3D-BLAST: 3D PROTEIN STRUCTURE ALIGNMENT, COMPARISON, AND CLASSIFICATION USING SPHERICAL POLAR FOURIER CORRELATIONS

LAZAROS MAVRIDIS

DAVID W. RITCHIE

INRIA Nancy – Grand Est, LORIA, 615 rue du Jardin Botanique
54506 Vandoeuvre-lès-Nancy, France

This paper presents a novel sequence-independent method of aligning protein structures using three-dimensional spherical polar Fourier (SPF) representations of protein shape. The approach is demonstrated by clustering subsets of the CATH database for each of the four main CATH fold types, and by searching the entire CATH database of some 12,000 structures using several protein structures as queries. Overall, the automatic SPF clustering approach agrees very well with the expert-curated CATH classification, and ROC-plot analyses of the database searches show that the approach has very high precision and recall. Database query times can be reduced considerably by using a simple rotationally-invariant pre-filter in tandem with a more sensitive rotational search with little or no reduction in accuracy. Hence it should soon be possible to perform on-line 3D structural searches in interactive time-scales.

1. Introduction

The BLAST and FASTA sequence alignment programs are probably considered by most biologists as the standard tools for searching genomic nucleotide or amino acid sequence databases. However, it is known that in nature protein structure is more conserved than protein sequence. Hence, structural alignments can provide significant insights about protein function and can help classify protein families into functional super-families [1]. Considering the large and rapidly growing number of three-dimensional (3D) protein structures available in the Protein Databank (PDB [2]), it is important to develop new methods to align and compare protein structures [3]. Current protein structure alignment methods such as SSM [4], CE [5], VAST [6], SSAP [10], and DALI [1] typically use graph-matching and dynamic programming techniques to identify and align cliques of backbone Cα atoms or secondary structural features. However, these approaches are slow compared to conventional sequence alignment methods. Furthermore, finding the best way to perform 3D structural alignments remains an open question [7]. The most widely used protein structure classifications are the CATH [8] and SCOP [9] databases, both of which are curated by human experts. In CATH, the classification is initially performed using SSAP, whereas SCOP relies more on visual inspection by the curators. In both cases, it would be desirable to be able to assemble and update structural classifications in a more automated way.

In a significant step towards performing protein structure comparisons more efficiently, Mak *et al.* [11] and Sael *et al.* [12] showed that the 3D shapes of large protein molecules could be compared and classified very rapidly using special 3D pose-invariant descriptors derived from spherical harmonic (SH) and Zernike polynomials [13]. Hence this kind of approach offers the possibility of being able to search a 3D database of protein structures very rapidly. Similarly, it has been shown previously that the related spherical polar Fourier (SPF) representation provides a fast way to perform protein-protein docking correlations [14,15]. However, until now, the SPF approach has not been used to superpose and compare protein shapes. Conceptually, the use of Zernike descriptors has close parallels with the SPF representation, although Mak *et al*. and Sael *et al*. did not exploit the special rotational properties of the SH basis functions. Hence their approaches can detect similar protein shapes, but cannot superpose them. There is growing interest in the use of SH-based shape representation techniques [16-25]. The novelty of our SPF representation compared to other SH-based approaches is its use of orthonormal Laguerre-Gaussian (GL) radial functions to give a 3D density-based representation of molecular shape. This removes the requirement that molecular surfaces should be star-like with respect to the chosen coordinate origin [26], and it allows both rotational an translational correlation expressions to be calculated analytically [27]. In this paper, we demonstrate that SPF correlations may be used to superpose and compare protein structures in an efficient and completely sequence-independent manner. We present results obtained by clustering multiple protein structures selected from the CATH database, and we demonstrate the utility of our approach by performing queries of single protein structures against the entire CATH database of some 12,000 protein structures.

2. Methods

2.1. *Spherical Polar Fourier Shape Density Functions*

In the SPF approach, protein shapes are represented as 3D density functions expressed as expansions of orthonormal basis functions:

$$R(\mathbf{r}) = \sum_{n=1}^{N}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} a_{nlm}R_{nl}(r)y_{lm}(\vartheta,\varphi) \qquad (1)$$

where N is the order of the expansion, $R_{nl}(r)$ are Laguerre-Gaussian radial functions, $y_{lm}(\vartheta,\varphi)$ are real spherical harmonics, and a_{nlm} are the expansion coefficients which are calculated numerically as described previously [14]. Figure 1 shows the SPF representations of a pair of similar nitrogenase domains at several expansion orders.

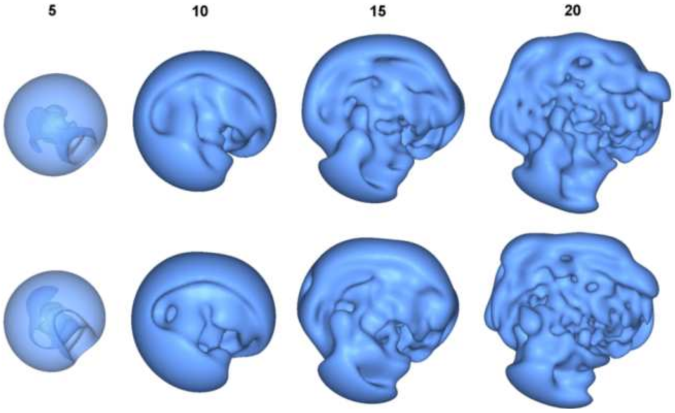

Figure 1. Two similar nitrogenase proteins, represented as SPF expansions at expansion orders N=5, 10, 15, and 20. The protein in the top row is from azotobacter vinlandii (PDB code 2MIN). The protein in the bottom row is from clostridium pasteurianum (PDB code 1MIO).

In order to superpose a pair of protein structures we calculate a rotation-dependent Carbo-like similarity score S_{ROT} using:

$$S_{ROT} = \frac{\sum_{nlm}^{N} a_{nlm}b_{nlm}}{\left(\sum_{nlm}^{N} a^2{}_{nlm}\right)^{1/2}\left(\sum_{nlm}^{N} b^2{}_{nlm}\right)^{1/2}} \qquad (2)$$

Conceptually, one protein is held fixed and a six-dimensional (6D) rotational/translational search over positions of the second protein is performed. However, in practice it is more efficient to implement the search using one translational and five Euler angle rotational coordinates [14].

