A Fourier Fingerprint-Based Method for Protein Surface Representation

Martin J. Bayley,[†,‡] Eleanor J. Gardiner,[†] Peter Willett,[†] and Peter J. Artymiuk*[,‡]

Krebs Institute for Biomolecular Research, Department of Information Studies and Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received November 24, 2004

A crucial enabling technology for structural genomics is the development of algorithms that can predict the putative function of novel protein structures: the proposed functions can subsequently be experimentally tested by functional studies. Testable assignments of function can be made if it is possible to attribute a putative, or indeed probable, function on the basis of the shapes of the binding sites on the surface of a protein structure. However the comparison of the surfaces of 3D protein structures is a computationally demanding task. Here we present four surface representations that can be used locally to describe the global shape of specifically bounded local region models. The most successful of these representations is obtained by a Fourier analysis of the distribution of surface curvature on concentric spheres around a surface point and summarizes a 24 Å diameter spherically clipped region of protein surface by a fingerprint of 18 Fourier amplitude values. Searching experiments using these fingerprints on a set of 366 proteins demonstrate that this provides an effective and an efficient technique for the matching of protein surfaces.

INTRODUCTION

At present, a revolution is taking place in structural molecular biology. Knowledge of the sequences of whole genomes means that many new proteins are now potentially available for structural analysis. It is most fortunate that this explosion in our knowledge of gene sequences has been at the same time accompanied by major methodological advances in the processes of structure solution, most notably at present in protein crystallography. The process of solving a crystal structure has been immeasurably facilitated by new methodologies for protein overexpression, crystallization screening, cryogenic crystal cooling, the use of synchrotron radiation, SeMet incorporation, phasing by multiwavelength anomalous dispersion, direct methods for heavy atom location, and automated map interpretation and refinement.[1] Major opportunities are therefore available to crystallographers, NMR spectroscopists, and electron microscopists in determining the structures of a vast array of novel, biologically important proteins, many of whose functions are, however, at present unknown.

These advances have led to a massive acceleration in the number of structures being solved. In particular, in the area of structural genomics (or structural proteomics) an increasing number of proteins of unknown function are being solved.[2,3] Knowledge of the structure can be crucial in assigning tentative and experimentally testable functions to these proteins, some of which may prove to be of great importance. At present such an assignment is typically derived either from a generalized assignment based on a similarity of fold to other proteins[4] or even from the chance observation of ligand binding.[5]

A crucial enabling technology for structural genomics will therefore be the development of algorithms that can predict the putative function(s) of a novel protein structure: the proposed functions can subsequently be experimentally tested by functional studies. As stated above, in many cases a generic function or mechanism can be assigned on the basis of overall similarities of the protein fold, and we ourselves have made significant methodological innovations[6,7] and findings in this area.[8,9] However much more specific, testable assignments of function can be made if it is possible to attribute a possible, or indeed probable, function on the basis of the chemical properties of groups displayed in the binding sites on the surface of a protein structure or on the basis of the 3D shape of binding site cavities, which is the problem considered here. In addition to proposing possible protein functions, such methods may also be of use in the design of drugs and inhibitors by examination of small molecules that bind to similar sites in other molecules.

Several different definitions of what constitutes the surface of a protein have been described in the literature,[10,11] and Via et al.[12] have provided a helpful and wide-ranging survey of methods for describing and comparing protein surfaces. Of the various approaches that have been described, the most widely used is that described by Connolly.[13,14] The Connolly surface and its derivatives are very widely employed in the representation of the shapes of protein surfaces and also as a vehicle to represent important surface features such as electrostatic fields and hydrophobicity.[15−17] However, a disadvantage in applications such as surface-matching or docking is that a large number of points are necessary to permit a full representation of a surface. Calculation of such surfaces at lower densities of points inevitably results not only in a much less precise description but also in potential discrepancies in the way that other, similar surfaces may be sampled. Recently several new attempts at reduced local surface representation have been reported,[18−22] including graph theoretical approaches (e.g., the Cavbase program).[23] Most involve partitioning the surface into smaller (possibly overlapping) patches in some reasonably canonical manner,

* Corresponding author phone: +0114 222 4190; e-mail: p.j.artymiuk@sheffield.ac.uk.

[†] Department of Information Studies.

[‡] Department of Molecular Biology and Biotechnology.

10.1021/ci049647j CCC: $30.25 © 2005 American Chemical Society
Published on Web 04/12/2005

based on local properties such as curvature,[18,21] molecular surface properties,[19,20] or local geometry defined by ray tracing.[22]

Other attempts have been made to represent surfaces through the use of Fourier transforms for the surface shape definition of whole proteins[24] and spherical harmonics for docking.[25] Although effective in many respects, these have a number of potential disadvantages for local shape comparison that include not only problems describing re-entrant/invaginated surfaces but also more fundamentally the fact that the global nature of the Fourier description makes it very difficult to detect purely local similarities between otherwise very different surfaces without regeneration of overlaid surface data for comparison. This does, however, provide some improvements on traditional surface matching methods in terms of speed of operation, reducing the size of the matching problem from 6 to 3 degrees of freedom. Moreover, these representations can be used locally to describe the global shape of specifically bounded local region models. It is precisely such similarities that we wish to detect: localized sites in completely different proteins that bind the same ligand are clearly such local, nonglobal similarities. In this paper, we discuss four different types of surface representation that can be used for the characterization of protein surfaces.

METHODS

Among the factors that need to be taken into account when considering methods for the representation and matching of protein surfaces are alignment, sampling, and bounding. First, if we wish to score the similarity of the shape of two 3D surface models, it is frequently necessary to overlay the two surfaces so as to maximize the common features of the chosen representation: for traditional 3D sample point-based models, this is a highly computationally intensive task involving the use of iterative, optimization methods (see, for example, ref 17). The methods here, however, are alignment-free, using representations that are derived from the basic coordinate data and that do not require the generation of a superimposition. Next, some decision must be made as to the level of sampling that is to be adopted in the generation of the representation: less detail introduces the potential for greater noise in scoring how well the surfaces can match. Finally, it is necessary to define a boundary (preferably by automated means) to the surface regions being matched so as to allow a sensible comparison to be made between two surfaces when the similarity is computed. The various approaches that we have studied, as detailed below, provide a range of solutions to these three, linked problems.

Our initial studies sought to generate a sparse set of richly descriptive surface points to aid efficient comparison using techniques such as genetic algorithms, graph theory, etc.[6,17,26] Methods A and B reported below were developed to provide two such reduced representations. The testing procedure for these methods involved analysis of a set of active sites overlaid by optimal superimposition of the bound ligands. This allowed a direct comparison to be made between different sets of calculated reduced points that have been found in the region of the active site.

We then sought to summarize the local surface shape surrounding each sample point by an orientation-independent, geometrically canonical submodel (i.e., a set of rules or a standard geometrical framework that describes the geometric surface variations over a defined local region about each surface point) in a consistent format. The parameters of these submodels could then be compared directly one with another to gain a fast, orientation-independent measure of local surface shape (in a manner reminiscent of the fast, fingerprint-based methods that are widely used for the computation of small molecule similarities[27]). In essence, this reduces the local surface matching problem to a simple fingerprint comparison, with a once-only computational overhead in calculations to convert from global point surface sample models to local parameter format files. Methods C and D reported below fall into this category; both were tested by analyzing the position and spread in 3D space of the most similar points to a known site parameter-set.

The starting place for all of the methods was a Connolly sample model of the protein surface, sampled at 2 points per Å^2 and consisting of surface point coordinates and unit vectors normal to each point. In each case, coordinate sets were selected from the Protein Data Bank (PDB) and used without further energy refinement. Ligand molecules including water and metal ions were removed. The aim was to develop a method capable of identifying similar surface shapes, initially irrespective of their chemical characteristics. Therefore, in the few cases where they were present, hydrogen atoms were also removed in order to ensure that the protein was characterized solely by its geometric shape without consideration of the chemical nature or the ionization state of the surface atoms or the ambiguities associated with undetermined hydrogen torsion angles. We have chosen, as have previous workers, to consider only a single, fixed conformation (i.e., that stored in the PDB) as it is our aim to first identify effective and efficient methods for rigid macromolecules; only when this has been achieved will it be appropriate to consider the extension of these methods to the more challenging flexible case.

