J Mol Model (2002) 8:199–207
DOI 10.1007/s00894-002-0083-z

ORIGINAL PAPER

Shin-ichi Nakayama · Peter Willett

A sphere-based descriptor for matching protein structures

Received: 1 November 2000 / Accepted: 3 April 2002 / Published online: 7 June 2002

Abstract This paper describes the use of a descriptor based on the number of α-carbon atoms within a sphere centered on the α-carbon in each amino acid residue in a protein. The descriptor can be used instead of the residue types in a dynamic programming algorithm, thus providing an efficient way of aligning protein structures. The method is applied to the alignment of protein families and to database searching. The results indicate that the method can quickly align protein sequences considering the 3D structure and can find proteins that are 3D structurally similar and dissimilar to the target protein.

Keywords Similarity score · Residues numbers in a radius · 3D structure · Dynamic programming

Introduction

The identification of alignments between pairs of proteins is of importance for applications such as the evolution of proteins, the active sites of enzymes, and protein engineering. Many methods have hence been developed for the comparison of protein sequences and protein structures. The classical approach is based on the dynamic programming algorithm of Needleman and Wunsch. [1] This uses a similarity matrix, such as the PAM-250 matrix (Dayhoff et al. [2]), in which the *IJ*th element describes the mutation probability between amino acids *I* and *J*. The Needleman–Wunsch algorithm aligns two protein sequences so as to maximize the sum of the similarity scores between matching pairs of amino acids. In this alignment, the sum is decreased when the insertion or deletion of amino acid residue(s) occurs. Increases in the efficiency of the basic algorithm have been described by Lipman and Pearson [3] and Murata, [4] inter alia.

There has also been much interest in methods for scoring alignments of protein structures. The three-dimensional (3D) structures of proteins have been characterized in many ways: the difference distance matrix of α-carbon atoms (Nishikawa and Ooi [5]); vectors (Rossman and Argos [6]); the root-mean-square deviation (RMSD) of α-carbon atoms in a segment (Remington and Matthews [7]); and torsional angles (Karpen et al. [8]). These methods are, again, all demanding of computational resources, although improved approaches have been described (Holm and Sander; [9] Lessel and Schomburg [10]). Descriptors based on the secondary structure of proteins have proved to be very popular (Murthy [11]), with the vectorial representation of secondary structures suggested by Abagyan and Maiorov [12] forming the basis for several efficient matching algorithms based on graph theory (Grindley et al.; [13] Mitchell et al. [14]). Efficient dynamic programming procedures that use local α-carbon interatomic distances have been reported by Taylor and Orengo. [15] Recently, Kawabata and Nishikawa [16] proposed the hierarchical alignment method using secondary structure elements, environmental states, and residue–residue distance.

Nishikawa and Ooi [17] have reported the prediction of protein structures from their sequences using a descriptor that is based on the number of α-carbon atoms located within a sphere of a user-defined radius centered on the α-carbon atom of each amino acid residue. This descriptor characterizes the local structural environment of each α-carbon and can be calculated very rapidly. Here we discuss the use of the descriptor for the calculation of similarity scores by means of a dynamic programming approach; the effectiveness of the descriptor is illustrated by its use for the alignment of protein families and for database searching. This method is similar to that of Taylor and Orengo [15] in the methodology. Their descriptor is based on the sum of differences of distances

S.-i. Nakayama (✉)
University of Library and Information Science,
1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
e-mail: nakayama@ulis.ac.jp

P. Willett
Krebs Institute for Biomolecular Research
and Department of Information Studies,
University of Sheffield, Western Bank,
Sheffield S10 2TN, UK

of two α-carbons aligned by amino acid type. The number of α-carbon atoms within a sphere could describe a local 3D structure considering not only similar sequences but also unrelated ones.

Calculation of the similarity score and alignment

For each residue in a protein structure, a count is made of the number of α-carbon atoms within a sphere of radius r Å centered on the α-carbon of the chosen residue. Let the values of this number for the ith residue in one protein, a, and the jth residue in another protein, b, be N_{ai} and N_{bj}, respectively. Then the similarity score for these two residues, S_{ij}, is given by

$$\mathrm{IF}|N_{ai} - N_{bj}| > \mathrm{minimum}\{N_{ai}, N_{bj}\}\,\mathrm{THEN} S_{ij} = 0$$

$$\mathrm{ELSE} S_{ij} = 1 - \left(|N_{ai} - N_{bj}|/\mathrm{minimum}\{N_{ai}, N_{bj}\}\right)$$

This equation normalizes the absolute difference of the numbers by using the smaller of the two values at each position. These scores act as the input to the fast dynamic programming algorithm described by Murata, [4] which then calculates the alignment and the similarity score for the pair of proteins a and b. A C implementation of the method takes about 0.3 CPU seconds (when implemented in C on a Unix workstation with an R4000SC processor) to process a pair of 300-residue proteins.

The method, hereafter referred to as the *sphere method*, requires the specification of the radius, r, for the calculation of the scores, S_{ij}, and experiments were hence carried out to determine an appropriate value for r. These experiments used the proteins deoxyhemoglobin and carboxyhemoglobin, which have the same amino acid sequences but slightly different structures (4HHB and 1HCO, respectively). The similarity values were calculated at each position in the sequence using values for r of 6, 10 and 14 Å. These values are plotted in Fig. 1 against the corresponding sequence position, and it will be seen that there are noticeable differences in the scores, i.e., in the numbers of adjacent α-carbons, as r is varied. The similarity values at 6 Å radius are very low at the terminal residues and at about residues 40 and 90. These latter locations represent places where carbon monoxide is attached in carboxyhemoglobin, this attachment causing a slight distortion of the structure that is detected when a radius of 6 Å is employed. The similarity values at the larger radii are noticeably larger and less variable in magnitude than when 6 Å is used. This suggests that, as would be expected, scores based on larger radii are more sensitive to differences in the overall shapes of entire proteins than they are to local differences.

