Eurographics Workshop on 3D Object Retrieval (2010)
I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp (Editors)

SHREC'10 Track: Protein Model Classification

L. Mavridis[†,1] V. Venkatraman,[†,1] D. W. Ritchie,[†,1] N. Morikawa,[2] R. Andonov,[3] A. Cornu,[3] N. Malod-Dognin,[3]

J. Nicolas,[3] M. Temerinac-Ott,[4] M. Reisert,[4] H. Burkhardt,[4] A. Axenopoulos,[5] P. Daras[5]

[1]ORPAILLEUR / INRIA Nancy - Grand Est, France
[2]GENOCRIPT, Japan
[3]SYMBIOSE, IRISA / INRIA Rennes, France
[4]Albert-Ludwig University Freiburg, Germany
[5]Informatics & Telematics Institute Thessaloniki, Greece

Abstract

*This paper presents the results of the 3D Shape Retrieval Contest 2010 (SHREC'10) track Protein Models Classification. The aim of this track is to evaluate how well 3D shape recognition algorithms can classify protein structures according to the CATH [CSL*08] superfamily classification. Five groups participated in this track, using a total of six methods, and for each method a set of ranked predictions was submitted for each classification task. The evaluation of each method is based on the nearest neighbour and area under the curve(AUC) metrics.*

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Curve, surface, solid, and object representations—Geometric algorithms, languages, and systems

1. Introduction

The specific shapes of protein molecules are central to their biological function. Conventional approaches to compare and classify proteins usually work with their amino acid sequences (e.g. BLAST [AGM*90] and FASTA [LP85]). However, in Nature, the 3D structures of proteins are often more conserved than their sequences. Hence, structural alignments can provide significant insights about protein function and can help classify protein families into functional super-families [HS95].

Currently, the most widely used protein structure classification systems are CATH [CSL*08] and SCOP [MBHC95], both of which are curated by human experts. In CATH, the classification is initially performed using the SSAP [OT96] structural alignment tool, whereas SCOP relies more on visual inspection by the curators. However, with the rapid growth of the three-dimensional (3D) protein structures in the Protein Data Bank (PDB [BWF*00]), it would be desirable to be able to assemble and update structural classifications in a more automated way.

2. Task

The task of this track is to classify protein structures according to their CATH superfamilies. Five groups (listed in the order in which they registered) participated in this track, and each group was intially provided with a data set of 1000 proteins, selected by the track organisers, with which they could train or prepare their algorithms. All information about the nature of the proteins and their primary amino acid sequence information was masked to prevent the participants from using such knowledge in conventional protein sequence analysis software. Five days before the deadline for the track, 50 further protein structures (Figure 1) were made available to be used as queries against the initial set. The participants were asked to rank the initial dataset in order of similarity to each of the query proteins. Thus each group was asked to submit 50 ranked lists for each similarity method used.

3. Data

Using CATH version 3.3, the track organisers assembled a dataset of 1000 protein structures from 100 CATH superfamilies, where each superfamily consisted of at least 10 structures, and where each structure contained at least 50 amino acids. From 50 of the superfamilies, one additional member

† Organizer of this SHREC track.

