Eurographics Workshop on 3D Object Retrieval (2018)
A. Telea, T. Theoharis, and R. C. Veltkamp (Editors)

SHREC 2018 – Protein Shape Retrieval

Florent Langenfeld[1,*], Apostolos Axenopoulos[2], Anargyros Chatzitofis[2], Daniela Craciun[1], Petros Daras[2], Bowen Du[3], Andrea Giachetti[4], Yu-kun Lai[5], Haisheng Li[3], Yingbin Li[3], Majid Masoumi[6], Yuxu Peng[7,5], Paul L. Rosin[5], Jeremy Sirugue[1], Li Sun[3], Spyridon Thermos[2], Matthew Toews[6], Yang Wei[3], Yujuan Wu[3], Yujia Zhai[3], Tianyu Zhao[3], Yanping Zheng[3], Matthieu Montes[1,*,@]

[1] Laboratoire GBA, EA4627, Conservatoire National des Arts-et-Métiers, 2 rue Conté, 75003 Paris, France

[2]Information Technologies Institute, Centre for Research and Technology Hellas, Greece

[3]School of Computer Science and Information Engineering, Beijing Technology and Business University, No. 33 Fucheng Road, Haidian District, Beijing 100048, P.R. China

[4]Department of Computer Science, Università di Verona, Strada le Grazie 15, 37134 Verona

[5]School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 3AA, UK

[6]École de Téchnologie Supérieure, University of Québec, Montréal, QC, Canada

[7]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan Province, China

[*]Track organizers

[@]Corresponding author: Matthieu Montes, matthieu.montes@cnam.fr

Abstract

Proteins are macromolecules central to biological processes that display a dynamic and complex surface. They display multiple conformations differing by local (residue side-chain) or global (loop or domain) structural changes which can impact drastically their global and local shape. Since the structure of proteins is linked to their function and the disruption of their interactions can lead to a disease state, it is of major importance to characterize their shape. In the present work, we report the performance in enrichment of six shape-retrieval methods (3D-FusionNet, GSGW, HAPT, DEM, SIWKS and WKS) on a 2 267 protein structures dataset generated for this protein shape retrieval track of SHREC'18.

1. Introduction

The goals of structural biology include developing a comprehensive understanding of the molecular shapes and forms embraced by biological macromolecules and extending this knowledge to understand how different molecular architectures are used to perform most biological processes. Among these macromolecules, proteins are critical effectors involved in most processes and display a dynamic and complex surface. They can be composed of hundreds of thousands atoms and display multiple conformations differing by local (residue side-chain) or global (loop or domain) structural changes at the atomic scale which can drastically impact their global and local shape. Since the structure of proteins is linked to their function and the disruption of their interactions can lead to a disease state, it is of major importance to characterize their shape as it will allow the identification of potential binders such as other proteins, drugs or nucleic acids.

Since most shape-retrieval methods are not dedicated to protein shape comparison, we generated two version of the dataset for the participants: original Protein Data Bank files [BWF*00] (which describe the atomic coordinates of a protein) and the mesh of the Solvent Excluded Surface (SES) of the protein [Con83] in OFF (Object File Format) format. All data was extracted from high resolution structures to stay as close as possible to a real-life case study.

The dataset included identical, structurally similar and structurally different proteins. The dataset is composed of 2 267 unique structures distributed into 107 classes. The participants were asked to compare the 2 267 structures for their surface dissimilarity. The number of classes was not provided to best match a real-world blind study. Six groups using six different methods returned their results that are reported in the present work.

2. Data Set

To reflect the ability of the methods to retrieve the different surfaces representing the same protein domain, we relied on the reference database of protein structures, the Protein Data Bank (PDB) [BWF*00], and on the Structural Classification of Proteins - extended (SCOPe) database [FBC14, CFB17] to build relations between distinct PDB entries.

The Protein Data Bank is the world-wide repository for experimental biological macro-molecules. In February 2018, it comprised 137 917 entries, describing 42 193 distinct protein sequences. Version 2.06 of the SCOPe database contained 77 439 PDB entries distributed over 244 326 domains, the lowest-level of the SCOPe classification tree. Highest levels (class and fold) are discriminated according to structure/shape while lowest levels (superfamily, family, protein and species) are built on evolutionary concerns. We defined the dataset classes as domains with the same parent at the species level of the SCOPe database, ensuring that domains from

DOI: 10.2312/3dor.20181053

the same class were identical. Thus, intra-class relations were established if and only if two SCOPe domains displayed the same species parent, while all other relations were considered as extra-class. Below, is an extensive description of our protocol to build up the dataset.

First, the SCOPe database tree was built. Consequently, the same domains found in different PDB entries were gathered into the same leaves of the tree, allowing the selection of PDB entries while keeping the intra-class information. Since 244 326 domains were implemented in the SCOPe database, we applied the following filters to restrict the size of the dataset to a manageable order of magnitude for the participants.

1. To reflect the experimentally observed variability of protein conformations, we selected only Nuclear Magnetic Resonance (NMR) structures [Wüt01] that usually contain several conformations of the same protein.
2. The dataset was limited to protein domains with no more than 200 residues.
3. "Artifacts", "Low resolution protein structures" and "automated matches" branches of the SCOPe tree were not retained.
4. Structures in complex with small molecules or displaying modified residues were not retained.
5. Highly homologous domains from 7 PDB structures, namely 1ed7, 1f40, 1j6y, 1qnz, 2gri, 2kn5 and 2rr9 [IOH*00, SIC*00, LWW*02, TZL*00, SJA*07, FLG*09, SJI*11] were added.
6. We separated individual domains of multi-domain structures, individual chains from multi-chain structures, and individual conformers.

In total, from the 79 PDB structures describing 88 domains, we retained 2 267 individual structures separated in 107 classes. 18 out of the 107 classes were populated by only one conformer while the biggest class displayed 110 conformers. The average class size was 21.18.