Method A: Curvature-Based Point Reduction/Summarization. The first approach sought to reduce the data encoded in a Connolly surface by categorizing each point using the local biaxial curvature of its surrounding points. This has been the subject of previous work by a number of researchers[18,21] and has shown some promise as a method. For each subject point (i.e., the point to be represented by a local canonical model), we first located all the other Connolly points between a distance of r1 and r2 from that point. The resulting subset, which is shown in Figure 1a, contains all points within the distance range r1 to r2, albeit being unevenly distributed throughout space and in no particular order. It is then necessary to order these points relative to the central subject point and its surrounding surface. This was done using a polar coordinate system, setting the pole vector direction as the mean surface normal of all Connolly points found within an r2 radius of the central subject point. Thus (unless the local surface region is highly re-entrant), the pole vector placed at the central subject point will always pass through the previously isolated subsampled loop of points. Each loop sample point was then transformed into 3D polar coordinate space relative to the origin (the original subject point) and the polar axis (Figure 1b). The points were then binned in θ (longitude angle) by dividing θ into a number of equal sectors, and each bin was represented by

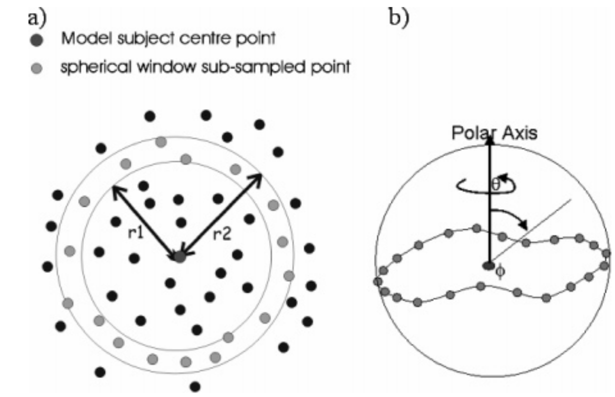
a)
b)
Model subject centre point
spherical window sub-sampled point
r1
r2
Polar Axis
θ
ϕ

Figure 1. Method A: canonical point representation. (a) Sphere clipping is used to obtain a loop of points around central point. (b) The points are set in a polar coordinate system and then represented by their ϕ angles.

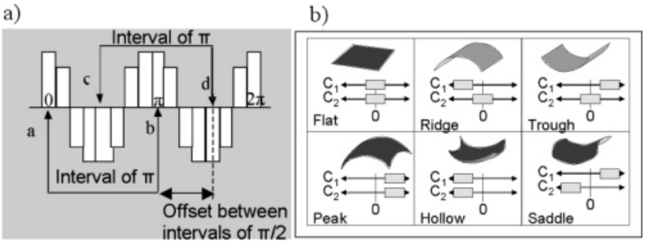

Figure 2. Method A: biaxial curvature generation. (a) Opposing biaxial curvatures can be extracted from a 16-point sampled polar signal of surrounding surface curvature. (b) Six different categories of biaxial curvature with associated biaxial curvature.

the maximum ϕ (latitude angle) value for its constituent points. This means that each sample point was described by a 1D signal encoding the curvature distribution around it. The signal values were divided by $\pi/2$ in order to transform from radians to a possible value range of 1 to -1. For biaxial curvature calculations, 16 sectors were used.

The calculation of curvature is shown in Figure 2a where a sphere-clipped signal has been labeled with the phase positions required to calculate a single pair of biaxial curvature coefficients. By analyzing the signal at opposing (180°) phases, the principal curvature coefficient (C1) across the sphere was then the mean of these two values (for example the signal values at phase positions labeled a and b in Figure 2a). The secondary, orthogonal curvature coefficient (C2) was then the mean opposing signal at 90° to the phase of the signal values used to calculate C1 (i.e., using the signal values at phase positions labeled c and d in Figure 2a). All possible combinations of signal curvature were generated by repeating the calculation four times, shifting the signal through the first four discrete signal phase internals. The dominant biaxial curvature pair was then taken to be the pair with the greatest positive or negative value of either C1 or C2. Each curvature coefficient was then partitioned into one of three categories: up (mean opposing signal values above $+ 0.1$), down (mean opposing signal values below -0.1), or approximately flat (mean opposing signal values between $+0.1$ and -0.1). This defines the six possible biaxial curvature categories: peak, hollow, saddle, trough, ridge, or flat (as shown in Figure 2b). The radius of the sphere-clip method used defined the "localness" of the biaxial curvature and was found to be an important factor in the success of the method: too large a radius and an entire cavity would be mapped to a single cluster; too small a radius, and a cluster would represent just a single atom. The best results were obtained with radii of 2.5 and 3.5 Å for r1 and r2, respectively.

The initial tests were performed upon a subtilisin molecule (PDB code 2SIC chain E, with the inhibitor removed, and containing no hydrogen atoms). The points calculated to be of each particular curvature type (hollow, saddle, etc.) were tagged and output to a 3D VRML model. Visual inspection of the clusters of points of each of the different types of curvature showed that the clusters of points obtained for the hollow and saddle categories were most indicative of the 3D surface landscape in and around active sites. The other categories of curvature tended either to cluster separately on each atom, making the representation no more indicative than an atomic level PDB model or in tangled linear formations that would be poorly represented if summarized by clustering methods.

A crucial characteristic of any representation method is that it is unique (i.e., invariant to model sampling). Initial testing to determine that method A was invariant to model sampling was performed using a set of Connolly models that had the same orientation but differing sample patterns. This was achieved by taking a PDB structure and reorienting it by specific rotations applied about the *X*, *Y*, and *Z* axes. The surface was next sampled using the Connolly program to create a different sample pattern model of the same protein, and this sample point model rotated back into its original structure orientation through reverse order, negative application of the previous rotation transformations. The tests here used a region of points of radius 6 Å centered on the subtilisin catalytic triad (PDB code 2SIC). Clusters of the same curvature type and size were found to occur in the same positions around the active site, in the 10 differently sampled models tested (data not shown), demonstrating the robustness of the method to the initial Connolly sampling.

Unfortunately, despite this invariance of the representation to sample pattern, further testing revealed significant weaknesses in the method. Specifically, searches were performed on a set of eight similar ATP binding proteins from the same CATH family[28] (PDB codes 1AUX, 1B6S, 1C30, 1E4E, 1EZ1, 1GSA, 1GSH, 1IOV) for which the active site had been optimally overlaid to see whether these exhibited the same patterns of hollows and saddles. Biaxial curvature assessment was performed on the Connolly models, and VRML wire-frame models were developed to enable visual assessment of the overlay. Results from the 28 pairwise comparisons of the overlaid wire-frame models showed that in only six cases were there good correlation between the positions of the clusters identified by the procedure, with even subtle differences in the 3D surface shape yielding much larger changes in the positions of the clusters that were identified. It was hence decided not to develop method A any further.

Method B: Cavity Shape Summarization. The second method aimed to overcome the sensitivities of the previous curvature-based method by providing a more general summary of the presence and the size of individual surface cavities. Specifically, the method used a set of geometric rules informed by the relative local position and surface normal vectors of local sample points to search for specific regions of clearance between locally opposing walls of protein surface, as illustrated in Figure 3. The Connolly

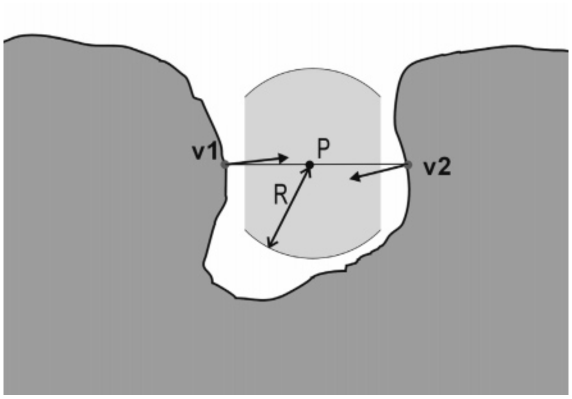

Figure 3. Method B: cavity detection. The dark gray area represents a protein region bounded by the Connolly surface. v1 and v2 are surface points whose normals are approximately opposed, at the required distance. P is their midpoint. The light gray region represents the wheel-shaped cavity found by bounding a sphere of desired radius *R*, centered at *P*, by two planes perpendicular to the line joining v1 and v2. For details, see the text.

surface was analyzed to find pairs of points (v1 and v2) that were within a given distance threshold ($D \pm \delta$) of each other and that had approximately opposing surface point normal vectors. (The angle between the two normal vectors was defined as opposing if in the range 180° ± 30°). The pairs of points were further analyzed to calculate the maximum clearance of the cavity; this was done by checking whether any surface points could be found within a radius (*R*) of the center (*P*) of the chosen pair of points. This test bump region was further truncated by excluding those points in the sphere subset found outside two parallel planes, perpendicular to the line between the chosen pair of points and positioned at 20% and at 80% of the distance along that line. This truncation ensured that the method could find cavities with minor protrusions occurring into the cavity as a result of potential differences in the "knobbliness" of cavity walls. Thus, the method searched for the presence of wheel-shaped regions of clearance between two opposing walls of protein surface, as illustrated in Figure 3.