2.2. *Rotationally Invariant Fingerprints*

Protein superpositions can be calculated relatively quickly by using FFT techniques to accelerate the 6D search [14]. However, it is necessary to develop even faster comparison techniques in order to search very large 3D structural databases. Noting that expansion coefficients with the same values of m transform amongst themselves under rotation, it is natural to use the vector interpretation of SH coefficients to construct rotationally invariant (RI) fingerprints (RIFs) as:

$$A_{nl} = \sqrt{\sum_{m=-l}^{l} a_{nlm}^2} \qquad (3)$$

If the coefficients a_{nlm} define the shape density of a protein, then the rotation-invariant descriptors A_n together encode the protein's radial mass distribution. By analogy to Eq 2, the RIF similarity score is written as:

$$S_{RIF} = \frac{\sum_{n=1}^{N} A_n B_n}{\left(\sum_{n=1}^{N} A_n\right)^{1/2} \left(\sum_{n=1}^{N} B_n\right)^{1/2}} \qquad (4)$$

2.3. *Implementation*

Our approach has been implemented in a C program called 3D-Blast. We have used 3D-Blast to calculate and store SPF coefficients for all the proteins of the CATH database. However, the approach could equally be used to store SPF descriptions of other protein structure databases such as SCOP, for example. Given a protein query structure in PDB format, 3D-Blast will calculate its SPF representation and then use only the resulting SPF expansion coefficients to search its database.

2.4. *Data Preparation*

In order to evaluate the SPF approach, clustering experiments were performed on selected proteins from the CATH database [8,10], and the entire CATH database was searched using SPF density functions as queries. In CATH, proteins are assigned to a super-family according to their fold class, architecture, topology, and homology. This classification scheme is essentially hierarchical, with the top-level class consisting of four possible fold types: All-α, All-β, α+β, and irregular. Each of the four levels in the CATH hierarchy is identified by a numeric code. Additionally, CATH names each protein according to its four-letter PDB code, its chain letter, and the number of domains (e.g. 1IOMA02). These naming conventions are followed here. For each clustering experiment, five or six super-families with the same architecture were selected in such a way as to give around 30 protein structures for each CATH fold class. Hence the aim of these experiments is to assess how well our approach can identity proteins with the same topology and homology within a given fold architecture. The details of the super-families used here are shown in Table 1.

For the initial database search experiment, asparagine synthetase (PDB code 12AS, CATH super-family 3.30.930.10) was selected as the query structure. The 3.30.930.10 super-family has 27 members, and these were treated as "positives" while the remaining proteins in the database were treated as "negatives" with respect to the query. If a scoring function were to reproduce exactly the CATH classification, the 27 positives would appear at the top of the ranked list. However, such an ideal outcome is seldom observed in practice. Hence Receiver-Operator-Characteristic (ROC) curves [28] are used to analyse objectively the precision/recall characteristics of the scoring functions. In a ROC analysis, each element of the ranked list is considered in turn, and the number of positives and negatives in the sublists to each side of the current element are counted. Here, we call the high similarity sublist the "hit list". A true positive (TP) is assigned when an element in the hit list contains an original positive, and a false positive (FP) is assigned if that element contains a negative. Conversely, a true negative (TN) is assigned when an original negative falls outside the hit list, and a false negative (FN) is assigned if that position is occupied by a positive member. ROC plots are produced by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis, where TPR and FPR are given by:

$$TPR = \frac{TP}{TP + FN} \qquad (5)$$

$$FPR = \frac{FP}{FP + TN} \qquad (6)$$

The quality of a scoring function can quickly be assessed from the shape of a ROC plot. For example, a random scoring function would give a diagonal line (TPR=FPR), whereas a perfect scoring function would give a horizontal line (TPR=1) with a maximum value for the area under the curve (AUC=1).

Table 1. The 23 CATH super-families used in the four clustering experiments, grouped according to class and architecture. For each super-family the name of each representative protein is provided according to the CATH naming convention.

Class + Architecture	Topology + Homology	Protein Name and Function	Representative member
All-α Orthogonal Bundle (1.10)	230.10	Cytochrome p450-Terp; Domain 2	1iomA02
	120.10	Bifunctional Trypsin/Alpha-Amylase Inhibitor	1beaA00
	225.10	NK - Lysin	1l9lA00
	167.10	Regulator of G-Protein Signalling 4; Domain 2	1dk8A02
	30.10	DNA Binding (I) , subunit A	1qrvA00
All-β Ribbon (2.10)	109.10	Umud Fragment, subunit A	1jhfB00
	150.10	Urease, subunit B	1ejxB00
	110.10	Cysteine Knot Cytokines, subunit B	1g47A00
	77.10	Hemagglutinin; Chain A, domain 2	1jsdA02
	160.10	Vascular Endothelial Growth Factor - 165, Heparin - binding Domain	1kmxA00
	10.10	Seminal Fluid Protein PDC - 109 (Domain B)	1h8pA02
α+β Roll (3.10)	130.10	P-30 Protein	1dy5A00
	170.10	Elastate; Domain 1	1u4gA01
	150.10	DNA Polymerase III; Chain A, Domain 2	1ok7A01
	110.10	Ubiquitin Conjugating Enzyme	2grrA00
	120.10	Flavocytochrome B2; Chain A; Domain 1	1cyoA00
Irregular (4.10)	280.10	MYOD Basic-Helix-Loop-Helix Domain, subunit B	1nlwE00
	290.10	Bacteriorhodopsin Fragment	1bctA00
	410.10	Factor Xa Inhibitor	1g6xA00
	490.10	High-Potential Iron-Sulfur Protein; Chain A	1iuaA00
	400.10	Low-density Lipoprotein Receptor	2fcwB02
	320.10	Dihydrolipoamide Transferase	1w85I00

3. Results

3.1. *Expansion Resolution and Protein Superposition*

It was shown previously that SPF expansions of order $N \geq 25$ are required for protein-protein docking correlations [14]. However, from our previous results on small-molecule virtual screening [29], we expected that considerably lower order expansions would be sufficient to calculate satisfactory protein shape-density superpositions. In order to test this hypothesis, we selected four example protein pairs with sequence identities ranging from 28% to 43% previously identified by Levitt and Gerstein [30], and we superposed them using different expansion orders. The root mean square deviation (RMSD) for each superposed pair was calculated between corresponding Cα atoms identified by ProFit [31]. Figure 2 shows the RMSD as a function of the expansion order *N*.

