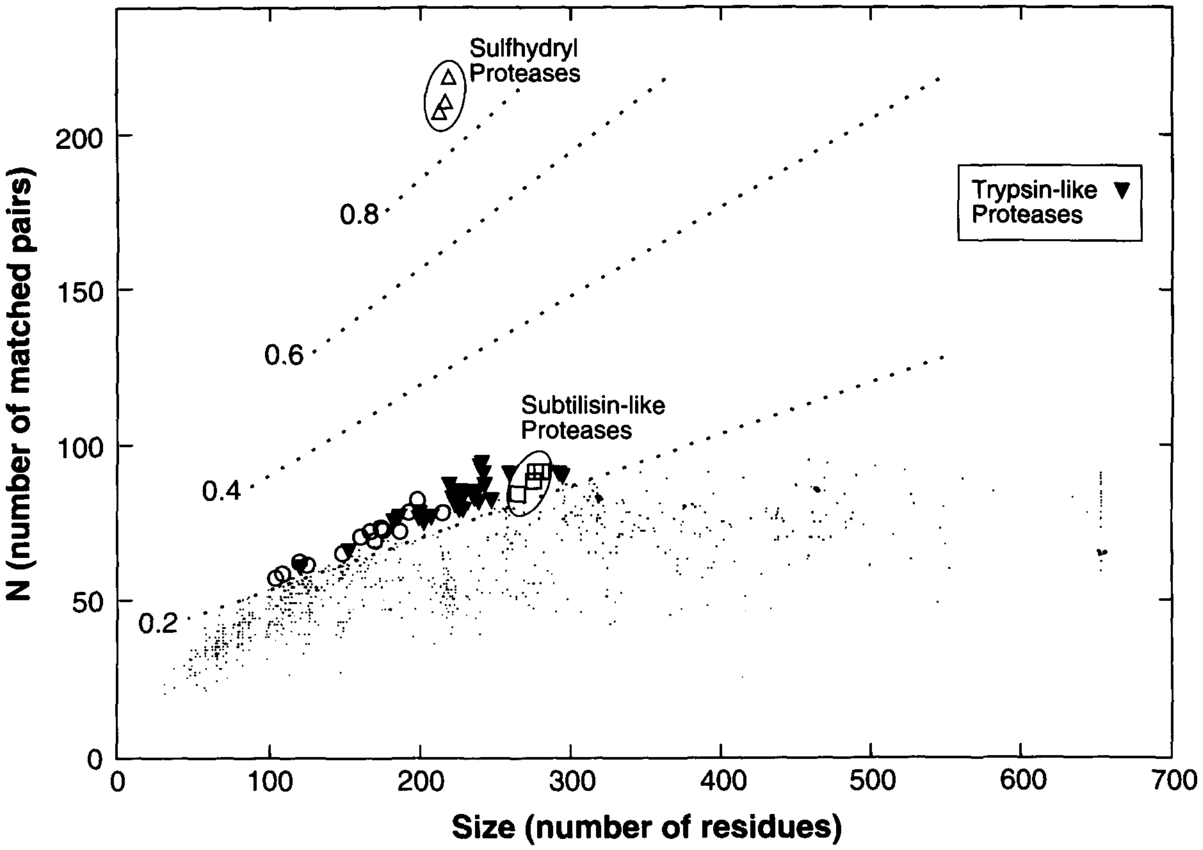

Fig. 4. The results obtained from the comparison of actinidin (2act) against the structural database containing 1,191 selected proteins from the PDB. Only 1 entry was selected if several almost identical chains exist for the same structure. This data set does not contain all the entries included in the 170-protein data set used in the other scans. Each dot in the figure represents 1 comparison between actinidin and one of the data set proteins. The x-axis is the size (in number of residues) of the protein compared and the y-axis the number of matched pairs found in the comparison with actinidin. A similarity score is computed for each comparison (see legend of Fig. 1). Similarity levels (dotted lines) are plotted at the values 0.2, 0.4, 0.6, and 0.8. Only 877 out of the 1,191 proteins in the data set are shown in this figure. Lower similarity scores were obtained for the remaining 314 proteins. The highest ranking scores correspond to the other sulfhydryl proteases (marked by empty triangles) in the data set. Following these, many trypsin-like proteases are found (noted by inverted triangles). The subtilisin-like proteases are also within the top scores (depicted by empty squares). Among the highest scores, 15 nonrelated proteins are also found (shown as empty circles). These are listed below. All other proteins are shown by small dots. Within the 40 top-ranking proteins, we find the 3 sulfhydryl proteases in the database, 24 trypsin-like and 4 subtilisin-like serine proteases. Sizes above 650 are shown at the 650 mark. The scan of 1,191 proteins required 38 min CPU time on a Silicon Graphics Indigo workstation, or about 2 s per pairwise comparison. The 15 nonrelated proteins ranking among the trypsin and subtilisin like proteases are (listed in order of decreasing score): 1cola, 1grcb, 5p21, 2gcr, 1gpla, 1ovb, 2fcr, 1rslc, 1akea, 4gcr, 2fx2, 1ofv, 1bbc, 2tir, 1rnla.

ity between evolutionary convergent or divergent structures, matching isolated residues regardless of their sequence order, and the speed with which these comparisons are carried out. We foresee that our method will enable routine comparisons of any new structure—whether determined crystallographically, by NMR, or computationally—against the database of 3D structures, much in the same manner as investigators today compare a newly determined protein or DNA sequence with the sequence database. During the last decade, sequence comparisons have provided us with a wealth of information and an insight into evolutionary and functional aspects of biological macromolecules. Structural comparisons advance us considerably further. Different sequences may result in similar folds. Ultimately, it is the structure that is recognized and that plays a critical role in carrying out the necessary biological functions.

Implications for protein folding and rational drug design

The implications of the availability of such a tool are numerous. They range from applications to the protein folding problem to searches for pharmacophoric patterns and thus to computer-aided drug design. Investigations that may potentially aid in studies of protein folding include both novel analyses of the structural database as well as evaluations of test structures (e.g., Bowie et al., 1991).