Different cavity sizes were categorized in terms of the opposing point distance range (*D*) and the bump test region radius (*R*) with a cavity being defined by its size category and the position of the center point of the opposing point pair. Three cavity categories were defined: "small" ($D = 4$ Å ± 0.5 Å and $R = 1.5$ Å), "medium" ($D = 6$ Å ± 0.5 Å and $R = 2$ Å), and "large" ($D = 8$ Å ± 0.5 Å and $R = 2.5$ Å).

Again, our initial tests checked whether the method was independent of the sample pattern (as described for method A). This was performed using the same Connolly models as used in method A (i.e., PDB code 2SIC with 10 different Connolly sample patterns but the same global orientation). The results were output as VRML scatterplots of detected cavity summary points: the plots showed a good level of consistency in the overlays, with summary points for each category type occurring in the same localized regions of 3D space (data not shown).

We then took the set of eight ATP binding proteins that were used to test method A; these Connolly models were processed and VRML models generated to assess the relative positioning of the different types of cavity size. In all resultant models, a region of cavity type medium was found within the region of the bound adenine molecule. In most of the test set there was also a large amount of cavity type large detected in the triphosphate region of the bound ligand. However, despite the presence of these specific cavity types, there was again far too much inconsistency in the size and shape of different clusters of representations to think that they might the form the basis for a robust searching procedure.

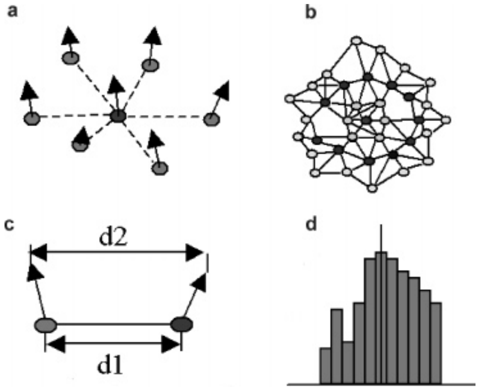

Figure 4. Method C: schematic of the mesh curvature fingerprint method. (a) Nearest neighbor points around a given point. (b) Patch containing all points within three mesh connections of a given point. (c) Calculation of curvature between neighboring points. (d) Resultant probability density function.

Method C: Local Mesh-Based Curvature Summarization. The third approach investigated the encoding of the distribution of the intensity of curvature over a patch of surface so as to give a discretely binned probability density function (PDF) that summarized the relative local curvature around a sample point. The values stored in the PDF thus defined a parameter set that could be used for local similarity comparisons. Another feature of this method used the neighbor relationships between proximal points to provide an efficient means for bounding and referencing local curvature information.

At each Connolly surface point, six nearest neighbor points were isolated, as shown in Figure 4a, which corresponds to the nodes and edges in a tessellating pattern of triangles. Then, for each point in turn, a patch was defined that consisted of all of the points within a defined number of mesh connections of the subject point; this was done by recursive referencing of the previously calculated, nearest neighbor relationship lists. Figure 4b shows all points within three mesh connections of the subject point. For each neighbor relationship within this patch, a curvature coefficient was calculated as a function of the difference in the Connolly unit normal vectors. If d1 is the distance between the two neighbor relationship points and d2 is the distance between the ends of the Connolly unit vectors at those two points (Figure 4c), then (d2 − d1)/d1 gives a measure of the curvature between those two points. The values of curvature were then binned into a discrete PDF that, after assessment of all patch relationships, was standardised to a PDF of sum unit area (Figure 4d).

Testing of this method involved scoring the similarity between pairs of parameter sets as the dot product of their vector forms. The parameter set for a given test point on the surface of a protein Connolly model was compared against all parameter sets extracted from the points of an entire Connolly surface, with the test point being chosen from within a cavity to attempt to mimic an active site. Points of high similarity were isolated by application of a threshold to the similarity score and their distribution in space noted, with the hope that a number of points at and around the position of the original test point would be found among the most similar points.

However when the invariance of the method to the sample pattern was investigated (as described above, using in this case bovine pancreatic trypsin inhibitor, PDB code 1BTI), it was unfortunately found that the most similar points did not necessarily come from around the cavity that was the origin of the original test point (data not shown). Attempts to alleviate the problem by varying the sampling rate, the PDF binning scheme, and the type of mesh were all unsuccessful; it was hence concluded that while this representation provided an easily calculated and concise summary of an individual protein surface, it was not suitable for the comparison of different surfaces.

Method D: Surface Sphere-Clipped Fourier Parameter Summary Model. We sought to overcome the sample pattern invariance problem noted above (methods A and C) by summarizing the local surface shape around a sample point in a more rigidly defined geometrically canonical form. In essence this is a modified version of method A, the canonical point representation of which was illustrated in Figure 1 and which was described previously. The principal modification here involves taking the curvature distribution signal resulting from the method A representation and carrying out a Fourier analysis of it. Each sample point was described by a 1D signal encoding the curvature distribution around it. A Fourier analysis of this signal yielded a vector of amplitudes and a vector of related phases, and the six lowest order harmonic values of amplitudes (higher order amplitudes → 0) were then used as the local parameter set to represent the surface shape around each sample point. We have then summarized more of the local surface shape information about a point by using multiple Fourier representations, specifically by using a nested set of three, evenly spaced sphere-clipping radii. Since the surface is continuous, using rings of approximately atomic diameter separation ensured that no significant surface shape change was missed. In the work reported here, three sets of six Fourier parameters were generated in each case, representing spheres of radii 4 ± 0.5, 8 ± 0.5, and 12 ± 0.5 Å around each central point; in effect, the representation summarized a 24 Å diameter spherically clipped region of protein surface by a fingerprint of 18 Fourier amplitude values.

Initially, we used only the Fourier amplitudes as this enables the sets of values to be compared directly, without the need to reorientate signals for optimal overlay. This does, of course, introduce the possibility of finding false positives as two Fourier series of similar amplitudes but with very different phases would be identified as denoting similar surface points. This can be overcome by using a second stage process, whereby the signal data (i.e., the binned values of latitude angle prior to Fourier decomposition) can be compared in a single degree of freedom optimization, filtering out false positives (which will have a relatively poor optimal overlaid similarity score). Although this second stage requires greater computational resource, it need only be done for a subset of the most similar patches of protein found in the initial stage.

It was noted that if a parameter set for a specific point on a given protein surface was compared to all other point parameter sets on that protein, the most similar points typically occurred adjacent to the original parameter set point. This "similarity well" effect gives the method a degree of robustness in that there is always a group of points (rather than just a single solution) that is available for searching.

Again, our initial tests checked whether the method was independent of the sample pattern (as described for method A). This was performed using the same Connolly models as used in method A (i.e., PDB code 2SIC chain E, with 10 different Connolly sample patterns but the same global orientation). The Fourier fingerprint representation for each sample point in each model was then generated, and a known test point fingerprint was isolated at the center of the protein's active site in one of the models. The similarity of the test point Fourier fingerprint to all of the other points' Fourier fingerprints was calculated, and the 20 most similar points were identified, with the similarity calculated by the dot product of the pairs of 18 element vectors. We found that the most similar points clustered in and around the same position on the protein surface for all of the cases that we tested and that this position was that of the chosen test point. This is demonstrated in the 2D plots that are shown in Figure 5 for 4 of the 10 different orientations that were tested. While not completely identical, it is clear that there is a very high degree of commonality, thus demonstrating that we have been able significantly to reduce the sampling dependencies that we have noted previously (method C).

We next conducted an initial study on a small data set of serine proteases containing 5 subtilisins, 5 trypsins, 4 elastases, 2 chymotrypsins, and 4 non-serine proteases as potential false positives. For each of the four classes of protein, a representative single query point fingerprint was selected from a position in the active site of a single protein (2SIC for the subtilisins, 5CHA for the chymotrypsins, 1EAU for the elastases, and 1AVW for the trypsins). Each class was searched for as described above. Two values were recorded for similarity as follows: for each target protein the 20 most similar points to the query were found using the dot product of the Fourier fingerprint vectors. The mean score was then the mean of these dot products and the mean spread score was the mean of all pairwise inter-point distances between the 20 matched target points. Both measures showed a similar trend, which was amplified through data fusion, by taking the product of the mean score and the mean spread score. The rankings from the fused similarity scores are shown in Table 1. In the subtilisin and elastase searches, proteins from the class being searched for occurred at or near to the top of the rankings; in the case of the trypsin and chymotrypsin searches, these two classes of protein bunched together, which is not unexpected as they possess a very similar active site.

