Eurographics Workshop on 3D Object Retrieval (2017)
I. Pratikakis, F. Dupont, and M. Ovsjanikov (Editors)

SHREC'17 Track: Protein Shape Retrieval

Na Song[1], Daniela Craciun[2], Charles W. Christoffer[3], Xusi Han[4], Daisuke Kihara[3,4], Guillaume Levieux[5], Matthieu Montes[2], Hong Qin[6], Pranjal Sahu[6], Genki Terashi[4], Haiguang Liu[1]*

1. Complex Systems Division, Beijing Computational Science Research Center, Z-Park II, Haidian, Beijing, China 100193
2. Laboratoire GBA EA4627, Conservatoire National des Arts et Métiers, 2 rue Conté, 75003 Paris, France
3. Department of Computer Science, Purdue University, 305 N. University St., West Lafayette, IN 47907, USA
4. Department of Biological Sciences, Purdue University, 249 S. Martin Jischke Dr., West Lafayette, IN 47907, USA
5. Laboratoire CEDRIC EA4647, Conservatoire National des Arts et Métiers, 2 rue Conté, 75003, Paris, France
6. Computer Science Dept., Stony Brook University, Stony Brook, NY 11794, USA
 Track organizers

Abstract

The large number of protein structures deposited in the protein database provide an opportunity to examine the structure relations using computational algorithms, which can be used to classify the structures based on shape similarity. In this paper, we report the result of the SHREC 2017 track on shape retrievals from protein database. The goal of this track is to test the performance of the algorithms proposed by participants for the retrieval of bioshape (proteins). The test set is composed of 5,854 abstracted shapes from actual protein structures after removing model redundancy. Ten query shapes were selected from a set of representative molecules that have important biological functions. Six methods from four teams were evaluated and the performance is summarized in this report, in which both the retrieval accuracy and computational speed were compared. The biological relevance of the shape retrieval approaches is discussed. We also discussed the future perspectives of shape retrieval for biological molecular models.

Categories and Subject Descriptors (according to ACM CCS): Computing methodologies~Shape modeling • Computing methodologies~Shape analysis

1. Introduction

Protein structures provide critical information for the understanding of protein functions. Large databases of proteins are rapidly accumulated in recent years[BKW*77], effective and efficient methods for protein retrieval and classification are required for detailed structural and functional analysis. The Protein Challenge in SHREC 2007 included a new track in biology field for the Shape Retrieval Contest which gives the participants a chance to explore the protein structures and shapes [VH07]. However, the conventional structural comparison is based on amino acid sequence alignment to identify the paired atoms from two structures. A translocation matrix is optimized by translation and rotation operations to find the minimum distance between two structures. This difference metric is named as Root-Mean-Square-Deviation (RMSD). In spite of wide usage of this structure comparison method, it has severe limitations: (1) The structures to be compared must have high sequence similarity in order to find the one-to-one correspondence of atoms in two models; (2) the structures to be compared should have relative high resolutions, such that atomic position assignment is correct. There are developments of model comparison methods that do not depend on sequence alignment, for example, the DALI program does not require sequence information for model alignment [HKRS08]. Nonetheless, new algorithms and programs are needed to reveal structure relations within the database, or to search functional relations between a newly determined structure and the existing ones. Recently, a new challenge has arisen due to the high throughput structure determination at low resolutions, in which case protein structure comparison by amino acid sequence is impractical. For example, the small angle X-ray scattering data only provides information that is sufficient to build 3D models at nanometer resolutions, which cannot be used to assign atomic positions of the molecules [HMH*09, LHZ12, SdVS*16]. Through this SHREC track, we hope to stimulate the development of shape-based methods and the application of these methods in the research of structure biology.

DOI: 10.2312/3dor.20171055

To get connected to the community of shape studies, we intentionally removed the biological information of the protein structures. This is accomplished in three steps: (1) scale the models to the same size such that they perfectly fit in the sphere with a radius of 30Å; (2) map the atoms to the 3D grid with grid size of 1Å; (3) remove the sequence information by using Carbon atoms to represent the shape in the form of point clouds. In future contest, we plan to include the consideration of biological relevant information, such as the actual molecular size or sequence information.

Four teams submitted results using six shape retrieval methods before the deadline of the track. The specific task of retrieval is to find the top 200 models from 5,856 shapes in the database for each of the 10 query shapes. The performance of each method is subject to the evaluation of several criteria, including Discounted Cumulative Gain (DCG), Nearest Neighbor (NN), First Tier, Second Tier and Mean Average Precision (MAP). The execution time and memory requirements of the methods are reported as well.

