Global-to-Local Protein Shape Similarity System driven by Digital Elevation Models

Daniela Craciun*, Jeremy Sirugue* and Matthieu Montes*
*Conservatoire National des Arts et Métiers, Laboratoire GBA, EA 4627
2 rue Conté, 75003 Paris, France, Email: firstname.lastname@cnam.fr

***Abstract*—We are currently developing a bio-shape similarity system for supplying high-throughput protein shape similarity applications within massive datasets. The proposed system is powered by a global-to-local shape similarity system which exploits shape elevation and local convexity attributes. In the first step, a global similarity is computed between the shape descriptors associated to each protein input. The procedure outputs best N similarities chosen by the user, within a *query-to-cluster* approach. The second stage is a patch-based local similarity computation method which is designed to find the best similar target from the cluster for supplying *query-to-target* protein retrieval applications. The local patch-based similarity comparison benefits of a multi-CPU implementation, offering thus fast query search capabilities within massive datasets. Experimental results on the SHREC 2017 BioShape dataset [4] composed of 5484 models, illustrate the effectiveness of the proposed system.**

I. Introduction and Motivation

Structural biologists face a rapid growing of the number of protein structures in the Protein Data Bank [1] (130000 structures in 2017) which induces a considerable need for fast protein similarity search methods able to screen such large databases in a reasonable amount of time. Since protein 3D structures are more conserved during evolution than their respective amino acid sequences [7], sequence-based protein similarity search methods fail to retrieve structural homologs that share low sequence homology. Protein structural similarity search methods such as DALI [11], CE [12], FAST [13] or FATCAT [14] that use different methods to align protein structures and compute their similarity, notably using the Root Mean Square Deviation of their alpha carbons. Most of the existing protein comparison methods share the major limitation that they are too computationally expensive to search a large protein structure database in a reasonable amount of time [15].

In this paper, we introduce a shape-based method which allows to perform high-throughput protein classification without relying on human operator intervention. In addition, the proposed method is designed to perform fast query search within massive datasets. The present research work introduces a global-to-local framework designed in a complementary fashion, along with a multi-CPU implementation, in order to cope with rapidity constraints. Our paper is organized as follows: Section II describes the proposed global-to-local **P**rotein **S**hape **S**imilarity **S**earch **S**ystem (**PS4**), followed by Section III which presents experimental results and the performance evaluation on the SHREC 2017 (SHape REtrieval Contest) BioShape dataset [4]. Finally, Section IV concludes the present research work and gives main perspectives.

II. Proposed Protein Shape Similarity Search System (PS4)

The proposed **P**rotein **S**hape **S**imilarity **S**earch **S**ystem (**PS4**) is composed of two main stages: the first stage is performed for each shape and consists in the **M**acromolecular **S**hape (**MS**) representation as a **D**igital **E**levation **M**odel (**DEM**), encoded over a 2D grid. The second stage corresponds to the shape comparison phase which is supplied via a global-to-local framework relying on the MS-DEMs descriptors. Both stages are summarized through the following two sections.

A. Representing Macromolecular Shapes as Digital Elevation Models

Macromolecular triangular surface computation. The shape representation algorithm applies the EDTSurf [2] 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 descriptor computation. The present work exploits the DEM concept traditionally employed in cartography for representing Earth's surface from terrain elevation data [8]. The algorithm starts by applying the mesh flattening procedure introduced in [3], which maps the mesh onto the unit sphere using the Laplace-Beltrami operator [10]. The spherical mapping provides a valid solution for any genus-0 triangle meshes, being adapted in our current research work. 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 shape's elevation, while providing topology and fast comparison over a 2D grid space. The DEM descriptor is stable under rotations and translations variations of the input mesh and varies in presence of scale transformations. A detailed description of the algorithm can be found in [5]. The final output is the digital elevation model associated to the macromolecular surface, noted MS-DEM. Figures 1 (a)-(d) illustrate the results obtained for a target belonging to the protein pool of the SHREC 2017 BioShape track [4].