DOI: 10.2312/3DOR/3DOR10/117-124

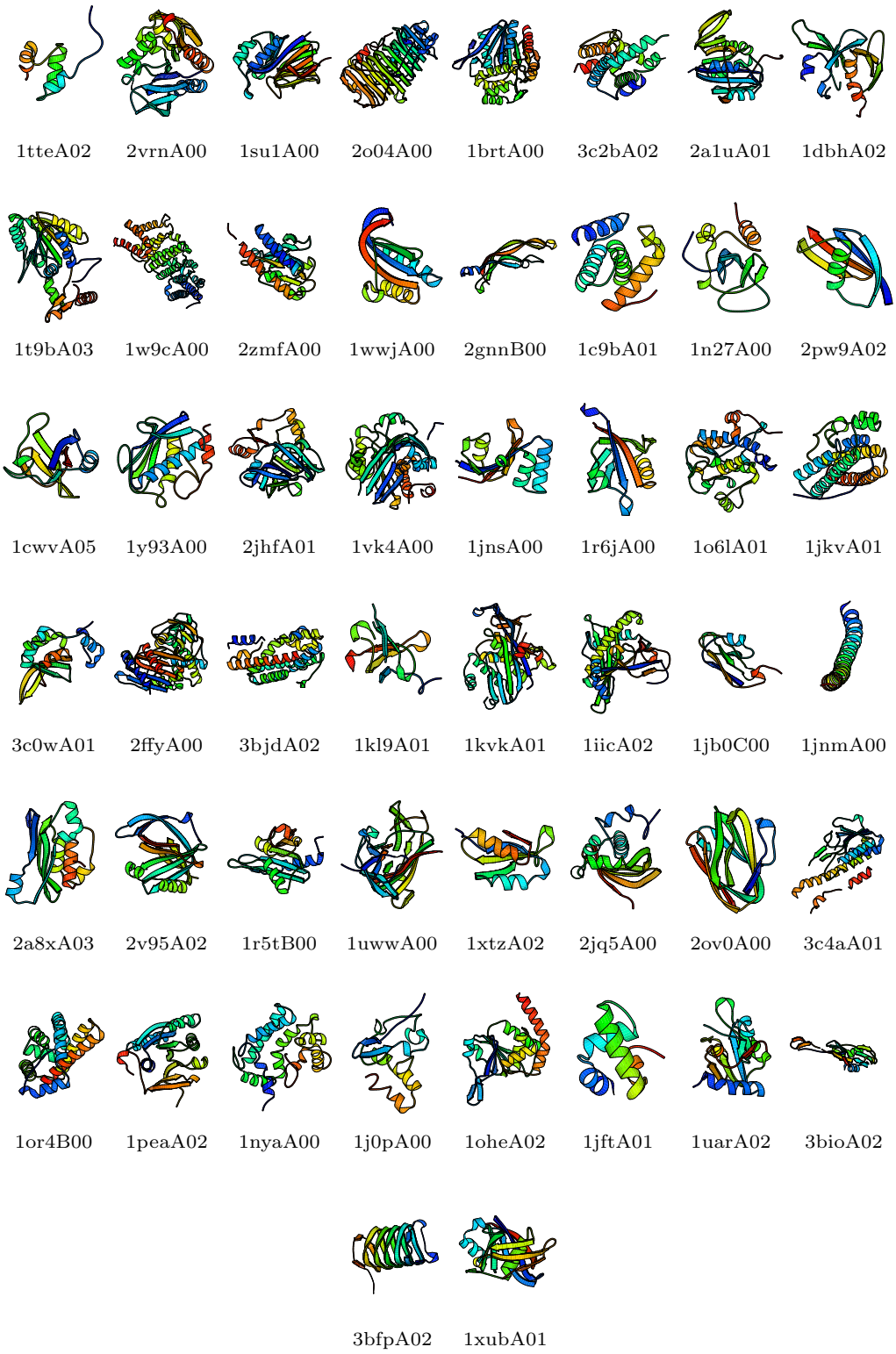

Figure 1: *"Ribbon cartoon" representations of the 50 protein structures used as queries in the evaluation (in numerical query order from top left to bottom right). Each protein is labelled according to the CATH naming scheme.*

was selected at random to serve as a query structure. The protein file names and protein sequence information were masked to try to prevent the participants from using conventional protein sequence matching techniques. Hence, the supplied data files included only the x, y and z coordinates and radii for the atoms within each protein. A simple table was also provided which associates each given protein structure file with a synthetic superfamily name (e.g. "F001"), as shown in Table 1.

Protein File	Protein Family
P0001.pdb	F001
P0002.pdb	F001
...	...
P0500.pdb	F050
...	...
P0999.pdb	F100
P1000.pdb	F100

Table 1: *An extract of the classification file, which was provided with the initial data set.*

4. Evaluation

For the evaluation, the participants were asked to provide a ranked list for each of the query proteins against the 1000 protein dataset. Using these ranked lists, the perfomance of each method was measured in two different ways.

- Nearest neighbour: If the first protein of each ranked list was found to be a member of the same CATH superfamily as the query, this was counted as a correct prediction. The overall percentage of correct predictions was calculated over the 50 queries submitted by each group.
- ROC plot (Receiver Operating Characteristic [Ega75]): By construction, each ranked list contained 10 true positives (TPs) and 990 true negatives (TNs). In order to measure the overall ability of each method to distinguish the TPs from the TNs, the list was traversed sequentially and the rate of TPs (TPR) against the rate of FPs (FPR) was plotted. The area under the curve (AUC) of each ROC plot was calculated to give a single numerical performance measure. A perfect prediction would consist of a list of 10 TPs followed by 990 TNs, giving an AUC of 1.0.

