Computational and Structural Biotechnology Journal 26 (2024) 1–10

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method Article

PLO3S: Protein LOcal Surficial Similarity Screening

Léa Sirugue *, Florent Langenfeld, Nathalie Lagarde, Matthieu Montes *

Laboratoire GBCM, EA7528, Conservatoire National des Arts et Métiers, Hesam Université, 2, rue Conté, Paris, 75003, France

ARTICLE INFO

Keywords:
Protein surface
Protein structure
Surface similarity
Spectral geometry
Surface comparison

ABSTRACT

The study of protein molecular surfaces enables to better understand and predict protein interactions. Different methods have been developed in computer vision to compare surfaces that can be applied to protein molecular surfaces. The present work proposes a method using the Wave Kernel Signature: Protein LOcal Surficial Similarity Screening (PLO3S). The descriptor of the PLO3S method is a local surface shape descriptor projected on a unit sphere mapped onto a 2D plane and called Surface Wave Interpolated Maps (SWIM). PLO3S allows to rapidly compare protein surface shapes through local comparisons to filter large protein surfaces datasets in protein structures virtual screening protocols.

1. Introduction

Proteins are central in most biological processes. Proteins can be described through their sequence, structure, surface and/or function(s). The protein surface is an abstract, geometric representation of the protein potential interactions, structure, fold, and sequence [35,18,12]. Proteins sharing a related function display similar surfaces that can be independent of their sequence and/or structure similarity [44,18,27]. Different methods based on protein surface comparison have been developed for protein-protein interactions prediction [45,39,12], protein structural alignment [29] or protein shapes classification [18,15,43,28,9,14,27,32].

Surface comparison methods can be classified into different categories depending on their shape descriptor computed from the surface. (1) The methods based on spectral geometry establish a relationship between the surface shape and the spectra of the Laplace-Beltrami operator. A spectrum of the Laplace-Beltrami operator is a fingerprint composed of the eigenvalues obtained using the differential Laplace-Beltrami operator [37,50,2]. (2) The methods based on histograms summarize local or global geometrical or topological features of the surface [20,51,41,40,55]. (3) Projection-based methods use the projection(s) of the protein topography in the 2D space [34,8]. (4) Zernike-based methods use the moments of 2D or 3D Zernike polynomials [23,6,54,30]. They have been widely used on protein surfaces and display high performances in retrieval [23,21,6,17]. (5) The last category comprises methods based on geometric learning using convolutional neural networks [31,55,12].

Surface shapes can be described globally or locally. A global surface shape descriptor describes the surface shape of the whole object [20,42,36] which allows direct comparisons of the whole surface shape of different objects whereas a local surface shape descriptor is defined over a surface patch and allows comparisons with other surface patches.

To our knowledge, no spectra-based method has yet been specifically designed for global protein surface comparison. In the present work, we describe Protein LOcal Surficial Similarity Screening (PLO3S), a fast, global protein surface shapes comparison method based on a local spectral descriptor, Surface Wave Interpolated Maps (SWIM). SWIM is a wave kernel signature (WKS) [2] conformally projected on a 2D plane. In PLO3S, the values of SWIM are processed using a dense point-to-point comparison. PLO3S is designed to blindly screen large protein surfaces datasets in order to discard protein surfaces that do not share high local surficial similarity to the query. This allows for (1) further protein surface shapes screening with finer local surface shape comparison methods that cannot handle large datasets, and (2) reducing the number of false positive potential binding sites in a drug discovery pipeline.

2. Materials and methods

2.1. PLO3S

Protein LOcal Surficial Similarity Screening (PLO3S) is a screening method for finding surficial similarities of proteins, independently of the sequence of the proteins. PLO3S is described in two steps: (1) the

* Corresponding authors.
E-mail addresses: sirugue.lea@gmail.com (L. Sirugue), matthieu.montes@cnam.fr (M. Montes).

https://doi.org/10.1016/j.csbj.2023.12.002
Received 14 November 2022; Received in revised form 1 December 2023; Accepted 3 December 2023
Available online 11 December 2023

