Calculation of Protein Domain Structural Similarity Using Two-Dimensional Representations

Benjamin C. P. Allen, Guy H. Grant, and W. Graham Richards*

Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QH, United Kingdom

Received August 2, 2002

By reducing protein structures to two-dimensional representations, it is possible to speed up the alignment of the structures and hence calculate similarity indices faster that using three-dimensional representations. Using amino acid based representations gives much better discrimination between proteins and faster calculations. Taking into account the relative similarity of the amino acids involved allowed improved accuracy at very little time cost.

1. INTRODUCTION

Proteins are the fundamental units of living systems. Understanding the structure and function of proteins is crucial for the design of new and effective drugs. Thus determining the functionality of proteins directly is the focus of much biochemical research.[1] However, discovering the function of a given protein from biochemical studies is an extremely time intensive process, so it is necessary to seek alternative strategies. By calculating the similarity between a novel protein and a range of proteins of known functionality, it will hopefully be possible to predict the functionality of the unknown protein. This work develops a rapid computational technique for calculating protein similarity. Proteins consist of one or more polymeric chains of amino acids, generally containing between 50 and 450 amino acids per chain. A protein can be described by its amino acid sequence or by its three-dimensional structure. Sequence data can be obtained from genetic data, and the Human Genome project has provided sequences for an estimated 30 000 genes, giving approximately 100 000 proteins.[2] Structural data can only be obtained by X-ray crystallography, NMR techniques, or other forms of crystalographic analysis, all of which are much more time-consuming, and hence there are only 18 244 structures in the Protein Data Bank, many of which may be almost identical.[3] Several classification schemes have been developed to organize systematically this array of proteins.[4] In this work, the CATH system is used as a basis for analyzing the similarity results.[5] CATH stands for Class, Architecture, Topology, Homologous Superfamily. Proteins are automatically assigned a Class on the basis of their secondary structure content. They are then visually classified into one of about forty Architectures, which include such well-known structures as the beta- propellor and alpha four-helix bundle. A program called SSAP is then used to divide members of a given architecture into Topology groups on the basis of the connectivity of the major structural elements.[6] Finally the Homologous Superfamilies are assigned according to levels of sequence identity. To validate the new technique, it was shown that the calculated similarity values could be used to assign proteins accurately to the correct Homologous Superfamily groups.

A number of techniques have been developed for the calculation of protein similarity.[4] There are a wide range of statistical techniques for assessing the sequence similarity between proteins, and these methods are used for such applications as calculating genetic distances between species.[2] Unfortunately, it has been demonstrated that protein sequence similarity is of little use in determining functional homologies, since proteins with highly divergent sequences can have very similar structures.[7] Structural features are a much better guide to function. A smaller number of techniques have been developed recently for calculating structural similarity. Many of the earlier techniques required that the sequences of the two proteins be prealigned, which is frequently neither possible nor practical, and limits their application to small sets of genetically similar proteins.[8] Standard similarity techniques developed for small molecules have been scaled up for use with proteins.[9] Unfortunately such methods are extremely computationally intensive, with time scales of order 25 s per pairwise alignment, which is not practical for large data sets.[9] These tests have shown that overall protein similarity provides a fairly good guide to protein function.[9] Several heirarchical approaches, such as SCOP[10] and CATH, have been used to classify proteins, but they all require human input to classify the overall structure.[11] A method based on distance matrix overlap, called FSSP, has been developed.[12] It finds similar structural elements between a pair of proteins and assesses the degree of topological similarity between the two. This provides a pairwise similarity measure for use with proteins, with calculation times of about 11 s per pairwise calculation.[13] These timings are obtained from publications from the mid 1990s and have certainly been outdone using more modern computer hardware.

Dimensionality reduction is a mathematical technique originally developed for data analysis.[14] Using Sammon's mapping, a 3D structure can be reduced to 2D while keeping the atomic distance matrix almost unchanged.[15] The most time-consuming step in the calculation of similarity indices is the alignment of the two molecules to be compared, and this can be dramatically speeded up by performing the

* Corresponding author phone: (0)1865 275908; fax: (0)1865 275905; e-mail: graham.richards@chem.ox.ac.uk.

10.1021/ci020275t CCC: $25.00
Published on Web 12/13/2002

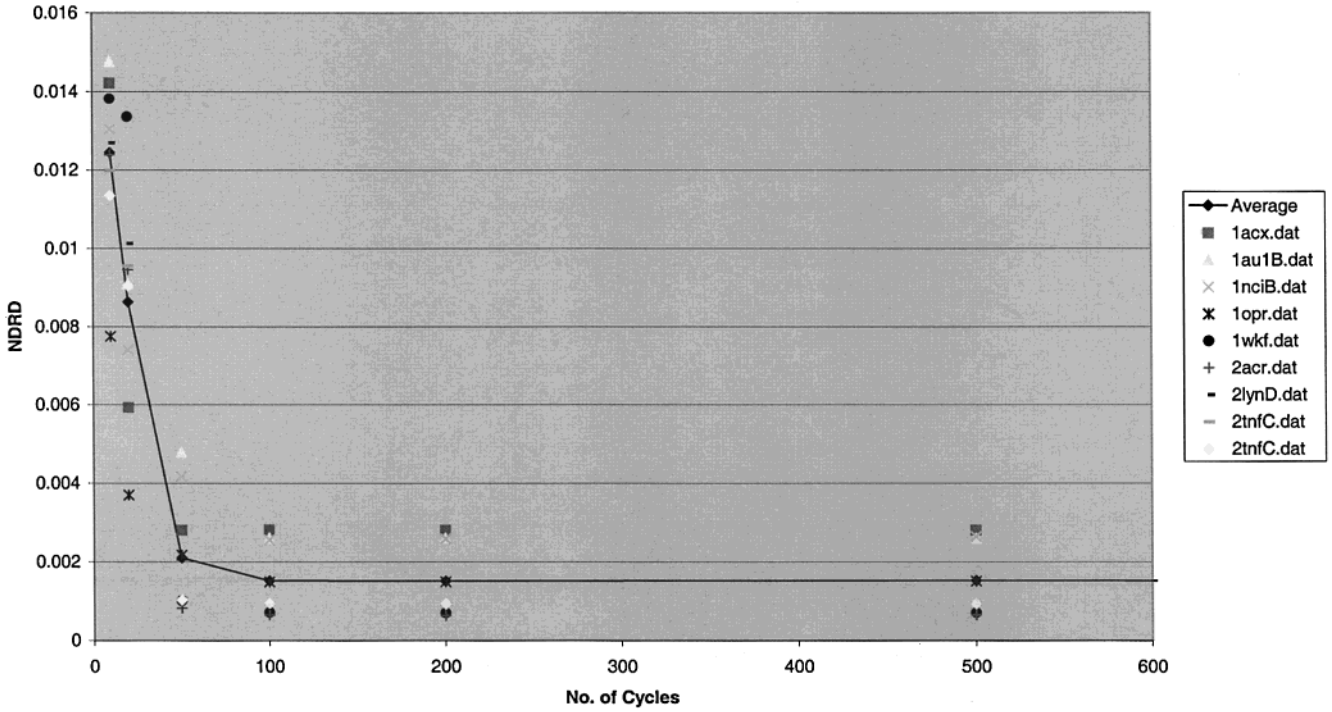

Figure 1. Distance matrix errors versus no. of algorithm cycles for a test set of Sammon mapped proteins.

operation in 2D. In general the use of 2D representations can provide a genuine increase in calculation speed, all else being equal.[16] The technique has been shown to perform well for small molecules.[16] Using this method, whole molecule similarity indices for proteins were calculated, and the accuracy of the similarity values was estimated by comparison with CATH classifications.