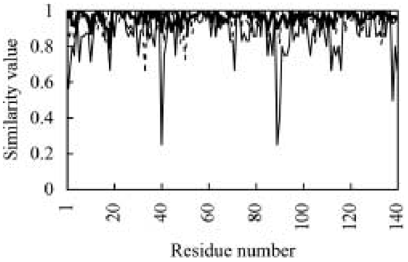

Fig. 1 Similarity values of each residue number for comparison of human deoxyhemoglobin to carbonmonoxyhemoglobin. The solid line, broken line, and bold solid line are for a radius of 6, 10, and 14 Å, respectively

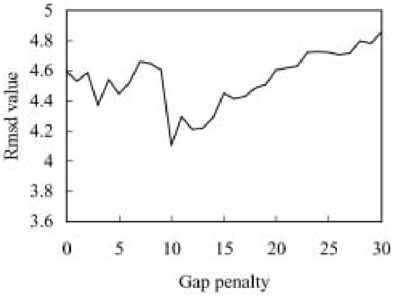

Fig. 2 Effect of gap penalty on RMSD value in pepsin and penicillopepsin by alignment using the mutation-matrix method

Aligning protein families

The RMSD is widely used for comparing the 3D structures of similar proteins (Schulz [18]). Its calculation requires the two protein structures to be aligned, but the available methods for this purpose (e.g., May and Johnson [19]) are very time-consuming. The method described here provides an effective and an efficient way of generating such alignments. This is demonstrated by consideration of alignments for the acid proteases, which include porcine pepsin (5PEP), penicillopepsin (3APP), endothiapepsin (4APE), chymosin b (4CMS), and rhizopuspepsin (6APR). In this investigation, RMSD values were calculated as the square root of the sum of the different distances between the aligned residues. The RMSD values based on the alignments resulting from our method were compared with alignments resulting from the dynamic programming algorithm described by Murata [4] using a conventional mutation matrix, PAM-250. [2] The latter alignments, which consider only the sequence of a protein, are restricted to the identification of evolutionary similarity and cannot take account of the structural similarity that is explored by the sphere method.

The alignments resulting from both methods, the sphere procedure and the mutation-matrix procedure, depend on the gap penalty that is used in the dynamic programming algorithm. The gap penalty means the decreasing value for the insertion or deletion of amino acid residue(s) in the alignment. We hence calculated RMSD values for the two types of alignment using a range of gap penalties, as shown in Fig. 2 (for the mutation matrix) and Fig. 3 (for the sphere method), with the latter runs also involving changes in the radius, r. For this pair of proteins, the minimum in the mutation-matrix plot is at a gap penalty of 10; other pairs of proteins gave differ-

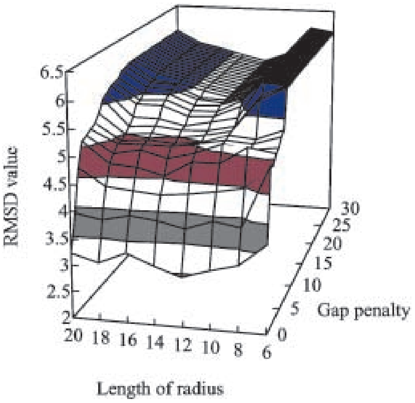

Fig. 3 Effects of length of radius and gap penalty on RMSD value in pepsin and penicillopepsin by alignment using the sphere method

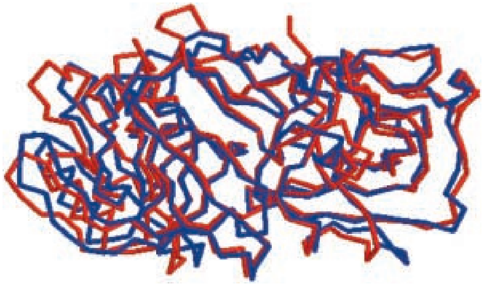

Fig. 4 Superposition of endothiapepsin (*blue*) and rhizopuspepsin (*red*)

Table 1 RMSD values for aligning acid protease structures. The first value in each element of the table is the value obtained with the sphere method described here and the second value that obtained with a conventional mutation matrix. The parenthetic value after the first value is the *r* value (Å) that gives the RMSD value

	4CMS	6APR	4APE	3APP
5PEP	1.50(16) 1.85	2.58(10) 4.10	3.42(14) 4.03	2.94(12) 4.10
4CMS		2.27(10) 3.38	2.92(16) 3.55	2.54(10) 3.38
6APR			2.41(16) 4.17	2.50(12) 3.73
4APE				1.84(12) 1.88

ent values. The sphere method gave the smallest RMSD values when the gap penalty was set to 0, for all values of *r*, with the overall minimal value being obtained with *r* set to 12 Å.

Table 1 lists the RMSD values for the sphere and mutation-matrix alignment methods when applied to the pairs of acid protease. The RMSD values quoted for both methods are the smallest that were obtained as the gap penalty was varied in the mutation-matrix method and as the *r* value was varied in the sphere method. The *r* values of the sphere method are shown in parentheses after the RMSD values. It will be seen that the sphere method yields a smaller RMSD, i.e., a better alignment, than does the mutation-matrix method for all pairs of acid proteases. The *r* values of overall minimal value place a range of 8 to 16 Å in the above protein pairs. Nishikawa and Ooi [17] obtained 14 Å for the radius in their protein-folding study. Our results give a similar *r* value to their value.