In a further step, we also calculated similarity based on the polar signal fingerprints (which had been generated for the entire database as a byproduct of the Fourier fingerprint generation). These signal fingerprints represent how surface latitude angle varies with radially segmented longitude angle in the local polar canonical model (Figure 1b), used to derive the Fourier fingerprint representation. To compare two signal fingerprints, one is fixed and the other undergoes cyclic permutation to determine the best match, again calculated

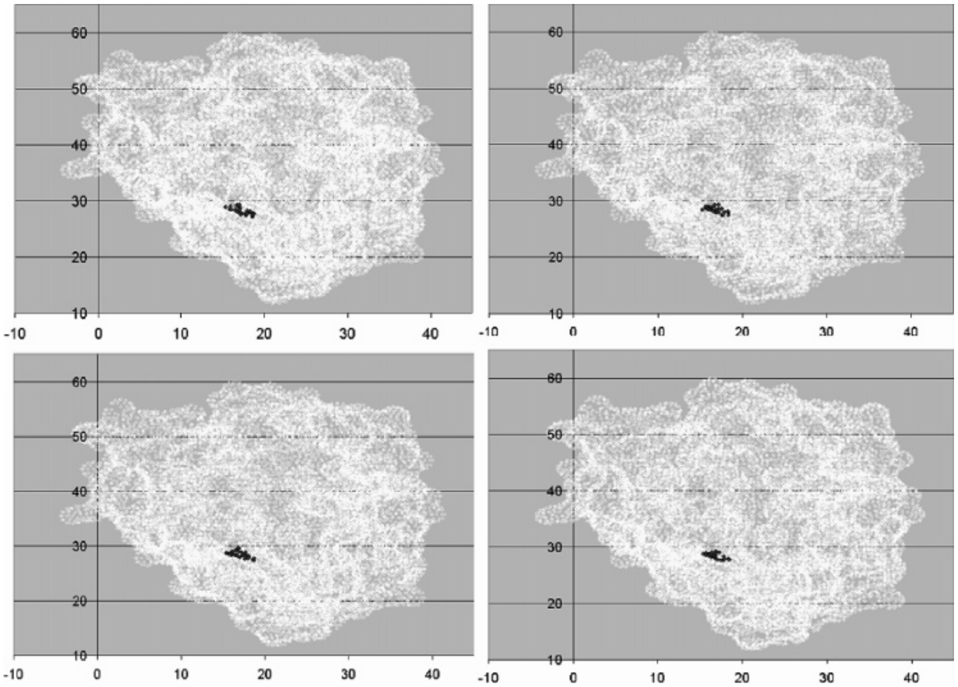

Figure 5. Method D: Fourier fingerprint sample-pattern invariance testing. Results for four differently sampled, similarly orientated Connolly surfaces of subtilisin, (PDB code 2SIC). All points found to be closer than a threshold of 1.4 similarity units are shown.

Table 1. Fourier Fingerprint Results for the Serine Protease Test[a]

subtilisin 2SIC		chymotrypsin 5CHA		elastase 1EAU		trypsin 1AVW	
subtilisin	**2SIC**	**chymotrypsin**	**5CHA**	**elastase**	**1EAU**	**trypsin**	**1AVW**
subtilisin	**1SBN**	*trypsin*	*1AUJ*	**elastase**	**1ELT**	**trypsin**	**1AUJ**
subtilisin	**1CSE**	*trypsin*	*1AVW*	**elastase**	**1B0F**	*chymotrypsin*	*5CHA*
subtilisin	**1GNS**	*trypsin*	*1ANE*	subtilisin	1SBN	*chymotrypsin*	*1EQ9*
subtilisin	**1A4F**	other	1TCB	other	1FDW	**trypsin**	**1A0J**
other	1FDW	*trypsin*	*1A0J*	trypsin	1AUJ	**trypsin**	**1ANE**
other	1TCB	**chymotrypsin**	**1EQ9**	subtilisin	1GNS	**trypsin**	**1FY4**
trypsin	1ANE	elastase	1B0F	trypsin	1AVW	other	1TCB
chymotrypsin	1EQ9	*trypsin*	*1FY4*	trypsin	1ANE	elastase	1ELT
elastase	1EAU	elastase	1ELT	subtilisin	1CSE	other	1B4X
elastase	1ELT	other	1I8T	subtilisin	1A4F	elastase	1EAU
other	1I8T	subtilisin	1GNS	trypsin	1A0J	elastase	1B0F
trypsin	1FY4	subtilisin	1SBN	other	1I8T	other	1I8T
elastase	1B0F	subtilisin	2SIC	other	1B4X	subtilisin	1CSE
chymotrypsin	5CHA	subtilisin	1CSE	chymotrypsin	5CHA	other	1FDW
other	1B4X	other	1FDW	other	1TCB	subtilisin	1GNS
trypsin	1A0J	subtilisin	1A4F	trypsin	1FY4	subtilisin	1A4F
trypsin	1AVW	elastase	1EAU	chymotrypsin	1EQ9	subtilisin	1SBN
trypsin	1AUJ	other	1B4X	subtilisin	2SIC	subtilisin	2SIC

[a] The column headings show the query protein. The columns have been ranked by similarity score, with the proteins most similar to the query at the top. Proteins with similar activity to the query are in boldface type, and chymotrypsin/trypsin proteins with related activity are in italic type.

using the dot product of the two vectors, as above. Figure 6 shows the superposed signals for the subtilisin query 2SIC chain E and the three top-ranked subtilisin targets for each of the 4, 8, and 12 Å rings (as described above) found by the signal matching process. As each signal represents the change in curvature on moving around a local sphere, it is clear that the local surface, as measured by change in curvature, is indeed very similar for each of the four proteins.

The Fourier fingerprint match identifies similar surfaces but does not provide a superposition of the corresponding surfaces. Figure 7a shows two views of the query subtilisin protein 2SIC chain E (above) with the positions of the catalytic triad residues Asp32, His64, and Ser221 shown as green, cyan, and yellow spheres, respectively. The top hits found when searching using a single 2SIC fingerprint from the center of the catalytic triad are shown as small red spheres and are seen to cluster in the center of the catalytic triad (as expected from the invariance testing previously described). Figure 7b shows the subtilisin 1SBN, in the same orientation (based on the superposition of the catalytic triad atoms). The top hits found when matching 1SBN fingerprints with the same query fingerprint from 2SIC are again shown as small red spheres. It is evident that they cluster in a very similar position in relation to the catalytic triad, showing that the method is finding hits in the correct position in the active site.

It is important to note that the method does not give an alignment of the two surfaces but merely locates regions of

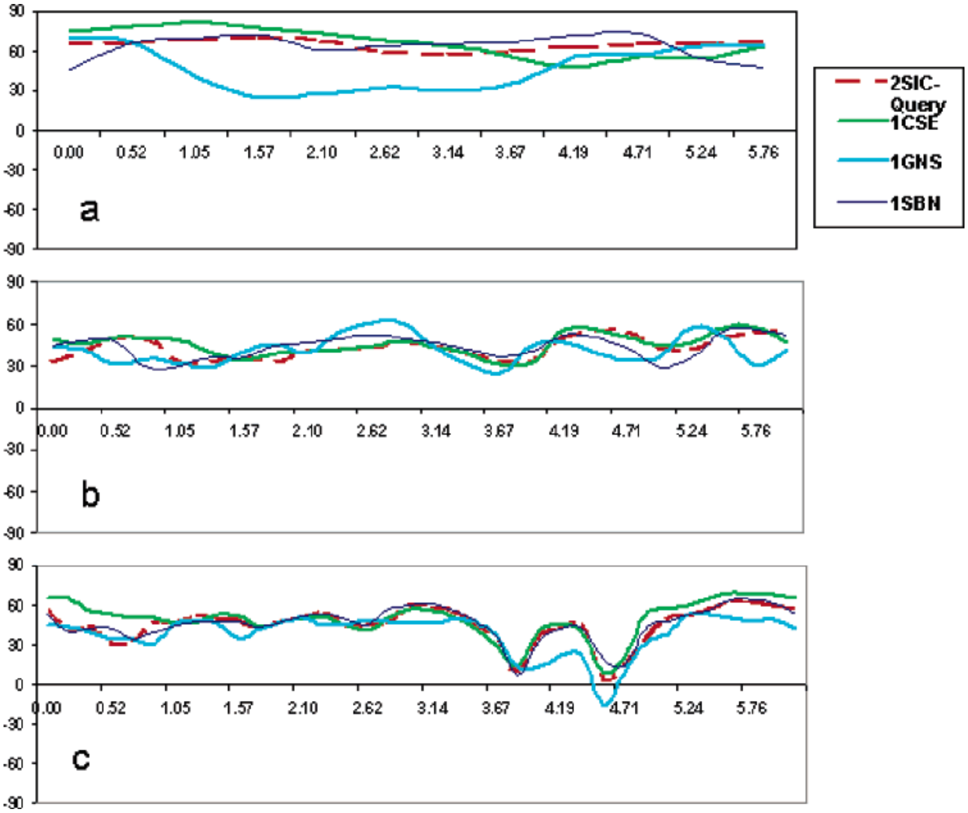

Figure 6. Method D: signal overlay for subtilisin target and three top subtilisin hits from Table 1. (a) Signals from 4 Å sphere. (b) Signals from 8 Å sphere. (c) Signals from 12 Å sphere.

similarity indicated by clusters of matching fingerprints. The proteins in Figure 7 have therefore been aligned manually. This is further discussed in the concluding section.