2. METHODS

2.1. Sammon Mapping. Sammon mapping was developed to reduce higher dimensional data sets to two dimensions to allow analysis by eye.[14] Here it is used to produce two-dimensional molecular representations from the three-dimensional atomic coordinates of a protein. The Sammon mapping is a nonlinear algorithm that takes an arbitrary set of two-dimensional coordinates and minimizes the error measured between the distance matrices of the two- and three-dimensional data sets by adjusting those coordinates, using the steepest decent algorithm.[14] This results in a set of two-dimensional coordinates with a highly similar distance matrix to the three-dimensional atomic coordinates from which they are derived. This makes it well suited to simplifying any problem in which interatomic distances are important, such as the calculation of molecular similarity indices. The algorithm is prone to errors when dealing with spherical systems.[15] However testing showed that the problem is insignificant when dealing with large complex systems such as proteins.

A set of eight proteins, covering a spectrum of sizes from 40 to 500 amino acids in length, was used to test the accuracy and consistency of the method. First a 2D mapping was calculated from each protein at a range of optimization parameter values. The resultant distance matrix errors calculated between the 2D and 3D distance matrices are shown in Figure 1, measured using the normalized distance matrix root-mean-squared deviation (NDRD) given by

$$E_{nm} = \frac{1}{\sum_{i \neq j}^{N} d_{ij}} \left(\sum_{i \neq j}^{N} (d_{ij}^{n} - d_{ij}^{m})''\right)^{1/2} \quad (1)$$

This indicates that at an optimization time of 100 cycles all the protein representations had reached acceptable levels of accuracy. Next, eight mappings of each protein at each of a smaller range of optimization parameter values were calculated. The average deviations from the mean for each set of replicates were all less than 2% at 100 cycles, confirming that 100 cycles provided consistent and accurate two-dimensional maps. Timings for the Sammon mapping at 100 cycles for all 127 test proteins are shown in Figure 2 and indicate that for the largest proteins the calculations took approximately 20 min. The same tests were repeated for the amino acid based representations. They showed that accurate and consistent results could be obtained for systems of that size with a 1000 cycle optimization time and that the largest proteins took less than 1 min to map. All timing calculations refer to a PIII-450 processor with 64Mb RAM.

2.2. Protein Similarity. Protein similarity is a natural extension of the techniques used to calculate similarity in sets of smaller molecules.[16] Similarity calculations consist of optimally superimposing the target molecules, and then calculating the degree of overlap between them.[17] The most common technique is to rotate one of the molecules with respect to the other, calculating the similarity index for each relative orientation, and taking the best calculated value as the optimum.[18] For three-dimensional molecular representations there are six degrees of freedom; *x*, *y*, and *z* translations and the three Euler angles. By using two-dimensional representations the degrees of freedom are reduced to *x*, *y*,

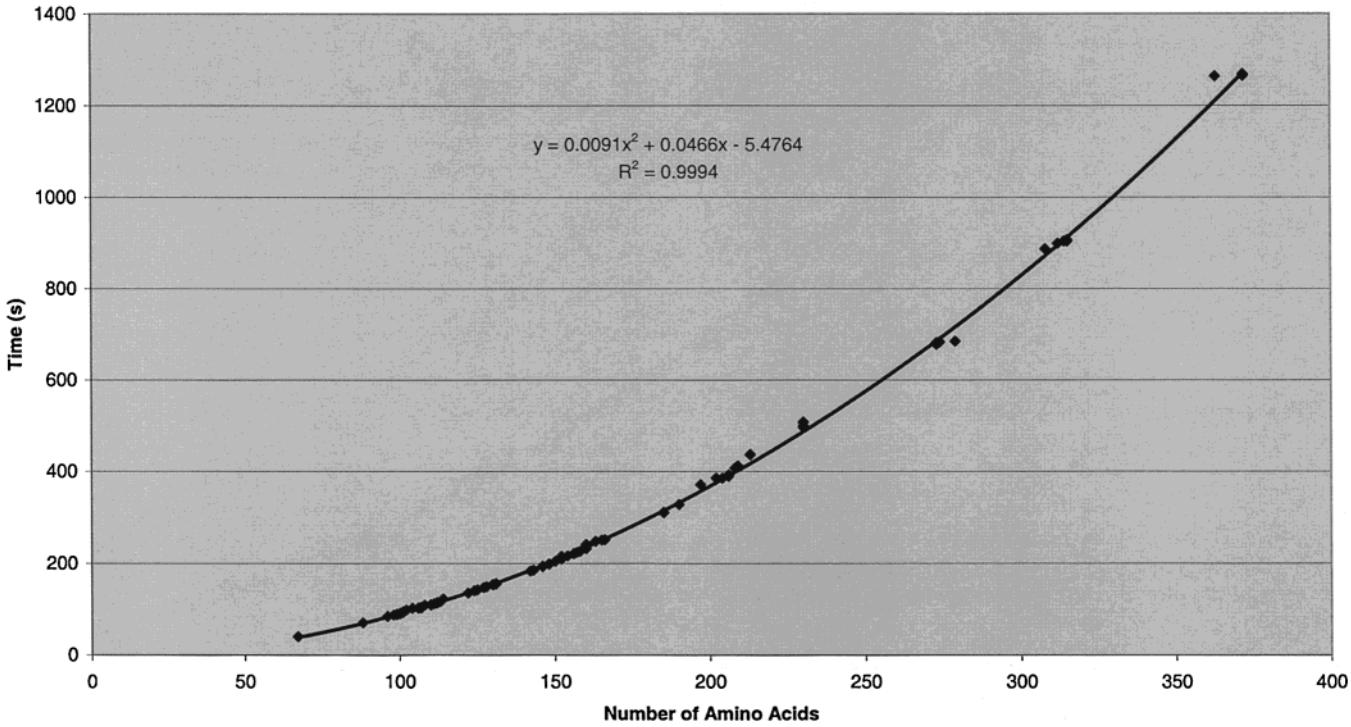

Figure 2. Sammon mapping times versus no. of amino acid for all atom representations of the full set of proteins.

and θ. By aligning the centers of mass of the two molecules, the translational degrees of freedom can be neglected, leaving three rotational degrees of freedom in the 3D case and only one for 2D systems. It has been shown for small molecules that using 2D representations reduces the calculation times considerably and that the reduction is greater for larger systems, making the technique particularly valuable for proteins.

Any form of spatially varying property can be used to calculate the degree of overlap between a pair of superimposed molecules. Commonly used properties include electron density and measures of shape such as the overlap of van der Waals spheres.[17] The general form of the calculations is given by

$$R_{AB} = \frac{\int F_A(v)F_B(v)dv}{(\int F_A(v)2dv)^{1/2}(\int F_B(v)2dv)^{1/2}} \qquad (2)$$

where $F_A(v)$ describes the property of interest for molecule A. The 2D representations produces by Sammon mapping consist of atom positions, hence only properties whose overall function can be produced by summation of individual atom functions are of interest. Such functions can be described by

$$F_A(v) = \sum_{a=1}^{n} f_a(v) \qquad (3)$$

where molecule A has n atoms, and each individual atom has a property function $f(v)$. This gives an overall similarity function:

$$R_{AB} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\int f_a^i(v) \times f_b^j(v)dv}{(\sum_{i=1}^{n}\sum_{j=1}^{n}\int f_a^i(v) \times f_a^j(v)dv)^{1/2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\int f_b^i(v) \times f_b^j(v)dv)^{1/2}} \qquad (4)$$

For computational purposes, it is important that $f(v)$, or at least its integrals, can be easily calculated, so in this work all functions were replaced by least-squares fitted Gaussians. The use of Gaussian functions to represent atoms has been shown to provide good results in the 3D case, both for proteins and small molecules.[9,19] The time taken to perform a similarity calculation using this formula with a single Gaussian on a PIII 450 with 64 Mb RAM was measured to be 0.14 μs for a system with 1 atom center.