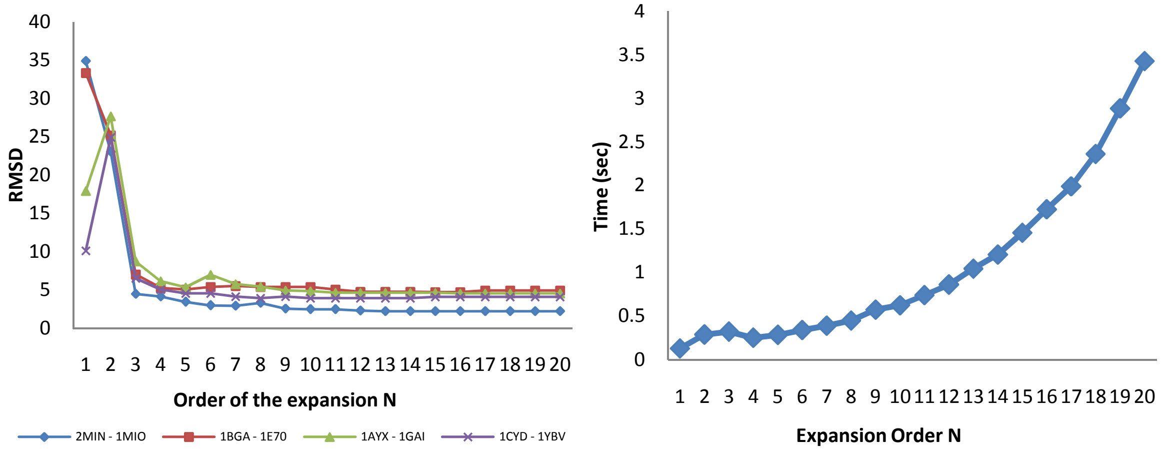

Figure 2. On the left, the RMSD values of four protein pairs are plotted against the expansion order *N*. On the right, the time per pair-wise search is plotted against the expansion order *N*.

Figure 2 shows that using an expansion order of $N \geq 3$ is generally sufficient to obtain good superpositions (with RMSDs ranging from 2Å to 5Å), requiring about 0.25 seconds per superposition. However, we choose to use expansions to order *N*=6 in order to disambiguate cases where flipping about an axis might give similar scores [29]. Figure 2 shows that calculating *N*=6 correlations is only marginally more expensive than using *N*=3. Figure 3 shows the *N=6* superposition of the nitrogenase proteins shown in Figure 1.

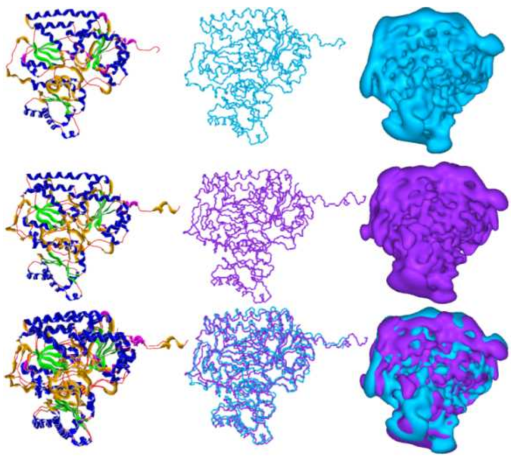

Figure 3. The superposition of a pair of nitrogenase proteins calculated using *N*=6 SPF correlations, shown as ribbon cartoons (left), backbone traces (middle), and as 3D SPF density expansions to order *N*=25 (right). The protein in the top row is from azotobacter vinlandii (PDB code 2MIN). Top row: PDB code 2MIN; middle row PDB code 1MIO; bottom row their superposed orientation. The two proteins have a sequence identity of 43%.

3.2. *Clustering Protein Structures*

For the All-α clustering experiment, five super-families were selected, as listed in Table 1. For each pair of proteins in this set, a correlation search was performed to find the orientation that gives the maximum Carbo similarity score (Eq 2). Ward's agglomerative clustering [32] was then applied to the resulting table of pair-wise similarity scores. The clustering results in Figure 4 show that the 1.10.230.10 and 1.10.167.10 super-families are correctly assigned to two separate groups. Although there exist some differences in the clusters produced for the remaining super-families, there is still a very good agreement between the calculated SPF clusters and the CATH hierarchy. The most notable exception is that suposin (PDB code 1N69) is grouped with the 1.10.30.10 super-family. From visual inspection of Figure 4 it can be seen that the overall fold of suposin is much closer to that of the 1.10.30.10 super-family than the CATH assignment of 1.10.225.10. This suggests that the automatic SPF classification could potentially help the CATH curators resolve unusual or ambiguous cases.

For the All-β class, six super-families were selected. As can be seen in Figure 5, SPF clustering correctly distinguishes all six groups, but two proteins are misplaced according to the CATH classification. These are the carboxy-terminal LIM domain (PDB code 1CTL) and the influenza virus hemagglutinin (PDB code 2VIR) which are grouped with the singleton heparin-binding domain (PDB code 1KMX). This seems to occur because 1CTL and 2VIR are calculated to be less similar to the other members of their respective CATH super-families, and they are grouped with 1KMX largely because all three proteins have similar steric bulk.

For the α + β class, five super-families were selected. From these five super-families the 3.10.130.10 and 3.10.120.10 super-families are correctly assigned into two groups, as shown in Figure 6. The other three super-families (3.10.110.10, 3.10.150.10, and 3.10.170.10), present a similar case to the All-α results, whereby one super-family group (3.10.110.10) is split into two sub-groups and two super-family groups (3.10.170.10 and 3.10.150.10) are merged into one. Nonetheless, despite these differences, the overall consistency of the SPF clustering with the CATH hierarchy is clearly very good.