978-1-5386-0706-0/17/$31.00 ©2017 European Union

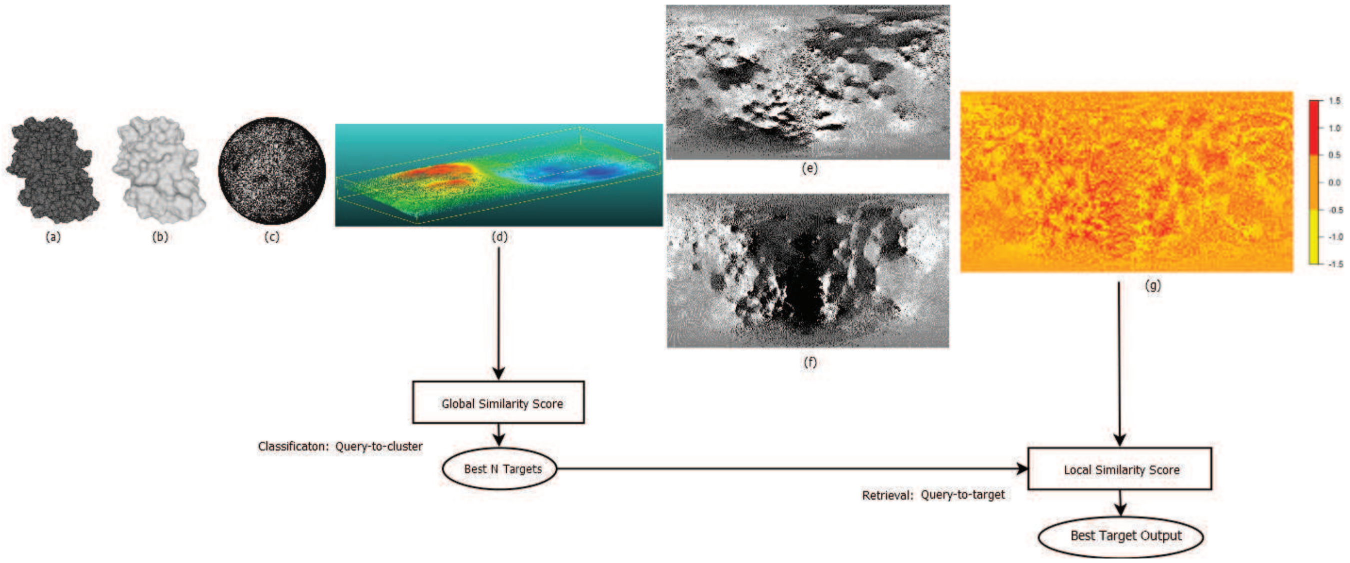

Fig. 1. Overview of the global descriptor computation stage (input model: $m10001$) and the integration within the global-to-local similarity computation framework of the **PS4** prototype: (a) input data $N_p = 187866$ points, $N_t = 357840$ triangles; (b) macromolecular mesh generated by EDTSurf [2]: $N_p = 86079$ points, $N_t = 172154$ triangles; (c) spherical mapping output [3]: $N_p = 86079$ points, $N_t = 172154$ triangles; (d) MS-DEM output: $N_p = 86089$ points, bounding box dimensions: $[472, 257, 36.112]$, (e) Gradient G_x computed for the MS-DEM illustrated in Figure (d); (f) Gradient G_y computed for the MS-DEM illustrated in Figure (d); (g) Convexity map result corresponding to the MS-DEM illustrated in Figure (d).

B. *Global-to-Local Protein Shape Similarity Computation*

The proposed protein similarity search system relies on a global-to-local shape similarity search framework designed in a complementary fashion. The global similarity stage allows to perform fast comparisons, being therefore employed as a first stage to search for best similar candidates w.r.t. the query. In exchange, the local stage provides a finer comparison, being suitable for selecting the best rank similarity among targets clustered at the global comparison stage.

Global Comparison of MS-DEMs. The MS-DEM shape descriptor is used along with different global distances for supplying the protein shape similarity computation stage. The present research work evaluates the Mean Absolute Differences (d_{MAD}) and the Root Mean Square Deviation (d_{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. In absence of scale variation, the similarity score is valid for meshes with a similar number of points, belonging to the same class. In this configuration, the global comparison stage was compared w.r.t. state-of-the-art shape retrieval algorithms and a detailed peformance evaluation can be found in [5]. In the present research work, the shape comparison stage outputs the dissimilarity matrix which is exploited for extracting the best N similarities chosen by the user w.r.t. the query. Figure 2 illustrates an example of the *query-to-cluster* procedure output which provides the best $N = 4$ similar targets w.r.t. the query q_{11}.