The alignments of the proteases resulting from the sphere method are shown in Table 2. The alignments were generated in the order defined by the RMSD values, i.e., chymosin b with porcine pepsin, then endothiapepsin with penicillopepsin, then rhizopuspepsin with chymosin b, and endothiapepsin with rhizopuspepsin. The alignments were very similar to those resulting from the standard mutation matrix, with the exception of some proline and glycine residues that were aligned using the mutation matrix. Prolines and some glycines generally occur in turns, which are relatively flexible and which are thus sometimes placed in different positions. The sphere method could find neighboring residues in turns as structural information is used instead of sequence information. Even so, the alignment was successful in matching residues that are important for enzyme activity, such as Asp-32 and Asp-215 in porcine pepsin. The superposition of endothiapepsin and rhizopuspepsin is determined from this alignment and is shown in Fig. 4. The superposition obtaining is fitted for overlapping of the two structures. The calculation time is less than 1 s (R4000SC processor) although the method of May and Johnson [19] required about 30 min (R3000 processor) for protein pairs with half the number of amino acid residues. It is shown that the sphere method could quickly align two protein sequences considering their 3D structures.

Database searching

We now describe the use of the sphere method to match several different target proteins against a set of protein structures taken from the Protein Data Bank (PDB) (Berman et al. [20]). The recent PDB contains more than 10,000 structures, but it is too large for the following investigations. Thus, we use the old version of the PDB (April 1992) and select a set of 890 structures (containing a total of 1,360 chains) for which fully determined sequences are available. The first searches used azurin (1AZU), which is a moderate-size protein containing 126 residues, and porcine pepsin (5PEP), which is a large protein containing 326 residues, as the target proteins for which matching structures were required. The efficiency of the sphere method is demonstrated by its requiring ca. 10 CPU minutes on the SONY NEWS-5000 workstation, which uses an R4000SC processor, to calculate the similarity scores between porcine pepsin and all the proteins in the 890-member subfile.

In the search system, only one *r* value should be determined to calculate the similarity score quickly. From Table 1, the *r* value seems to be suitable between 8 to 16 Å. We investigated other protein pairs, two immunoglobulins, two hemoglobins, and immunoglobulin and hemoglobin. The results are shown in Fig. 5. The results

Table 2 Alignment of the acid proteases. 5PEP: porcine pepsin, 4CMS: chymosin b, 6APR: rhizopuspepsin, 4APE: endothiapepsin, 3APP: penicillopepsin. Underlines mean the aligned amino acid by using the mutation matrix

```
5PEP --I-GDEPLENYL--DTEYFGTIGIGTP-AQD-FTVIFDTGSSNLWVPSVYCSSLACSDHNQFNPDD
4CMS -GEVASVPLTNYL--DSQYFGKIYL-GTPPQE-FTVLFDTGSSDFWVPSIYCKSNACKNHQRFDPRK
6APR AGVGTVPMTDYGN--DIEYYGQVTIGTP-GKK-FNLDFDTGSSDLWIASTLCT-NCGSGQTKYDPNQ
4APE --STGSATTTPIDSLDDAYITPVQI-GTPA-QTLNLDFDTGSSDLWVFSSETTASEVDGQTIYTPSK
3APP -AASGVATNTPTAN-DEEYITPVTI-GG-T-T-LNLNFDTGSADLWVFSTELPASQQSGHSVYNPSA
-----------------------------------------------------------------------
SS-TFEATSQELSIT-Y-G-TGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGILGLAYPSI
SS-TFQNLGKPLSIH-Y-G-TGSMQGILGYDTVTVSNIVDIQQTVGLSTQEPGDVFTYAEFDGILGMAYPSL
SS-TYQADGRTWSIS-Y-GDGSSASGILAKDNVNLGGLLIKGQTIELAKREAA-SFASGPNDGLLGLGFDTI
STTAKLLSGATWSISYG-D-GSSSSGDVYTDTVSVGGLTVTGQAVESAKKVSSSFTEDSTIDGLLGLAFSTL
-T-GKELSGYTWSIS-YGD-GSSASGNVFTDSVTVGGVTAHGQAVQAAQQISAQFQQDTNNDGLLGLAFSSI
-----------------------------------------------------------------------
S-AS--G-ATPVFDNLWDQGLVS-QDLFSVYLSSNDDSGSV-VLLGGIDSSYYTGSLNWVPVS-VEGYWQIT
A-SE--Y-SIPVFDNMMNRHLVA-QDLFSVYMDRNGQ-ESM-LTLGAIDPSYYTGSLHWVPVT-VQQYWQFT
T-TV--RGVKTPMDNLISQGLISRPIFGVYLGKAKNG-GGGEYIFGGYDSTKFKGSLTTVPIDNSRGWWGIT
NTVSPTQ-QKTFFDNAKAS-L-D-SPVFTADLGY-HA-PGT-YNFGFIDTTAYTGSITYTAVSTKQGFWEWT
NTVQPQS-QTTFFDTVKSS-L-A-QPLFAVALKH-QQ-PGV-YDFGFIDSSKYTGSLTYTGVDNSQGFWSFN
-----------------------------------------------------------------------
LDSITMDGETIACSGGCQAIVDTGTSLLTGPTSAIANIQSDI--GASE-NSDG-EMVISCSSIASLPDIVFT
VDSVTISGVVVACEGGCQAILDTGTSKLVGPSSDILNIQQAI--GATQ-NQYG-EFDIDCDNLSYMPTVVFE
VDRATVGTSTVA-S-SFDGILDTGTTLLILPNNIAASVARAY--GASD-NG-D-GTYTISCDTSAFKPLVFS
STGYAV-GSGTFKSTSIDGIADTGTTLLYLPATVVSAYWAQVSGAKSSSS--V-GGYVFPCSAT-LPSFTFG
VDSYTA-GSQS--GDGFSGIADTGTTLLLLDDSVVSQYYSQVSGAQQD-S—NAGGYVFDCSTN-LPDFSVS
-----------------------------------------------------------------------
INGVQYP-LSPSAY-ILQ-DD-DSCTSGF--EGMDVPTSSGELWILGDVFIRQYYTVFDRAN-NKVGLAPVA
INGKMYP-LTPSAY-TSQ-DQ-GFCTSGF--QS----EQ--KW-ILGDVFIREYYSVFDRAN-NLVGLAKAI
INGASFQVSPDSLV-FEE-FQ-GQCIAGFG-YG----NWG-FA-IIGDTFLKNNYVVFNQGV-PEVQIAPVA
VGSARIV-IPGDYIDFGPISTGSSSCFGGIQSSA---GIG-IN-IFGDVALKAAFVVFNGATTPTLGFASK-
ISGYTAT-VPGSLINYGPSGD-GSTCLGGIQSNS---GIG-FS-IFGDIFLKSQYVVFDSDG-PQLGFAPQA
-----------------------------------------------------------------------
----
----
EVVA
----
----
```