For the irregular class, six super-families were selected. As in the All-β example, SPF clustering is completely consistent with the CATH hierarchy. However, two proteins are misplaced with respect to the CATH classification, namely bikunin from the inter-alpha-inhibitor complex (PDB code 1BIK) and the tick anticoagulant peptide (PDB code 1D0D), which are both grouped with the 4.10.490.10 super-family. This seems to be due to the difference in size between those proteins and the rest of their super-family of factor XA inhibitors. For example, Figure 7 shows that bikunin has a repeat of the same motif as the other factor XA inhibitors. Hence, it is sterically too large to be clustered with the other XA inhibitors, and is instead placed with the larger proteins of the 4.10.490.10 super-family.

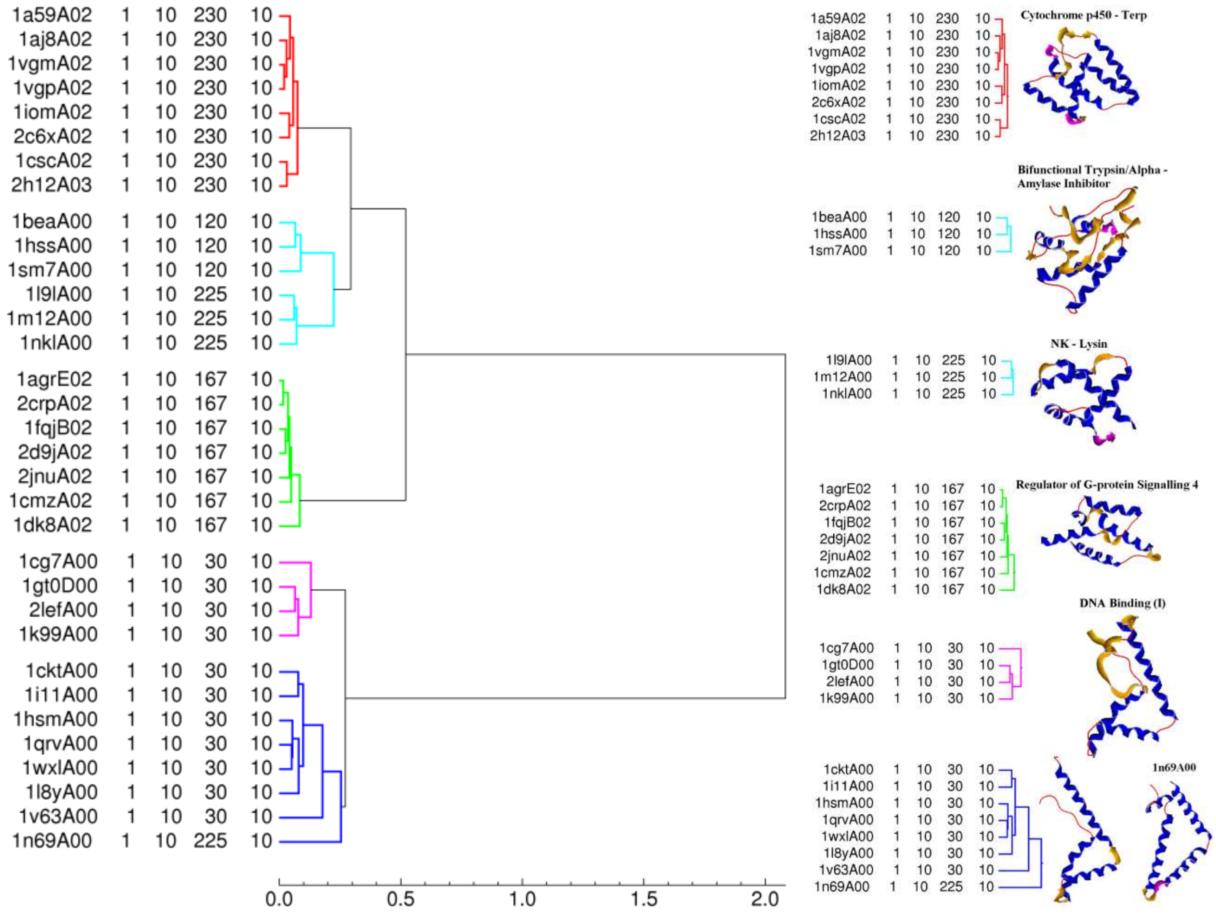

Figure 4. SPF clustering results of the All-α class. Left: the dendrogram obtained using *N=6* with five clusters; Right: the corresponding groups and the representative proteins for each group.

3.3. *Database Searching*

Because it is relatively time-consuming to calculate high order comparisons, different rotational and rotation-invariant parameters were tested to explore the extent to which the speed of queries on a large database can be improved while still performing accurate searches. In each case, the asparagine synthetase structure (PDB code 12AS, CATH super-family 3.30.930.10) was superposed onto each protein in the database using $N=6$ SPF correlation searches, and the database structures were ranked in order to similarity to the query. Figure 8 shows the resulting ROC curves obtained for a range of expansion orders when querying the CATH database, from which it can be seen that our approach gives very good precision and recall. Figure 9 shows the 27 members of this super-family, which were treated as true positives with respect to the query. To analyse the results further, the TPs were clustered into 5 groups. The query belongs to group 1 and when using an expansion order of $N \geq 6$, all members of this group were found in the top 10 hits. groups 2 and 3 have similar β-sheet structures to group 1, but different arrangements of α-helices. All proteins in groups 2 and 3 are ranked in the top 20% of the database. All proteins in group 4 are ranked in the top 30%. The singleton group 5 is an obvious outlier due to its extra α-helical domain.

Figure 8 also shows results for the RIF scoring function. Compared to the rotation-dependent scoring function, the RIF function generally performs remarkably well. However, the two functions behave rather differently on the first percentages of the database. For example, the rotational searches give a TPR of around 40% on the first 0.1% of the database, whereas the RIF searches give a TPR of only around 10%. Hence, the RIF function is not sufficiently sensitive to be used on its own but it could usefully be used as a fast pre-filter on the database so that the more expensive rotation-dependent function is only applied to the most promising candidates. In order to test the notion, the CATH database was searched using the RIF and rotational scoring functions in tandem using several protein structures as queries: asparagine synthetase, ALF4-activated Giα1 protein (PDB code 1AGR), chicken cysteine-rich protein (PDB code 1B8T), dihydrolipoyllysine-residue acetyl transferase (PDB code 1W4E), and UbcH7 (PDB code 1C4Z). Using a RIF pre-filter similarity threshold of 0.99, which selects from 2% to 15% of the database for rotational re-scoring, each tandem search takes less than 10 minutes compared to 75 minutes for full rotational searches on a 2.3GHz Pentium Xeon processor. Figure 10 shows the resulting ROC plots for the rotational, RIF, and the tandem searches. This figure shows that tandem searches achieve the same high level of performance as the rotational searches. The very high precision/recall values further support the utility of the SPF scoring functions.