5. Methods

5.1. Spherical Polar Fourier Shape Density Functions (SPF) by L. Mavridis and D.W. Ritchie

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

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

Method Name	Participants
3DBlast	L. Mavridis and D.W. Ritchie
3DZernike	V. Venkatraman
GENOCRIPT	N. Morikawa
Contact Maps	R. Andonov, A. Cornu, N. Malod-Dognin, and J. Nicolas
Group Integration	M. Temerinac-Ott, M. Reisert, and H. Burkhardt
Spherical Trace Transform	A. Axenopoulos and P. Daras

Table 2: *Participating groups and methods.*

where N is the order of the expansion, $R_{nl}(r)$ are Laguerre-Gaussian radial functions, $y_{lm}(\vartheta, \varphi)$ are spherical harmonics, and a_{nlm} are the expansion coefficients which are calculated numerically as described previously [RK00]. Figure 2 shows the SPF representations of a pair of similar nitrogenase domains at several expansion orders. For this track, we used expansions to order $N = 25$ for all calculations.

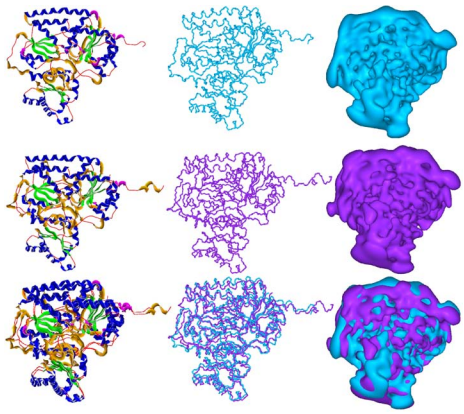

Figure 2: *The superposition of a pair of nitrogenase proteins, shown as ribbon cartoons (left), backbone traces (middle), and as 3D SPF density expansions to order* N=25 *(right). The protein in the top row is from azotobacter vinlandii (PDB code 2MIN). Top row: PDB code 2MIN; middle row PDB code 1MIO;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

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

$$S_{ROT} = \frac{\sum_{nlm}^{N} a_{nlm} b_{nlm}}{[\sum_{nlm}^{N} a_{nlm}^2]^{\frac{1}{2}} [\sum_{nlm}^{N} b_{nlm}^2]^{\frac{1}{2}}} \qquad (2)$$

Conceptually, one protein is held fixed and a six-dimensional (6D) rotational/translation 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 [RK00].

5.2. 3DZernike by V. Venkatraman

3D Zernike descriptors [NK04, LLL*08], an extension of spherical harmonics, have been used for molecular shape retrieval and more recently for protein-protein docking [VLYK09]. A key point in favour of this representation is that of rotational invariance while allowing for a compact shape representation to an arbitrary expansion order. Mathematical and implementation details can be found in the papers by Novotni and Klein [NK04] and Mak *et al.* [MGM08]. For the current protein classification task, the following procedure was used:

Surface Generation Molecular surfaces for the proteins were generated using the MSMS software [SOS96].

Binary Voxelization The program *binvox* [Min] was used to produce a binary voxel grid (voxel dimension set to 128).

Zernike moments Software provided by Novotni and Klein [Nov] was used to calculate Zernike moments upto an expansion order $N = 20$. Each protein is thus represented by a vector of 121 coefficients.

Similarity Measurement Two protein shapes A and B represented by their respective Zernike moments were compared using the Euclidean metric $d = \sqrt{\sum_{i=1}^{121}(A_i^2 - B_i^2)}$.

5.3. GENOCRIPT / D2 Encoding by N. Morikawa

We performed the retrieval of the dataset of 1000 protein structures for structurally similar proteins of 50 query structures in the following three steps. First, the "CA" (or α-Carbon atom) traces of the proteins were extracted from the supplied data files by considering the pattern of atom radii. Next, the D2 codes of the 1000 protein structures were computed by program "ProteinEncoder" and saved in a ".code" file (392 KB): target_SHREC2010.code. Also computed were the D2 codes of the 50 query structures: query_structure.code. Then, retrieval of the dataset was carried out with program "ComSubstruct," which computes the length of the longest common subsequence of two D2 codes. For example, the top 100 D2 code-similar fragments are obtained by typing the following command: "ComSubstruct -l -o1 -s -w1.1 -b100 query_structure.code target_SHREC2010.code." Because more than one fragment may correspond to a protein, the top-most fragment was chosen for each protein to obtain the ranked list of protein names. See below for more detail. The programs ProteinEncode and ComSubstruct are available from `http://www.genocript.com`.