All PDB files generated were further cleaned and prepared using the pdb4amber routine of AmberTools [CCC*17]: water molecules were removed while missing atoms, if any, were added. The resulting structures were submitted to participants in PDB format.

The EDTSurf program [XZ09] was used to generate the Solvent Excluded Surface [Con83] of each structure. Standard parameters were used. The inner surface was not computed. An in-house script was used to convert PLY files to OFF files. These 2 267 OFF files were submitted to participants.

3. Evaluation

Normalized Discounted Cumulative Gain

The Discounted Cumulative Gain (DCG) is a weighted statistics assuming that correct results associated with a higher rank should imply a gain in the performance rating as users are more likely to consider these results. For a list R of correct results, a list G is generated, where G_i is 1 if element R_i is in the correct class (the ground truth class associated with element i GT_i), or 0 otherwise.

The Discounted Cumulative Gain is then computed using the following:

$$DCG_i = \begin{cases} G_1, & \text{if } i = 1 \\ DCG_{i-1} + \frac{G_i}{log_2(i)}, & \text{otherwise} \end{cases}$$

This value is then divided by the maximal value possible (*i.e.* the value obtained by the ground truth) as follows:

$$DCG = \frac{DCG_k}{1 + \sum_{j=2}^{|C|} \frac{1}{log_2(j)}}$$

where k is the number of objects in the dataset and C the size of the classes. This value is a good summary for a comparative evaluation of the performance of different methods performance. A normalized value *nDCG* of the *DCG* is therefore computed over all methods, and compared to the average value *aveDCG*:

$$nDCG_{algo} = \frac{DCG_{algo}}{aveDCG} - 1$$

where a negative value indicated that the performance of the method is under the average while a positive value indicated that the performance of the method is over the average. The norm of the value indicates the gap to the average performance.

Nearest Neighbor, First-tier and Second-tier

These parameters check the ratio of models that belong to the same class as the query. For Nearest Neighbor, the first match only is considered, while the $|C| - 1$ and $2 * (|C| - 1)$ first matches are considered for First-tier and Second-tier parameters.

Precision-Recall plot and E-measure

Precision P represents the ratio of models from class C retrieved within all objects attributed to class C, while Recall R represents the ratio of models from class C retrieved compared to $|C|$.

The $E - measure$ is a composite parameter of both Precision and Recall:

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

All analyses were done using the Princeton Shape Benchmark utilities [SMKF04].

4. Participants & Methods

4.1. 3D convolutional framework for protein shape retrieval (3D-FusionNet), by S. Thermos, A. Chatzitofis, A. Axenopoulos and P. Daras

Problem Definition

The idea behind the proposed framework is to combine state-of-the-art hand-crafted descriptors that effectively represent the 3D molecular shape with the features extracted using a deep Neural Network (NN). The NN has been trained on a different dataset of flexible molecules, the MOLMOVDB [EMG03]. The input 3D model is the Solvent Excluded Surface (SES) of a protein molecule,

which has been created from the molecule's tertiary structure (PDB format) using the EDTSurf software. This software produces a high-resolution watertight triangulated mesh. The triangulated mesh is simplified and used as input to the algorithm that extracts the hand-crafted features, while for the deep NN architecture a $32 \times 32 \times 32$ voxel model is created.

A Shape Descriptor Based on Diffusion Distances

Extraction of hand-crafted features is based on the combined DDMR shape descriptor, which has been introduced in [ARP*16], and it is invariant to protein conformations. At a pre-processing stage, the high-resolution mesh is simplified resulting in a set of N_S uniformly sampled points that provide a coarse representation of the 3D molecule. At the descriptor extraction step, the Modal Representation of the Diffusion-Distance Matrix (DDMR descriptor) is extracted. DDMR is a global shape descriptor, which is produced by applying Singular Value Decomposition on the Diffusion Distance Matrix of all N_S oriented points, keeping the first n singular values ($n = 40$ in our experiments). The diffusion distance between two points on a surface is considered as an average length of paths connecting the points in a sense of inner distances and it is able to capture topological changes in molecular shapes [LLZ*10].

Volumetric Binary Grid

Based on the approach of Nooruddin and Turk [NT03], we rasterize the protein 3D model to a binary voxel grid. The 3D models of the proteins are watertight, thus the parity count method is applied for binary voxelization. To this end, a voxel v is classified by counting the number of times that a line crossing the center of the voxel intersects polygons of the 3D model surface. Ray-casting the 3D model with parallel rays, all of the voxels along the ray are classified. For an odd number of intersections, voxel v is considered interior to the model, while for an even number, outside. For a $N \times N \times N$ voxel grid resolution, where $N = 32$, we cast $N \times N = 1024$ rays, with each ray passing through N-voxel centers.

Fusion Architecture

The proposed architecture, depicted in Fig. 1, consists of a convolutional neural network (CNN), utilized for 3D shape representation learning, followed by a Multi-Layer Perceptron (MLP), which fuses the CNN-extracted features with the hand-crafted descriptors presented in Section "A Shape Descriptor Based on Diffusion Distances". In detail, the VoxNet CNN [MS15] is used, which consists of 2 volumetric convolutional layers, 1 max pooling layer and 3 fully connected layers, and efficiently encodes the spatial structures such as planes and corners at different scales and orientations. The VoxNet processes the voxel inputs and the features after its last fully connected layer are concatenated with the corresponding hand-crafted ones. The latter are processed by the MLP followed by a Softmax layer used for classification.

For the protein shape retrieval, a transfer learning approach is adopted. At first, the fusion architecture is trained on the MOLMOVDB dataset [EMG03] which consists of over 200 classes of proteins. Subsequently, the Softmax layer is dropped and the architecture is used for feature extraction. For each previously unseen input, a feature vector is extracted. After the completion of the feature extraction, the Euclidean distance metric is used to measure the distance between the evaluated input models. Small distance values indicate that the corresponding feature vectors represent the same protein class.