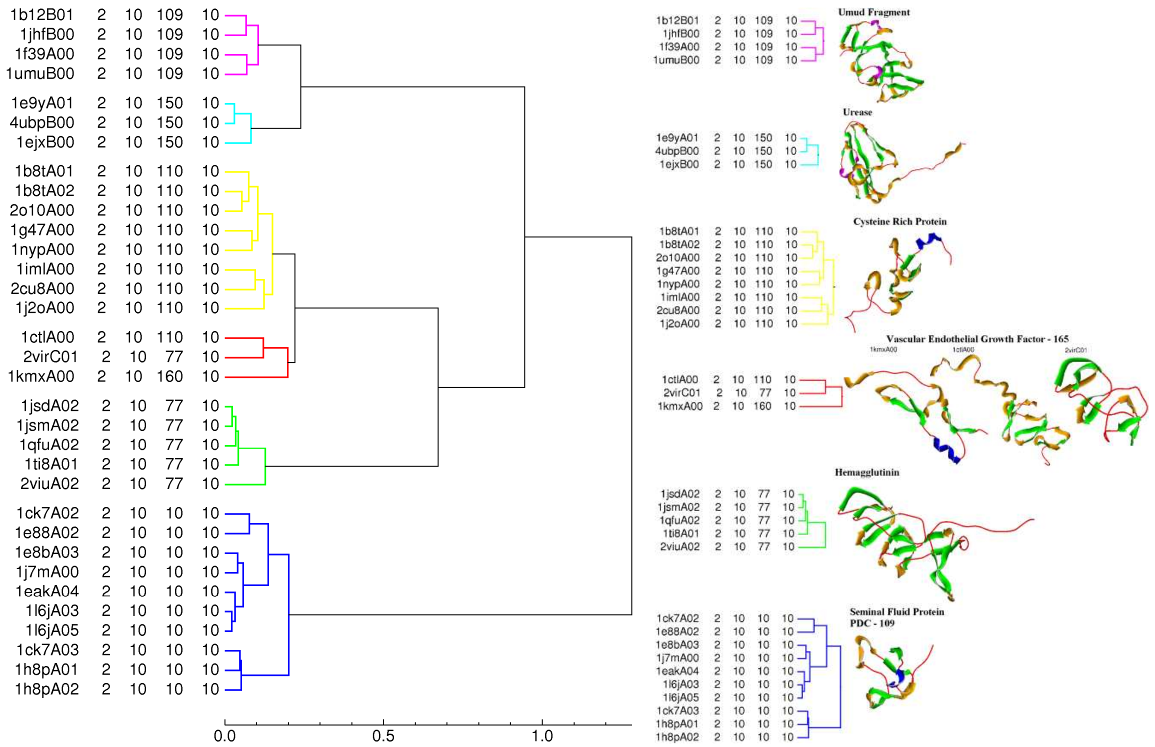

Figure 5. SPF clustering of the All-β class. Left: the dendrogram obtained using $N=6$ with six clusters; Right: the corresponding groups and the representative proteins for each group.