RESULTS AND DISCUSSION

More Detailed Experiments. Of the four methods described above, method D—that based on Fourier descriptions of a sphere—was the one that was deemed worthy of further investigation. Although the fingerprint tests reported above showed promise, it is clear that the method is subject to a "good" choice of search point and hence is quite dependent upon the original Connolly surface. It was therefore decided to represent an active site by a set of Fourier fingerprints chosen from the center of the active site. In the tests reported here, each query active site was represented by 10 such fingerprints.

The main purpose of the test was to determine whether the method was capable of ranking proteins of similar activity to the query above a large body of other proteins. Using a multipoint representation also gave an opportunity to consider different similarity measures. Three such measures were considered. They were calculated as follows:

For each of the 10 query fingerprints:

(1) The top 15 hits in the target protein were recorded, giving 15 similarity scores where similarity was calculated as the dot product of the vector form of the two fingerprints being assessed. The mean of these similarity scores was then calculated.

(2) The mean spatial spread was calculated as the mean distance between every possible pair of points in the top 15 hits.

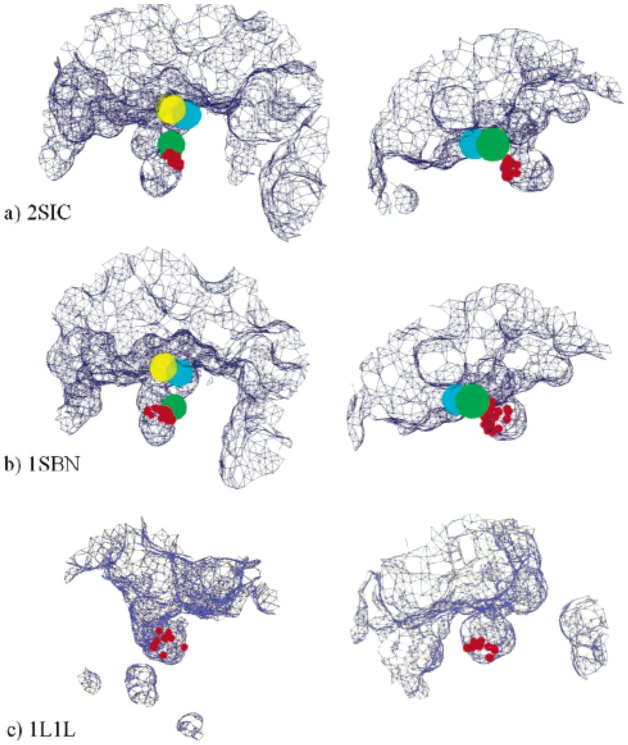

Figure 7. Views of (a) subtilisin query (2SIC) and two hits: (b) 1SBN and (c) 1L1L. In all cases the protein interior is in the lower part of the picture and the outer surface above. Red spheres correspond to the most closely matching fingerprints. In panels a and b green, cyan, and yellow spheres represent the positions of the residues Asp32, His64, and Ser221 of the catalytic triad of subtilisin. (c) 1L1L is a false positive and is discussed later in the text.

The mean similarity measure was then calculated as the mean of the 10 individual fingerprint mean scores, and the mean spatial spread was calculated as the mean of the 10 individual fingerprint mean spreads. The third measure, the product fusion similarity, was then calculated as the product of the mean similarity and the mean spatial spread. The sets of signal fingerprints corresponding to the chosen Fourier fingerprints were also used: the signal similarity score was then calculated as the product of the mean signal similarity score and mean signal spatial spread, exactly as described for the Fourier fingerprints.

Timings were carried out on a Silicon Graphics 180 MHz Origin200 workstation with an R10000 processor. It took approximately 2 min to encode a 300 residue protein with a Connolly surface of 2 dots/Å^2, outputting both the Fourier fingerprint and the signal fingerprint. This is a once and for all initial cost. A single-point fingerprint search of a single protein took about 3−15 s, and a signal fingerprint search is approximately 12 times slower. But it would be sensible only to perform such a search on proteins that passed an initial Fourier screening search, thus reducing the search time required. However it is clear that, using a cluster of fast Linux processors, a search of a database of several hundred proteins could easily be performed in a few minutes.

For these experiments, the initial small data set was augmented by a set of nonhomologous proteins, chosen on the basis that the proteins were monomeric and had no substrate molecules bound, giving a database of 366 proteins. Of these, five were subtilisins (PDB codes 2SIC, 1SBN, 1CSE, 1GNS, 1AF4) of which 1GNS and 1AF4 were unliganded; five were trypsins (PDB codes 1AVW, 1AUJ, 1A0J, 1ANE, 1FY4) of which 1FY4 was unliganded; two were chymotrypsins (PDB codes 5CHA and 1EQ9) of which 5CHA was unliganded; and three were other unliganded members of the trypsin family, namely, α-tryptase, the elastase-like fibrinolytic enzyme component A, and progranzyme K (PDB codes 1LTO, 1M9U, and 1MZA, respectively). One of each of the subtilisin, trypsin, and chymotrypsin proteins was chosen as a query, and the database was searched using its active site, which was represented by the 10 fingerprints closest to the center of its catalytic triad. To see those proteins that might be expected to be retrieved, BLAST searches[29] were carried out against the entire test database to find all proteins homologous to each query protein, and these are listed in Table 2. These proteins are in boldface type (very similar) or italic type (similar) in Tables 3−5.

Search Analysis. The results for subtilisin (query PDB code 2SIC, chain E), trypsin (query PDB code 1AVW), and chymotrypsin (query PDB code 5CHA) are shown in Tables 3−5. The top 60 ranked hits (16% of the database) are shown for each similarity measure described above. In each case proteins whose activity is similar to the query are in bodlface type. Since trypsins and chymotrypsins are closely related, for these searches the proteins with related activity are shown in italic type. The first column gives the mean similarity score; the second column gives the mean spatial spread. We did not necessarily expect this measure to rank similar surfaces at the top, but we were investigating any improvement in rank when this measure was fused with the mean similarity to give the product fusion similarity (column 3). The fourth column shows the signal similarity score.

Table 2. Table of Proteins with Measurable Sequence Similarity to the Query Proteins 2SIC, 1AVW, and 5CHA[a]

PDB code	expectation value (BLAST)	sequence identity (%)	ligand
	2SIC Query		
2SIC	e-154	100	protein inhibitor
1SBN	e-154	100	protein inhibitor
1GNS	e-140	93	none
1CSE	e-104	69	protein inhibitor
1AF4	e-104	69	none
	1AVW Query		
1AVW	e-132	100	protein inhibitor
1AUJ	e-112	82	inhibitor
1ANE	e-109	81	inhibitor
1A0J	e-100	73	inhibitor
5CHA	5e-45	43	none
1LTO	2e-37	38	none
1FY4	4e-36	39	substrate
1MZA	2e-36	36	none
1EQ9	3e-28	33	inhibitor
1M9U	2e-26	31	none
	5CHA Query		
5CHA	e-141	100	none
1AVW	6e-45	43	protein inhibitor
1ANE	3e-45	43	inhibitor
1AUJ	3e-44	42	inhibitor
1A0J	6e-42	41	inhibitor
1LTO	3e-41	38	none
1M9U	4e-33	34	none
1FY4	7e-32	36	substrate
1EQ9	2e-28	37	inhibitor
1MZA	2e-27	35	none

[a] The table gives the BLAST[29] expectation value and sequence identity between each protein and the respective query protein. The type of ligand (if any) present in the coordinate set is also indicated. None of the other proteins in the search database could be retrieved by a BLAST search; therefore, they have no detectable homology (expectation values > 5e+00) to the query proteins.

The results were very encouraging. The subtilisin search (Table 3) ranked all of the subtilisins among the top 60 using each of the mean similarity score, the product fusion similarity score, and the signal similarity score, with all but one being in the top five for the mean similarity score and the signal similarity score. It was particularly encouraging that the unliganded subtilisin 1GNS was ranked in the top five by the mean similarity score, the product fusion similarity score, and the signal similarity score. The other unliganded subtilisin 1AF4 was also found among the top 14% of hits for each of the different similarity measures. It is of particular interest to examine high-scoring false positives in these searches. An example is shown in Figure 7c where a non-serine protease protein (1L1L, ribonucleoside triphosphate reductase) with no sequence homology to the search protein 2SIC is ranked highly on two criteria (Table 3). As can be seen, a small internal surface cavity of somewhat similar shape is identified. However, chemically it is dissimilar, having a relatively hydrophobic character.