2. The Data Set

2.1 Query Set

The query set is composed of 10 proteins which are obtained from the Protein Data Bank [BKW*77]. In order to ensure structural diversity, the 10 proteins were manually picked from the special collection, the molecule of the Month, which summarizes the structure and function of one important protein molecules each month by David Goodsell[GDZ*15]. Table 1 shows the source of each protein. For details about the molecules, the readers are referred to the online documents at each URL. Figure 1 shows the surface representations for 10 query shapes.

Table 1: *The list of the query molecules. The set is selected from the "molecule of the month".*

Protein	URL	PDB ID
Lysozyme	http://pdb101.rcsb.org/motm/9	2LYZ
Antibody	http://pdb101.rcsb.org/motm/170	4MMV
HIV reverse Transcriptase	http://pdb101.rcsb.org/motm/33	3HVT
Insulin	http://pdb101.rcsb.org/motm/14	2HIU
HSP90	http://pdb101.rcsb.org/motm/108	2CG9
Bacteriphage	http://pdb101.rcsb.org/motm/2	1CD3
G protein	http://pdb101.rcsb.org/motm/58	1GG2
Ribosome	http://pdb101.rcsb.org/motm/121	4V4J
Penicillin-binding Protein	http://pdb101.rcsb.org/motm/29	1HVB
.Zika virus	http://pdb101.rcsb.org/motm/197	5IRE

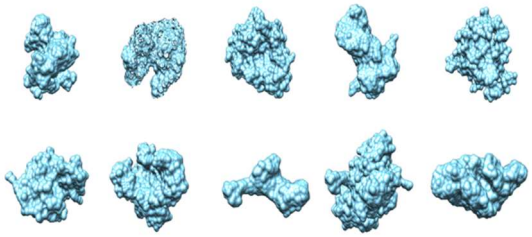


Figure1: *Surface representation for the query molecules. The 10 queries (with query IDs: q_11 to q_20) are arranged from left to right, top to bottom.*

2.2 Target Set

For the target set, a random selection from protein database may result high redundancy (i.e., multiple similar structures of the same molecule might be included), which increases the computation time with no practical benefits. In order to remove the redundancy, we construct the target set from a subset of PDB models, composed of 13,182 protein structure that are non-redundant in terms of the amino acid sequence. These structures are abstracted to shape models that fit in a sphere of the same radius (30Å). The atoms are mapped to 3D grid points separated by 1Å to form a point cloud representation of the corresponding protein molecule. These operations ensure that the biological features are removed to a good extend, so that the technology in computational shape comparison can be applied.

In the second stage, for each query structure, we used the exhaustive model matching method (see Method section) to rank the 13,182 structures, then chose the top 1,000 shape models. As a result, 10,000 models were collected for ten queries. Because same model can be present in the top 1000 for different queries, the final dataset contains 5,854 unique models after removing the repeats. This is the target set for shape retrieval.

3. Methods

Six protein retrieval methods have been proposed by four different research groups. In the following, these methods are briefly described.

3.1 3D Zernike polynomial based method: the choice of ground truth and the exhaustive model matching

The ground truth was not available for this track. We have to use model matching approach to rank the models by comparing a query with a target at discretized orientations that finely span the SO(3) rotational space. 3D Zernike moments (3DZM) are used as the protein shape descriptors [Can99, LPJZ12, NK03].

A brief summary about 3D Zernike polynomial and moments are provided. The 3D Zernike polynomial at order (n,l,m) is represented as:

$$Z_{nlm}(X) = R_{nl}(r) \cdot Y_l^m(\vartheta, \phi) \quad (1)$$

where $R_{nl}(r)$ and $Y_l^m(\vartheta, \phi)$ are the Zernike radical function and spherical harmonic functions, respectively. The order parameters need to satisfy the following: $n \geq l$, $(n - l)$ is even, and $-l \leq m \leq l$.

$\{Z_{nlm}(X)\}$ are orthonormal and complete within the unit sphere.

Therefore, 3D Zernike moments Ω_{nl}^m of an object described by $\rho(X)$ can be defined as:

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|X|<1} \rho(X) Z_{nlm}^*(X) dX \quad (2)$$

Any object in 3D space can be described using a function $\rho(R)$, which can be scaled to fit in a unit sphere, to obtain a scaled representation, $\rho(X)$.

Noticing that the 3D Zernike moments Ω_{nl}^m are not invariant under rotations. Novotni & Klein collect the moments into $(2l + 1)$-dimensional vectors,

$$\Omega_{nl} = \left(\Omega_{nl}^{-l}, \Omega_{nl}^{-l+1}, \Omega_{nl}^{-l+2} \ldots., \Omega_{nl}^{l}\right)^t \quad (3)$$

whose norm, $F_{nl} = \|\Omega_{nl}\|$, is rotational invariant, so that the {Fnl} is defined as 3D Zernike descriptors (3DZD) for shapes [NK*03].