5.3.1. Extraction of the CA Trace of a Protein

To identify the main-chain fragments of N-CA-C atoms, the supplied data files were examined for the atom radius pattern of 1.70-2.00-1.74. Only the CA atoms of the N-CA-C fragments are considered in our method.

5.3.2. D2 Encoding of Local Protein Structures

We used a discrete differential geometrical technique called "D2 encoding" to analyse local protein structures, where the conformation of all five-CA fragments (i.e. fragments of five CA atoms) of a protein are encoded using a five-tetrahedron sequence [Mor07]. First, the conformation of each five-CA fragment is represented by a folded sequence of five tetrahedrons. Next, the corresponding (0,1)-valued sequence of length five, which are denoted as a base-32 number, are assigned to the center CA atom of the fragment. Then, we obtain a description of the conformation of a protein by arranging base-32 numbers in the order that the corresponding CA atoms appear in the CA trace. The base-32 number sequence is called the D2 code of a protein.

5.3.3. Dataset Search by ComSubStruct

One of the simplest measures of sequence similarity is the length of the longest common subsequence (LCS). We used the length of the LCSs of two D2 codes to quantify the differences between two protein backbone conformations. The width of compare window was set to the product of "1.1" and the length of the shorter sequence using the "-w" option. The width of slide step was then the product of 0.1 and the length of the shorter sequence.

5.3.4. Sorting Structures

Protein names are ranked based on the length of LCS. The length of a protein sequence is used for tie-break purposes (the shorter, the better). The similarity scores are obtained by dividing the LCS-length by the protein-length. More precisely, the maximum value is (protein-length - 4) / protein-length.

5.4. Contact Map Overlap Maximization by R. Andonov, A. Cornu, N. Malod-Dognin and J.Nicolas

5.4.1. Principle

This approach compares protein structures based on common inter-atomic contacts. Formally, the contact map of a protein is a graph, $CM = (V,E)$, with vertices V associated to the amino-acids of the protein and contact edges E associated to close amino acids (Euclidean distance between CA atoms, which are known to form the backbone of the protein) smaller than a given threshold. The similarity between two proteins is then determined by the maximum overlap of their contact maps (equivalent to their maximum Number

of Common Contacts (*NCC*)). Finding this number and the associated alignment between the amino-acids of both proteins, known as Contact Map Overlap maximization (CMO), is an NP-hard problem [GIP99] and has been extensively studied in the bioinformatics and computer science communities [CCI*04, XS07].

5.4.2. The A_purva Solver

To classify the queries in the context of SHREC_10 we used the solver A_purva which has been recently proposed in [AYMD08]. A detailed description can be found in [MD10]. A_purva is able to solve CMO in an exact manner in the framework of a classical branch and bound approach (B&B) where upper (UB) and lower (LB) bounds are generated by Lagrangian relaxation. When an instance is optimally solved, we have the relation $LB = NCC = UB$. Otherwise $UB > LB$ and the so called relative gap $\frac{UB-LB}{UB}$ gives an idea of the precision of the results. This property was very useful in the context of SHREC_10 where, because of the time limitation, we were forced to limit the search process on the root of the B&B only.

A_purva was launched without branch and bound, with a limit of 10 000 subgradient descent iterations (i.e. about 20sec per instance). For most query instances (with less than 700 CA atoms) a limit of 2 000 iterations and 4 sec gave the same results.

5.4.3. Extraction of the Backbone and Generation of the Contact Map

In order to adapt A_purva to SHREC_10 conditions where only the coordinates of the atoms have been provided, without identifying their names, we proceeded as follows. Interesting atoms have been filtered on the basis of stable distances that could correspond to the protein backbone (in a PDB file, consecutive atoms N, CA, C and O exhibit N-CA, CA-C and C-O bonds with relatively fixed distances of 1.45Å, 1.53Å and 1.24Å, respectively). Note that we did not use atom radii for this purpose. Globally, the procedure tends to filter all CA and a few other carbon atoms in each protein that we consider as CA in the rest of the treatment.

The contact maps were generated with a distance threshold of 7.5Å between two CA atoms, excluding natural contacts between consecutive amino-acids.