2.3. Amino Acids. For proteins, in addition to the all atom representations previously used for smaller molecules, it is possible to use amino acid based representations, where the protein is treated art the residue level. Instead of using an atom based property function, average properties can be calculated for each individual amino acid. Each amino acid was described by a single Gaussian function, centered at the position of the alpha carbon, of width equal to the radius of gyration of the amino acid side chain in its extended form, and height equal to the molecular mass of the amino acid, as this provides a reasonable assessment for the likely radius at which the amino acid will interact with its neighbours, and a crude assessment of the strength of the van der Waal's interaction. Tests showed that using these properties provided better discrimination than using identical Gaussians for all

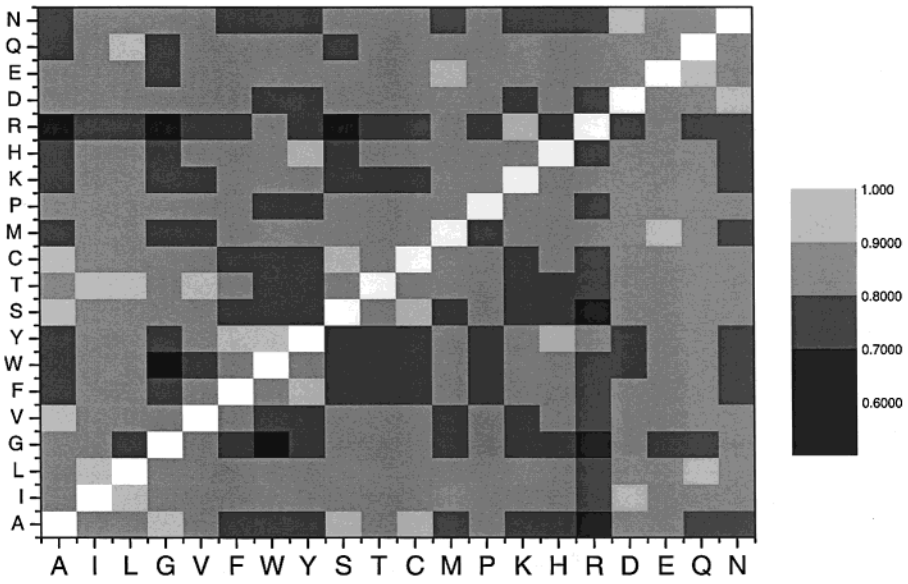

Figure 3. Amino acid similarity values.

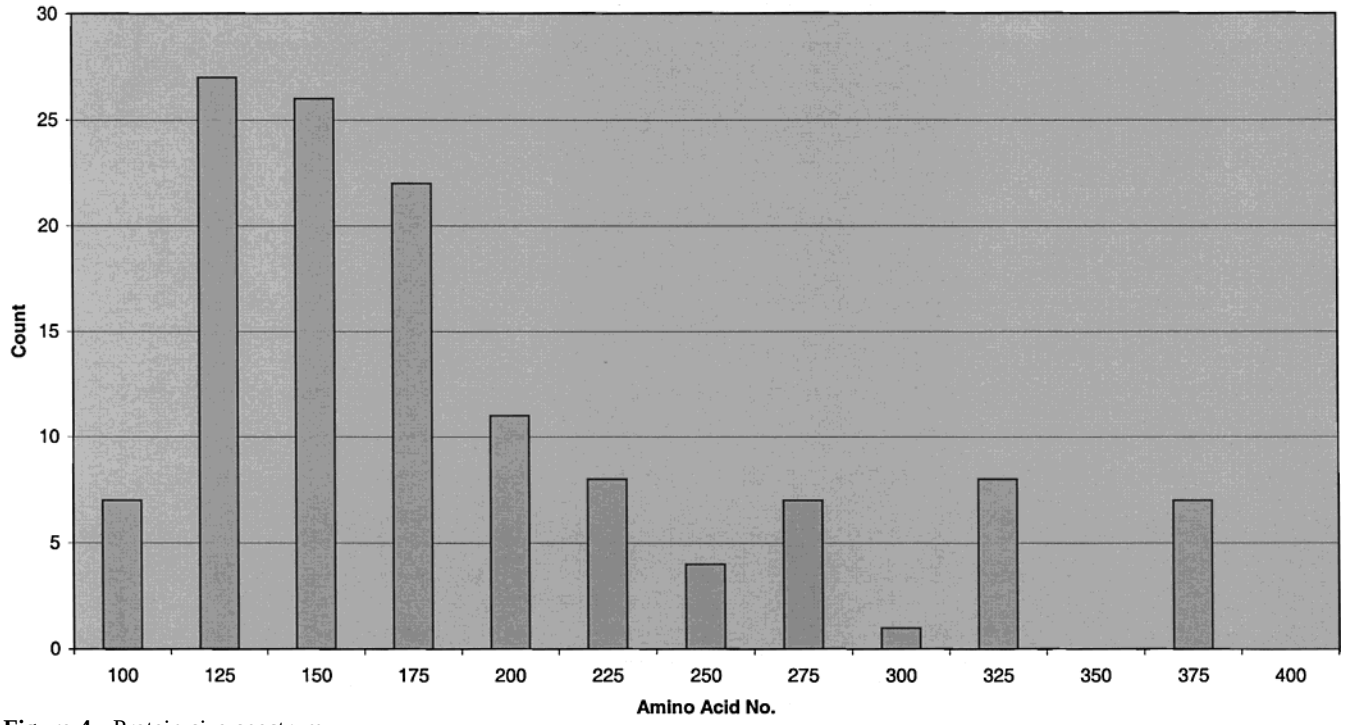

Figure 4. Protein size spectrum.

the amino acids. This significantly reduces the number of property centers used, thereby significantly reducing the calculation time. To increase discrimination further, a set of small molecule similarities were calculated for the set of amino acids, and used to modify the overlap terms. The similarity values were calculated using rigid rotation at 1 degree increments of the energy optimized conformers. The overlap between a given pair of amino acid centers became

$$O_{ab} = f_a^i(v) \times f_b^j(v) \times S(a^i, b^j) \qquad (5)$$

where $S(a, b)$ is the similarity between the two amino acids. The amino acid similarity values were calculated using the ASP module within the program TSAR.[21,22] They are shown in Figure 3.