4.2. Global Spectral Graph Wavelet framework (GSGW), by M. Masoumi and M. Toews

Our global spectral graph wavelet (GSGW) framework [MRH18] is based on the eigensystem of the Laplace-Beltrami operator that are invariant to isometric transformations. GSGW is a multi-resolution descriptor that incorporates the vertex area into the definition of spectral graph wavelet [MLH16, MH17] in a bid to capture more geometric information and, hence, further improve its discriminative ability. GSGW also provides a general and flexible interpretation for the analysis and design of spectral descriptors. For a vertex j of a triangle mesh, spectral graph wavelet is defined as [MLH16]:

$$\mathbf{s}_L(j) = \{W_{\delta_j}(t_k, j) \mid k = 1, \ldots, L\} \cup \{S_{\delta_j}(j)\}, \tag{1}$$

where $W_{\delta_j}(t_k, j)$ and $S_{\delta_j}(j)$ are the spectral graph wavelet and scaling function coefficients at resolution level L, respectively. We then represent a shape $\mathbb{M}$ by a p-dimensional vector

$$\mathbf{x} = \mathbf{S}\mathbf{a} = \sum_{i=1}^{m} a_i \mathbf{s}_i, \tag{2}$$

where $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_m)$ is a $p \times m$ matrix of local spectral graph wavelet signatures and $\mathbf{a} = (a_1, \ldots, a_m)^\mathsf{T}$ is an m-dimensional vector of mesh vertex areas (i.e. each element a_i is the area of the Voronoi cell at mesh vertex i).

We refer to the p-dimensional vector $\mathbf{x}$ as the global spectral graph wavelet (GSGW) descriptor of the protein surface. The GSGW descriptor enjoys a number of desirable properties including simplicity, compactness, invariance to isometric deformations, and computational feasibility. Moreover, GSGW combines the advantages of both band-pass and low-pass filters.

Our proposed protein shape retrieval algorithm consists of four main steps. In the first step, we represent each protein in the dataset by a spectral graph wavelet signature matrix, which is a feature matrix consisting of local descriptors. More specifically, let $\mathcal{D}$ be a dataset of n proteins modeled by triangle meshes $\mathbb{M}_1, \ldots, \mathbb{M}_n$. We represent each surface $\mathbb{M}_i$ in the dataset $\mathcal{D}$ by a $p \times m$ spectral graph wavelet signature matrix $\mathbf{S}_i$, whose columns are p-dimensional local signatures and m is the number of mesh vertices. In the second step, we compute the p-dimensional global spectral graph wavelet descriptor $\mathbf{x}_i = \mathbf{S}_i\mathbf{a}_i$ of each protein $\mathbb{M}_i$, for $i = 1, \ldots, n$. Subsequently, the feature vectors $\mathbf{x}_i$ of all n shapes in the dataset are arranged into a $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\mathsf{T}$. In the third step, we calculate volume v and surface area a of each 3D model $\mathbb{M}_i$ and then aggregate them to $\mathbf{X}$ to provide further discrimination power for GSGW. Finally, we compare a query $\mathbf{x}$ to all data points in $\mathbf{X}$ using ℓ_1-distance to find the most relevant shapes to the query. The lower the value of this distance is, the more similar the shapes are.

The experiments were conducted on a laptop with an Intel Core i7 processor running at 2.00 GHz and 16 GB RAM; all the algorithms were implemented in MATLAB. In our setup, a total of 31

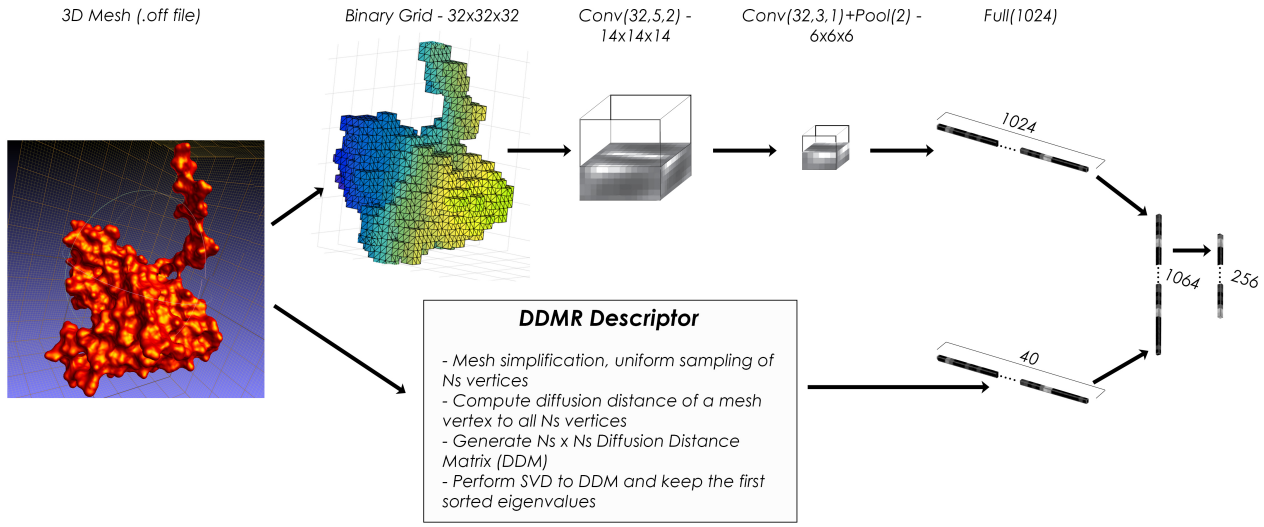

Figure 1: *The proposed fusion architecture that consists of a VoxNet [MS15] CNN (top information stream), followed by a MLP model (right-most). The latter fuses the VoxNet-extracted and the hand-crafted features (bottom information stream), respectively.*

eigenvalues and associated eigenfunctions of the LBO were computed. For the proposed approach, we set the resolution parameter to $R = 30$ (i.e. the spectral graph wavelet signature matrix is of size $495 \times m$, where m is the number of mesh vertices).