5.4.4. Scoring Scheme

Based on the obtained values, two scoring functions were tested in order to detect the similarity between a query Q and each protein P of CATH superfamilies. The first one was first proposed in [XS07]:

$$\mathrm{SIM}(Q,P) = \frac{2 \times LB}{|E_Q| + |E_P|}. \tag{3}$$

Once results known, this default score appeared to be the best one for the classification task. The nearest neighbour score with it reaches 88%.

The second index used the confidence in the results of A_purva, $C = \frac{LB}{UB}$, and was finally retained for the contest. The score is given by:

$$\mathrm{Cscore}(Q,P) = \frac{C \times UB + (1-C) \times LB}{|E_Q| \times (1 + \frac{\mathrm{abs}(|E_P| - |E_Q|)}{\max(|E_P|, |E_Q|)})} \tag{4}$$

A final step used the knowledge of superfamily labels: the mean rank of the three best scores for each superfamily was computed. It allowed classifying all proteins of a same superfamily together if they got a good rank.

All Contact Maps computations were done on the Ouestgenopole bioinformatics platform `http://genouest.org`.

5.5. Group Integration for Protein Structure Description by M. Temerinac-Ott, M. Reisert and H. Burkhardt

Group Integration (GI) is a powerful tool for describing three dimensional structures [BS01]. The main idea is to average the representatives of a transformation group (e.g. Euclidean) in order to obtain group invariant descriptors, which can be compared in order to determine similarities. Group integration can be extended by Spherical Harmonics [RB06] in order to obtain more robust descriptors. The details of our method are explained in [TRB07].

5.5.1. Modelling Protein Shape

Proteins can be described by the position of the atoms of the protein and their order in the amino acid sequence. In order to apply group integration to proteins, the proteins are modelled as superpositions of Gaussian distributions centered at the positions of the atoms.

In the SHREC'07 protein track [MTB07], only CA atoms were used, whereas here we now use all atoms to compute GI features. However, we did not use the provided atom radius data.

5.5.2. Classifying Proteins Based on GI Features

The result of group integration is a multidimensional histogram $H_{\alpha,\beta,\gamma,\Delta,\mu,\mathfrak{l}}$ with 2048 bins. Through concatenation of the histogram dimension, we obtain one feature vector for each protein. The similarity measure $s(x,y)$ between two feature vectors is obtained using the χ^2_2 distance.

5.6. 3D Protein Classification Using the Spherical Trace Transform by A. Axenopoulos and P. Daras

Our 3D shape-based approach is presented for the efficient search, retrieval, and classification of protein molecules. The method relies on the geometric 3D structure of the proteins, which is produced from the corresponding PDB files. After proper positioning of the 3D structures, in terms of translation and scaling, the Spherical Trace Transform is applied

to them so as to produce geometry-based descriptor vectors, which are completely rotation invariant and perfectly describe their 3D shape.

5.6.1. Preprocessing

Since the exact 3D position and radius of the protein's atoms is known from the available PDB file, the protein can be represented as a set of spheres. Then, the Solvent Excluded Surface is computed using the MSMS algorithm [SOS95].

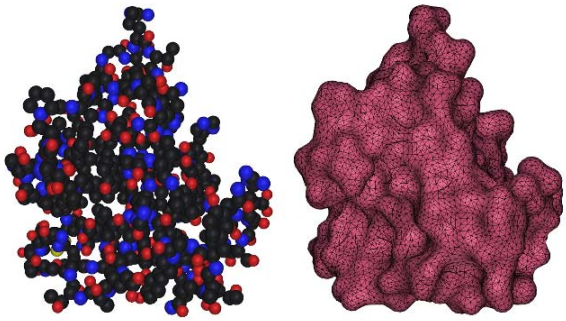

Figure 3: *3D representation of a protein with a) spheres and b) Solvent Excluded Surface.*

The protein is now represented as a triangulated mesh which provides a sufficient approximation of the protein's 3D shape. As a next step, a voxelization process, similar to the one presented in [DZA*06] takes place. More specifically, the 3D mesh is placed into a bounding cube, which is partitioned in equal cube shaped voxels. Voxels that lie inside the 3D model or on the surface are assigned non-zero values.

5.6.2. Descriptor Extraction

Every 3D object is expressed in terms of a binary volumetric function. In order to achieve translation invariance, the center of mass of the 3D object is calculated and the model is translated so that its center of mass coincides with the coordinate system origin. Scaling invariance is also accomplished, by scaling the object in order to fit inside the unit sphere. Then, a set of concentric spheres is defined. For every sphere, a set of planes which are tangential to the sphere is also defined. Further, the intersection of each plane with the object's volume provides a spline of the object, which can be treated as a 2D image.