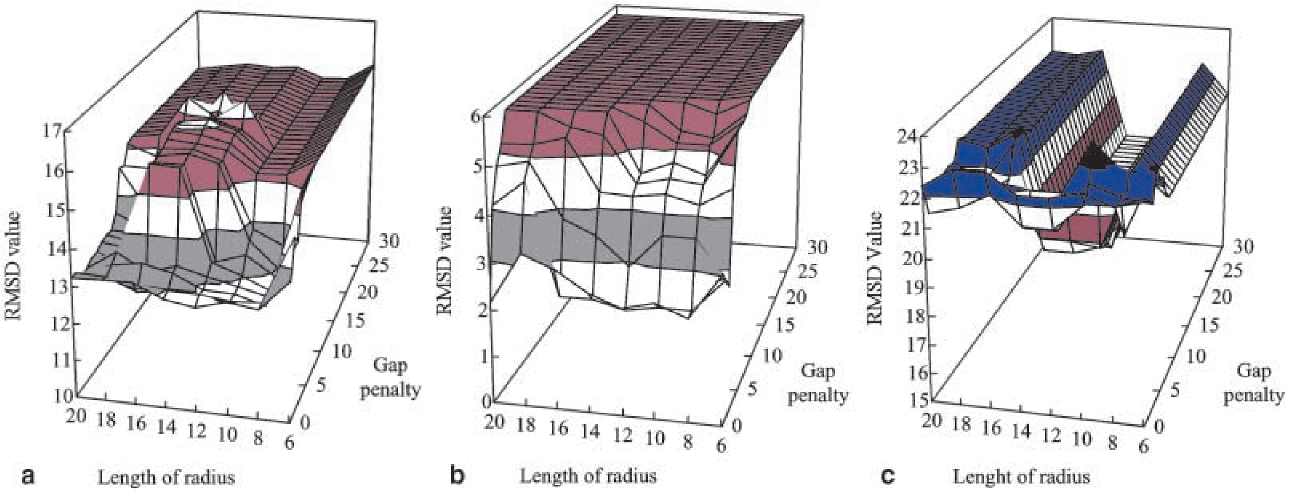

Fig. 5 Effects of length of radius and gap penalty on RMSD value in (**a**) human immunoglobulin G1 Fc fragment A chain and mouse immunoglobulin MC/pc603 Fab fragment L chain, (**b**) human carbonmonoxyhemoglobin α chain and β chain, and (**c**) human immunoglobulin G1 Fc fragment A chain and human carbonmonoxyhemoglobin α chain by alignment using the sphere method

