3D-Blast: Fast 3D Protein Shape
Superposition, Comparison, and Classification

Lazaros Mavridis and Dave Ritchie
INRIA Nancy - Grand Est, LORIA, 615, rue du Jardin Botanique
54506 Vandoeuvre-lès-Nancy, France

Introduction
In biology, it is axiomatic that a protein's amino acid sequence determines its 3D molecular structure, and that a protein's 3D structure determines its specific function. Thus, protein structure and function are intimately related. Often, biologists calculate sequence alignments as a first step in determining the function and evolutionary origin of an unknown protein. However, this typically requires that there be at least 25% similarity between a known sequence and the query sequence. Nonetheless, proteins can have much lower sequence similarities than this, yet still share the same overall fold. In other words, in Nature, protein folds are more highly conserved than protein sequences. Therefore, structural alignment can be a useful tool for analysing the evolutionary or functional relationships of proteins that have very low sequence similarity. However, whereas symbolic techniques such as the Smith-Waterman and Needleman-Wunsch algorithms for aligning protein sequences have become standard tools in bioinformatics, it remains an open question as to how best to align the 3D structures of proteins (Sippl and Wodak, 2008). Most existing structural alignment algorithms are based on comparisons of the Ca backbone traces or vectors formed by the Co-Cb atoms of each non-glycine amino acid, for example (Kolodny et al., 2005). However, these approaches are significantly more computationally expensive than the symbolic techniques because they typically entail the calculation of multiple least-squares rotation matrices, and this is very expensive in the context of dynamic programming alignment algorithms.

Here, we show that spherical polar Fourier (SPF) representations may be used to superpose and compare protein structures in an efficient and completely sequence-independent manner. We present results which illustrate the accuracy with which protein shapes can be encoded, reconstructed, and classified using SPF expansions. We also demonstrate the utility of our approach by performing queries of a single protein structure against the entire CATH database (DB). We have implemented this approach in the program 3D-Blast.

Polar Fourier Expansion of Shape Density Functions
In the SPF approach, protein shapes are represented as 3D density functions expressed as expansions of orthonormal basis functions:

\[ p(r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_n^{(m)} R_n(r) \phi_m(\theta, \phi) \]

where \( p(r, \theta, \phi) \) are the real spherical harmonics, \( N \) is the order of the highest polynomial power of the expansion, \( R_n(r) \) are radial functions, and \( a_n^{(m)} \) are the expansion coefficients which are calculated as described previously (Ritchie and Kemp 2000). Figure 1 shows a pair of nitrogenase proteins (Levitt and Gerstein 1998) at various expansion orders.

Figure 1. Two nitrogenase proteins, represented as SPF expansion at selected expansion orders. The protein in the top row is from autotrophic rhodanese (PDB code 2D4Q). The protein in the bottom row is from Nitrosospira (PDB code 1M60).

Superposing Protein Structures
In order to superpose a pair of protein structures we maximize the Carbon similarity score \( S_{C_{\text{CD}}} \) by performing a 6D rotational/translational search over all positions of the 2nd protein, where \( S_{C_{\text{CD}}} \) is given by:

\[ S_{C_{\text{CD}}} = \sum_{n=0}^{N} \sum_{m=-n}^{n} \left( \sum_{i=1}^{I} a_i^{2(n-m)} \right) \]

and where \( b_n^{(m)} \) are the rotated and translated coefficients of protein b. Figure 2 shows the superposition of the example proteins obtained using expansions to order N=6.

Figure 2. The two nitrogenase proteins from Figure 1 are in their superposed orientations (top row). The proteins have 43% sequence identity. The density functions (right) are shown using an expansion order of N=25.

Rotationally Invariant Descriptors
Although we can perform very fast rotational superpositions (~0.3 sec/pair) using the above approach, we need to develop even faster comparison techniques in order to search large 3D structural databases. It is therefore natural to use the vector interpretation of SPF coefficients to construct rotationally invariant fingerprints (RIF) as described below.

\[ A_{i} = \sum_{n=1}^{N} a_{n}^{2} \]

If the coefficients \( a_{n}^{(m)} \) define the protein shape density, then the rotation-invariant descriptors \( A_{i} \) are related to the mass of the protein.

In order to compare proteins using their RIFs we calculate a new Carbo-like similarity function \( S_{C_{\text{CD}}} \) using:

\[ S_{C_{\text{CD}}} = \sum_{i=1}^{I} a_i \]

Clustering CATH Families
In order to evaluate our approach we performed clustering experiments on selected proteins from the CATH DB v3.2 with sequence identity to 69% (Oregano et al. 1997). CATH classifies proteins according to class, architecture, topology and homology and gives them a superfamily classification. For our experiments we use a clustering study for each of the four classes of CATH was performed. For each of the clusters an arbitrary number of superfamilies were selected with the same architecture.

Figures 3 and 4 show the clustering achieved using 3D-Blast for all four CATH classes. As can be seen, all clusters generally agree with the CATH classification, with a few exceptions. For example, in the All-\( \alpha \) clustering results, the 1N9A4D protein is assigned to a different group because of the topology of the n-\( \alpha \) helices.

Figure 3. The columns on the left shows the 3D-Blast clustering results. The columns on the right shows the results for clustering CATH v3.4 (for secondary structures).

Searching the CATH Database
We searched the entire CATH DB of 12,000 proteins using asparagus synthetase (PDB code 12AS, CATH superfamily 3.3.90.10) as the query molecule. Figure 5 shows the results of the search in the form of receiver operating characteristics (ROC) plots. This figure also shows the 27 members of this superfamily, which were treated as true positives with respect to the query. These plots show that our approach gives very good precision and recall.

To further analyze the results, the DB hits were clustered into 5 groups. The query belongs to group 1, and all members of this group were found in the top 10 DB hits. Groups 2 and 3 have similar \( \beta \)-sheet structures to Group 1, but different arrangements of \( \alpha \)-helicels. All proteins in Groups 2 and 3 are ranked in the top 25% of the DB. All proteins in Group 4 are ranked in the top 30%. Finally, the singleton Group 5 is an obvious outlier due to its extra-\( \alpha \)-heliical domain.

The DB search results for the RIF scoring function show that this function can be used as a fast pre-filter. We therefore performed a RIF search and re-accepted the top 30% using the ROT function. Figure 5 shows that this gives similar or better results compared to full rotational searches in one third of the time.

Conclusions
It has been shown that low resolution SPF expansions provide a reliable and fast way to superpose and compare protein structures. Our clustering results are generally consistent with the CATH standard. The database search results suggest it will be possible to perform real-time on-line database searching using 3D shape-based queries.

References
This project is funded by the ANR, grant reference ANR-08-CEXC-01701.