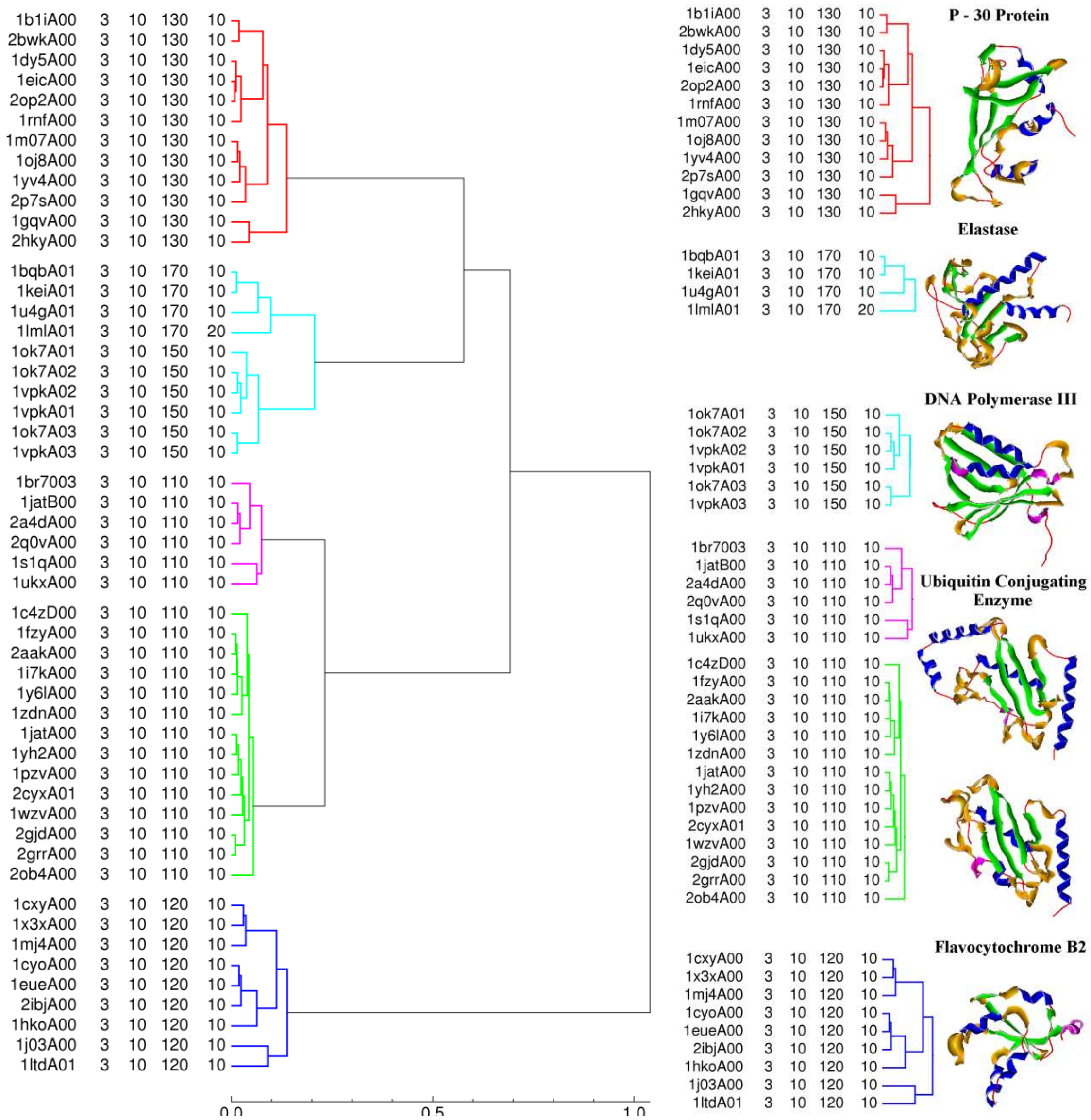

Figure 6. SPF clustering of α + β class. Left: the dendrogram obtained using *N=6* with five clusters; Right: the corresponding groups and the representative proteins for each group.

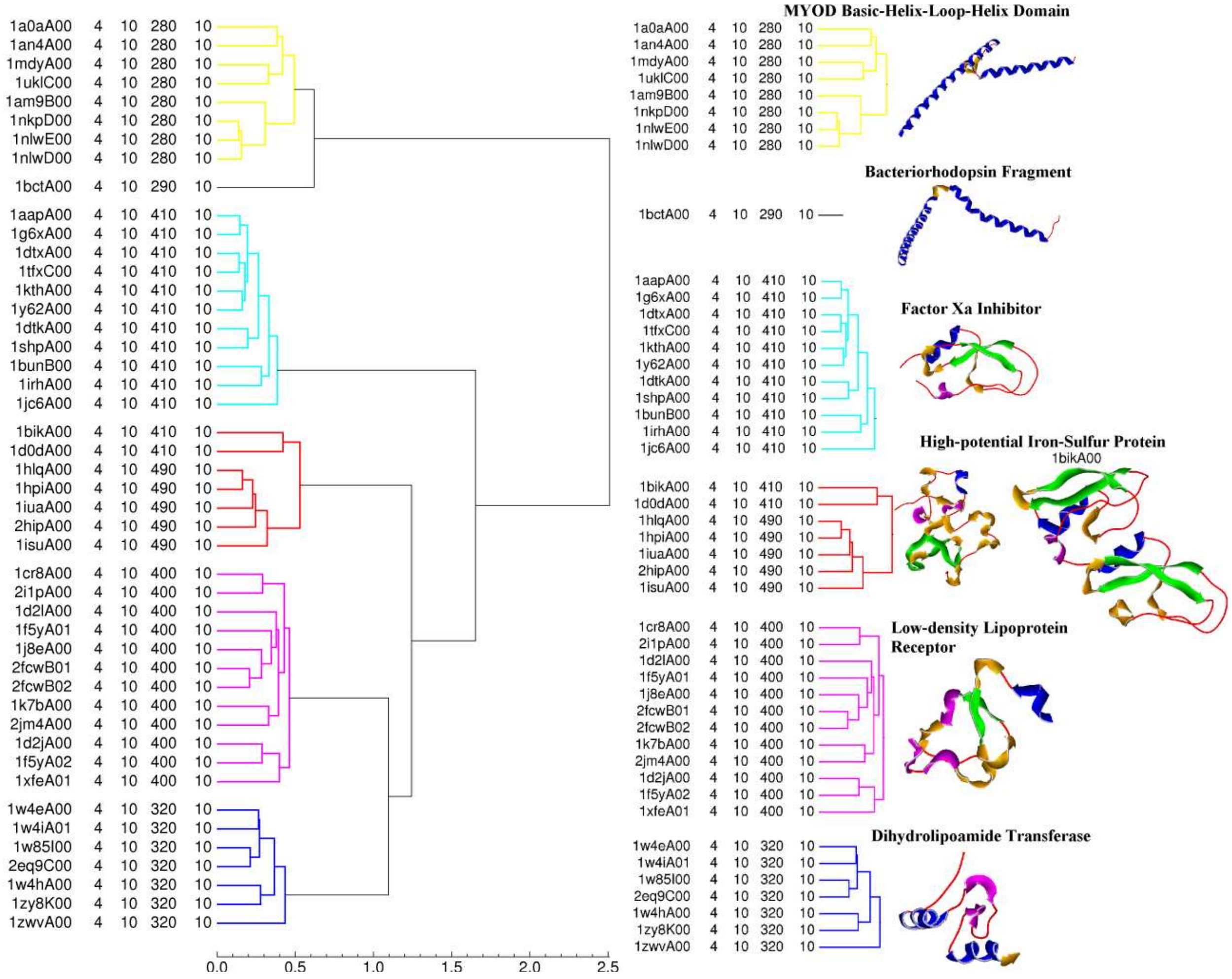

Figure 7. SPF clustering of the irregular class. Left: the dendrogram obtained using *N=6* with six clusters; Right: the corresponding groups and the representative proteins of each group.