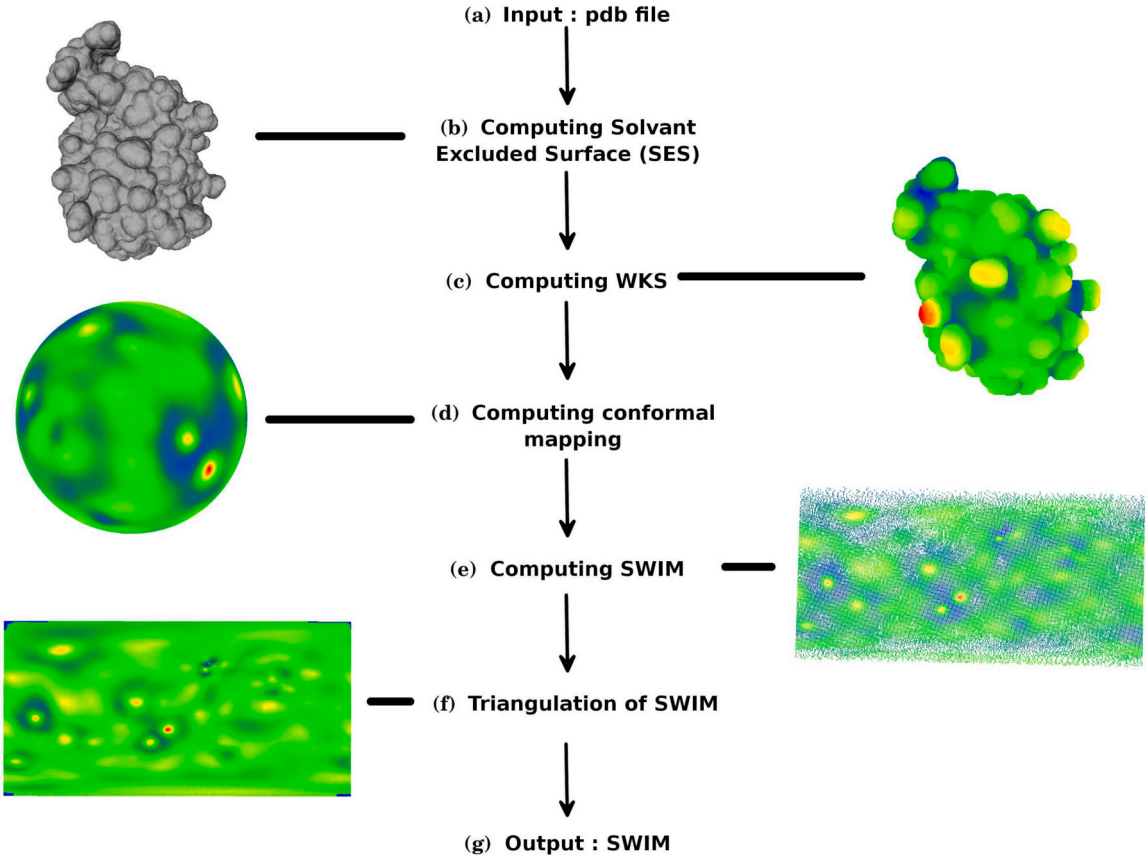


Fig. 1. Diagram for the generation of a SWIM. The Solvent Excluded Surface is computed using EDTSurf [53] (a-b). In a second step, the Wave Kernel Signature (WKS) [2] is calculated for each point of the surface (c). The surface is then projected on a unit sphere [1] (d). The sphere is mapped onto the 2D plane [9] (e). The points of the map are interpolated (f) to form the final descriptor called SWIM (g).

computation of the Surface Wave Interpolated Maps (SWIM) descriptor, and (2) the computation of the surface shape similarity.

2.1.1. Computation of the SWIM descriptor

The Solvent Excluded Surface (SES) of the all atoms protein structure (Fig. 1a) retrieved from the PDB [3] is computed using EDTSurf [53] (Fig. 1b). The Wave Kernel Signature (WKS) is then computed on the resulting 3D mesh M [2] (Fig. 1c). For each point of the surface of the 3D mesh, a vector of size N, representing the WKS descriptor is computed. This descriptor, based on the eigenvalues of the Laplace-Beltrami operator, has the property of invariance to isometry and is robust to perturbations [2].

On a second step, the 3D mesh is projected on a unit sphere S using the ITK algorithm [13] based on the method described by Angenent et al. [1] (Fig. 1d). The frame of the unit sphere is defined using a reference point on its surface called the *pole*. The pole of the unit sphere is selected arbitrarily, and the triangles of the mesh are projected while preserving their angles. This transformation is conformal and bijective. It is to note that, even if the distances and surface areas are not preserved in the projection, they are only modified by a scaling factor.