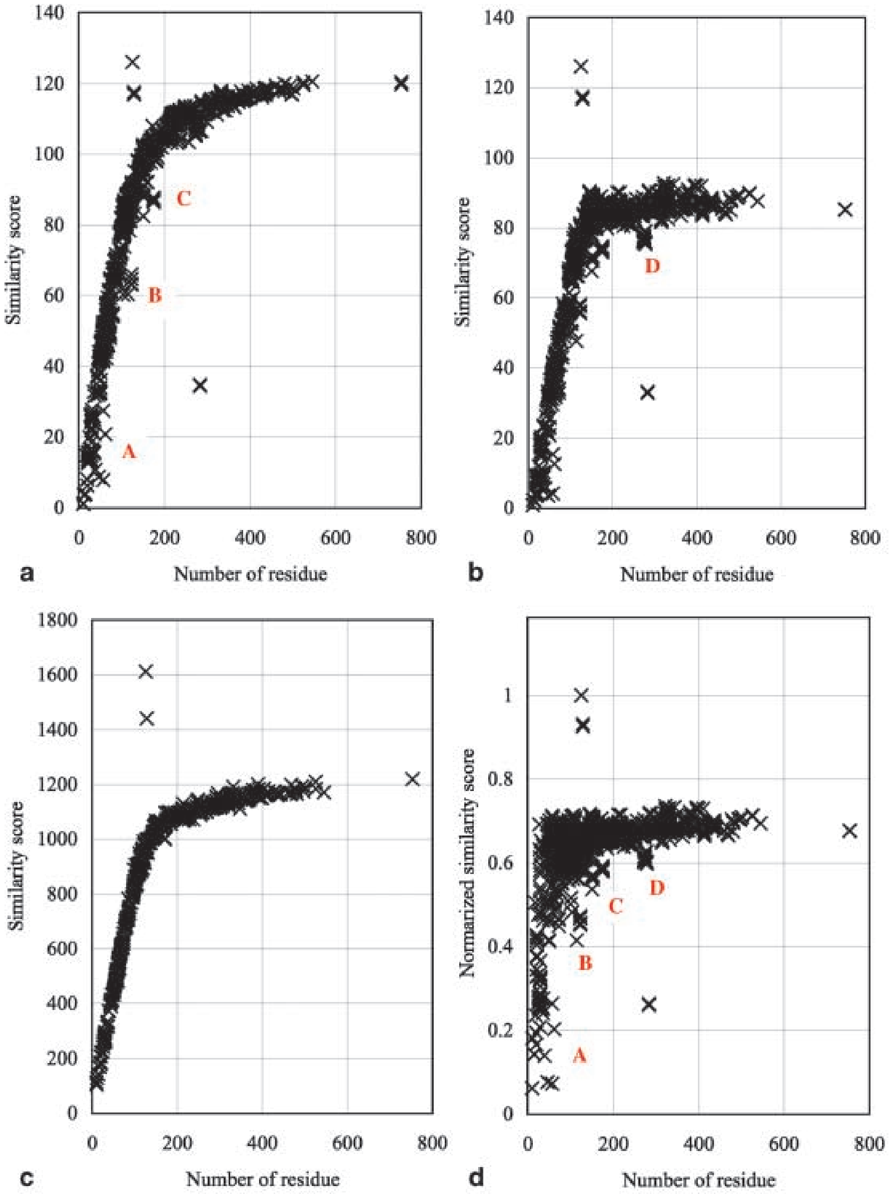

Fig. 6 Similarity scores of proteins in PDB to azurin using the sphere method with a radius of 12 Å and with the gap penalty set to (**a**) 0, and (**b**) 4, (**c**) using the mutation-matrix method with the gap penalty value is set to 6, and (**d**) normalized version of (**b**)

for same family proteins (Fig. 5a and b) are similar to those in Fig. 3 and give small RMSD values in the range of 8 to 16 Å of *r* value. The results for different family proteins (Fig. 5c) are quite different to the other figures and give a small RMSD value in the range of 10 to 14 Å of *r* value. We determined the r value of 12 Å as the central value of the above ranges.

The results of experiments with different values for the gap penalty are shown in Figs. 6 and 7, which plot the similarity score between the target protein (azurin and porcine pepsin, respectively) and a PDB structure against the number of residues in that structure. The plots have a characteristic shape, with the score rising rapidly as larger and larger database structures are considered and then leveling off. The fact that the scores vary with the length means that we cannot use the scores to rank the structures in decreasing order of shape match, as can be done with similarity measures such as the maximal common subgraphs used by Grindley et al.; [13] instead, the scores here are used to highlight proteins with scores that are notably different from those of other proteins of a similar length.

Figure 6a summarizes a search with azurin with the sphere radius set to 12 Å and the gap penalty set to zero. The interesting proteins are those that are well separated from the primary curve, and it will be seen that the first such proteins are at about 150 residues: these represent the structure of azurin itself and two chains of its oxidized form, and it is thus hardly surprising that these are more similar to the target than most of the other structures in the search file. Tropomyosin (2TMA) is the protein lying well below the curve at 284 residues. This contains only helix secondary structure elements and has a long, thin shape that is radically different from the short, column-like shape of azurin. Other proteins lying below the main curve, but much closer to it than tropomyosin, are found at lengths of about 50, 100 and 150 residues. The first group of proteins (marked A in

Fig. 7 Similarity scores of proteins in PDB to pepsin using the sphere method with a radius of 12 Å and with the gap penalty set to (**a**) 0, and (**b**) 4, (**c**) using the mutation-matrix method with the gap penalty value is set to 6, and (**d**) normalized version of (**b**)