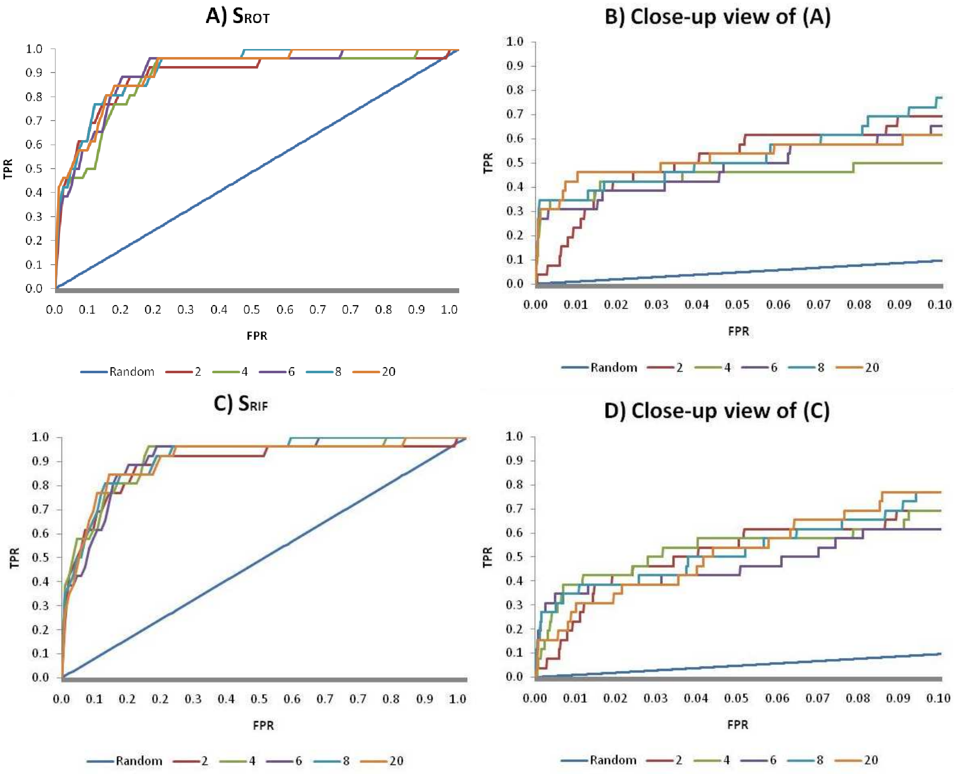

Figure 8. ROC plot analyses obtained when querying the entire CATH database with the 12asA00 structure. A: rotation-dependent scoring function (Eq 2); B: close-up view of (A); C: RIF scoring function (Eq 4); D: close up view of (C).

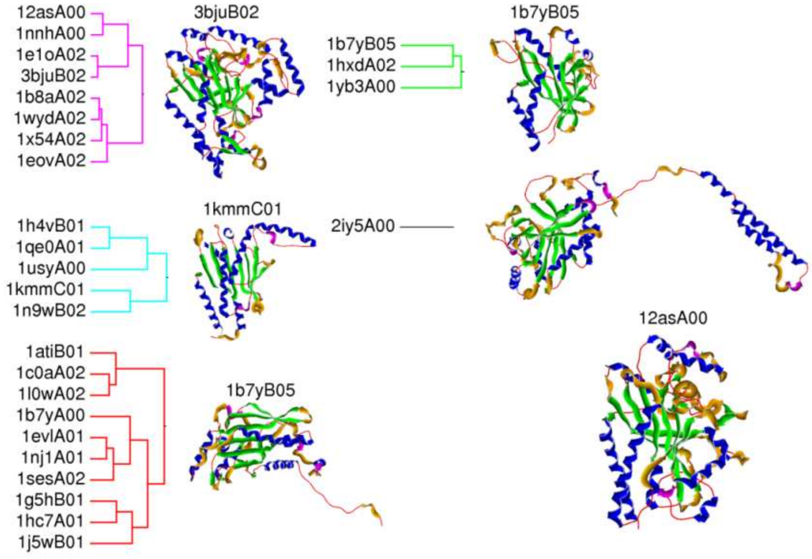

Figure 9. Clustering the CATH super-family 3.30.930.10 into five groups. The representative members of each group are illustrated along with the query protein 12asA00.