Then, the unit sphere is transformed onto the 2D plane using the two spherical coordinates of the angles (θ, ϕ) [8,9] (Fig. 1e). A map of size $(\theta_{max} - \theta_{min})/\delta, (\phi_{max} - \phi_{min})/\delta$ is created. θ_{max} and θ_{min} are the maximum and minimum values of θ and ϕ_{max} and ϕ_{min} are the maximum and minimum values of ϕ. δ is the step for dividing the sphere in areas, where each area is represented by a point on the map in the discrete plan. The value associated to each point on the map is the WKS descriptor represented by a vector (Fig. 2). These maps are called Surface Wave Interpolated Maps (SWIM).

The final step is the interpolation of each point of the map (Fig. 1f). The map is encoded as an image and the projection on the 2D plane is not filling each pixel with a value. This can create an imbalance if a map has large areas with no value, or if the neighborhood of a point of the map is considered for comparison. Therefore, for each pixel with no value, the three points on the map defining the triangle with the smallest area containing the pixel are used to interpolate the value of this pixel.

The main issue with this representation is the deformation in the neighborhood of the poles while passing from the unit sphere to the 2D plane. To handle this issue, we use a multiview approach where the pole axis is rotated by an angle α in the planes perpendicular to each of the three Cartesian axes of the unit sphere (Fig. 3). Then, a SWIM is created as mentioned above (Fig. 1g). We used a multiview approach to generate a set of seven SWIMs by using seven projections with a $\frac{2\pi}{3}$ rotation (Fig. 3). This approach offers an optimal balance between minimizing the impact of the initial arbitrary pole selection and maintaining the computational efficiency of the method. As the SWIMs are compared locally (see section 2.1.2 below), the multiview approach helps us avoid the arbitrary choice of the pole of the unit sphere at the second step of the workflow. This set of seven SWIMs is the final descriptor used for comparing protein surfaces.

2.1.2. Computation of the surface shape similarity

A dense, exhaustive (*i.e.* point by point) comparison is performed between the generated sets of SWIMs.

Two shapes T and V are compared by searching the best matches of the vectors of their respective SWIM C_T and C_V. To compare the two vectors H_{k_T} and H_{k_V} at points k_T and k_V from the maps C_T and C_V, respectively, the Earth Mover's Distance (EMD) [38] is used. EMD is a metric that measures the distance between two probability distributions, taking into account the proximity of the bins in a histogram. The sequence of WKS eigenvalues can be represented as a histogram in

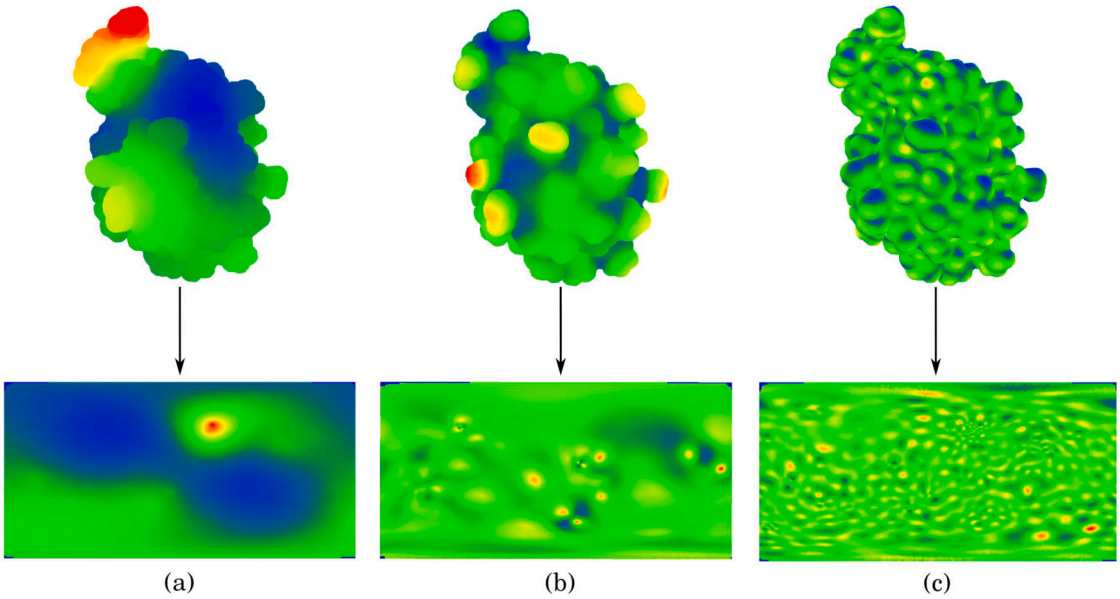


Fig. 2. WKS on the surface of one conformation of ubiquitin (5xbo_A_1) (top) and its corresponding SWIM (bottom) for the 1st (a), 50th (b) and 100th (c) value of the WKS.