Next, 2D rotation invariant functionals, F, are applied to this 2D image, producing a single value. Thus, the result of these functionals when applied to all splines, is a set functions defined on every sphere whose range is the results of the functional. Finally, a rotation invariant transform, T, is applied on these functions, in order to produce rotation invariant descriptors. For the needs of the SHREC, the implemented functionals F are the 2D Krawtchouk moments, and the Polar Fourier Transform, while the T function is the Spherical Fourier Transform.

A more detailed description of the extraction of these descriptors is available in [DZA*06]. The dimension of descriptor vectors is $N_{Fourier} = 1080$ for the descriptors based on the Polar-Fourier 2D functional and $N_{Krawtchouk} = 1080$ for the descriptors based on the Krawtchouk 2D functional.

5.6.3. Matching

Firstly, the descriptors are normalized so that their absolute sum is equal to 1. Then, the Minkowski L1 distance is computed for a pair of descriptor vectors. The L1 distance is a measure of dissimilarity between two descriptor vectors. In order to transform this dissimilarity into a similarity metric, a decreasing sigmoid function was applied so that low-dissimilarity values are closer to 1 and high-dissimilarity values are closer to 0.

6. Results

In this section, we present the perfomance evaluation results of the track. Each participating group submitted one set of results based on their selected set of parameters. This was a blind experiment and each group could only submit one set of results. Therefore, it was not possible for partipants to tune the parameters of their algorithms.

Nearest neighbour : Table 3 summarizes the retrieval rates for all the methods. There were five cases in which none of the methods found the nearest neighbour. These were: Q12 (1wwjA00), Q30 (1iicA02), Q40 (3c4aA01), Q43 (1nyaA00), and Q48 (3bioA02). In a further seven cases, only one method found the nearest method as the top match. However, there were 11 additional cases in which several methods found the nearest neighbour as the second hit (i.e. 4 for GENOCRIPT, 3 for Group Integration, 3 for 3DBlast and 1 for 3DZernike).

Method	Correct Predictions
3DBlast	68%
3DZernike	8%
GENOCRIPT	56%
Contact Maps	80%
Group Integration	52%
Spherical Trace Transform	0%

Table 3: *Nearest neighbour results.*

ROC plots : For each of the submitted result lists, a ROC plot and its corresponding AUC was calculated. Figure 4 shows the resulting AUC of all methods for each target. Because early recognition of TPs is at least as important as obtaining a good overall AUC score, we also calculated another set of AUC values which correspond to the first part, up to 10% of the database, of the ROC curves. An aggregate

ROC plot was also calculated to summarize the overall performance of each method as a single ROC curve, as shown in Figure 5.

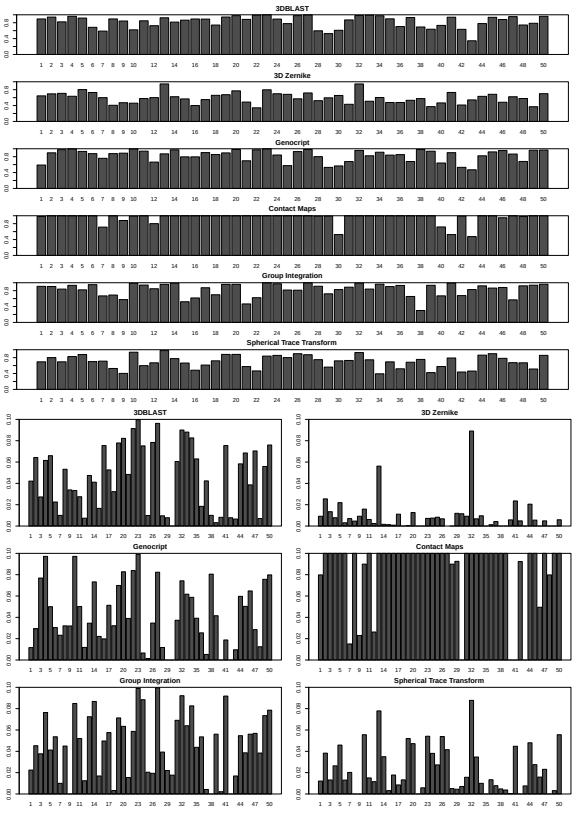

Figure 4: *Bar chart analyses for each method showing the calculated AUC for each of the 50 query proteins. The upper bar charts show the total AUC, whereas the lower bar charts show the AUCs calculated for the top 10% of the database.*