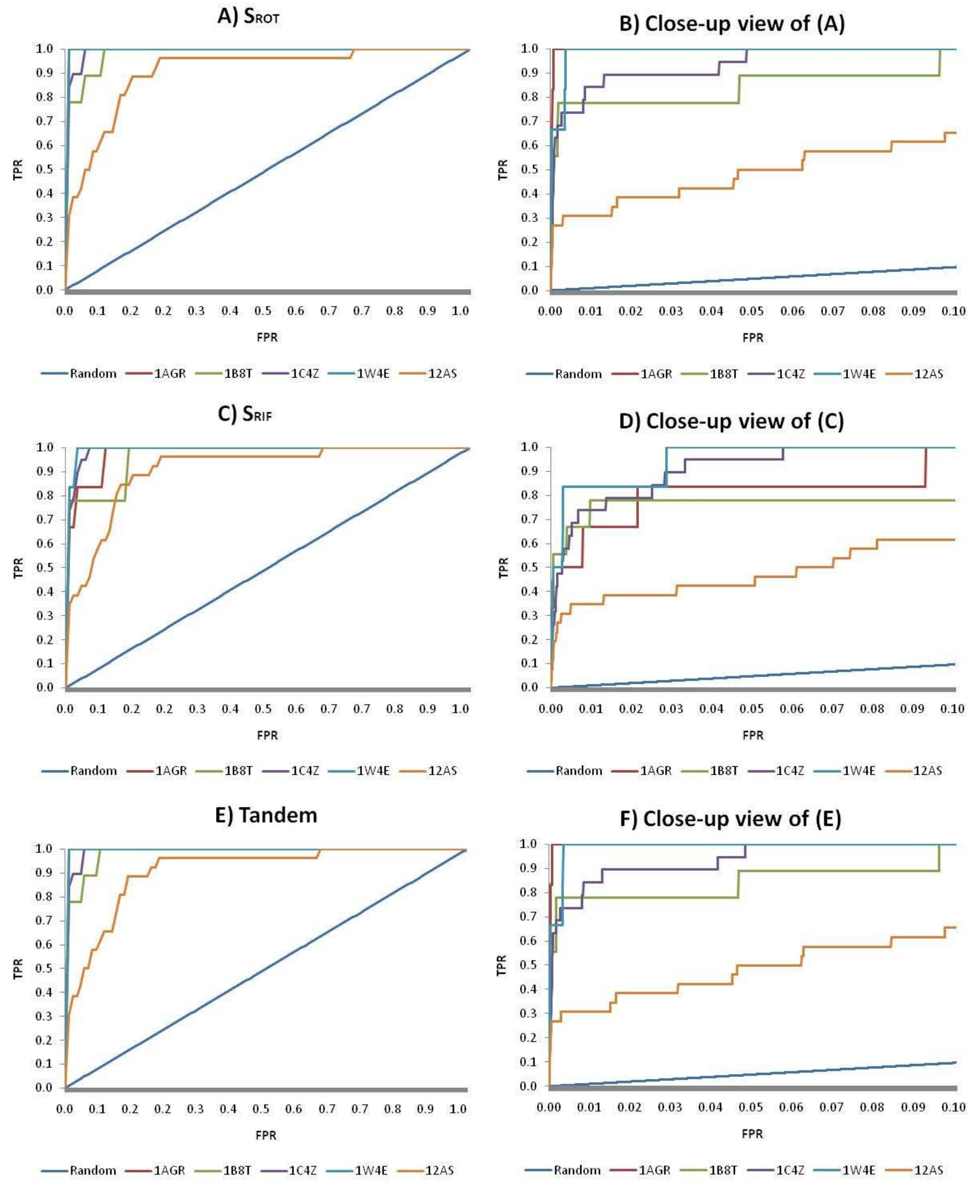

Figure 10. ROC plot analyses obtained when querying the entire CATH database using the 1AGR, 1B8T, 1W4E, 1C4Z and 12AS, structures. A: rotation-dependent scoring function (Eq 2); B: close-up view of (A); C: RIF scoring function (Eq 4); D: close up view of (C); E: Tandem scoring; F: close up view of (E).

4. Discussion and Conclusions

It has been shown that low resolution SPF expansions provide a reliable and fast sequence-independent way to superpose and compare protein structures. The clustering results of the four CATH super-families show that SPF clustering generally agree with the CATH classification. The database search results show that a large protein structure database can be queried accurately using SPF expansions to order $N=6$. In order to accelerate further the scoring calculation, the use of a simple rotationally invariant scoring function was investigated. Our RIF scoring function is significantly faster to calculate, but it is less precise than a rotational search. Nonetheless, the RIF function may still be used effectively as an initial filter when searching large databases. Using the RIF function in tandem with the rotational function to re-score only the top matches of the database gives very promising results.

Thus tandem searches are much faster than full rotational searches, yet the recall/precision is almost equivalent. Currently a limitation of our approach is that proteins larger than a typical domain of around 100-150 residues are not represented well due to the exponential decay of the GL functions at large radial distances. However, we are working to extend the resolution range of our approach to be able to describe proteins of any size, and we are enhancing the approach to be able to detect the presence of symmetrical domains and repeated motifs, for example. We also aim to develop a web-based interface for general use by the biological community. As well as being able to calculate rapidly sequence-independent protein structure superpositions, we believe that the SPF approach could provide an automatic and objective way to enhance the quality of protein structure classifications.

Acknowledgements

This work is funded by the Agence Nationale de la Recherche, grant reference: ANR-08-CEXC-017-01.

References

1. L. Holm and C. Sander. *Trends in Biochemical Sciences.* **20**, 478 (1995).
2. H.M. Berman *et al. Acta Cryst.* **D58**, 899 (2002).
3. R. Kolodny, P. Koehl, and M. Levitt. *J. Mol. Biol.* **346**, 1173 (2005).
4. E. Krissinel and K. Henrick. *Proceedings of the Fifth international Conference on Molecular Structural Biology, Vienna*. 88 (2003).
5. I.N. Shindyalov and P.E. Bourne. *Protein Engineering* **11**, 739 (1998).
6. T. Madej, J.F. Gibrat and S.H. Bryant. *Proteins: Struct, Func. Bioinf.* **23**, 356 (1995).
7. M.J. Sippl and M. Widerstein. *Bioinformatics.* **24**, 426 (2008).
8. A.L. Cuff, I. Sillitoe, T. Lewis, O.C. Redfern, R. Garratt, J. Thornton, and C.A. Orengo. *Nucleic Acids Research.* **37,** 310 (2008).
9. A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. *J. Mol. Biol.* **247**, 536 (1995).
10. C.A. Orengo and W.R. Taylor. *Methods Enzymol.* **266**, 617 (1996).
11. L. Mak, S. Grandison, and R.J. Morris. *J. Mol. Graph. Model.* **26**, 1035 (2008).
12. L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara. *Proteins: Struct, Func. Bioinf.* **72**, 1259 (2008).
13. M. Novotni and R.J. Klein. *J. Comp.-Aided Mol. Des.* **36**, 1047 (2004).
14. D.W. Ritchie and G.J.L. Kemp. *Proteins: Struc., Funct., Genet.* **39**, 178 (2000).
15. D.W.Ritchie, D. Kozakov, and S. Vajda. *Bioinformatics.* **24**, 1865 (2008).
16. S.E. Leicester, J. L. Finney and R.P. Bywater, *JMG.* **6**, 104 (1988).
17. S.E. Leicester, J. Finney and R. Bywater. *J. Math. Chem.* **16**(3-4), 315 (1994).
18. S.E. Leicester, J. Finney and R. Bywater. *J. Math. Chem.* **16**(3-4), 343 (1994).
19. N.L. Max and E.D. Getzoff. *IEEE Comput. Graphics Appl.* **8**(4), 42 (1988).
20. N.L. Max. *JMG.* **6**(4) 210 (1988).
21. B.S. Duncan and A.J. Olson. *Biopolymers*. **33**, 219 (1993).
22. B.S. Duncan and A.J. Olson. *Biopolymers*. **33**, 231 (1993).
23. B.S. Duncan and A.J. Olson. *JMG.* **13**, 258 (1995).
24. B.S. Duncan and A.J. Olson. *JMG.* **13**, 250 (1995).
25. A. Gramada and P.E. Bourne, *BMC Bioinformatics*, **7**, 242 (2006).
26. R.J. Morris, R.J. Najmanovich, A. Kahraman and J.M. Thornton, *Bioinformatics*, **21**(10), 2347 (2005).
27. D.W. Ritchie. *J. Appl. Cryst.* **38**, 808 (2005).
28. J.P. Egan. *Academic Press, New York* (1975).
29. L. Mavridis, B.D. Hudson and D.W. Ritchie, *J. Chem.Inf. Model.,* **47**(5), 1787 (2007).
30. M. Gerstein and M. Levitt, *Protein Science*. **7**(2), 445 (1998).
31. A.D. McLachlan. *Acta Cryst.* **A38**, 871 (1982).
32. J.H. Ward. *J. Am. Stat. Assoc.* **58**(301), 236 (1963).