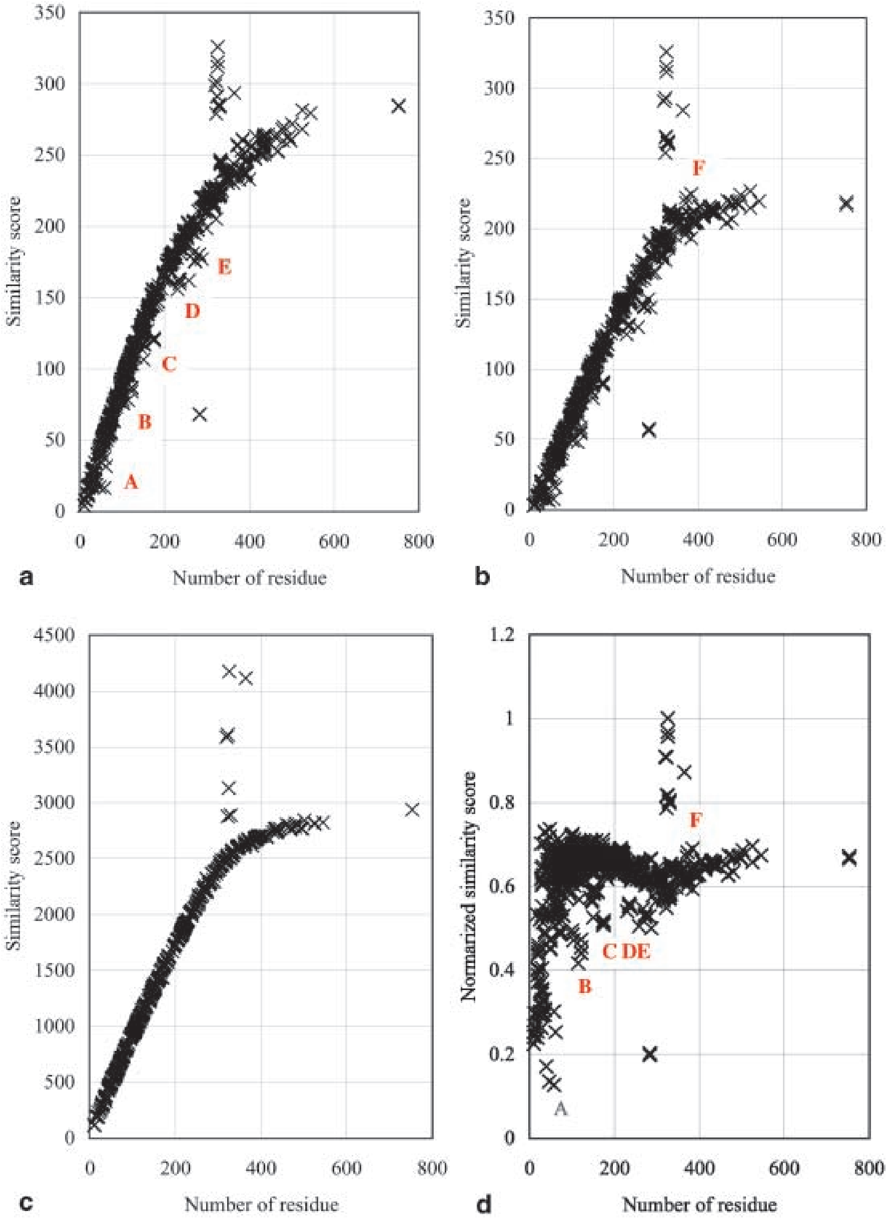

Fig. 6a) are the fourth chains of virus coat proteins and of polio virus, which have an extended structure, the second group (B) are trp aporepressor and interferons, which have a loose globular shape, and the third group (C) consists of kallikreins and hemagglutinnes, which again have an extended structure. All of these thus have shapes that are, again, very different from that of azurin, thus suggesting that the sphere method is able to distinguish between the overall shapes of proteins.

Figure 7a summarizes the corresponding search, i.e., a sphere radius of 12 Å and a zero-valued gap penalty, with porcine pepsin. A large, well-separated cluster of structures is observed at around 300 residues. This contains porcine pepsin itself, together with other acid proteases including pepsinogen and chymosin, and virus acid proteases. Tropomyosin is again placed well below the curve, and other sub-curve proteins are found at about 50, 100, 150, 250, and 300 residues. The first three groups of proteins (marked A–C) are similar to those found in the azurin search, which is to be expected as both porcine pepsin and azurin are globular proteins. The last two groups (D and E) contain the third and first chains of virus coat proteins and of polio virus, all of which have long, partly extended shapes. However, they all also contain a large globular segment, containing about 100 residues, that matches well with azurin, and thus places them on the main curve of Fig. 6a, rather than being clear of it, as occurs with porcine pepsin.

Figures 6b and 7b describe the azurin and porcine pepsin searches with the sphere radius still at 12 Å but with the gap penalty set to 4. Here, the curve tends to flatten off above the number of residues in the target protein. A group of dissimilar proteins is clearly visible (marked D) at around 250 residues in Fig. 6b. The group comprises subtilisins, thermitase, and mesentericopeptidase, all of which are typical globular proteins and thus noticeably dissimilar from the column-like shape of azurin. However, Fig. 7b shows that they are not found to be markedly dissimilar from porcine pepsin. Here, a similar protein (F) is visible, corresponding to the 385- and 372-residue chains of ovalbumin (1OVA). These chains have sequences and conformations that are different from porcine pepsin but the overall shapes of the two proteins are broadly similar. The superposition of ovalbumin and porcine pepsin obtained by our alignment is shown in Fig. 8.

For comparison with these results, Fig. 6c and Fig. 7c show the similarity score plots (for azurin and porcine pepsin, respectively) when the scores are calculated by the system using a conventional mutation-matrix method using PAM-250 matrix with a gap penalty of 6. In these figures, high-scoring clusters were seen containing members of the target protein's family, as with the sphere method; however, no other clusters were detected, such as that containing tropomyosin and denoting proteins with different 3D structures.

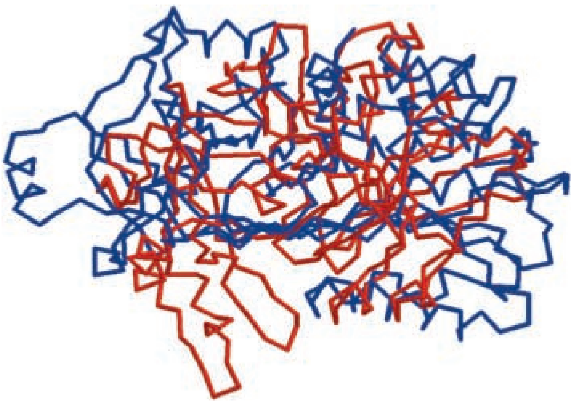

Fig. 8 Superposition of ovalbumin A chain (*blue*) and pepsin (*red*)

Figures 6d and 7d show the data in Figs. 6b and 7b after each similarity score has been normalized by dividing the score for each database protein by the number of residues in whichever is the smaller of the target protein and of the database protein. The curves now contain a large, reasonably flat section, although dipping at around the length of the target protein in the porcine pepsin searches. Some randomized numerical sequence experiments (results not shown here) demonstrated that the normalized similarity score tends to have a minimum value at around the number of residues in the target protein, and Fig. 6d thus shows several proteins with both a similar structure and a similar number of residues.

The groups of proteins seen in Fig. 6a and b, and in Fig. 7a and b, are clearly seen in Fig. 6d (marked A–D) and Fig. 7d (marked A–F), respectively. There are many low-scoring proteins containing less than 50 residues in both Figs. 6d and 7d: this is a reflection of the fact that it is difficult for such small proteins to adopt a globular shape. Proteins with similar structures to the target protein typically have similarity scores in excess of 0.8.