4.3. Histograms of Area Projection Transform (HAPT), by A. Giachetti

In our runs, we characterized the protein shapes with the Histograms of Area Projection Transform (HAPT) [GL12]. The method, usually well suited for nonrigid shape retrieval, is based on a spatial map (Multiscale Area Projection Transform) that encodes the likelihood of the 3D points inside the shape of being centres of spherical symmetry. This map is obtained by computing, for each radius of interest, the value:

$$APT(\vec{x},S,R,\sigma) = Area(T_R^{-1}(k_\sigma(\vec{x}) \subset T_R(S,\vec{n}))) \qquad (3)$$

where S is the surface of the object, $T_R(S,\vec{n})$ is the parallel surface of S shifted along the normal vector $\vec{n}$ (only in the inner direction) and $k_\sigma(\vec{x})$ is a sphere of radius σ centred in the generic 3D point $\vec{x}$ where the map is computed. Values at different radii are normalized in order to have a scale-invariant behaviour, creating the Multiscale APT (MAPT):

$$MAPT(\vec{x},R,S) = \alpha(R)\, APT(\vec{x},S,R,\sigma(R)) \qquad (4)$$

where $\alpha(R) = 1/4\pi R^2$ and $\sigma(R) = c \cdot R \quad (0 < c < 1)$.

A discretized MAPT is easily computed, for selected values of R, on a voxelized grid including the surface mesh, with the procedure described in [GL12]. The map is computed in a grid of voxels with side s on a set of corresponding sampled radius values $R_1, \ldots, R_n$.

For the proposed task, discrete MAPT maps were quantized in 8 bins and histograms computed at the selected scales (radii) were concatenated creating an unique descriptor. Voxel side and sampled radii were fixed set for each run and chosen to represent the approximate radii of the spherical symmetries visible in the models. We tested two different options, in the first (runs 1 and 2) we put $s = 1$ and we computed the MAPT histograms for 8 increasing radii starting from $R_1 = 1$ iteratively adding a fixed step of 1 for the remaining values $\{R_2, \ldots, R_8\}$. For the second (runs 3 and 4), we put $s = 0.5$ and we computed the MAPT histograms for 8 increasing radii starting from $R_1 = 0.5$ iteratively adding a fixed step of 0.5 for the remaining values $\{R_2, \ldots, R_8\}$.

The procedure for model comparison then simply consists in concatenating the MAPT histograms computed at the different scales and measuring distances between shapes by evaluating the Euclidean distance (runs 1 and 3) and the Jeffrey divergence (runs 2 and 4) of the corresponding concatenated vectors.

The estimation of the descriptors took on average 1.4 sec per model with the first discretization option, 2.4 with the second on a laptop with an Intel® Core™ i7-4720HQ CPU running Ubuntu Linux 16.04. The descriptor comparison time was negligible.

4.4. Protein Shape Retrieval driven by Digital Elevation Models (DEM), by D. Craciun, J. Sirugue and M. Montes

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 supplied *via* global distance measures computed over the DEMs.

Representing Macromolecular Shapes as Digital Elevation Models

The shape representation algorithm applies the EDTSurf [XZ09] technique to generate the macromolecular surface (MS) from the input data. The descriptor computation stage starts by applying the mesh flattening procedure used to map the mesh onto the unit

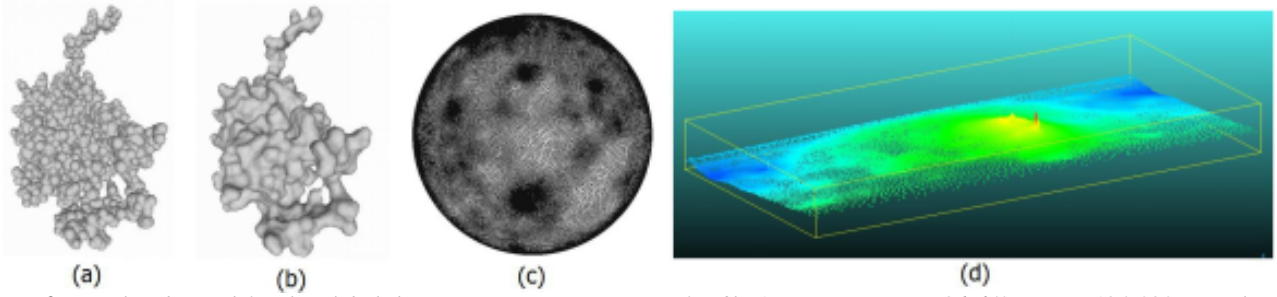

Figure 2: *Results obtained for the global descriptor computation stage for file 1: (a) PDB input: 96 642 points, 184 080 triangles, (b) macromolecular mesh generated by the EDTSurf method [XZ09]: 95 505 points, 191 022 triangles, (c) spherical mapping output: 95 505 points, 191 022 triangles; (d) MS-DEM descriptor output: 95 505 points.*

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 the file 1 belonging to the protein pool of the SHREC 2018 track.

Global Comparison of MS-DEMs

The present research work evaluates the Mean Absolute Differences (MAD) which is measured over the points belonging to the 2D grids. The MAD distance is computed over the minimum number of points belonging to the overlapping area computed between the query and the target meshes.

Runtime Evaluation

The MS-DEM descriptor computation for file 1 (shown in Figure 2) is performed in 8 seconds on a 64b Linux machine, equipped with 32Gb of RAM memory and an Intel Xeon @ 3.40 GHz. The mean runtime for MS-DEM Comparison is $7.1396 \cdot 10^{-4}$sec corresponding to a mean number of compared points in the overlapping area of $7.9795 \cdot 10^{4}$ points.