Patch-based Local Comparison of MS-DEMs via Convexity Maps. The local shape comparison stage takes as input the best $N = 4$ similarities output by the global comparison stage and finds the best rank similarity w.r.t. the query. The local similarity computation is performed through the use of convexity maps which are computed from each MS-DEM descriptor. The MS-DEMs are exploited for computing the gradient along X and Y directions, noted G_x and G_y, respectively. The convexity coefficients are computed by identifying gradient extremum values (minimas and maximas) and by assigning convexity labels to each point belonging to the MS-DEM descriptor. Figures 1 (e), (f) and (g) illustrate an example of the G_x, G_y and the associated convexity map, noted C_{map}, respectively.

The pairwise patch-based comparison is performed by extracting patches from the convexity maps corresponding to the query and the target meshes. For each patch extracted in the query, a local dissimilarity measure is computed w.r.t. each patch extracted from the target mesh. The d_{MAD} distance is computed between each query patch and all patches belonging to the target. Similar patches selected w.r.t a threshold value are further considered for computing the overall dissimilarity score between the query and the target meshes. In our experiments, it was observed that a patch ray of 7 pixels provides accurate results in a reasonable amount of time. Moreover, the results let us concluded that higher patch ray values lead to a computationally expensive framework without improving considerably the accuracy. Selected similar patches have less than 20% dissimilarity compared to the maximum d_{MAD} distance computed over all the compared patches.

In order to avoid the computational burden of the local comparison stage, the convexity maps are computed from MS-DEMs with a reduced resolution (by a factor of 2). In addition,

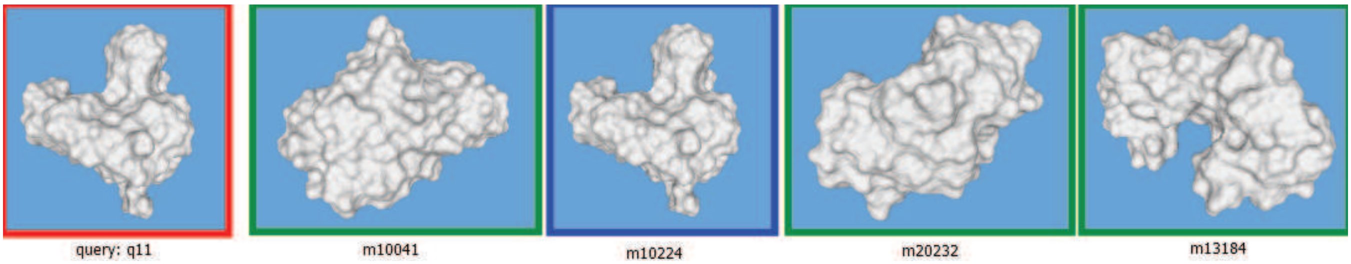

Fig. 2. Output generated by the *query-to-cluster* procedure for query q_{11}: Global similarity output obtained using the d_{MAD} distance. The procedure outputs $N = 4$ best similarities (illustrated from left to right), the identity query (model: $m10224$) is found as the 2^{nd} top similarity (contour emphasized in blue color).

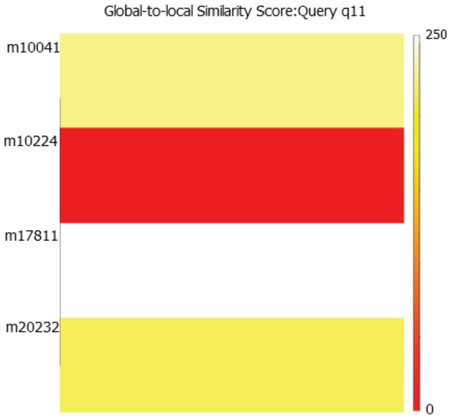

Fig. 3. Output generated by the *query-to-target* procedure for query q_{11} via the patch-based local similarity comparison of convexity maps. Input: $N = 4$ top similarities generated by the global comparison stage (*query-to-cluster*) via the d_{MAD} distance. Output: best similar target obtained for q_{11}: $m10224$.

the local patch-based comparison benefits of a multi-CPU implementation. Figure 3 illustrates an example of the best similarity output found at the local comparison stage for the first query, q_{11}. As shown in Figure 2, the global similarity phase outputs the best $N = 4$ similar targets, with the identity query found as the 2^{nd} top-rank similarity (blue contour). Figure 3 illustrates that the identity target of query q_{11}, noted $m10224$, was found as the best similar target by the local similarity comparison phase.