The non-globular human immunoglobulin G1 Fc fragment (1FC1) A chain has a characteristic bent shape. Figure 9 a and b shows the similarity scores obtained with the sphere method (using a sphere radius of 12 Å and a gap penalty of 4) and the mutation-matrix method (with a gap penalty of 6), respectively. Tropomyosin is also placed well below the main curve. The proteins beneath the curve (marked A in Fig. 9a) are mainly serine proteases, which are typical globular proteins. None of these proteins are observed as outliers in Fig. 9b. Immunoglobulin fragments are seen above (B, C and D) both of the main curves at around 200 and 300 residues; a fur-

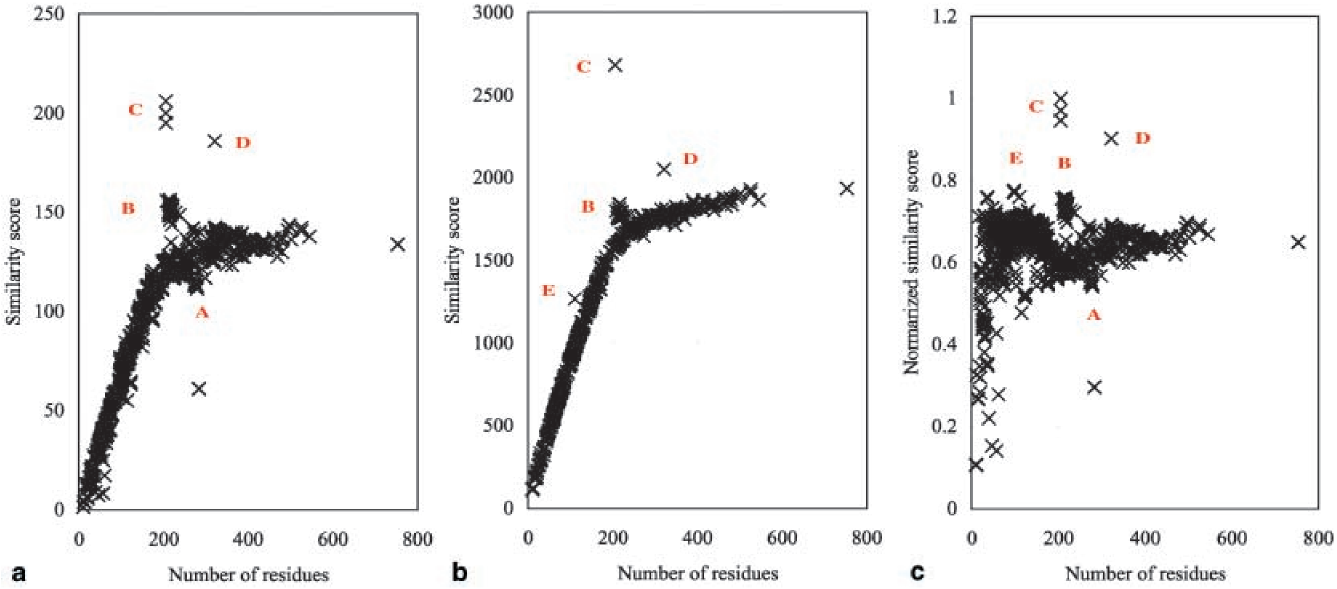

Fig. 9 Similarity scores of proteins in the PDB to human immunoglobulin Fc fragment (**a**) using the sphere method with a radius of 12 Å and with the gap penalty set to 4, (**b**) using the mutation-matrix method with the gap penalty value is set to 6, and (**c**) normalized version of (**a**)

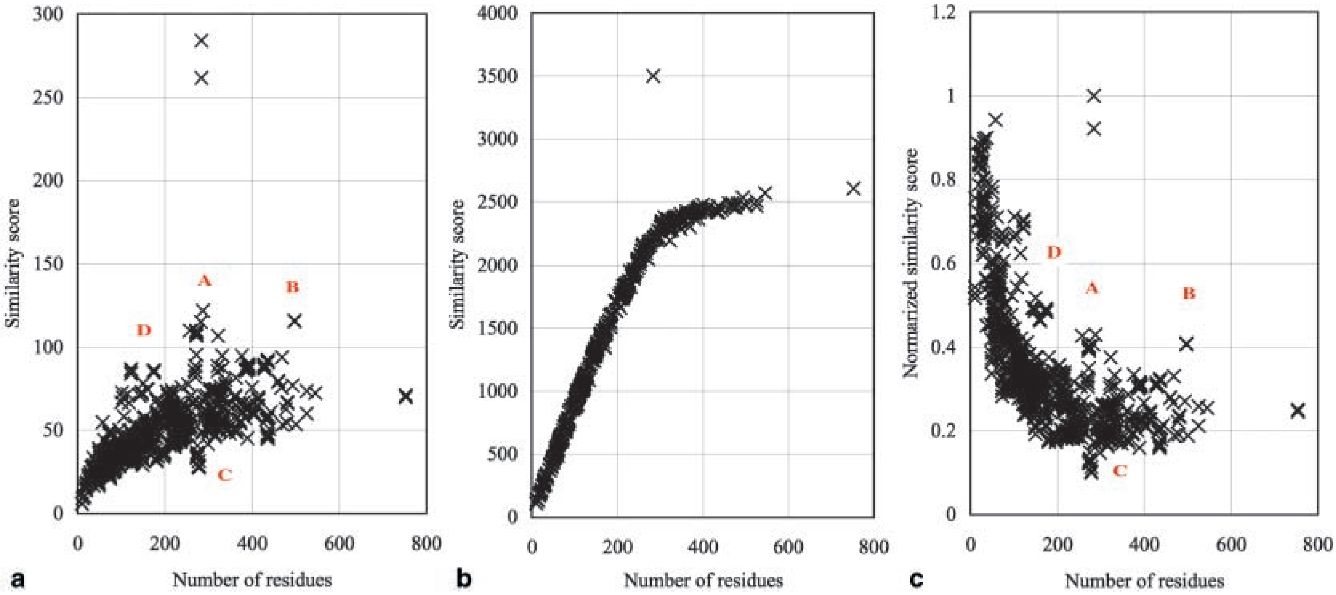

Fig. 10 Similarity scores of proteins in PDB to tropomyosin (**a**) using the sphere method with a radius of 12 Å and with the gap penalty set to 4, (**b**) using the mutation-matrix method with the gap penalty set to 6, and (**c**) normalized version of (**a**)