Scale-invariant wave kernel signature (SIWKS), by Y. Wu, Y. Zheng, L.i Sun, Y. Li, T. Zhao, B. Du, Y. Zhai, Y. Wei and H. Li

In this track, we propose a new feature for 3D shape retrieval called scale-invariant wave kernel signature(SIWKS). The process can be described as follows. Firstly, the WKS represents the average probability of measuring a quantum mechanical particle at a specific location. By letting vary the energy of the particle, the WKS encodes and separates information from various different Laplace eigenfrequencies. Based on the Schrödinger equation each point on an object's surface is associated with a Wave Kernel Signature. Then, we found that WKS is the sensitivity to scale transformation. We bring the spirit of eigenvalue normalization based methods to construct a scale invariant wave kernel signature. Finally, the scale factor in WKS is removed.

$$WKS \begin{cases} WKS(x,\cdot) : R \to R \\ WKS(x,e) = C_e \sum_i \phi_i^2(x) e^{\frac{-(e_i - log\lambda_i)^2}{2\sigma^2}} \end{cases} \quad (5)$$

$$SIWKS \begin{cases} SIWKS(x,\cdot) : R \to R \\ SIWKS(x,e) = C_e \sum_i \frac{\phi_i^2(x)}{\lambda_i} e^{\frac{-(e_i - log\lambda_i)^2}{2\sigma^2}} \end{cases} \quad (6)$$

where

$$C_e = \left(\sum_i e^{\frac{-(e_i - log\lambda_i)^2}{2\sigma^2}} \right)^{-1}$$

in equations (5) and (6).

4.5. Wave Kernel Signature (WKS), by Y. Peng, Y. Lai and P. L. Rosin

In recent years, significant attention has been devoted to descriptors obtained from the spectral decomposition of the Laplace-Beltrami operator associated with the shape. Notable examples in this family are the Heat Kernel Signature (HKS) and the recently introduced Wave Kernel Signature (WKS) [ASC11]; the latter is described in the previous section. They are computationally efficient, isometry-invariant by construction, and can gracefully cope with a variety of transformations.

The eigenvalues and the eigenvectors are obtained from protein mesh files in our experiment. We use the MeshLP package to compute the eigenvalues and the eigenvectors of the Laplace operator on the mesh. The time axis is sampled logarithmically. The code is modified from `https://github.com/areslp/matlab/tree/master/HKS/` and uses default values.

WKS is computed from the eigenvalues and the eigenvectors. We use the code from `https://github.com/ChunyuanLI/spectral_descriptors` [LB13]. We set the number of features to 100 and the variance is 100×5.

The vocabulary of the WKS feature is created. The size of vocabulary is chosen as 1 000 according to our experiments on the subset

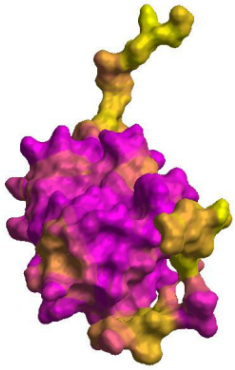

Figure 3: *Protein mesh coloured according to the WKS feature*

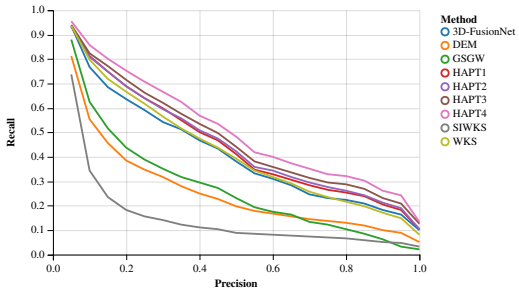

Figure 4: *Precision-Recall curves. Each curve corresponds to a dissimilarity matrix provided by a participant to the track*

of the FSSP database [ARP*16, HS96]. The size of the feature vector at each vertex on the mesh is 50 and normalized. Then we randomly select 10% of the mesh points and apply Ovsjanikov's improved k-means algorithm (`http://www.lix.polytechnique.fr/~maks/code/shapegoogle_code.zip`) [BBGO11] to generate the vocabulary.

We compute the Bag of Features (BoFs) for each protein using hard Vector Quantization. The feature is a 1000×1 vector and normalized.

The distance between the BoFs of any two proteins is computed using the L1 distance $\|X - Y\|_1$. For a given query shape, the shapes from the dataset are retrieved based on this distance.

5. Results

Each team submitted one to four 2267×2267 dissimilarity matrices resulting in 9 methods to evaluate. The following section summarizes the performance in retrieval for each method, and insights are given regarding the number of conformers in each class. Three types of classes were defined: small classes contained less than 20 conformers, medium classes contained 20 to 40 conformers, and large classes contained at least 41 conformers.

5.1. Overall results

The results summarized in Table 1 were computed for each dissimilarity matrix (3D-FusionNet, HAPT1, HAPT2, HAPT3, HAPT4, SIWKS, DEM, WKS and GSGW) over all classes and over each type of class (small, medium or large). Overall, the Histograms of Area Projection Transform (HAPT) method displayed the best results, especially for the run 4 which showed the best results for all statistics. Wave Kernel Signature based Shape Descriptor (WKS) and 3D convolutional framework for protein shape retrieval (3D-FusionNet) displayed similar overall results, close to the HAPT runs. Digital Elevation Models (DEM) and Global Spectral Graph Wavelet framework (GSGW) methods followed in performance, and the Scale-invariant Wave Kernel Signature (SIWKS). Similar trends are observed with the Precision-Recall curves (Figure 4).

5.2. The number of conformers impacts the methods' performance

We computed the Nearest Neighbor, First-tier, Second-tier, E-measure, Discounted Cumulative Gain statistics and normalized Discounted Cumulative Gain statistics over the whole dataset for all methods. The values of all statistics for all methods were then computed for the small, medium and large classes (Table 1).