The trypsin search (Table 4) found three of the four target trypsin structures using both the product fusion similarity score and the signal similarity score, with the signal similarity score ranking the unliganded 5CHA third. The chymotrypsin search (Table 5) found the single other chymotrypsin target (1EQ9) only when using the signal similarity score (ranked 24). However, considering trypsins and chymotrypsins as a single group, both chymotrypsin and trypsin searches each ranked at least five of the nine target proteins in the top 60

Table 3. Results Gained from Searching for the 2SIC (Subtilisin) Active Site in a Database of 366 Protein Molecules[a]

2SIC mean similarity		2SIC mean spatial spread		2SIC product fused rankings		2SIC signal similarity score		
PDB	value	PDB	value	PDB	value	PDB	value	rank
2SIC	**1.54**	1HZ6	44.3	**2SIC**	**86.4**	**2SIC**	**1207**	1
1SBN	**1.65**	1MWP	46.6	**1SBN**	**112**	**1SBN**	**1565**	2
1L1L	1.82	1H75	46.7	1MWP	161	**1CSE**	**1911**	3
1CSE	**1.99**	1PKO	47.1	1FCQ	164	1L1L	1962	4
1GNS	**2.08**	1I2T	47.5	**1GNS**	**172**	**1GNS**	**1998**	5
1M7X	2.16	1QAD	48.4	1VHS	172	1LCI	2105	6
1E9L	2.16	1NCN	50.0	1NCN	172	1HZF	2132	7
1LCI	2.20	1PHT	51.4	1QAD	174	1M7X	2150	8
1MI1	2.20	1KAF	52.9	1E69	178	1EUR	2174	9
1N7O	2.21	1MN8	54.5	1JFU	180	1H09	2189	10
1HZF	2.23	2GPR	54.8	1UFK	180	1UOK	2194	11
1IQ0	2.24	**2SIC**	**56.2**	1PKO	180	1BCO	2219	12
2PLC	2.24	1KW4	60.2	2GPR	192	1N7O	2220	13
1CIY	2.25	1HYP	61.0	1OHU	200	1E9L	2228	14
1E69	2.27	1JFU	61.5	1H75	213	1E69	2257	15
1FCQ	2.27	1VHS	61.8	2BAA	215	1PGS	2264	16
2BCE	2.31	1BKR	63.5	1KAF	216	1BQC	2276	17
1PGS	2.31	1UOY	67.2	**1CSE**	**216**	1QYI	2286	18
1RI6	2.31	**1SBN**	**67.7**	1MZG	223	1XYZ	2287	19
1E8Y	2.33	1JQQ	68.4	1BKR	223	1MI1	2295	20
1IJQ	2.34	2END	70.0	**1AF4**	**224**	1GSO	2310	21
1NM8	2.35	1V7O	71.6	1WER	227	1RI6	2312	22
1UOK	2.35	1OHU	71.6	2END	238	2PLC	2315	23
1QHT	2.37	1PTX	71.6	1MN8	238	1IQ0	2318	24
1EH9	2.37	1QAU	71.9	1V7O	240	1EH9	2319	25
1BCO	2.40	1FCQ	72.2	1K6K	249	1IJQ	2327	26
1F1S	2.40	1NH9	73.2	1H4A	250	2BCE	2327	27
1EPU	2.40	1MZG	73.4	1K0M	254	1M53	2328	28
3SEB	2.41	1UFK	73.6	1KW4	254	1CIY	2334	29
1QCX	2.41	1C4R	74.6	1HZF	256	1I5P	2341	30
1EG3	2.42	1THX	75.3	1M9U	257	1FEP	2345	31
1GKU	2.42	1TEN	75.6	1C4R	261	3SEB	2349	32
1PJR	2.43	2IGD	76.7	1I2T	265	2BAA	2357	33
1K0M	2.44	1HST	76.8	1C44	271	1GQE	2367	34
1FEP	2.45	1FAS	76.8	1KOE	273	2ENG	2374	35
1UFK	2.45	1E69	78.4	1THX	274	1F1S	2377	36
1M53	2.45	1AGI	78.9	1HYP	276	1GKU	2380	37
2HVM	2.45	1C44	79.3	1PBW	276	2HVM	2382	38
1JY1	2.45	1WER	82.2	1PHT	284	1CEO	2385	39
1NY1	2.45	**1GNS**	**82.3**	1IO2	285	1FCQ	2391	40
1QYI	2.47	1BFG	82.8	1M7X	288	1QCX	2393	41
1I5P	2.47	1H4A	83.4	1KU1	290	1QHT	2395	42
1PW4	2.48	1IIB	84.6	1JQQ	290	1E8Y	2395	43
1EUR	2.48	1Q1H	84.7	1CEO	291	1NM8	2399	44
1O14	2.49	2BAA	85.9	1AGI	294	1JY1	2408	45
2PTD	2.50	1K6K	86.0	1OW1	294	1NIJ	2410	46
1B4X	2.50	1B79	86.6	2ENG	298	1NY1	2416	47
2BAA	2.50	1A62	86.8	1O0X	298	1O0X	2418	48
1I8T	2.52	1R69	87.3	1A62	301	1UFK	2420	49
1JIH	2.52	1I1J	87.4	1BFG	306	1IXK	2420	50
1MTZ	2.52	**1AF4**	**88.1**	1CNU	306	**1AF4**	**2422**	51
1AF4	**2.54**	1JOS	88.8	1VMO	306	1MTZ	2423	52
1MHS	2.54	1R9W	90.1	1RI6	307	1B4X	2429	53
1O7F	2.54	1VMO	90.7	1LCI	308	1MLA	2430	54
1IAL	2.55	1CNU	90.8	1DBX	311	1EG3	2431	55
1CEO	2.56	1M9U	91.0	1YPR	313	1M9U	2436	56
1H7S	2.57	1L8R	91.2	1HZ6	316	2PTD	2442	57
1GQE	2.58	1CQM	92.1	1TEN	316	1O14	2443	58
1GSO	2.58	1C2A	92.6	1AGJ	318	1HYQ	2446	59
1MLA	2.58	1TMY	92.7	1R9W	325	1WER	2455	60

[a] For each fingerprint search performed, the top 15 most similar hits were used to compile measures of mean similarity score and mean spatial spread of the cluster of hits. Each of the above scores represent the mean results of 10 such searches, each using a Fourier fingerprint representing a point close to the center of the searched for active site. (NB signal-based fingerprint scores give higher figures as a result of the binary 2 byte per value data storage method used to compress signal output files. 1° angular dissimilarity is 364.1 in this scoring system.) The top 60 ranked hits are shown (≈16% of the database) with those known to be similarly active in boldface type.

using both the product fused ranking and the signal similarity score, with the chymotrypsin search finding all except two of the targets in the top 60, including three in the top five using the signal similarity score. Only one target was not found in any of the trypsin/chymotrypsin searches. This was pro-granzyme K (1MZA), in which the molecule crystallized

Table 4. Results Gained from Searching for the 1AVW (Trypsin) Active Site in a Database of 366 Protein Molecules[a]