Specifically, (1) using this methodology, a nonredundant, 3D structural database is already in the process of being constructed. (2) Using this structural database we can compile and catalogue recurring 3D motifs. Such motifs may represent particularly stable folding units. (3) In the derivation and analysis of the 3D motifs, the exchangeability of amino acid types at analogous positions in space in different proteins may be noted. (4) Interresidue potential functions, routinely used in the evaluation of test structures, have been derived from the frequencies of occurrence of neighboring amino acid pairs. The nature of their environment was disregarded, owing to the large combinatorial complexity that such a task would have entailed. This information is straightforwardly obtained from analysis of spatial motifs. (5) Because our technique can handle all atoms, regardless of their connectivity, side-chain packing within these motifs can be included as well. In such studies we can focus on a particular atom—or group of atoms—type. (6) Furthermore, inclusion of information regarding amino acid side chains is expected to improve the quality of the motifs obtained. Here, side-chain orientation, size, and information pertaining to its environment may be a fruitful direction to consider. Indeed, such considerations are likely not only to ameliorate the quality of the 3D motifs, but to speed up their detection as well. The inclusion of such descriptors is expected to facilitate detection of motifs between further diverged—evolutionarily or functionally—proteins. First steps in this direction are already being taken. (7) Test-folded structures may be scanned, examining their potential similarity to 3D motifs found in the database. The presence of a functional or stable folding unit, may be indicative of the "goodness" of the test structure. (8) Classes of different structures generated using energy minimization calculations may share a common 3D substructure. The existence of such a substructure may provide an insight into the process of protein folding. (9) The availability of such a fast technique for examination of protein structures affords intensive comparisons of structures generated from a variety of random sequence types. The significance of observed motifs and of residue pairs can be better gauged statistically as well.

Our tool also enables fast searches of databases of drugs. Analysis of small molecules binding to the same (or similar) receptor(s) is expected to result in (nonpredefined) pharmacophoric patterns. Also, conversely, analysis of the surfaces of receptors that bind similar ligands may detect a surface motif (Fischer et al., 1993a). However, because surface atoms (residues) are much more flexible than those located in the interior of the protein molecules, detection of a surface motif is a more difficult problem. Still, our applications of this computer vision-based technique to the surface comparison problem (Fischer et al., 1993a) and to the protein-protein or protein-small molecule ligand docking problem (Lin et al., 1994; Norel et al., 1994) have been successful.