As expected, all methods performed better on large classes (more than 40 conformers in the class) than on medium classes (20-40 conformers) or small classes (less than 20 conformers). The performances for small classes were significantly lower for all methods compared to other classes, except for HAPT methods 2-4 whose First-tier statistics were better for the small classes than for the other classes. Each method displayed distinct Nearest Neighbor, First-tier and Second-tier performances depending on the type of class (small, medium or large). The mean Nearest Neighbor value for small classes was 0.278 (the highest value was 0.418 for HAPT4) and First-tier mean value was 0.392 (only HAPT2-4 runs displayed First-tier values above 0.5). On the contrary, for medium and large classes, Nearest Neighbor parameters were significantly increased compared to small classes, whereas First-tier and Second-tier statistics decreased. This behavior is particularly marked for large classes where Nearest-Neighbor displayed a mean value of 0.706 while First-tier mean value is 0.254, which is lower than the First-tier mean value for small classes (0.392).

Last, depending on the number of conformers in each class, the respective performances of the methods varied. As an example, the normalized Discounted Cumulative gain (nDCG) of HAPT4 for small classes was 24.89% while it was 17.24% and 6.91% for medium and small classes respectively. Conversely, the DEM method display a normalized Discounted Cumulative Gain of -31.51%, -15.93% and -8.20% for small, medium and large classes respectively. Thus, the data set composition could influence the choice of the method: for large classes, 3D-FusionNet, HAPT4 and WKS displayed similar performances, while for small classes, HAPT4 outperformed the other methods.

A comparable analysis was performed based on the size of the proteins (number of atoms in the PDB files) and the size of the meshes (number of vertices in the OFF files). No clear pattern was extracted from this analysis, except for the Nearest Neighbor statistics that is inversely correlated to the size of the system (either ex-

Table 1: *Results summary by method and by size of the classes. Normalized DCG for small, medium and large classes are computed with respect to the average DCG of small, medium and large classes, respectively. NN = Nearest Neighbor, Tier1 = First-tier, Tier2 = Second-Tier, DCG = Discounted Cumulative Gain, nDCG = normalized Discounted Cumulative Gain.*

Method	Class	NN	Tier1	Tier2	E-measure	DCG	nDCG (%)
3D-FusionNet	All	0.689	0.404	0.459	0.366	0.681	3.92
	Small	0.297	0.394	0.255	0.253	0.496	-0.94
	Medium	0.672	0.468	0.524	0.397	0.693	3.52
	Large	0.822	0.287	0.323	0.319	0.697	6.08
HAPT1	All	0.713	0.413	0.534	0.409	0.719	9.72
	Small	0.390	0.456	0.306	0.288	0.573	14.53
	Medium	0.716	0.503	0.630	0.462	0.748	11.71
	Large	0.782	0.280	0.310	0.301	0.676	2.82
HAPT2	All	0.703	0.439	0.541	0.415	0.72	9.87
	Small	0.357	0.522	0.335	0.314	0.592	18.34
	Medium	0.693	0.509	0.632	0.466	0.746	11.44
	Large	0.799	0.283	0.315	0.304	0.678	3.15
HAPT3	All	0.712	0.459	0.56	0.433	0.734	12.00
	Small	0.276	0.549	0.339	0.330	0.590	17.92
	Medium	0.714	0.542	0.663	0.490	0.766	14.38
	Large	0.813	0.278	0.308	0.311	0.685	4.21
HAPT4	All	**0.77**	**0.493**	**0.584**	**0.462**	**0.755**	**15.21**
	Small	**0.418**	**0.613**	**0.358**	**0.373**	**0.625**	**24.89**
	Medium	**0.768**	**0.578**	**0.688**	**0.515**	**0.785**	**17.24**
	Large	**0.854**	**0.300**	0.315	**0.338**	**0.702**	**6.91**
SIWKS	All	0.199	0.109	0.189	0.114	0.452	-31.03
	Small	0.112	0.111	0.067	0.078	0.333	-33.40
	Medium	0.183	0.102	0.190	0.118	0.432	-35.56
	Large	0.257	0.142	0.207	0.123	0.534	-18.73
DEM	All	0.421	0.238	0.319	0.231	0.555	-15.31
	Small	0.088	0.158	0.079	0.113	0.343	-31.51
	Medium	0.428	0.277	0.370	0.262	0.563	-15.93
	Large	0.551	0.196	0.250	0.205	0.603	-8.20
WKS	All	0.717	0.41	0.49	0.377	0.701	6.97
	Small	0.288	0.417	0.281	0.264	0.522	4.24
	Medium	0.718	0.473	0.561	0.416	0.720	7.53
	Large	0.805	0.294	**0.347**	0.313	**0.702**	6.78
GSGW	All	0.514	0.261	0.35	0.247	0.581	-11.34
	Small	0.272	0.306	0.197	0.166	0.430	-14.08
	Medium	0.476	0.281	0.373	0.264	0.574	-14.34
	Large	0.674	0.230	0.302	0.223	0.637	-3.03

pressed as the number of atoms or the number of vertices). The smaller the system is, the better the methods performed in terms of Nearest Neighbor retrieval.

6. Discussion

Proteins are linear chains of amino-acids who fold into a broad variety of 3D shapes. They are intrinsically dynamic objects whose motions, hence their surface and their shape, are directly related to their activity. Many parameters influence the protein dynamics and/or folding: the amino-acids composition of the protein or the presence of a given substance (ion, ligand, nucleic acid or protein) drive the protein conformation to another and therefore may modulate its function. To date, high-resolution experiments have been able to determine the structure of more than 130 000 proteins [BWF*00].

Here, we designed a dataset to evaluate the ability of shape retrieval methods to identify protein shape variations, and more specifically to distinguish between variations due to the protein dynamics and variations due to the protein sequence (the amino-acids composition). The ground truth was designed with this aim, and therefore relied on the protein sequence only. Other considerations (active vs inactive states of proteins for example) may constitute a new viewpoint to propose new datasets for a future track.