1AVW mean similarity		1AVW mean spatial spread		1AVW product fused rankings		1AVW signal similarity score		
PDB	value	PDB	value	PDB	value	PDB	value	rank
1AVW	**1.71**	1H6T	53.8	**1AUJ**	**105.6**	**1AVW**	**1345**	1
1I5P	1.83	**1AUJ**	**55.3**	1H6T	131.3	**1AUJ**	**2014**	2
1BCO	1.87	1LWB	58.5	1TOL	142.3	*5CHA*	*2192*	3
1F1S	1.88	1TOL	58.9	**1AVW**	**143.5**	1M7X	2195	4
1CIY	1.89	2CPL	59.3	2CPL	149.6	1UEK	2269	5
1L1L	1.89	1PGV	60.2	1PGV	150.9	1I5P	2287	6
2BCE	1.90	1NO1	61.0	1KNG	172.3	1IQ0	2299	7
1AUJ	**1.91**	1KNG	63.7	1DBX	174.8	1PGS	2299	8
1PGS	1.91	1KAF	63.8	1LWB	179.4	**1A0J**	**2355**	9
1M7X	1.91	1KW4	67.1	1C2A	180.5	1FEP	2356	10
1OXW	1.94	1C2A	69.5	1JHS	185.9	1O9G	2361	11
1I8T	1.95	1CFY	70.1	1KAF	185.9	1LFP	2373	12
1B4X	1.96	1J48	72.3	1B0F	189.9	1OXW	2384	13
1M53	1.99	1FAS	73.7	*5CHA*	*191.3*	1GX3	2384	14
1GS0	1.99	1F32	74.8	2PTH	192.9	1BCO	2384	15
1GKU	1.99	1DBX	76.5	1CFY	194.2	1GKU	2385	16
5CHA	*2.00*	1JFU	77.0	1JFU	197.3	1L1L	2397	17
3SEB	2.01	1M4J	77.5	2LIS	199.4	1EH9	2400	18
1N7O	2.03	2LIS	78.6	1UFK	221.1	**1ANE**	**2402**	19
1LCI	2.03	1JHS	81.8	1NO1	234.1	1E8Y	2407	20
1LRZ	2.04	**1AVW**	**83.9**	1DVO	234.8	3SEB	2408	21
1VJ1	2.06	1GH2	85.5	1Q2Y	236.2	1EG3	2417	22
1O7F	2.07	1DVO	85.5	1GH2	237.9	1UOK	2421	23
1NM8	2.07	1PTX	85.6	1M4J	243.1	1EUR	2426	24
1LTO	*2.07*	1M1S	86.0	1F32	245.1	1M53	2428	25
1ET9	2.08	1LU4	87.3	1ELT	247.4	1LRZ	2428	26
1MHS	2.08	1ONL	89.6	1ANE	248.0	1CIY	2429	27
1G8P	2.08	1B0F	90.6	1KW4	252.4	1N7O	2433	28
1UOK	2.09	2PTH	91.6	1LU4	254.8	1AGJ	2437	29
1E8Y	2.09	1VCA	92.2	**1FY4**	**256.2**	1GS0	2439	30
1B0F	2.10	1MN8	92.3	1ES5	258.1	1F1S	2442	31
1JIH	2.10	1AGI	92.5	1QO2	259.6	1QHT	2444	32
1MI1	2.10	1HZ6	93.7	2FCB	260.6	1E9L	2448	33
2PTH	2.11	1ROW	93.9	1L1N	260.7	*1M9U*	*2451*	34
1QHT	2.11	2FCB	94.5	1LKF	261.0	1O7F	2455	35
1HZF	2.11	1Q2Y	95.6	1KZF	265.2	1IXV	2463	36
1EG3	2.12	*5CHA*	*95.7*	1VCA	265.9	1EPU	2468	37
1G4M	2.12	1NOA	99.2	1Q88	265.9	1G8P	2474	38
1ES5	2.12	1BM8	101	1J3A	266.8	1R88	2476	39
1ELT	2.12	1AHO	101	1JL1	267.7	1IJQ	2476	40
1XYZ	2.12	1CDY	101	1ONL	271.1	1TCB	2478	41
1GQE	2.13	1J3A	103	*1LTO*	*274.0*	1GSO	2482	42
1H1N	2.13	1UFK	103	1J48	275.5	*1EQ9*	*2485*	43
1EPU	2.13	1MIL	103	1AGI	278.5	1RI6	2488	44
1CRZ	2.13	1JL1	104	1ET9	285.9	1Q42	2490	45
1GSO	2.14	1H3L	105	2GPR	288.6	1O14	2491	46
1ANE	**2.14**	1BFG	106	1NAR	303.6	1KZF	2493	47
1FDW	2.14	1EW4	109	1O9G	310.5	1JIH	2494	48
1UFK	2.14	1QO2	110	1M1S	310.8	1XYZ	2500	49
1E9L	2.15	2GPR	111	1PTX	312.1	1OTK	2503	50
1L1N	2.15	1IIB	111	1BFG	312.4	1G61	2505	51
1JY1	2.17	1MZG	113	1CDY	312.5	1IXK	2506	52
1P4X	2.17	1Q88	114	1KOE	313.2	1NM8	2508	53
1K47	2.17	1KOE	114	1H1N	318.1	1K47	2510	54
1B04	2.18	1FY4	115	1PRZ	319.4	1IAL	2515	55
1EDG	2.18	1ANE	116	1FAS	323.2	1LKF	2516	56
1NAR	2.18	1ELT	117	1MZG	323.2	1PW4	2517	57
1ND7	2.18	1R69	117	**1A0J**	**323.8**	1NAR	2527	58
1O14	2.19	1LKF	118	1MIL	334.6	1LCI	2530	59
1OTK	2.19	1HQZ	119	*1M9U*	*336.3*	1ARB	2531	60

[a] For each fingerprint search performed, the top 15 most similar hits were used to compile measures of mean similarity score and mean spatial spread of the cluster of hits. Each of the above scores represent the mean results of 10 such searches, each using a Fourier fingerprint representing a point close to the center of the searched for active site. The top 60 ranked hits are shown (≈16% of the database) with those known to be similarly active in boldface type and those of related activity in italic type.

Table 5. Results Gained from Searching for the 5CHA (Chymotrypsin) Active Site in a Database of 366 Protein Molecules[a]

5CHA mean similarity		5CHA mean spatial spread		5CHA product fused rankings		5CHA signal similarity score		
PDB	value	PDB	value	PDB	value	PDB	value	rank
5CHA	**1.56**	1M1S	55.4	**5CHA**	**131**	**5CHA**	**1322**	1
1L1L	1.85	1GH2	62.9	1M1S	163	*1AUJ*	*2071*	2
1PGS	1.87	1H3L	73.4	1GH2	163	1L1L	2214	3
1GKU	1.90	1KW4	79.9	1QO2	173	*1AVW*	*2215*	4
1CRZ	1.91	1QO2	81.9	2LIS	202	1PGS	2220	5
1LCI	1.92	2LIS	84.5	1JSS	206	*1ANE*	*2228*	6
1I5P	1.93	**5CHA**	**82.9**	1CFY	223	1OXW	2236	7
1M7X	1.95	1CFY	88.7	1L1N	229	1BCO	2249	8
1E9L	1.95	1JSS	93.4	1PGV	246	1CIY	2265	9
1OXW	1.96	1C2A	98.7	1H3L	250	1I5P	2271	10
1CIY	1.97	1NO1	102	2CPL	266	1UOK	2272	11
1F1S	1.99	1L1N	102	1KW4	273	1IQ0	2282	12
1AVW	*1.99*	1PGV	105	1J3A	273	1UEK	2288	13
1BCO	1.99	1H75	109	1Q2Y	278	1M53	2291	14
1JIH	1.99	1Q2Y	111	1DBX	284	1LCI	2293	15
1GSO	2.00	2CPL	111	1C2A	281	1QHT	2293	16
1M53	2.01	1FAS	113	*1AUJ*	*283*	1E9L	2294	17
1IQ0	2.01	1TOL	112	1UFK	287	1M7X	2297	18
1EG3	2.02	1KNG	114	1KNG	292	2PLC	2315	19
1UOK	2.02	1DBX	114	1TOL	294	1EH9	2317	20
2BCE	2.02	1LU4	118	1ES5	294	1F1S	2319	21
1LKF	2.00	1NKO	118	1OHU	318	1N7O	2321	22
1N7O	2.03	1F32	120	1LU4	318	1GX3	2326	23
1NM8	2.00	1R26	120	1KXO	321	**1EQ9**	**2329**	24
1HZF	2.03	1KTE	120	1RI6	326	1GSO	2333	25
1AUJ	*2.04*	1H03	123	1JHS	331	1IXK	2334	26
1EH9	2.02	1J3A	122	1R26	336	1E8Y	2335	27
1EPU	2.03	1UFK	123	1EZK	339	1JIH	2341	28
1E8Y	2.06	1KXO	130	1NKO	347	*1A0J*	*2348*	29
1QHT	2.08	1OHU	131	1NO1	350	2HVM	2349	30
1RI6	2.06	1EZK	133	1F32	351	1XYZ	2349	31
1LRZ	2.09	1ES5	135	*1M9U*	*355*	1LRZ	2354	32
1I8T	2.07	1MIL	136	1KTE	356	1LKF	2361	33
2PLC	2.09	1AGI	139	*1AVW*	*354*	1GKU	2362	34
1O7F	2.10	1PTX	138	1AGI	355	1EUR	2362	35
1QCX	2.10	1PHT	139	1D2O	360	1LFP	2366	36
1MI1	2.10	*1AUJ*	*139*	1H75	364	1ET9	2369	37
1OI7	2.06	1GRJ	139	1QNT	368	1TCB	2374	38
1XYZ	2.10	1AMX	143	1AMX	371	1I8T	2375	39
1NY1	2.12	1JHS	144	1Q42	371	*1M9U*	*2375*	40
1FDW	2.12	1GD8	145	2GPR	377	1JLN	2376	41
1JY1	2.11	1FC3	145	1IXV	375	2BCE	2376	42
1ANE	*2.12*	1TEN	147	1JL1	385	1RI6	2378	43
1QO2	2.11	1ROW	147	1NAR	384	1IJQ	2379	44
1FVR	2.12	1ONL	147	1JFU	382	1ARB	2380	45
1NAR	2.10	1C44	149	1GRJ	393	1VJ1	2382	46
1G8P	2.13	2GPR	151	1EG3	400	1ELT	2387	47
1E4F	2.14	1BFG	152	1NGN	396	1QCX	2390	48
1PJR	2.14	1NGN	153	2PTH	395	1E4F	2391	49
1IJQ	2.14	1AHO	154	1BFG	401	1NM8	2393	50
1JLN	2.16	1UCS	156	*1FY4*	*405*	1EG3	2396	51
1QNT	2.16	1UCS	157	1MIL	405	1FEP	2396	52
1YFO	2.15	1HYP	156	1PBW	403	1GS0	2399	53
1VIN	2.16	1RI6	158	1JYH	416	1ES5	2403	54
1G4M	2.16	1JFU	158	1FC3	414	1K47	2406	55
1L2L	2.16	1P9I	159	1ICX	423	*1FY4*	*2408*	56
1K47	2.17	1JL1	158	1FAS	427	1CRZ	2411	57
1EDG	2.17	1CDY	160	1ONL	426	1O7F	2411	58
1A0J	*2.17*	1DVO	160	*1A0J*	*432*	1EPU	2412	59
1D2O	2.17	1NG6	161	1O9G	439	2SIC	2415	60