Table 1: Protein Homologous Superfamilies

class	architecture	topology	H.S.	no.
mostly α	bundle (0 or 180)	lysin	bmf	8
			lis	10
		ferritin	ick	4
			itf	4
		4helix	asr	10
	barrel	isoprenoid syn.	eas	3
		glycosyltransferase	cem	1
mostly *β*	sandwich (2 sheets)	immunoglobulin	akp	6
			ncg	8
		jelly roll	xnb	10
			tnf	9
	barrel (1 curved sheet)	thrombin	hay	7
			ucy	7
		oncogene	jsg	2
α/*β*	barrel	TIM barrel	ads	8
			wkf	6
		serine protease	cmv	9
	α*β*α-sandwich	Rossmann fold	opr	4
			thm	8

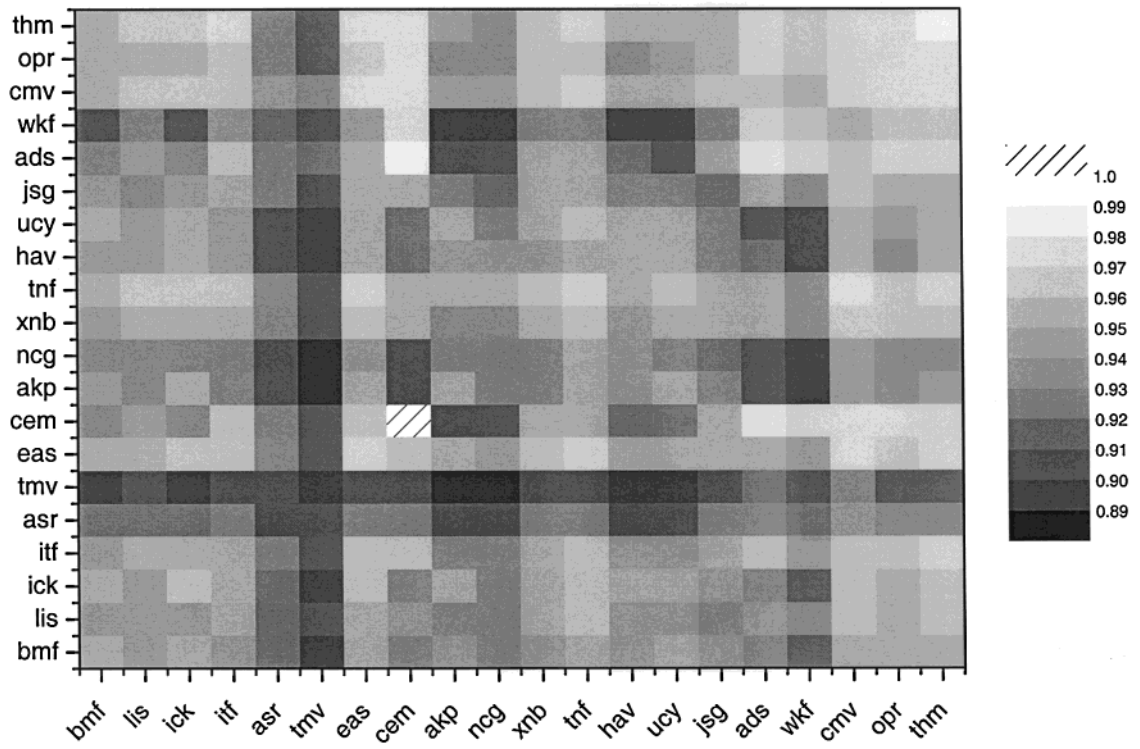

Figure 5. All atom protein similarity values, averaged over homologous superfamily groups.

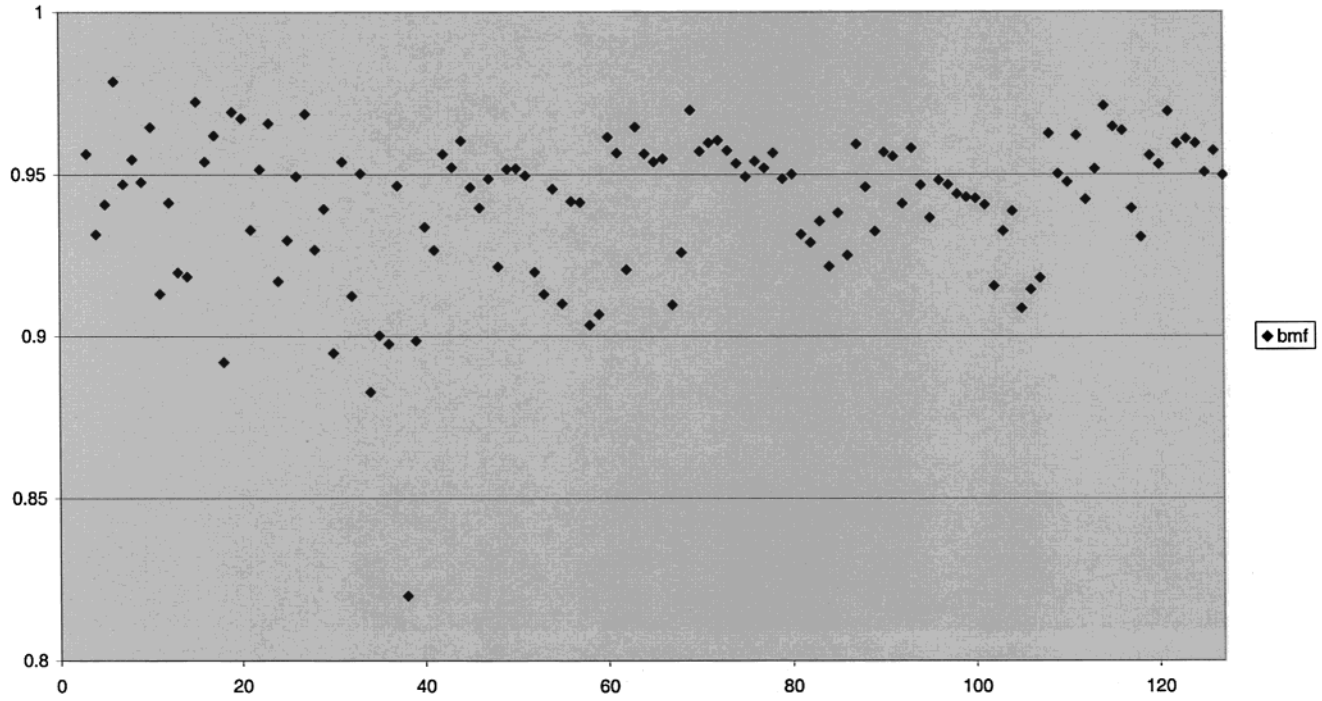

Figure 6. lbmfA3 all atom similarity spectrum.

3. RESULTS

3.1. Test Set. The test set consisted of 127 proteins, ranging in size from 100 to 400 amino acid residues. The spectrum of sizes is shown in Figure 4. The proteins were selected in groups according to the CATH database homologous superfamily assignments, as shown in Table 1. Only single domains were considered, including both single domain proteins and sections of multidomain proteins.

3.2. Atomic Orbital Overlap Results. Parameters for the atomic orbital Gaussian functions were found by fitting three Gaussians to the electron density functions calculated from the square of the STO3G atomic orbital wave functions, with the electron density set to zero beyond the van der Waals radius.[19] It has been shown that this modification is needed to represent the increase in hardness of molecular atoms.[19] The pairwise similarity values were calculated for each pair of molecules in the data set, using center of mass alignment followed by rotation at 10 degree intervals, then further rotation at 1 degree intervals about the best two 10 degree values. Tests on small molecules showed that this protocol