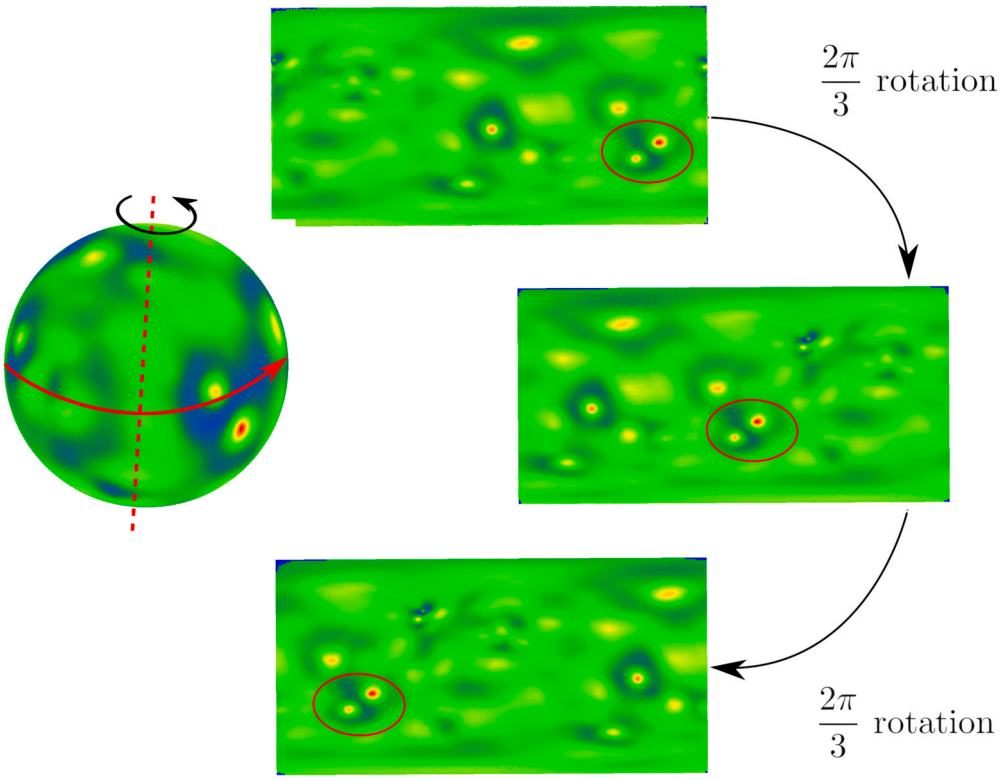


Fig. 3. Illustration of the multiview approach with a representation of one of the three rotations of the pole axis achieved in the planes perpendicular to the Cartesian axis of the Unit sphere (left panel). The corresponding SWIMs for the 50th value of the WKS with a rotation of $\frac{2\pi}{3}$ around a pole of the unit sphere of Ubiquitin (5xbo_A_1) are shown in the right panel.

descending order. Given this property, EMD is an appropriate distance measure for comparing protein surfaces using the WKS descriptor.

Since the WKS is represented as a 1D array, the EMD equation can be simplified as the sum of the absolute differences between the cumulative values of the WKS.

Given two WKS, two vectors with weights equal to one, $X = \{x_1, ...x_n\}$ and $Y = \{y_1, ...y_n\}$, $\forall\, i, j \in [1..n]\ d_{ij} = |x_i - y_j|$ the L_1 distance is defined and a function f_{ij} such as it minimizes the following equation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} d_{ij}, \tag{1}$$

with the constraints $f_{ij} > 0$, $\sum_{j=1}^{n} f_{ij} \leq 1$, $\sum_{i=1}^{n} f_{ij} \leq 1$ and $\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = n$.