[a] For each fingerprint search performed, the top 15 most similar hits were used to compile measures of mean similarity score and mean spatial spread of the cluster of hits. Each of the above scores represent the mean results of 10 such searches, each using a Fourier fingerprint representing a point close to the center of the searched for active site. The top 60 ranked hits are shown (≈16% of the database) with those known to be similarly active in boldface type and those of related activity in italic type.

was a site-directed mutant protein.[30] In this trypsin-like protease one of the catalytic triad residues (Ser195) is replaced by alanine, and furthermore a second triad side chain (His57) adopts a different conformation (χ1 rotated by 100°) to the other members of the family. The latter has been attributed not to the 195 mutation but to a different conformation of the nearby Gly215.[30] The surface shape in this region is therefore significantly different from other members of the family, and it is not surprising that it was not retrieved in the search. In general, the results of both searches are promising, and it is also encouraging that in all cases some unliganded coordinate sets were retrieved using liganded queries and vice versa.

There is some indication that fusing spread information improved the rankings for similarly active compounds that performed less well purely based on fingerprint similarity—this perhaps filtered false positives that were indicated by large spread values. In all cases the signal similarity score was at least as good and very often better than any other scoring method. This was not surprising since using just the Fourier amplitudes allowed matches between sets of points that did not have similar phases, thus introducing false positives. Using the phase information then eliminated this problem.

CONCLUSIONS

The comparison of the surfaces of 3D protein structures is a computationally demanding task. In this paper we have considered several ways of reducing the complexity of the representations that are needed to enable the identification of structurally similar surface patches. Our experiments have highlighted the need for localized canonical representations that are independent of the procedure used to sample the surfaces. One such representation is obtained by a Fourier analysis of the distribution of surface curvature on concentric spheres around a surface point. Searching experiments on a set of 366 proteins demonstrate that this provides an effective and an efficient technique for the matching of protein surfaces.

Both the Fourier fingerprints and signal fingerprints are purely shape-based representations, yet have been shown capable of locating similar active-site regions in globally quite different proteins. In addition, there is clearly much more local information available that can be used to enhance the description, such as the hydrogen-bonding capability, hydrophobicity, and electrostatic nature of the surface in the neighborhood of a surface point. In principle this enhancement can be achieved by augmenting the polar fingerprint vector with chemical descriptors of the nearby surface.

We have previously noted that the method does not give an alignment of the two surfaces but rather locates regions of similarity indicated by clusters of matching fingerprints. Therefore to produce a rigorous and optimal matching of two matched surfaces, full three-dimensional local comparison of the two subsurfaces would then be required. There are a number of possible methods already available, for example, those of Poirrette et al.,[17] Schmitt et al.,[23] or Cosgrove et al.[18]

In the tests described here, we have searched a database of proteins for a given active site in order to demonstrate the Fourier fingerprint method. To use the method to assign a putative function to an unknown protein, we envisage that the reverse procedure may be employed (i.e., that a database of active sites of known function will be established). Each active site will be represented by its sets of Fourier fingerprints, and then each of these may in turn be used as a query to search the unknown function (target) protein. This development is the subject of further investigation.

ACKNOWLEDGMENT

We thank the Biotechnology and Biological Sciences Research Council for funding.

REFERENCES AND NOTES

(1) Hendrickson, W. A. Synchrotron crystallography. *Trends Biochem. Sci.* **2000**, *25* (12), 637−643.

(2) Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J. R.; Booth, V.; Mackereth, C. D.; Saridakis, V.; Ekiel, I.; Kozlov, G.; Maxwell, K. L.; Wu, N.; McIntosh, L. P.; Gehring, K.; Kennedy, M. A.; Davidson, A. R.; Pai, E. F.; Gerstein, M.; Edwards, A. M.; Arrowsmith, C. H. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **2000**, *7* (10), 903−909.

(3) Schmid, M. B. Seeing is believing: The impact of structural genomics on antimicrobial drug discovery. *Nat. Rev. Microbiol.* **2004**, *2* (9), 739−746.

(4) Hwang, K. Y.; Chung, J. H.; Kim, S. H.; Han, Y. S.; Cho, Y. J. Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nat. Struct. Biol.* **1999**, *6* (7), 691−696.

(5) Zarembinski, T. I.; Hung, L. W.; Mueller-Dieckmann, H. J.; Kim, K. K.; Yokota, H.; Kim, R.; Kim, S. H. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (26), 15189−15193.

(6) Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **1993**, *229* (3), 707−721.

(7) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of techniques derived from graph-theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **1990**, *212* (1), 151−166.

(8) Artymiuk, P. J.; Rice, D. W.; Poirrette, A. R.; Willett, P. A Tale of 2 Synthetases. *Nat. Struct. Biol.* **1994**, *1* (11), 758−760.

(9) Artymiuk, P. J.; Poirrette, A. R.; Rice, D. W.; Willett, P. A polymerase I palm in adenylyl cyclase. *Nature* **1997**, *388* (6637), 33−34.

(10) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(11) Richards, F. M. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 15−176.

(12) Via, A.; Ferre, F.; Brannetti, B.; Helmer-Citterich, M. Protein surface similarities: A survey of methods to describe and compare protein surfaces. *Cell. Mol. Life Sci.* **2000**, *57* (13−14), 1970−1977.

(13) Connolly, M. L. Analytical molecular-surface calculation. *J. Appl. Crystallogr.* **1983**, *16* (Oct), 548−558.

(14) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 708−713.

(15) Nicholls, A.; Sharp, K. A.; Honig, B. Protein folding and association—insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **1991**, *11* (4), 281−296.

(16) Nicholls, A.; Bharadwaj, R.; Honig, B. GRASP—graphical representation and analysis of surface-properties. *Biophys. J.* **1993**, *64* (2), A166−A166.

(17) Poirrette, A. R.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Comparison of protein surfaces using a genetic algorithm. *J. Comput.-Aided Mol. Des.* **1997**, *11* (6), 557−569.

(18) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14* (6), 573−591.

(19) Cao, J.; Pham, D. K.; Tonge, L.; Nicolau, D. V. Predicting surface properties of proteins on the Connolly molecular surface. *Smart Mater. Struct.* **2002**, *11* (5), 772−777.

(20) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comput. Chem.* **2002**, *23* (12), 1176−1187.

(21) Pickering, S. J.; Bulpitt, A. J.; Efford, N.; Gold, N. D.; Westhead, D. R. AI-based algorithms for protein surface comparisons. *Comput. Chem.* **2001**, *26* (1), 79−84.

(22) Zauhar, R. J.; Moyna, G.; Tian, L. F.; Li, Z. J.; Welsh, W. J. Shape signatures: A new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003**, *46* (26), 5674−5690.

(23) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323* (2), 387−406.

(24) Leicester, S.; Finney, J.; Bywater, R. A quantitative representation of molecular-surface shape. 1. Theory and development of the method. *J. Math. Chem.* **1994**, *16* (3−4), 315−341.

(25) Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins* **2000**, *39* (2), 178−194.

(26) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A graph-theoretic approach to the identification of 3-dimensional patterns of amino-acid side-chains in protein structures. *J. Mol. Biol.* **1994**, *243* (2), 327−344.

(27) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.

(28) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH—a hierarchic classification of protein domain structures. *Structure* **1997**, *5* (8), 1093−1108.

(29) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.

(30) Hink-Schauer, C.; Estebanez-Perpina, E.; Wilharm, E.; Fuentes-Prior, P.; Klinkert, W.; Bode, W.; Jenne, D. E. The 2.2-Å crystal structure of human pro-granzyme K reveals a rigid zymogen with unusual features. *J. Biol. Chem.* **2002**, *277* (52), 50923−50933.

CI049647J