Similarity between two proteins is quantified by the overlap between two models. In the practice, we evaluate the overlap by correlation coefficient (abbreviated as c.c.) between two models.

For a given orientation, c.c. is defined as:

$$c.c. = \frac{<\rho_1(r)\rho_2(r)> - <\rho_1(r)><\rho_2(r)>}{\sigma(\rho_1(r))\sigma(\rho_2(r))} \quad (4)$$

where $\rho_1(r)$ and $\rho_2(r)$ is the descriptor of the two proteins, $r \in r^3$.

The maximum c.c. for all orientations that finely samples the orientations, is used as the measure of shape similarity,

$$c.c.(protein_{fixed}, protein_{rot}) = \max_{arg\, i} (c.c.(protein_{fixed}, protein_{rot,i})) \quad (5)$$

$i = 1, .., N_{ori}$ and N_{ori} is the number of orientations.

The 3DZM model representations allow one to speed up the model rotation calculation by using Fast Fourier Transformation (FFT) method as described in the Trapani and Navaza (2006) [TN06].

The value of c.c. is in range [0,1], where 0 means unrelated, and 1 can be obtained only when aligning a model to itself. Larger c.c. value means higher similarity between the models. The good retrieval methods should be able to pick the models with large correlations to the query model. The exhaustive model matching using 3DZM is named as the 3DZM method. The pearson correlation between query and target using 3DZD is named as the 3DZD method.

3.2 Kihara's method

This method is provided by Prof. Kihara and his team. They also used 3DZD as shape descriptors as part of the model representation. The Euclidian distance between 3DZD's are used to measure the model difference. The other consideration in this method is the biological relevance of the retrieved models, so they tried to guess the trace of the protein main chain and try to align the models. Furthermore, the Kihara's method carried out re-ranking using molecular size information computed from the point cloud representations.

For models whose sizes are between 0.8 and 1.25 times the size of the query, if the main chain comparison score, TM-score ≥ 0.5, the models will be added to the top of the retrieval list if they are not at the top already.

3.3 Shape Retrieval System driven by Digital Elevation Models

The molecular shape similarity system is composed of two main stages: the first stage is performed for each shape and consists in the global shape representation as a Digital Elevation Model (DEM), encoded over a 2D grid. The second stage corresponds to the shape comparison phase which is supplied via global distance measures computed over the DEMs.

Representing Macromolecular Shapes as Digital Elevation Models

Macromolecular triangular surface computation. The shape representation algorithm applies the EDTSurf [XZ*09] technique to generate the macromolecular surface (MS) from the input data. The algorithm exploits the Vertex Connected Marching Cubes and the Euclidean Distance Transform to generate the triangular mesh which is kept for further processing.

Digital Elevation Model computation. The present work exploits the DEM concept traditionally employed in cartography for representing Earth's surface from terrain elevation data. The algorithm starts by applying the mesh flattening procedure used to map the mesh onto the unit sphere using the Laplace-Beltrami operator, resulting in an isometry invariant shape representation [AHTK99,GGS03]. In the second step, the unit sphere is projected onto a 2D spherical panoramic grid and the elevation values of the input mesh are assigned to each 2D location of the panoramic grid. This results in a global descriptor which encodes elevation values, while providing topology and fast comparison over a 2D grid space.

The final output is the digital elevation model associated to the macromolecular surface, noted MS-DEM. Figure 2 illustrates the results obtained for a target belonging to the protein pool of the BioShape track.

Global Comparison of MS-DEMs

The MS-DEM shape descriptor is used along with different global distances for supplying the shape similarity computation stage. The present research work evaluates the Mean Absolute Differences (MAD) and the Root Mean Square Deviation (RMSD) distances. They are measured over the points belonging to the 2D grids. For input meshes with different number of points, distances are computed over the minimum number of points computed between the query and the target meshes. Two methods are originated from MS-DEM, namely **MS-DEM-MAD** and **MS-DEM-RMSD**.