Acknowledgments

We thank Drs. R. Jernigan, D. Covell, and in particular, J. Maizel, for discussions and suggestions. R.N. has been sponsored by the National Cancer Institute, DHHS, under contract 1-CO-74102 with Program Resources, Inc. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. H.W. has been supported in part by grant 83-00481 from the US-Israel Binational Science Foundation (BSF), Jerusalem, Israel. R.N. has been supported in Israel in part by grant 91-00219 from the BSF. H.W. and R.N. were supported by the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities. This work formed part of the Ph.D. thesis of D.F., Tel Aviv University.

References

Alexandrov NN, Takahashi K, Go N. 1992. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J Mol Biol 225*:5–9.

Bachar O, Fischer D, Nussinov R, Wolfson HJ. 1993. A computer vision based technique for 3D sequence independent structural comparison of proteins. *Protein Eng 6*(3):279–288.

Baker EN, Dodson EJ. 1980. Crystallographic refinement of the structure of actinidin at 1.7 Å resolution by fast Fourier least-squares methods. *Acta Crystallogr 36A*:559.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers GR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Betzel C, Pal GP, Saenger W. 1988. Synchrotron X-ray data collection and restrained least-squares refinement of the crystal structure of proteinase K at 1.5 Å resolution. *Acta Crystallogr 44B*:163.

Bowie JV, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Branden C, Tooze G. 1991. *Introduction to protein structure*. New York/London: Garland Publishing Inc.

Fischer D, Bachar O, Nussinov R, Wolfson HJ. 1992. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J Biomol Struct Dyn 9*(4):769–789.

Fischer D, Norel R, Wolfson H, Nussinov R. 1993a. Surface motifs by a computer vision technique: Searches, detection and implications for protein-ligand recognition. *Proteins Struct Funct Genet 16*:278–292.

Fischer D, Wolfson HJ, Nussinov R. 1993b. Spatial, sequence-order-independent structural comparison of α/β proteins: Evolutionary implications. *J Biomol Struct Dyn 11*(2):367–380.

Fujinaga M, Sielecki AR, Read RJ, Ardelt W, Laskowski M Jr., James MNG. 1987. Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 Å resolution. *J Mol Biol 195*:397–418.

Garavito RM, Rossmann MG, Argos P, Eventoff W. 1977. Convergence of active center geometries. *Biochemistry 16*:5065–5071.

Grindley HM, Artymiuk PJ, Rice DW, Willet P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol 229*:707–721.

Kamphuis IG, Kalk KH, Swarte MBA, Drenth J. 1984. Structure of papain refined at 1.65 Å resolution. *J Mol Biol 179*:233–256.

Kraut J, Robertus JD, Birktoft JJ, Alden RA, Wilcox PE, Powers JC. 1972. The aromatic substrate binding site in subtilisin/BPN and its resemblance to chymotrypsin. *Cold Spring Harbor Symp Quant Biol 36*:117–123.

Lamdan Y, Schwartz JT, Wolfson HJ. 1988. On recognition of 3D objects from 2-D images. In: *Proceedings of IEEE International Conference on Robotics and Automation, Philadelphia, Pennsylvania, April 1988*. pp 1407–1413.

Lin SL, Nussinov R, Fischer D, Wolfson H. 1994. Molecular surface representation by sparse critical points. *Proteins Struct Funct Genet 18*:94–101.

Matthews BW, Rossmann MG. 1985. Comparison of protein structures. *Methods Enzymol 115*:397–420.

Mitchel EM, Artymiuk PJ, Rice DW, Willet P. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol 212*:151–166.

Norel R, Fischer D, Wolfson H, Nussinov R. 1994. Molecular surface recognition by a computer vision based technique. *Protein Eng 7*:39–46.

Nussinov R, Wolfson HJ. 1991. Efficient detection of three-dimensional motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA 88*:10495–10499.

Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol 208*:1–22.

Tong L, Wengler G, Rossmann MG. 1993. Refined structure of sindbis virus core protein and comparison with other chymotrypsin-like serine proteinase structures. *J Mol Biol 230*:228–247.

Vriend G, Sander C. 1991. Detection of common three-dimensional substructures in proteins. *Proteins Struct Funct Genet 11*:52–58.