III. Results and Performance Evaluation

This section presents the performance evaluation of the proposed framework, **PS4**, on the dataset made available for the SHape REtrieval Contest (SHREC 2017) BioShape track [4]. We analyse the proposed system in terms of accuracy, runtime and memory usage.

Dataset. The dataset consists in 10 queries (q_{11},...,q_{20}) (selected from the molecule of the Month collection [9]) and 5484 targets. In order to allow accurate validation, for two queries (q_{11} and q_{14}), identic protein were included in the target set.

Evaluation measures. In the SHREC 2017 BioShape track, the protein similarity evaluation relies on the 3DZM method [6] which measures the accuracy by comparing the correlation coefficients computed between the query and each target. More details about the dataset and the evaluation protocol can be found in the SHREC 2017 BioShape track [4].

A. Accuracy Evaluation

As presented in the SHREC 2017 BioShape track [4], the accuracy of the proposed framework is evaluated w.r.t. the correlation coefficients obtained by the 3DZM method [6]. In order to analyse the behaviour of the local stage, we provide an evaluation of the **PS4** framework employed in both modes: global similarity computation and global-to-local computation mode. Figure 4 illustrates the results generated by the global stage (employing the d_{RMSD} distance) and the global-to-local framework. It can be observed that for queries q_{11} and q_{14}, the global-to-local approach retrieved successfully identity shapes as the best rank similarity.

The patch-based local comparison stage improves the average correlation coefficient, attaining 0.6598, compared to 0.6051 provided by the global comparison stage alone. In addition, while for some queries (i.e. $q_{12}, q_{13}, q_{17}, q_{18}$) the global minimum was lost by the local comparison stage, for the remaining queries ($q_{15}, q_{16}, q_{19}, q_{20}$), the global minimum was correctly maintained. While providing a rapid protein retrieval framework, there is still room for improving the local patch-based similarity computation stage by searching more stable and intrinsic features w.r.t. the dataset (shape, size, resolution).

B. Runtime and memory usage

The proposed algorithm is implemented in C/C++ and runs on a 64b Linux machine equipped with 32Gb of RAM memory and an Intel Xeon running at 2.3 GHz. Less computationally expensive stages, i.e. the descriptors' computation (MS-DEM, C_{map}) and the global comparison, are designed for simple-CPU implementation. The most expensive stage, i.e. the pairwise patch-based local comparison of convexity maps benefits of a multi-CPU implementation.

Simple-CPU global comparison of MS-DEMs. The first row of Table I resumes the average runtime for extracting the MS-DEM descriptor for one model belonging to the protein pool of the SHREC 2017 BioShape track [4]. The computation time for comparing one query against the entire protein pool

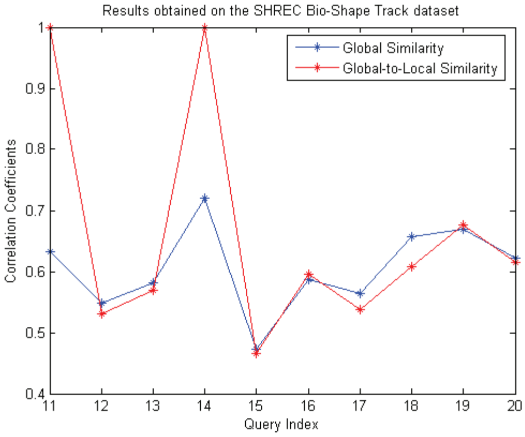

Fig. 4. Experimental results generated by the proposed system **PS4** on the SHREC 2017 BioShape dataset [4], results comparison: Global similarity vs. Global-to-Local similarity computations; Global similarity output generated by the d_{RMSD} distance computed between MS-DEM descriptors.

TABLE I
AVERAGE RUNTIME (SECONDS) OBTAINED FOR SIMPLE-CPU IMPLEMENTATION OF EACH MODULE COMPOSING THE SHAPE DESCRIPTOR EXTRACTION PROCEDURE ILLUSTRATED IN FIGURE 1.