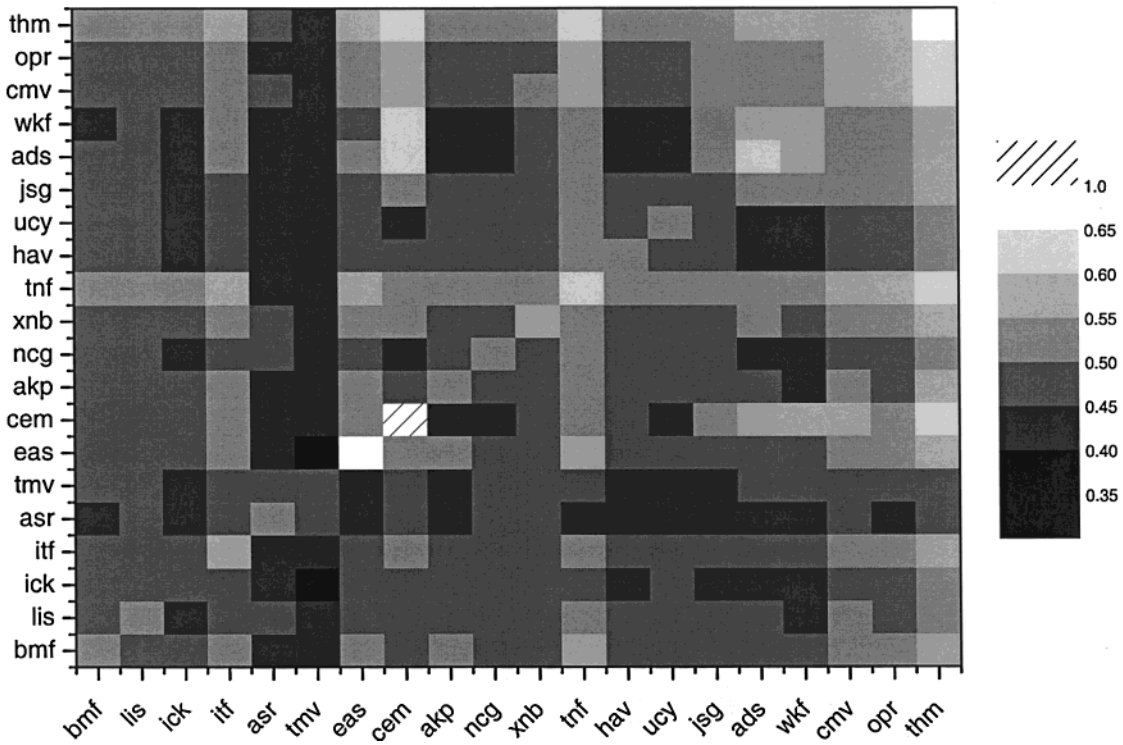

Figure 7. Alpha carbon centered Gaussian protein similarity values, averaged over homologous superfamily groups.

appeared to find the same optima as a full scan at 1 degree intervals. The results for the all atom atomic orbital based calculations are shown in Figure 5, averaged over the homologous superfamily groups, neglecting the self-similarity terms. The similarity spectrum for one particular protein domain, lbmfA3, is shown in Figure 6. The time taken for the complete run of 16 002 pairwise calculations was 155 h and 34 min, giving an average time per pairwise similarity computation of 35 s. The actual time for each protein pairing is proportional to the product of the number of atoms in each protein.

3.3. Alpha Carbon Centered Gaussian Results. The same set of proteins were used as for the all atom representation, but only the 2D alpha carbon positions were considered. The alpha carbon positions were reduced to 2D separately. The molecules were aligned by giving each alpha carbon a mass equal to the total mass of the corresponding amino acid and then aligning the centers of mass. Similarity values were calculated using the same 10 degree and then 1 degree protocol as in the first experiment. The results using this method are shown in Figure 7. The spectrum for lbmfA3 is shown in

Figure 8. The time for the entire set of calculations was 2 h and 41 min, giving an average time of 0.60 s.

3.4. Using Amino Acid Similarity. The previous calculations were repeated, but the amino acid similarity matrix shown in Figure 3 was used to modify the calculated inter amino acid overlap terms. The results are shown in Figure 9, and the overall time was 2 h and 50 min, giving an average timing of 0.64 s. The similarity spectrum for lbmfA3 obtained through this method is almost identical to that shown in Figure 8 and is therefore not included.

4. DISCUSSION

The atomic orbital overlap results are disappointing, as the universally high similarity values show little ability to discriminate between different protein topologies. The explanation is clearly shown by visualising the atomic orbital representation of any globular protein. At the atomic scale, the density of packing of the core of the protein makes it approximately homogeneous, which makes any two proteins highly similar when considered in this representation. This method may prove useful in the detection of active sites, where there exists a recognizable pocket in the atomic orbital field, but for whole molecule similarity all features of interest are swamped by core overlap. Coupled with the considerably longer time requirements, this method is not significantly useful.

The results from the alpha carbon based study are much better and show clear discrimination between homologous superfamilies (HS). Two-dimensional overlap plots of a similar pair of proteins, lbmfA3 and lmabA3 (both lysin bundle), with a similarity of 67% using amino acid centers and including amino acid similarity modifications, and a dissimilar pair, lbmfA3 and lncg (immunoglobbulin sandwich) with a score of 43%, are shown in Figures 10 and 11. They clearly show that improved similarity scores correspond well with structural similarity as assessed by eye. Most individual proteins were shown to be more similar to the other members of their HS than to any of the other proteins. To quantify this, T-test scores were calculated for each individual protein spectrum, to assess whether the within-group values could be identified as a distinguishable subset of the complete spectrum. The scores are shown in Figure 12. Most HS show values below 0.1, which indicates that they form a clearly identifiable subgroup of the entire set. Six groups, identified by HS representative names, show

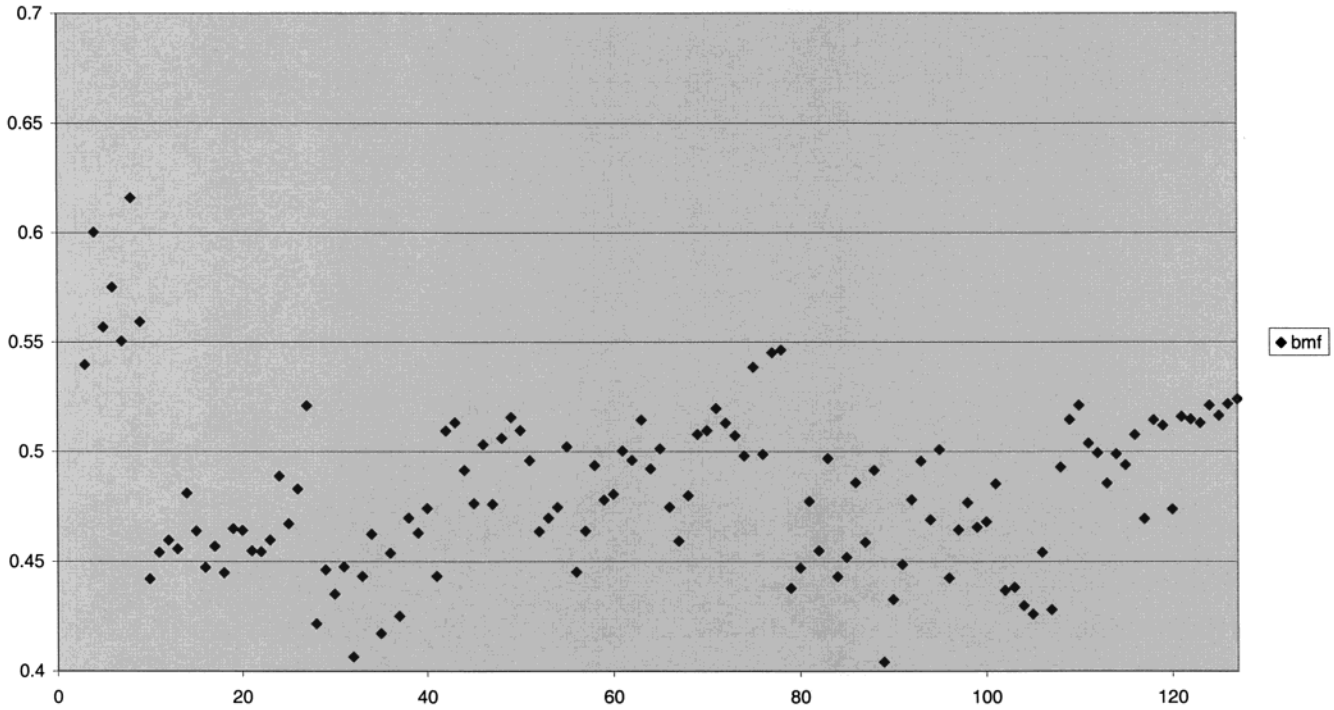

Figure 8. lbmfA3 alpha carbon based similarity spectrum.