The WKS is represented as a 1D vector with no weight, which means that the EMD equation can be simplified as the sum of the absolute differences between the cumulative values of the WKS. Given $D_X(i) = \sum_{k=1}^{i} x_k$ and $D_Y(j) = \sum_{k=1}^{j} y_k$, then the EMD L between two WKS X and Y is defined as:

$$L(X,Y) = \sum_{l=1}^{n} |D_X(l) - D_Y(l)|. \tag{2}$$

The surface shape dissimilarity score for the shapes T and V is the sum of the best distance of each point L. If C_T is composed of N_T points and C_V of N_V points, then the score $S(T,V)$ of dissimilarity between T and V is:

$$S(T,V) = \min\left(\sum_{k_T=1}^{N_T} \min_{k_V} L(X_{k_T}, Y_{k_V}) , \sum_{k_V=1}^{N_V} \min_{k_T} L(X_{k_T}, Y_{k_V}) \right). \tag{3}$$

The surface shape dissimilarity score is normalized between 0 (identity) and 255 (maximum dissimilarity of the shapes).

2.1.3. GPGPU optimization

The computation of the shape similarity is performed in parallel using the General Purpose GPU (GPGPU) sum reduction technique [10,7] where the data is divided into fixed sizes that can be accommodated by the internal memory of the GPU, thereby enabling parallel processing of the computations.

2.2. Spectral geometry based shape comparison methods

Heat Kernel Signature [50] is derived from the Heat Kernel which represents the diffusion of heat on an object as a function of time. The HKS is based on the spectrum of the Laplace-Beltrami operator. HKS is a local descriptor with the property of isometric invariance and is stable against perturbations [50].

Wave Kernel Signature [2] is based on the spectrum of the Laplace-Beltrami operator. It describes the energy of quantum particles on the surface of the shape based on the wave equation which is a solution of the Schrödinger's equation. A signature is created with the solution of the wave equation. WKS is a local descriptor with the property of isometry and scale invariance. Contrary to HKS, in WKS, time is not taken into consideration. It is replaced by the energy of the particle. For the WKS, the energy is related to the size of the geometry; a large energy represents a local geometric feature while a small energy represents a global geometric feature.

2.3. Protein structure comparison methods

In **Combinatorial Extension (CE)** [49], the protein is represented as a set of fragments of eight amino-acids. A fragment is aligned to another fragment composed of at least eight amino acids and forms an Aligned Fragment Pair (AFP). Using constraints on the maximum distance between AFPs, the AFPs are assembled to form a longer path. Then, an optimization is performed using the Z-score and dynamic programming [33].

In **TM-Align** [57], protein structures are aligned independently of their sequence length using TM-score [56]. In a first step, an alignment based on dynamic programming is proposed. It is decomposed into a residue-to-residue alignment and a secondary structure alignment. In a second step, the TM-score evaluates the matrix score of the alignment. The proteins are aligned again to increase the score, and this step is repeated until stability is found.

In **DeepAlign** [52], protein structures are aligned according to spatial proximity, evolutionary links and hydrogen bond similarity. The score is a combination of a value for amino-acids substitution [19], conformation substitution formed by the angles of the pseudo-bonds of the C_α atoms [58], the TM-score to estimate spatial proximity, and hydrogen bonds similarity.

2.4. Benchmarking dataset

The benchmarking dataset is a subset of the Protein Shape Retrieval Contest track dataset of SHREC19 [24] based on the protein level of the SCOPe classification [11]. For each of the 14 protein shape classes in the subset, only one RMN structure was retained to include side-chain conformational flexibility. All the RMN structures included in our benchmarking dataset present a similar number of conformations to obtain a balanced dataset. Therefore, each protein shape class contains from 20 to 30 conformations from a single PDB structure, for a total of 403 protein conformations. A preliminary study (unpublished results) highlighted that non-globular proteins were less adapted to a spherical projection such as the one used during the computation of the SWIM descriptor. We thus decided to include only globular proteins in the benchmarking dataset for this study. In the following, the IDs used are in the format: PDBID_A_X, where "PDBID" being the unique id of the PDB composed of four letters alphanumeric characters, "A" being the protein chain considered and "X" being the conformation number according to the PDB file. For each structure of the dataset, the Solvent Excluded Surface (SES) is computed using EDTSurf [53] with default parameters. EDTSurf output triangular meshes are stored as `.ply` file, converted to `.off` and `.pcd` formats, required by the different shape comparison methods.