Module	(b)	(c)	(d)	(e), (f)	(g)
CPU (s)	3.34	2.65	0.12	0.03	0.015
RAM (Mb)	8.08	6.6	**1.47**	2.3	**1.14**

takes is in average of 2.3502 seconds and 3.3518 seconds for d_{MAD} and d_{RMSD} distances, respectively.

Multi-CPU Patch-based Local Comparison of Convexity Maps. In order to avoid the computational burden, the pairwise local comparison procedure is implemented on $N_{CPU} = 24$ cores, taking in average 1 min 15 sec for providing the best similar target. When compared to the simple-CPU implementation, the parallelization allows to reduce the runtime by an average factor of 46. This result emphasizes the fast query search capability detained by the proposed protein shape similarity search system, **PS4**.

Memory usage. The average memory usage for storing the MS-DEM descriptor is 1.058 kb. The second row of Table I illustrates the memory usage for the target $m10001$ depicted in Figure 1. The overall memory usage required by the MS-DEM descriptor and the associated convexity map C_{map} is 2.61 Mb. When compared to the input mesh storage, (Figure 1 b)), the proposed descriptors reduce the memory usage by a factor of 3, being therefore suitable for processing massive datasets.

IV. Conclusions and Future Work

This paper presented the **P**rotein **S**hape **S**imilarity **S**earch System, (**PS4**), a global-to-local geometric-based protein similarity framework designed for supplying fast query search within massive datasets. The main features which ensure the rapidity of the proposed system operate at two levels: (i) software architecture: global and local shape similarity computation for fast *query-to-target* computation, and (ii) software implementation: multi-CPU optimization of local comparison. This gives rise to a protein similarity system designed in a complementary fashion: the global comparison stage allows rapid selection of best candidates, while the local stage provides best target selection among them. Experimental results on the SHREC 2017 BioShape dataset [4], containing 5484 models, demonstrate that our approach detains fast query search capabilities for supplying high-throughput *query-to-target* protein retrieval applications. Nevertheless, while solving the rapidity issue, there is still room for improving the accuracy of the local similarity stage. Research perspectives are concerned with the accuracy improvement of the local stage, in presence of various shape types, while still maintaining the fast query search capability detained by the proposed system, **PS4**.

Acknowledgment

The present research work is supported by the ERC Vi-DOCK Grant no. ♯640283 from the European Research Council Executive Agency.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N Shindyalov and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp. 235-242, 2000.

[2] D. Xu and Y. Zhang. Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform. *PLOS ONE*, 4(12), pp. 1–11, 2009.

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

[4] N. Song, D. Craciun, C. W. Christoffer, X. Han, D. Kihara, G. Levieux, M. Montes, H. Qin, P. Sahu, G. Terashi and H. Liu. SHREC'17 (SHape REtrieval Contest) BioShape Track: Protein Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, 2017.

[5] D. Craciun, G. Levieux and M. Montes. Shape Similarity System driven by Digital Elevation Models for Non-rigid Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, 2017.

[6] M. Novotni and R. Klein: 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the 8th ACM Symposium on Solid Modeling and Applications*, pp. 216-225, 2003.

[7] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. In *The EMBO Journal*, 5(4), pp. 823–826, 1986.

[8] C. L. Miller and R. A. Laflamme. The digital terrain model - theory and application. In *Photogrammetric Engineering*, pp. 433–442, 1958.

[9] D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman and S. K. Burley. The RCSB PDB Molecule of the Month: Inspiring a Molecular View of Biology. In *PLOS Biology*, 13(5), pp. 1–12, 2015.

[10] G. Craig, G. Xianfeng and A. Sheffer. Fundamentals of Spherical Parameterization for 3D Meshes. In *ACM SIGGRAPH*, pp.358–363, 2003.

[11] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. In *Journal of Molecular Biology*, 233(1), pp. 123–138, 1993.

[12] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. In *Protein Eng*, 11(9), 739–747, 1998.

[13] J. Zhu and Z. Weng. FAST: a novel protein structure alignment algorithm. In *Proteins*, 58(3), pp. 618–627, 2005.

[14] Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. In *Bioinformatics*, Suppl. 2, pp. 246–255, 2003.

[15] S. Mezulis, M. J. Sternberg and L. A. Kelley. PhyreStorm: A Web Server for Fast Structural Searches against the PDB. In *Journal of Molecular Biology*, 428(4), pp. 702–708, 2016.