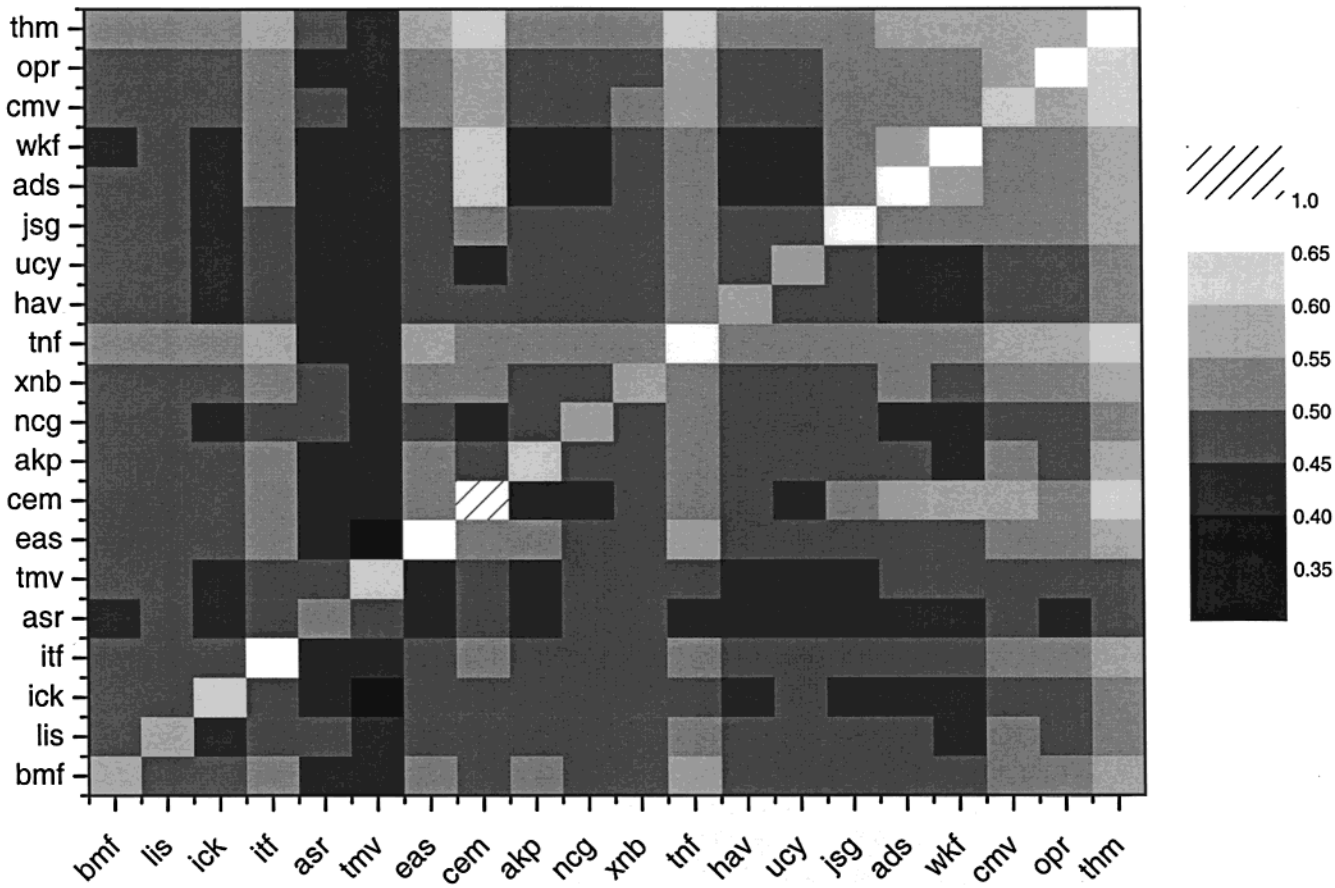

Figure 9. Amino acid similarity modified alpha carbon centered gaussian protein similarity values, averaged over homologous superfamily groups.

anomalous results: Ferritin (ick), 4Helix (tmv), Isoprenoid Synthetase (eas), Glycosyltransferase (cem), and Thrombin (ucy and hav). Members of the two thrombin groups have significant similarity with each other, making it difficult to separate out the HS in that particular topology. The two α-barrel architectures, eas and cem, also show significant similarity to each other. Furthermore only one example from the cem HS was included, indicating that the technique might have problems with assignment when insufficient examples are already known. The remaining two groups both have

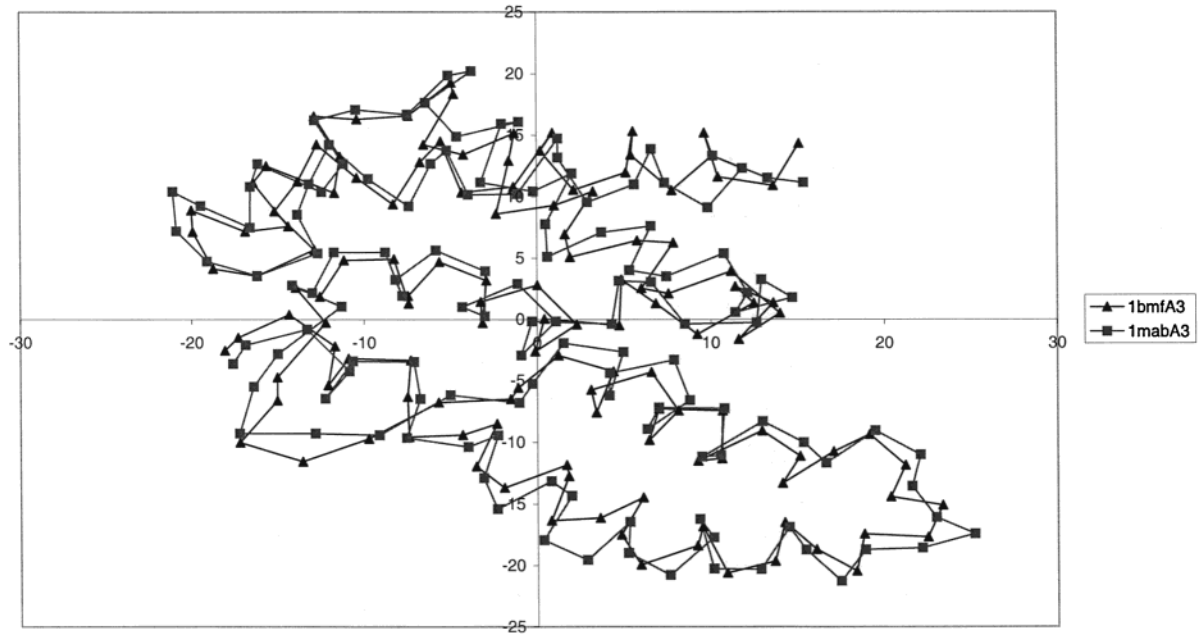

Figure 10. lbmfA3 amino acid modified similarity spectrum.