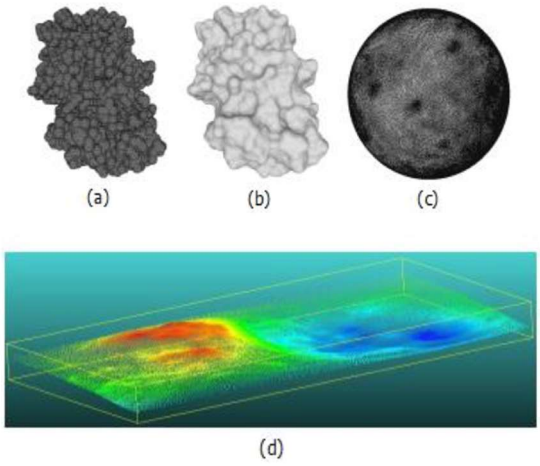


Figure 2: *Overview of the global descriptor computation stage for the input model m10001: (a) input data: 187866 points, 357840 triangles; (b) macromolecular mesh generated by EDTSurf [XZ*09]: 86079 points, 172154 triangles; (c) spherical mapping output: 86079 points, 172154 triangles; (d) MS-DEM output: 86 089 points, bounding box dimensions: [472, 257, 36.112].*

3.4 Principal component analysis based shell and sector (PCAS)

The surface for each protein molecule was generated using pymol and stored as obj _les. Using the obj _les, feature descriptors for each molecule was calculated. First each molecule was centered at the centroid and then aligned using PCA (Principal Component Analysis). After alignment a shell and sector based approach was adopted to describe the mass distribution of molecule using a histogram. The alignment step is necessary to work for this method since the feature descriptor is not rotation invariant.

Considering a set of points $p_1 \ldots p_n$ on the surface of molecule whose centroid is at origin, construct a matrix P whose i_{th} column is vector p_i.

$$P = \begin{bmatrix} p_1^x & p_2^x & p_3^x & \cdots & p_n^x \\ p_1^y & p_2^y & p_3^y & \cdots & p_n^y \\ p_1^z & p_2^z & p_3^z & \cdots & p_n^z \end{bmatrix}$$

Using P build the covariance matrix, $M = PP^T$, whose eigenvectors represent principal directions of shape variation. Using these vectors, the molecule can be aligned [AKKS99]. After alignment, the shell and sector approach is used for comparison. A sphere was divided into 11 concentric shells with 24 equally distributed sector directions along the sphere.

The radius of sphere is taken as the maximum radius of all the spheres which can cover all the molecules in the dataset. Number of points lying in each bin is then calculated by placing each point in its shell (by considering the points distance from centroid) and in each sector by taking the dot product of point's direction with the 24 equi-spaced points on the surface of the sphere.

The histogram is then normalized such that the values lie from 0 to 1 in each dimension. This in total generates a 11*24 bins vector for each molecule. Finally the distance between any two molecules is calculated using the euclidean distance of this feature vector.

4. Evaluation Measures

4.1 Consistency with the 'ground truth'

The correlation coefficient was calculated by the exhaustive model comparison. For each query, we compute the average c.c. between the query and the top N models, (N=1 to 200), for all methods.

$$\text{c. c.}_{\text{ave}} = \sum_{i=1}^{N} \text{cc}_i \qquad (6)$$

where cc_i is the correlation coefficient between the i-th retrieved model and the query model.

4.2 Average Discounted Cumulative Gain

The ranking of the retrieved models is used as weighting factor to evaluate the performance. If a correct model with higher similarity is ranked top, the performance of the method should be considered better. The Discounted Cumulative Gain criterion was designed to use this weighting factor.

For a query, with the top 200 models set $M = \{m_1, m_2, \ldots, m_{200}\}$, we compare it to the groundtruth $GT = \{gt_1, gt_2, \ldots, gt_{200}\}$, then, the flag set $G = \{x_1, x_2, \ldots \ldots\}$ can be calculated by

$$x_i = \begin{cases} 1 & \text{if } x_i \text{ in GT} \\ 0 & \text{if } x_i \text{ not in GT} \end{cases}$$

For the groundtruth (3DZM results), the set G:

$$IG = \{1,1, \ldots \ldots 1\}$$

So, for each method, the InitDCG can be calculated by the following function. The DCG can be obtained by the InitDCG divided by the groundtruch's InitDCG (IDCG).

$$\text{InitDCG[i]} = \begin{cases} \text{G[1]} & \text{i} = 1 \\ \text{InitDCG[i} - 1] + \dfrac{1}{\log_2 \text{G[i]}} & \text{others} \end{cases}$$

$$\text{DCG[i]} = \frac{\text{InitDCG[i]}}{\text{IDCG[i]}}$$

4.3 Other parameters

Nearest Neighbor (NN), First-tier (Tier 1) and Second-tier (Tier 2)

To check the ratio of models in the query's class that also appear within the top K matches, for Nearest Neighbour K=1, for the first tier K=|C|-1, and K=2*|C|-1 for the second tier, where |C| is the number of the query's class. Here we choose |C| = 100.