We only included proteins of small to medium size (less than 200 residues) to limit the size of the resulting high resolution meshes. Besides, we only selected 79 NMR structures which are generally

structures of small proteins. This is far from representing the protein size diversity contained in the Protein Data Bank. Introducing X-ray crystallographic structures or more recently deposited high resolution Cryo Electronic microscopy structures would be more representative of the diversity of the systems available in the PDB.

As presented in the results section, we observed various performances depending on the size of the classes: all methods displayed a better performance on lesser populated classes within the First-tier while they displayed a better Nearest-Neighbor performance on more populated classes. All methods did not perform uniformly over all the classes in terms of enrichment.

Finally, due to the size of protein structure databases (>130 000 structures in the PDB in 2018, roughly 10 000 to 15 000 new structures each year), it would be of importance to consider the computational time in the methods performances as well, with the aim to screen in a near future a datasets of hundreds of thousands protein structures.

Acknowledgements

D. Craciun, J. Sirugue and M. Montes research work is supported by the ERC ViDOCK Grant no. #640283 from the European Research Council Executive Agency. Y. Peng, Y. Lai and P. L. Rosin research work is supported by China Scholarship Council(CSC,20170080005).

References

[AHTK99] ANGENENT S., HAKER S., TANNENBAUM A., KIKINIS R.: On the Laplace-Beltrami operator and brain surface flattening. *IEEE Transactions on Medical Imaging 18*, 8 (Aug 1999), 700–711. doi:10.1109/42.796283. 5

[ARP*16] AXENOPOULOS A., RAFAILIDIS D., PAPADOPOULOS G., HOUSTIS E. N., DARAS P.: Similarity search of flexible 3D molecules combining local and global shape descriptors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 13*, 5 (September 2016), 954–970. doi:10.1109/TCBB.2015.2498553. 3, 6

[ASC11] AUBRY M., SCHLICKEWEI U., CREMERS D.: The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (Nov 2011), pp. 1626–1633. doi:10.1109/ICCVW.2011.6130444. 5

[BBGO11] BRONSTEIN A. M., BRONSTEIN M. M., GUIBAS L. J., OVSJANIKOV M.: Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Trans. Graph. 30*, 1 (Feb. 2011), 1:1–1:20. URL: http://doi.acm.org/10.1145/1899404.1899405, doi:10.1145/1899404.1899405. 6

[BWF*00] BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T. N., WEISSIG H., SHINDYALOV I. N., BOURNE P. E.: The Protein Data Bank. *Nucleic Acids Research 28*, 1 (2000), 235–242. URL: +http://dx.doi.org/10.1093/nar/28.1.235, arXiv:/oup/backfile/content_public/journal/nar/28/1/10.1093_nar_28.1.235/1/280235.pdf, doi:10.1093/nar/28.1.235. 1, 7

[CCC*17] CASE D., CERUTTI D., CHEATHAM T., III, DARDEN T., DUKE R., GIESE T., GOHLKE H., GOETZ A., GREENE D., HOMEYER N., IZADI S., KOVALENKO A., LEE T., LEGRAND S., LI P., LIN C., LIU J., LUCHKO T., LUO R., MERMELSTEIN D., MERZ K., MONARD G., NGUYEN H., OMELYAN I., ONUFRIEV A., PAN F., QI R., ROE D., ROITBERG A., SAGUI C., SIMMERLING C., BOTELLO-SMITH W., SWAILS J., WALKER R., WANG J., WOLF R., WU X., XIAO L., YORK D., KOLLMAN P.: Amber 2017. *University of California, San Francisco* (2017). 2

[CFB17] CHANDONIA J.-M., FOX N. K., BRENNER S. E.: SCOPe: Manual curation and artifact removal in the Structural Classification Of Proteins – extended database. *Journal of Molecular Biology 429*, 3 (2017), 348 – 355. Computation Resources for Molecular Biology. URL: http://www.sciencedirect.com/science/article/pii/S0022283616305186, doi:https://doi.org/10.1016/j.jmb.2016.11.023. 1

[Con83] CONNOLLY M. L.: Analytical molecular surface calculation. *Journal of Applied Crystallography 16*, 5 (Oct 1983), 548–558. URL: https://doi.org/10.1107/S0021889883010985, doi:10.1107/S0021889883010985. 1, 2

[EMG03] ECHOLS N., MILBURN D., GERSTEIN M.: MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Research 31*, 1 (2003), 478–482. URL: +http://dx.doi.org/10.1093/nar/gkg104, arXiv:/oup/backfile/content_public/journal/nar/31/1/10.1093_nar_gkg104/2/gkg104.pdf, doi:10.1093/nar/gkg104. 2, 3

[FBC14] FOX N. K., BRENNER S. E., CHANDONIA J.-M.: SCOPe: Structural Classification Of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research 42*, D1 (2014), D304–D309. URL: +http://dx.doi.org/10.1093/nar/gkt1240, arXiv:/oup/backfile/content_public/journal/nar/42/d1/10.1093/nar/gkt1240/2/gkt1240.pdf, doi:10.1093/nar/gkt1240. 1

[FLG*09] FRIEDLAND G. D., LAKOMEK N.-A., GRIESINGER C., MEILER J., KORTEMME T.: A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLOS Computational Biology 5*, 5 (05 2009), 1–16. URL: https://doi.org/10.1371/journal.pcbi.1000393, doi:10.1371/journal.pcbi.1000393. 2

[GGS03] GOTSMAN C., GU X., SHEFFER A.: Fundamentals of spherical parameterization for 3D meshes. *ACM Trans. Graph. 22*, 3 (July 2003), 358–363. URL: http://doi.acm.org/10.1145/882262.882276, doi:10.1145/882262.882276. 5

[GL12] GIACHETTI A., LOVATO C.: Radial symmetry detection and shape characterization with the multiscale area projection transform. *Comp.Graph.Forum 31*, 5 (2012), 1669–1678. URL: http://dx.doi.org/10.1111/j.1467-8659.2012.03172.x, doi:10.1111/j.1467-8659.2012.03172.x. 4