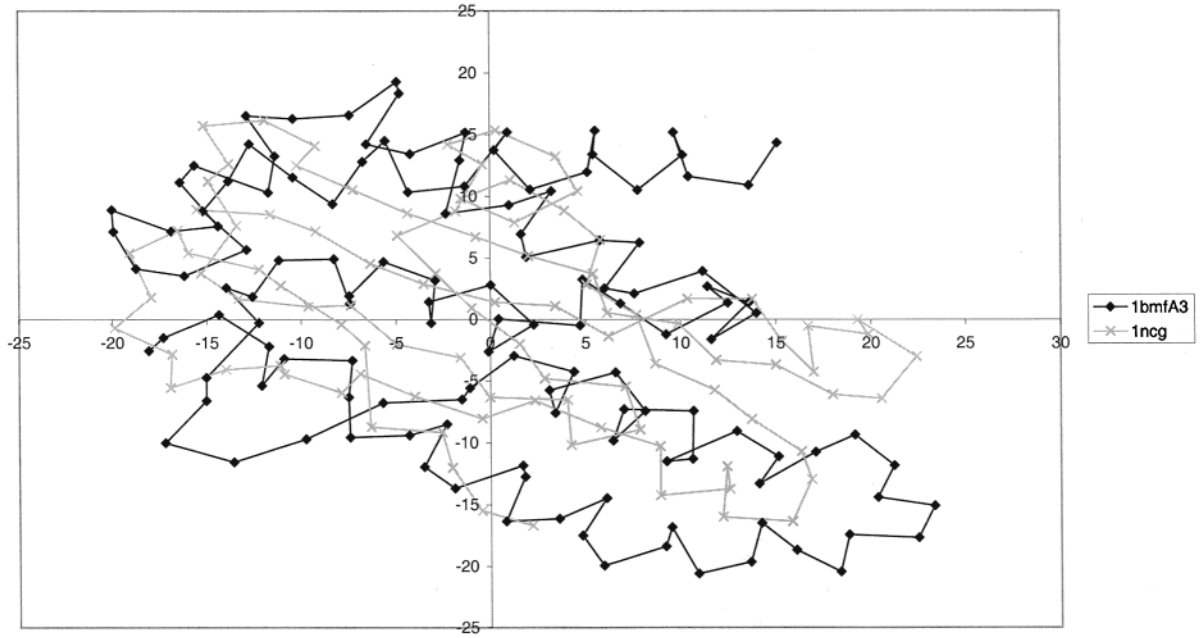

Figure 11. lbmfA3 and lmabA3 2D overlay.

rather low average self-similarities, with some members of their HS being significantly dissimilar to each other. Even these problematic groups are still reasonably identifiable.

Using the amino acid similarity values provided a slight but universal improvement to the discriminatory ability of the similarity values. Every protein was found to be more similar on average to its own HS than to any other. The T-test results show consistently slightly better values that for the unmodified calculations. While all the groups identified as giving anomalous results still show T-test scores above 0.1, the similarity values are now unambiguous. Using this method it would be possible to assign any member of the test set to its correct HS.

To further justify the method, the similarity data were used to produce a cluster diagram of the set of proteins. The set of similarity values for each protein was considered as a 127 dimensional vector, and Sammon's mapping was then used to reduce the set of vectors to a 2D form. The plot is shown in Figure 13. It can be seen that the majority of H.S. form fairly clear clusters. Those that do not are in most cases overlapping with other H.S.'s of the same topology. The Thrombin and 4Helix topologies show this behavior. A number of other groups have outliers, but are otherwise quite well clustered, and some of the smaller sets are hard to visulise as a cluster because they consist of insufficient data points. Overall, 10 groups are reasonably clearly clustered, and 4 more are overlapped only with their topological neighbors. The other 6 groups either have outliers, overlap with unrelated groups or are too small to visualize as clusters. This simple method of clustering shows reasonable agreement with the CATH classifications.

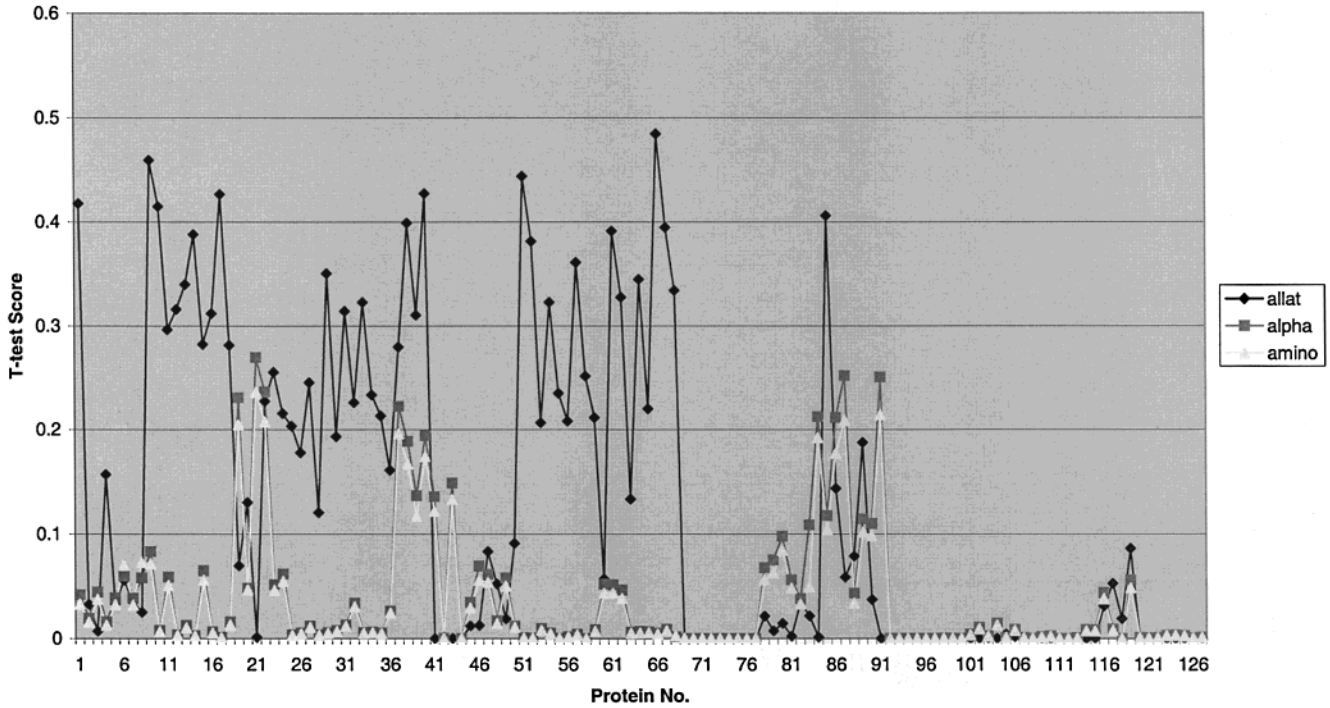

Figure 12. lbmfA3 and lncg 2D overlay.