Precision, Recall, E measure and MAP

Let A be the set of all relevant objects, and B be the set of all retrieved object then:

$$\text{precision} = \frac{A \cap B}{B}$$

$$\text{recall} = \frac{A \cap B}{A}$$

Recall evaluates how well a retrieval method finds the relevant models, while precision evaluates how well it weeds out irrelevant models.

$$E = 1 - \frac{2}{\frac{1}{p} + \frac{1}{R}}$$

E evaluates both the precision and recall performance.

MAP (Mean Average Precision) calculates average precision for every query, and counts the mean value of the average precision for all the classes, it gives the average precision accuracy for retrieval results of all queries.

4.4 Computational Cost

For the practical reasons, the speed of the algorithms is important for the users to choose which method to use. The total computational time includes the time it takes to extract the 3D shape descriptor for an object and to perform one query search on the database.

5. Results

Each team is instructed to submit the top 200 models (in descending rank based on model **similarity** to the query using their own measures). Four teams submitted results of six methods before the deadline of the contest. The following sections summarize the performance of the six methods.

5.1 A glance of the retrieval result

There are two special queries (q_11 and q_14), whose identical models are present in the target database. A basic retrieval test of each method is the retrieval of the query itself. Figure 3 shows the query q_11 and ranking order of itself in each method. It can be seen that 3DZM, 3DZD and Kihara's descriptor retrieval the query as the best matched model, the MS_DEM_MAD and the MS_DEM_RMSD ranked the query to be the second best. As for the PCAS descriptor, the query was ranked to be the 47th in the result. **All six** methods found query q_14 as the best matched model correctly. While it is clear that 3DZM, 3DZD and Kihara's method provide stable and effective performance to find a protein in a large database by using the Zernike descriptor.

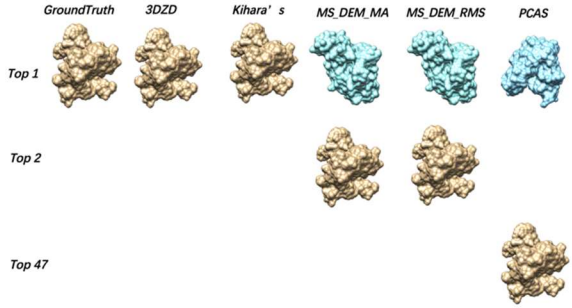


***Figure 3**: The results for query q_11. The query model was ranked to be the first by 3DZM (labeled as groundtruth), 3DZD and Kihara's methods, ranked as the second in the MS_DEM_MAD and MS_DEM_RMSD method, and ranked 47th in PCAS method.*

5.2 Overlaps between the query and models.

The c.c. values are computed using the exhaustive model comparison (3DZD method). Table 2 summarizes average correlation coefficients of the top 1, top 3 and top 5 models. The average c.c. for retrieved models up to the top 200 are plotted in Figure 4a for each method. As revealed in the Figure, the first model retrieved using 3DZM, 3DZD and Kihara's methods is more similar to query than the other methods, although the other methods also give the reasonable results. As more models with lower scores (bad matching) are included, the average correlation coefficients of different methods have larger separations.

***Table 2:** The average correlation coefficients between queries and their top N models.*

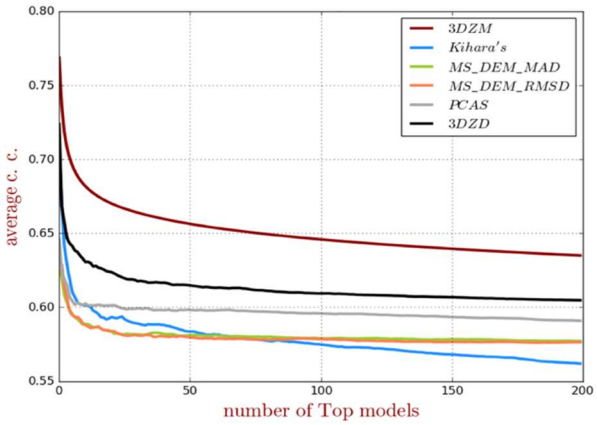
	Top 1	Top 3	Top 5
3DZM*	0.769	0.718	0.701
Kihara	0.715	0.641	0.619
MS-DEM-MAD	0.625	0.608	0.598
MS-DEM-RMSD	0.629	0.614	0.597
PCAS	0.646	0.619	0.607
3DZD	0.723	0.657	0.643