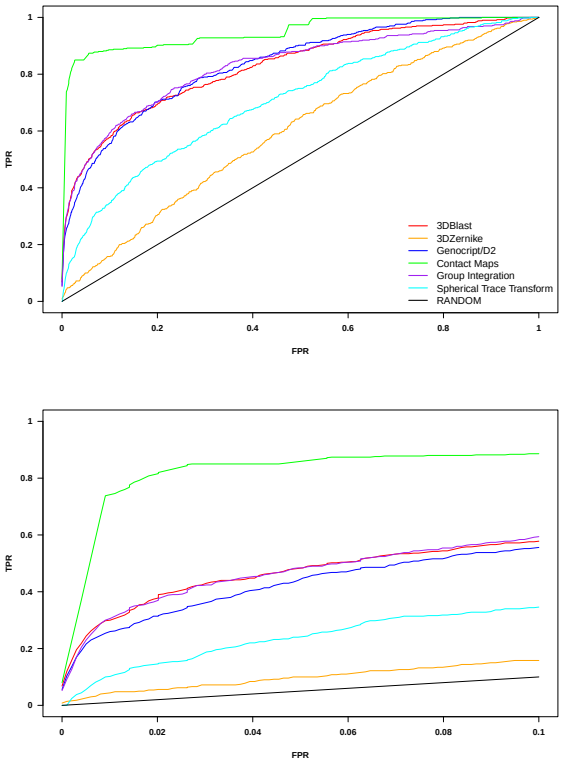

Figure 5: *The upper figure shows aggregate ROC plots for each method obtained when quering the 1000 protein dataset using the 50 query proteins. The lower figure shows an expanded view of the first 10% of the upper figure to highlight the early recognition behaviour of each method.*

7. Conclusions

In this paper, we have presented and compared the perfomance of six algorithms submitted by the five research groups who participated in this track.

Contact Maps and 3D-Blast were conceived specifically to compare proteins structures, and these approaches give the best results, although the Group Integration and GENOCRIPT/D2 approaches also perform very well. The Contact Maps and GENOCRIPT approaches both used a preselection step to try to infer the CA backbone structure of the corresponding proteins from simple geometrical invariants.

Contact Maps compares proteins on the basis of conserved proximities between atoms, where Genocript encodes the CA backbone structure of length N into a 16 valued-sequence of length (N-4).

Both 3D-Blast and 3D-Zernike compare shapes globally, but 3D-Blast uses FFT-based rotational comparisons, whereas 3D-Zernike uses a fast scale- and rotation-invariant scoring technique derived from a spherical harmonic plus Zernike polynomial expansion of each protein. The Spherical Trace Transform approach calculates scale- and rotation-invariant descriptors from 2D slices of the protein volumes using polar Fourier transforms. The Group Integration approach constructs and compares group invariant descriptors from the given atomic coordinates of each protein. With the exception of the 3D-Zernike approach, which gave unexpectedly disappointing results, the general shape classification approaches also gave very encouraging predictions when one considers the generic nature of those approaches and the very tight timetable under which this experiment was conducted.

Although in this experiment, some superfamilies may have been easier to identify than others, it is worth noting that no approach can reproduce the classification of the human experts in all cases. This suggests that protein model-

ing and classification is a difficult task for current 3D shape recognition methods. Therefore adopting a benchmark based on protein shape classification, such as the one presented here, will provide a challenging dataset with which to evaluate new 3D object recognition algorithms.

References

[AGM*90] ALTSCHUL S., GISH W., MILLER W., MYERS E., LIPMAN D.: Basic local alignment search tool. *J. Mol. Biol. 215* (1990), 403–410.

[AYMD08] ANDONOV R., YANEV N., MALOD-DOGNIN N.: An efficient lagrangian relaxation for the contact map overlap problem. In *WABI '08: Proc. of the 8th int. workshop on Algorithms in Bioinformatics* (2008), Springer-Verlag, pp. 162–173.

[BS01] BURKHARDT H., SIGGELKOW S.: Invariant features in pattern recognition – fundamentals and applications. *Nonlinear Model-Based Image/Video Processing and Analysis, edts. C. Kotropoulos and I. Pitas, John Wiley & Sons* (2001), 269–307.