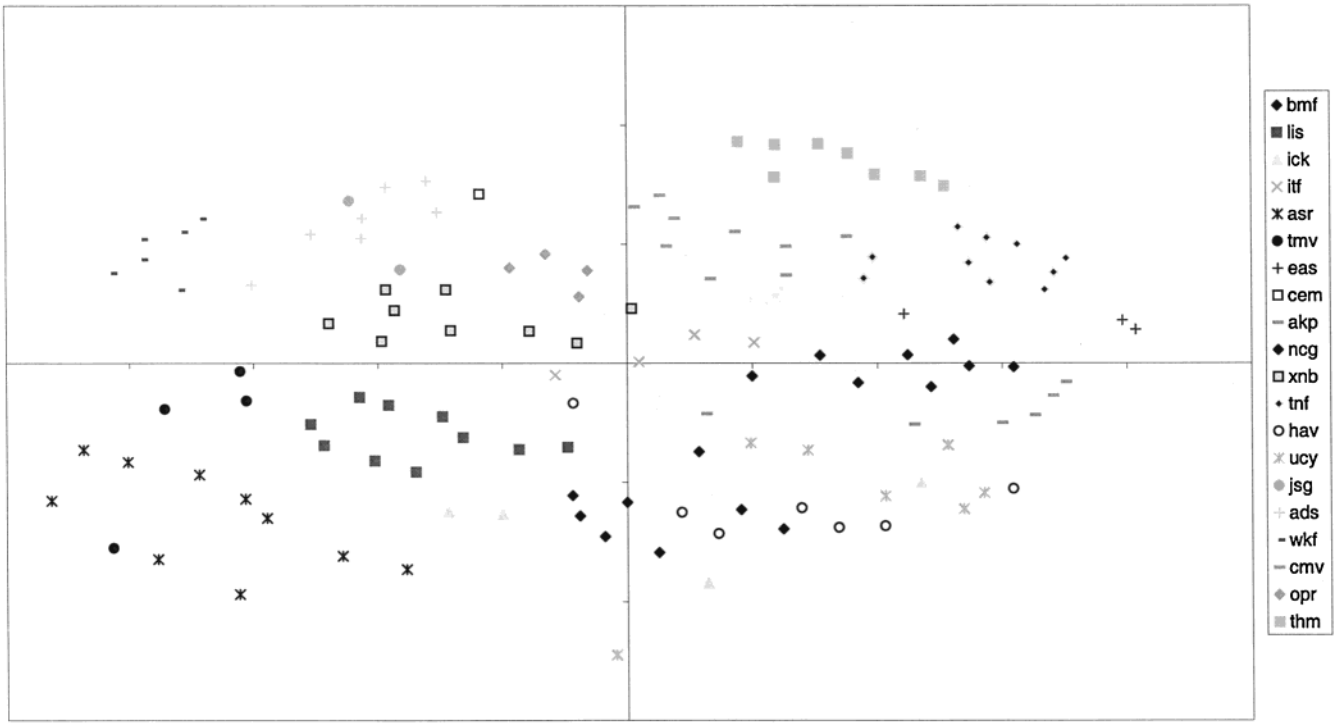

Figure 13. T-test scores for HS vs the complete set of proteins.

In comparison with other methods of calculating structural similarity, the use of 2D representations shows a number of benefits. First, it appears to be significantly faster than most comparable techniques, at about 0.6 CPU seconds per pairwise computation on a PIII 450. A fairly comprehensive review of structure comparison methods reports no other method of such speed; SSAP takes about 40 CPU seconds on a Sun 4/280 and FSSP takes 11 s on 'a fast computer worktation' in 1996.[13,23] The method has an advantage over distance matrix based algorithms such as FSSP in that it is not purely geometry based. The residue Gaussian parameters and the residue similarity matrix can be tuned to highlight chemical and physical features such as disulfide bonds.

Further Work. The basic technique could be improved by using some form of optimization rather than scanning to find the minima. The steepest decent algorithm was tested but did not work because the angular similarity spectrum is too irregular. Other methods of molecular alignment used for small molecules could also be applied.[20] The technique could be developed into a method for active site recognition,

either by pattern of amino acids, or by individual atoms, or by plotting activity centers. This might require a more sophisticated scanning algorithm, possibly neural net based pattern recognition software. An obvious and interesting application of the techniques would be to take advantage of the ability to apply alternative weighting schemes. Possiblities include schemes based on biochemical similarity of amino: i.e., hydrophobic/philics, charged; methods to highlight particular amino acids, e.g. cysteines; or methods to recognize and allow for common 19 point mutations.

5. CONCLUSIONS

By reducing protein structure data to 2D form, it is possible to calculate rapidly whole molecule similarity values for proteins. The values obtained by this method have been shown to be in broad agreement with the CATH system of protein structure classification. This method will allow the rapid and computerized assignment of new proteins to the most similar group of proteins based on their overall structure. It may be possible to use this to speed up considerably the determination of protein function. Timing measurements show that all 15 000 known protein structures could be reduced to 2D amino acid only form in about 1 week on a single PIII-450. When a new structure arrives, similarity values comparing it to every other known protein structure could be calculated in about 4 h on the same machine. That spectrum of similarity values could be used rapidly to assign the protein to a homologous superfamily or any other structure based categorization of interest.

Supporting Information Available: Proteins.list, the complete list of proteins by CATH identifiers. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bohm, G. New approaches in molecular structure prediction. *Biophys. Chem.* **1996**, *59*, 1−32.

(2) International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860−921.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(4) Thornton, J. M.; Orengo, C. A.; Todd, A. E.; Pearl, F. M. G. Protein Folds, Function and Evolution. *J. Mol. Biol.* **1999**, *293*, 333−342.

(5) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH−A Hierarchic Classification of Protein Domain. *Structure* **1997**, *5*(8), 1093−1108.

(6) Orengo, C. A.; Taylor, W. R. A local alignment method for protein-structure similarity. *J. Mol. Biol.* **1993**, *233*(3), 488−497.

(7) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. Gaussian-based approaches to protein-structure similarity. In *Molecular Modeling and Prediction of Bioactivity*; Gundertosste and Jorgensen., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 83−88.

(8) Burke, D. F.; Deane, C. M.; Nagarajaram, H. A.; Campillo, N.; Martin-Martinez, M.; Mendes, J.; Molina, F.; Perry, J.; Reddy, B. V. B.; Soares, C. M.; Steward, R. E.; Williams, M.; Carrondo, M. A.; Blundell, T. L.; Mizugnchi, K. An Iterative Structure-Assisted Approach to Sequence Alignment and Comparative Modeling. *Proteins: Struct., Funct. Genetics*; **1999**, Suppl. 3, 55−60.

(9) Maggiora, G. M.; Rohrer, D. C.; Mestres, J. Comparing protein structures: A Gaussian-based approach to the three-dimensional structural similarity of proteins. *J. Mol. Graphics Modeling* **2001**, *19*, 168−178.

(10) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(11) Taylor, W. R. A 'periodic table' for protein structures. *Nature* **2002**, *416*, 657−660.

(12) Holm, L.; Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **1993**, *233*, 123−138.

(13) Holm, L.; Sander, C. Mapping the protein universe. *Science* **1996**, *273*, 595−602.

(14) Sammon, J. S. A nonlinear mapping for data stucture analysis. *IEEE Trans. Comput.* **1969**, *C-18*(5), 401−409.

(15) Barlow, T. W.; Robinson, D. D.; Richards, W. G. Reduced dimensional representations of molecular structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*(5), 939−942.

(16) Allen, B. C. P.; Grant, G. H.; Richards, W. G. Similarity calculations using two-dimesional molecular representations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*(2), 330−337.

(17) Good, A. C.; Richards, W. G. Explicit Calculation of 3D molecular similarity. *Perspectives Drug Discovery Design* **1998**, *9*−*11*, 321−338.

(18) Leyda, L.; Carbo, R.; Arnau, M. An electron density measure of the similarity between two compounds. *Intl. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(19) Good, A. C.; Richards, W. G. Rapid evaluation of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*(1), 112−116.

(20) Barlow, T. W.; Robinson, D. D.; Richards, W. G. The utilization of reduced dimensional representations of molecular structure for rapid molecular similarity calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*(5), 943−950.

(21) A.S.P. (Automated Similarity Package). V3.11 Release Notes Oxford Molecular Limited: The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.

(22) T.S.A.R. (Tools for Structure Activity Relationships). Reference Guide 3.2 Oxford Molecular Limited: The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.

(23) Brown, N. P.; Orengo, C. A.; Taylor, W. R. A protein structure comparison methodology. *Comput. Chem.* **1996**, *20*(3), 359−380.

CI020275T