** The correlation coefficients are computed using exhaustive matching with 3DZM representations.*

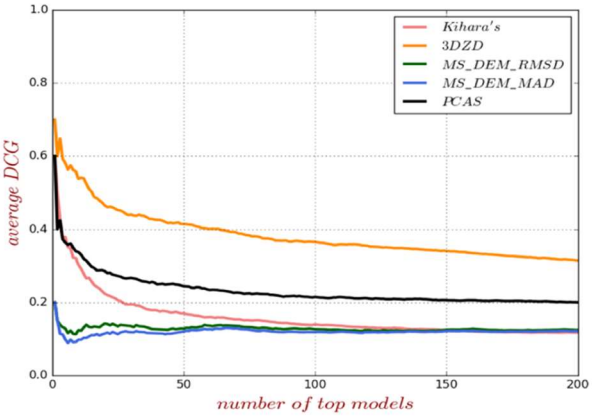
5.3 Evaluation curves

Figure 4 shows the evaluation curve of the retrieval result. The average DCG shows the consistency between the proposed methods and the groundtruth (here, defined using 3DZM ranking). Figure 4(a) shows the average correlation coefficients for all queries. Figure 4(b) shows that the Kihara, 3DZD and PCAS methods have good agreement with the ground truth. The MS_DEM methods are less consistent with the other three methods (as well as the ground truth). All methods are very consistent for the two queries that have exact models in the target set (q_11 and q_14), as shown in Figure 4(c). The PR curves in Figure4(d) indicate that the

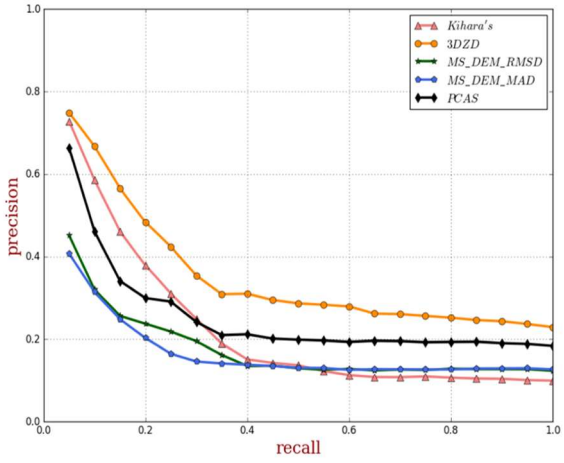
Kihara, 3DZD and PCAS methods have better performance in retrieving a few good match models. When more models are retrieved, the Kihara's method becomes similar to the MS-DEM approaches. It could be due to the fact that trace comparison starts play important roles in giving higher ranks for these models with lower similarity (as measured using 3DZM and 3DZD).



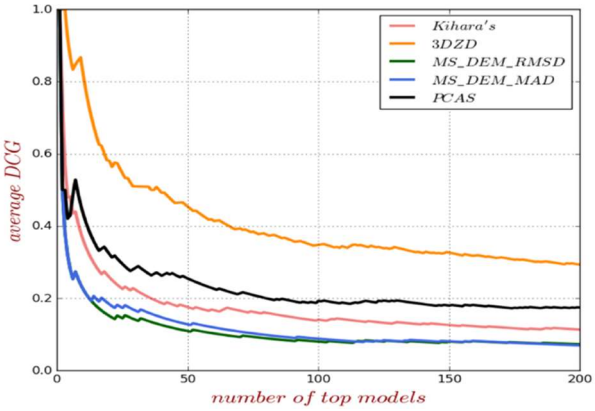
(a)



(b)



(c)



(d)

Figure 4: *Evaluation Curve. (a): The average correlation coefficients for all queries. The plots indicate that the top 10 models can be considered to be correct for most methods, indicated by the sharp transitions to the plateaus. (b): The average DCG for all queries. The ranking results from 3DZM method was used as the 'ground truth' for DCG calculation. (c): The average DCG for queries q_11 and q_14. The exact shape models for these two queries are present in the target set. (d): Precision-Recall Curve for all the queries.*

5.4 Statistics of evaluation parameters

Table 3: *Statistics of evaluation parameters*

Method	NN	Tier**1**	Tier**2**	MAP	E
Kihara	65.29	18.40	9.20	20.34	20.3
MS-DEM-MAD	31.40	24.35	12.15	15.16	26.21
MS-DEM-RMSD	34.92	23.92	11.95	16.02	25.85
PCAS	68.41	35.00	17.50	24.94	29.19
3DZD	74.44	26.70	22.20	34.44	38.30
3DZM	--	--	--	--	--

Among the five methods (besides the 3DZM), the 3DZD and the PCAS methods have the best performance in retrieving the top 200 models ranked using 3DZM (Table 3). The method from Kihara has the highest average consistency with the 3DZM method in the top 5 models.

The 3DZM method retrieves protein molecules based on the shape similarity. Especially for structures at low resolutions when the amino acid sequences are hardly to map to

the shapes, the performance of this 3DZM method remains accurate.