ther immunoglobulin fragment, pig immunoglobulin G1 pFc fragment (1PFC) is visible (E) at around 100 residues in Fig. 9b. Searches were also carried out using the A-chain of tropomyosin (2TMA) as the target protein, since it had appeared as an outlier in so many of the previous searches. Figure 10a and b shows a sphere-method search with a radius of 12 Å and a gap penalty of 4 and a mutation-matrix search with a gap penalty of 6, respectively. Most of the PDB-subset yielded low similarity scores, as would be expected from the fact that most proteins in the PDB are globular in character. The proteins with the highest scores, in excess of 250, were the A and B chains of tropomyosin. Proteins with relatively high scores were the first chains of virus coat proteins and polio virus (A) and catalase (B). Those with low scores, at a length of about 250 residues (C), were subtilisin, thermitase, and proteinase, which are all members of the serine protease family and which are typical globular proteins. There are a fair number of points above the curve (i.e., similar to tropomyosin) in the region 50–150 residues (D): these correspond to proteins such as those shown at A, B and C in Fig. 6a. There are no obvious outliers in Fig. 10b, except for the two chains in the target protein itself.

Figures 9c and 10c show the scores that were obtained when the sphere-method similarity scores were normalized. Figure 9c shows several proteins with scores well in excess of 0.8: these are chains from human immunoglobulin G1 Fc fragments (C) and the model structure of human immunoglobulin E Fc fragment (D). The other characteristic proteins in Fig. 9a and b are also seen in Fig. 9c. When the similarities were ranked in decreasing order of score, all but two of the top 20 structures were other immunoglobulins, the two exceptions being the C-terminal domains of cellobiohydrolases (1CBH and 2CBH); indeed, only four further non-immunoglobulins were found when the top 50 structures were inspected, the exceptions being neurotoxin (2SH1), fatty acid binding proteins (1IFB and 2IFB) and trypsin inhibitor (1TGS). Pig immunoglobulin G1 pFc fragment (1PFC), which is clearly marked in Fig. 9b, occurs at the 10th rank position. Figure 10c has a radically different shape from the other plots here. This is because the target protein has a near-linear structure that is different from nearly all of the other proteins in the PDB. It is difficult for a small protein to adopt a globular structure, and it is for this reason that they tend to have high similarity scores with the non-globular target structure used here. In all, there are 33 structures with similarity scores in excess of 0.8 and 16 of these have an obvious straight shape. When the similarities were ranked in decreasing order of score here, all but four of the top 20 structures had extended, straight shapes similar to that of tropomyosin itself, the only exceptions being gag polyprotein (2ZNF) and three insulin structures (3INS, 4INS and 2INS); the other proteins (most of which had separate structures for the two chains) were murein lipoprotein (1MLP), antifreeze polypeptide (1ATF), leucine zipper (2ZTA), delta hemolysin (1DHL, 2DHL, 3DHL), melittin (2MLT) and glucagon (1GCN). Despite the different shape of the curve, there are again four main groups of outliers, corresponding to those marked A–D in Fig. 10a.

Conclusions

In this paper, we have described the use of the sphere descriptor of Nishikawa and Ooi [17] for calculating similarity scores between pairs of proteins. The descriptor is calculated from the number of α-carbon atoms located within a sphere of a user-defined radius centered upon each α-carbon atom in turn of a protein's main chain. Initial tests using the acid proteases showed that the method produced inherently reasonable alignments with lower RMSD values than those obtained from a conventional mutation-matrix approach. Searches on a subset of the PDB demonstrated that the method is sufficiently fast for database-searching applications and that it is able to differentiate between proteins that are structurally similar and dissimilar to a user-defined target protein.

The Krebs Institute for Biomolecular Research is a designated center for biomolecular sciences of the Biotechnology and Biological Sciences Research Council.

References

1. Needleman SB, Wunsch CD (1970) J Mol Biol 48:443–453
2. Dayhoff MO, Schwartz RM, Orcutt BC (1975) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, D.C. pp 345–352
3. Lipman DJ, Pearson WR (1985) Science 227:1435–1441
4. Murata M (1988) Comput Chem 12:21–25
5. Nishikawa K, Ooi T (1974) J Theor Biol 43:351–374
6. Rossman MG, Argos P (1976) J Mol Biol 105:75–95
7. Remington SJ, Matthews BW (1980) J Mol Biol 140:77–99
8. Karpen ME, de Haseth PL, Neet KE (1989) Proteins Struct Funct Genet 6:155–167
9. Holm L, Sander C (1993) J Mol Biol 233:123–138
10. Lessel U, Schomburg D (1994) Protein Eng 7:1175–1187
11. Murthy MRN (1984) FEBS Lett 168:97–102
12. Abagyan RA, Maiorov VN (1988) J Biomol Struct Dynam 5:1267–1279
13. Grindley HM, Artymiuk PJ, Rice DW, Willett P (1993) J Mol Biol 229:707–721
14. Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1990) J Mol Biol 212:151–166
15. Taylor WR, Orengo CA (1989) J Mol Biol 208:1–22
16. Kawabata T, Nishikawa K (2000) Proteins Struct Funct Genet 41:108–122
17. Nishikawa K, Ooi T (1986) J Biochem 100:1043–1049
18. Schulz GE (1977) J Mol Evol 9:339–342
19. May ACW, Johnson MS (1994) Protein Eng 7:475–485
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242