[BWF*00] BERMAN H., WESTBROOK J., FENG Z., GILLILAND G., BHAT T., WEISSIG H., SHINDYALOV I., BOURNE P.: The protein data bank. *Nucleic Acids Research 28* (2000), 235–242.

[CCI*04] CAPRARA A., CARR R., ISRAIL S., LANCIA G., WALENZ B.: 1001 optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol. 11*, 1 (2004), 27–52.

[CSL*08] CUFF A., SILLITOE I., LEWIS T., GARRATT O., THORNTON J., ORENGO C.: The cath classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research 37* (2008), 310–314.

[DZA*06] DARAS P., ZARPALAS D., AXENOPOULOS A., TZOVARAS D., STRINTZIS M.: Three-dimensional shape-structure comparison method for protein classification. *IEEE/ACM transactions on Computational Biology and Bioinformatics 3(3)* (2006), 193–207.

[Ega75] EGAN P.: Signal detection theory and roc analysis. *Academic Press: New York* (1975).

[GIP99] GOLDMAN D., ISTRAIL S., PAPADIMITRIOU C.: Algorithmic aspects of protein structure similarity. In *FOCS '99: Proc. of the 40th Annual Symposium on Foundations of Computer Science* (1999), IEEE Computer Society, pp. 512–521.

[HS95] HOLM L., SANDER C.: Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences 20* (1995), 478–480.

[LLL*08] LEE S., LI B., LA D., FANG Y., RAMANI K., RUSTAMOV R., KIHARA D.: Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins: Structure, Function, and Bioinformatics 72(4)* (2008), 1259–1273.

[LP85] LIPMAN D., PEARSON W.: Rapid and sensitive protein similarity searches. *Science 227* (1985), 1435–1441.

[MBHC95] MURZIN A., BRENNER S., HUBBARD T., CHOTHIA C.: Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol. 247* (1995), 536–540.

[MD10] MALOD-DOGNIN N.: *Protein Structure Comparison: From Contact Map Overlap Maximisation to Distance-based Alignment Search Tool.* PhD thesis, University of Rennes 1, 2010.

[MGM08] MAK L., GRANDISON S., MORRIS R.: An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Journal of Molecular Graphics and Modelling 26(7)* (2008), 1035–1045.

[Min] MIN P.: Binary voxelation.

[Mor07] MORIKAWA N.: Discrete differential geometry of tetrahedrons and encoding of local protein structure, 2007.

[MTB07] M. TEMERINAC M. R., BURKHARDT H.: Shrec 2007: 3d shape retrieval contest, protein retrieval track. *Technical Report UU-CS-2007-015, R. C. Veltcamp and F. B. ter Haar (eds.)* (2007), 17–21.

[NK04] NOVOTNI M., KLEIN R.: Shape retrieval using 3d zernike descriptors. *Computer Aided Design 36(11)* (2004), 1047–1062.

[Nov] NOVOTNI M.: 3d zernike descriptors.

[OT96] ORENGO C., TAYLOR W.: Ssap: sequential alignment program for protein structure comparison. *Methods Enzymol 266* (1996), 617–635.

[RB06] REISERT M., BURKHARDT H.: Invariant features for 3d-data based on group integration using directional information and spherical harmonic expansion. *Proceedings of the ICPR'06, Hong Kong* (2006).

[RK00] RITCHIE D., KEMP G.: Protein docking using spherical polar fourier correlations. *Proteins: Struc. Funct. Genet. 39* (2000), 178–194.

[SOS95] SANNER M., OLSON A., SPEHNER J.: Fast and robust computation of molecular surfaces. In *In the 11th ACM Symposium on Computational Geometry* (1995).

[SOS96] SANNER M., OLSON A., SPEHNER J.: An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Biopolymers 38(3)* (1996), 305–320.

[TRB07] TEMERINAC M., REISERT M., BURKHARDT H.: Invariant features for searching in protein fold databases. *International Journal on Computer Mathematics , 'Special Issue on Bioinformatics' 84(5)* (2007), 635–651.

[VLYK09] VENKATRAMAN V., LEE S., YANG Y., KIHARA D.: Protein-protein docking using region-based 3d zernike descriptors. *BMC Bioinformatics 10* (2009), 407–428.

[XS07] XIE W., SAHINIDIS N.: A reduction-based exact algorithm for the contact map overlap problem. *Journal of Computational Biology 14*, 5 (2007), 637–654.