5.5 Computational demands

Table 4: *Execution time and storage requirement.*

Method	
3DZM	**Execution time:** Zernike moment computing for each model: 2 seconds Model search through 5,484 shapes: 700 seconds/query **Storage / Ram usage:** The storeage for one descriptor is about 0.1MB. **Remarks:** Mac OS with 3.5GHz Intel Core i5 processor and 16GB 1600MHz DDR3 memory
Kihara	**Execution time:** Feature extraction: 1.255 seconds/model**.** Search time using 3DZD: 5 seconds/query.MM-align for one query against all models: 12.5 hours. **Remarks:** Intel(R) Core(TM) i7-3820 using single CPU
DEM	**Execution time:** Total runtime for descriptor extraction: 6.1222 seconds/model Comparing one query against the entire protein pool takes in average 2.3502 seconds for MAD and 3.3518 seconds for RMSD. **Storage / Ram usage:** The average memory usage for storing the MS-DEM descriptor is 1.058 Kb. **Remarks:** 64b Linux machine equipped with 32Gb of RAM memory and an Intel Xeon running at 2.3 GHz.
PCAS	**Execution time:** Time required for (Feature extraction of 5,494 molecule and ranking of a test model with respect to 5484 models) = 16.84 seconds. **Storage / Ram usage:** 340 bytes/model. **Remarks:** RAM:8 GB
3DZD	**Execution time:** Zernike moment computing for each model: 2 seconds Model search through 5,484 shapes: **5 seconds/query** **Remarks:** Mac OS with 3.5GHz Intel Core

The computational speed and feature storage are summarized in Table 4. Although there are variations in the hardware used in different teams, the PCAS outperforms all other methods in terms of speed and feature storage space. The model alignment in Kihara's method takes significant amount of time. The DEM approach is also fast in model retrieval, but takes relative longer time for feature extraction.

6. Discussion

Unlike the 3D objects we encountered in daily life, protein or other biomolecule did not get enough attention in 3D Shape Retrieval Contest. Considering the different aspect of the function and the structure of the proteins, there will be different standpoint of the similarity measure metric for proteins. The conformation of the proteins would also change in different environment conditions. So that the definition of the similarity of proteins and the standard of groundtruth dataset is more complicated than the 3D objects. There is no widely used standard benchmark to evaluate the model retrieval performance yet. This needs to be changed to catch up with the rapid accumulation of 3D molecular structures. The development of structural determination methods using X-ray free electron lasers [SWC12] and Cryogenic electron microscopy [Che15] are rapidly expanding the number of protein structures, the shape-based protein structure research start to draw more attentions from various community. There are over 120,000 structures deposited in the Protein Data Bank, waiting for detailed analysis. Inspired by the beautiful structures of biological molecules, this bioshape track in the SHREC 2017, is aimed to encourage teams with different background to participate.

In this contest for bio-inspired shapes, we consider the geometric similarity more than biology relevance, and generated a subset of available structure as target set for model retrieval. Furthermore, the 6 methods all give the reasonable results, especially in the top 5 retrieval models, although from different standpoint of the similarity of proteins. Since this is the beginning of the protein retrieval task in 3D Shape Retrieval Context, the main consideration is the geometric similarity of the proteins. In the next context, more biology information would be considered into the retrieval conditions. This has been partially considered in the implementation of Kihara's method. The computational speed is also an important factor, considering that over 120,000 models might needs to be compared against each query. PCAS implementation can be fine tuned to further speed up the retrieval, which can serve as initial screening.

On the other side the successful criterion of retrieval and classification is not unique, although we stress that the shape similarity is the measure. Different retrieval method optimizes different aspects of the model similarity. It can be seen in the Appendix that the methods have varying performances for different queries. There are still spaces for improvement for all methods.

The lack of benchmark for bioshape retrieval is a problem. During the contest of SHREC 2017, we found one dataset randomly selected from FSSP database[HOS*92], which classifies 2,631 proteins into 27 classes using DALI algorithm. This database can be used for further evaluation for protein retrieval methods. We compared the 3DZM method against this dataset, and the results are very consistent with the FSSP classifications (the classification accuracy is 99.62% using the 3DZM method with $n_{max} = 20$).

The other more challenging task can be the datasets with flexible molecules that have multiple structures. The deformable shapes in molecular world are quite common and play critical roles in their functions. We plan to include such data in next SHREC.

Acknowledgement. The organizer would like to thank the SHREC'17 team for the help with track organizing. H. Liu acknowledges the funding from NSFC (award number: 11575021 and U1530401). CWC, XH, GK, DK are partially supported by National Science Foundation of the USA (IIS1319551). Daniela Craciun, Guillaume Levieux and Matthieu Montes (CNAM, Paris, France) research work is supported by the ERC ViDOCK Grant no. #640283 from the European Research Council Executive Agency.