[HS96] HOLM L., SANDER C.: The FSSP database: Fold classification based on Structure-Structure alignment of Proteins. *Nucleic Acids Research 24*, 1 (1996), 206–209. URL: +http://dx.doi.org/10.1093/nar/24.1.206, arXiv:/oup/backfile/content_public/journal/nar/24/1/10.1093_nar_24.1.206/2/24-1-206.pdf, doi:10.1093/nar/24.1.206. 6

[IOH*00] IKEGAMI T., OKADA T., HASHIMOTO M., SEINO S., WATANABE T., SHIRAKAWA M.: Solution structure of the chitin-binding domain of Bacillus circulans WL-12 chitinase A1. *Journal of Biological Chemistry 275*, 18 (2000), 13654–13661. URL: http://www.jbc.org/content/275/18/13654.abstract, arXiv:http://www.jbc.org/content/275/18/13654.full.pdf+html, doi:10.1074/jbc.275.18.13654. 2

[LB13] LI C., BEN HAMZA A.: A multiresolution descriptor for deformable 3D shape retrieval. *The Visual Computer 29*, 6 (Jun 2013), 513–524. URL: https://doi.org/10.1007/s00371-013-0815-3, doi:10.1007/s00371-013-0815-3. 5

[LLZ*10] LIU Y.-S., LI Q., ZHENG G.-Q., RAMANI K., BENJAMIN W.: Using diffusion distances for flexible molecular shape comparison. *BMC Bioinformatics 11*, 1 (Sep 2010), 480. URL:

https://doi.org/10.1186/1471-2105-11-480, doi:10.1186/1471-2105-11-480. 3

[LWW*02] LANDRIEU I., WIERUSZESKI J.-M., WINTJENS R., INZÉ D., LIPPENS G.: Solution structure of the single-domain prolyl cis/trans isomerase PIN1At from Arabidopsis thaliana. *Journal of Molecular Biology 320*, 2 (2002), 321 – 332. URL: http://www.sciencedirect.com/science/article/pii/S0022283602004291, doi:https://doi.org/10.1016/S0022-2836(02)00429-1. 2

[MH17] MASOUMI M., HAMZA A. B.: Spectral shape classification: A deep learning approach. *Journal of Visual Communication and Image Representation 43* (2017), 198–211. 3

[MLH16] MASOUMI M., LI C., HAMZA A. B.: A spectral graph wavelet approach for nonrigid 3D shape retrieval. *Pattern Recognition Letters 83* (2016), 339–348. 3

[MRH18] MASOUMI M., REZAEI M., HAMZA A. B.: Global spectral graph wavelet signature for surface analysis of carpal bones. *Physics in medicine and biology 63*, 3 (2018), 1–12. 3

[MS15] MATURANA D., SCHERER S.: VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Sept 2015), pp. 922–928. doi:10.1109/IROS.2015.7353481. 3, 4

[NT03] NOORUDDIN F. S., TURK G.: Simplification and repair of polygonal models using volumetric techniques. *IEEE Transactions on Visualization and Computer Graphics 9*, 2 (April 2003), 191–205. doi:10.1109/TVCG.2003.1196006. 3

[SIC*00] SICH C., IMPROTA S., COWLEY D. J., GUENET C., MERLY J.-P., TEUFEL M., SAUDEK V.: Solution structure of a neurotrophic ligand bound to FKBP12 and its effects on protein dynamics. *European Journal of Biochemistry 267*, 17 (2000), 5342–5355. URL: http://dx.doi.org/10.1046/j.1432-1327.2000.01551.x, doi:10.1046/j.1432-1327.2000.01551.x. 2

[SJA*07] SERRANO P., JOHNSON M. A., ALMEIDA M. S., HORST R., HERRMANN T., JOSEPH J. S., NEUMAN B. W., SUBRAMANIAN V., SAIKATENDU K. S., BUCHMEIER M. J., STEVENS R. C., KUHN P., WÜTHRICH K.: Nuclear Magnetic Resonance structure of the N-terminal domain of nonstructural protein 3 from the severe acute respiratory syndrome coronavirus. *Journal of Virology 81*, 21 (2007), 12049–12060. URL: http://jvi.asm.org/content/81/21/12049.abstract, arXiv:http://jvi.asm.org/content/81/21/12049.full.pdf+html, doi:10.1128/JVI.00969-07. 2

[SJI*11] SEKIYAMA N., JEE J., ISOGAI S., AKAGI K., HUANG T., ARIYOSHI M., TOCHIO H., SHIRAKAWA M.: PDBID: 2RR9, The solution structure of the K63-Ub2:tUIMs complex. 2

[SMKF04] SHILANE P., MIN P., KAZHDAN M., FUNKHOUSER T.: The Princeton Shape Benchmark. In *Proceedings of the Shape Modeling International 2004* (Washington, DC, USA, 2004), SMI '04, IEEE Computer Society, pp. 167–178. URL: https://doi.org/10.1109/SMI.2004.63, doi:10.1109/SMI.2004.63. 2

[TZL*00] TUGARINOV V., ZVI A., LEVY R., HAYEK Y., MATSUSHITA S., ANGLISTER J.: NMR structure of an anti-gp120 antibody complex with a V3 peptide reveals a surface important for co-receptor binding. *Structure 8*, 4 (Apr 2000), 385–395. URL: http://dx.doi.org/10.1016/S0969-2126(00)00119-2, doi:10.1016/S0969-2126(00)00119-2. 2

[Wüt01] WÜTHRICH K.: The way to NMR structures of proteins. *Nature Structural Biology 8* (Nov 2001), 923 EP –. URL: http://dx.doi.org/10.1038/nsb1101-923. 2

[XZ09] XU D., ZHANG Y.: Generating triangulated macromolecular surfaces by Euclidean Distance Transform. *PLOS ONE 4*, 12 (12 2009), 1–11. URL: https://doi.org/10.1371/journal.pone.0008140, doi:10.1371/journal.pone.0008140. 2, 4, 5