References

[AHTK99] Angenent S., Haker S., Tannenbaum A., Kikinis R.: On the Laplace-Beltrami operator and brain surface flattening. In: IEEE Transactions on Medical Imaging Bd. 18 (1999), Nr. 8, S. 700–711

[AKKS99] ANKERST M., KASTENMÜLLER G., KRIEGEL H. P., SEIDL T.: *3D Shape Histograms for Similarity Search and Classification in Spatial Databases* : Springer Berlin Heidelberg, 1999

[BKW*77] Bernstein F. C. , Koetzle T. F. , Williams G. J., Meyer J. E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., Tasumi M.: The Protein Data Bank: a computer-based archival file for macromolecular structures. In: J Mol Biol Bd. 112 (1977), Nr. 3, S. 535–542

[Can99] Canterakis N.: 3{D} Zernike Moments and Zernike Affine Invariants for 3{D} Image Analysis and Recognition. In: Scandinavian Conference on Image Analysis, 1999

[Che15] Cheng Y.-F.: Single-Particle Cryo-EM at Crystallographic Resolution. In: Cell Bd. 161, Elsevier (2015), Nr. 3, S. 450–457

[GDZ*15] Goodsell D. S., Dutta S., Zardecki C., Voigt M., Berman H. M., Burley S. K.: The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology. In: PLOS Biology Bd. 13, Public Library of Science (2015), Nr. 5, S. e1002140

[GGS03] Gotsman C., Gu X.-F, Sheffer A.: Fundamentals of spherical parameterization for 3D meshes. In: Acm Transactions on Graphics Bd. 22 (2003), Nr. 3, S. 358–363

[HKRS08] HOLM L., KÄÄRIÄINEN S., ROSENSTRÖM P., SCHENKEL A.: Searching protein structure databases with DaliLite v.3. In: *Bioinformatics* Bd. 24 (2008), Nr. 23, S. 2780

[HOS*92] Holm L., Ouzounis C., Sander C., Tuparev G., Vriend G.: A database of protein structure families with common folding motifs. In: Protein Science Bd. 1, Cold Spring Harbor Laboratory Press (1992), Nr. 12, S. 1691–1698

[HMH*09] Hura G. L., Menon A. L., Hammel M., Rambo R. P., Poole F. L., Tsutakawa S. E., Jenney F. E., Classen S., Frankel K. A., U. a.: Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). In: Nature methods Bd. 6, Nature Publishing Group (2009), Nr. 8, S. 606–12

[LHZ12] Liu H.-G., Hexemer A., Zwart P. H.: The Small Angle Scattering ToolBox (SASTBX): An open source software for biomolecular small angle scattering. In: Journal of Applied Crystallography Bd. 45 (2012), Nr. 3, S. 587--593

[LPJZ12] Liu, H.-G., Poon B. K., Janssen, A. J. E. M. ; Zwart P. H.: Computation of fluctuation scattering profiles via three-dimensional Zernike polynomials. In: Acta Crystallographica Section A: Foundations of Crystallography Bd. 68, International Union of Crystallography (2012), Nr. 5, S. 561–567

[NK03] Novotni M., Klein R.: 3D zernike descriptors for content based shape retrieval. In: Proceedings of the eighth ACM symposium on Solid modeling and applications, SM '03. New York, NY, USA : ACM, 2003 — ISBN 1-58113-706-0, S. 216–225

[SdVS*16] Schindler C. E. M., de Vries S. J., Sasse A., Zacharias M.: SAXS Data Alone can Generate High-Quality Models of Protein-Protein Complexes. In: Structure Bd. 24 (2016), Nr. 8, S. 1387–1397

[VH07] Remco C. Veltkamp, Frank B. ter Haar (eds.). SHREC2007: 3D Shape Retrieval Contest. (2007) Technical Report UU-CS-2007-015.

[SWC12] Spence J. C. H., Weierstall U., Chapman H. N.: X-ray lasers for structural and dynamic biology. In: Reports on Progress in Physics Bd. 75 (2012), Nr. 10, S. 102601

[TN06] Trapani S., Navaza J.: Calculation of spherical harmonics and Wigner d functions by FFT. Applications to fast rotational matching in molecular replacement and implementation into AMoRe. In: Acta Crystallographica Section A Bd. 62, Blackwell Publishing Ltd (2006), Nr. 4, S. 262–269

[XZ09] Xu D., Zhang Y.: Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform. In: PLOS ONE Bd. 4, Public Library of Science (2009), Nr. 12, S. e8140