2.5. Performance evaluation metrics

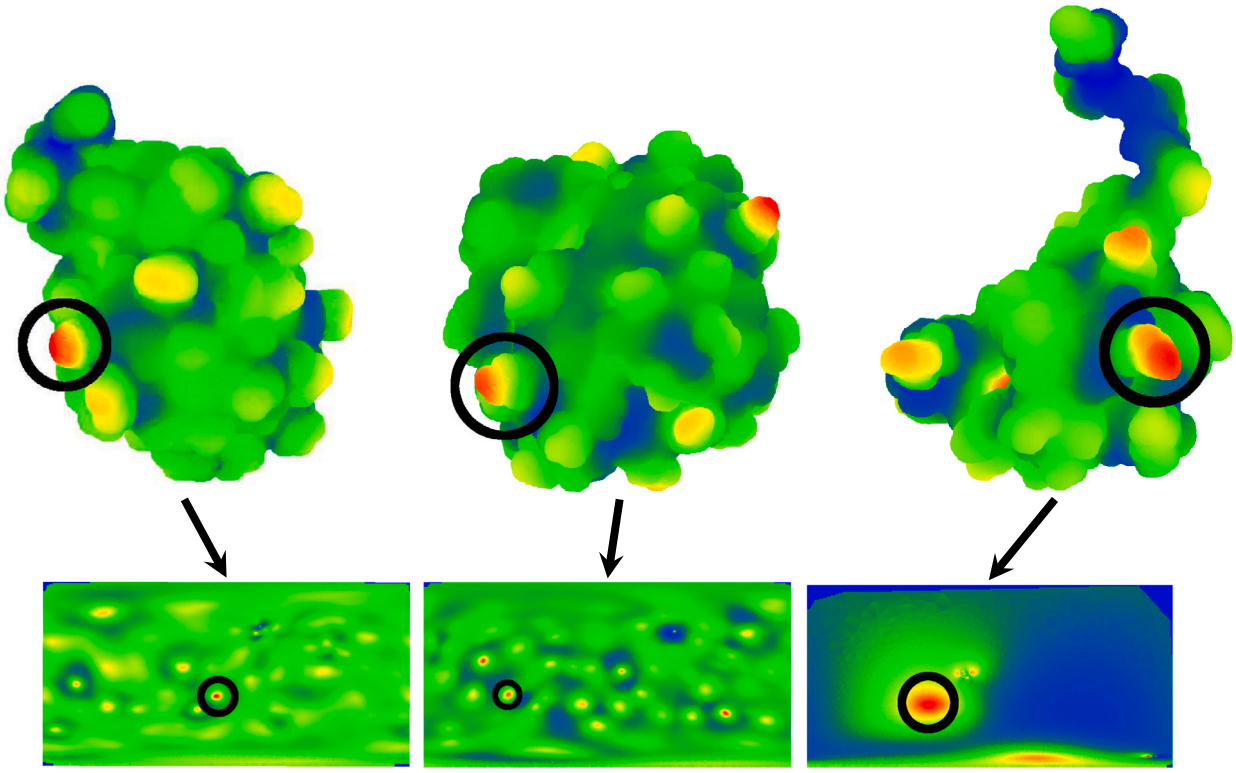
The performance of the PLO3S method to identify proteins presenting similar surfaces is evaluated. Within a protein class C of size $|C|$, the predictions realized using the PLO3S can then be classified into 4 cases: true positive (TP) for proteins presenting similar surfaces and correctly predicted as so by the PLO3S method; false negative (FN) for proteins presenting similar surfaces but not correctly predicted as so by the PLO3S method; false positives (FP) for proteins that are not surficial homologs to the query but not correctly predicted as similar by the PLO3S method; and true negative (TN) for proteins that are not surficial homologs to the query and are correctly predicted as not similar by the PLO3S method. The performance in retrieval of each method is evaluated using Precision-Recall and Negative Predicted Value (NPV) curves. The Precision-Recall curve plots the recall R as a function of the precision P. Precision P, or Positive Predicted Value, is the ratio of targets from class C (TP) retrieved within all objects attributed to class C (TP + FN). The recall R, also called Sensitivity, represents the ratio of retrieved targets from class C (TP) compared to $|C|$ (TP + FP), the size of class C. Precision-recall curves are computed using the Princeton Shape Benchmark tools [48]. NPV evaluates the percentage of objects rightfully classified as negative (TN) within all negatives (TN + FN).

2.6. Runtime

All calculations were performed on a 64-bit Linux Ubuntu desktop computer with an Intel Xeon 2.30 GHz CPU, 32 GB of RAM and a Quadro K4200 4 GB GPU.

3. Results

The evaluation of the performance of PLO3S in enrichment has been performed on a protein shapes dataset derived from the Protein Shape Retrieval Contest of the SHREC 2019 community benchmark [24]. The performance of PLO3S is compared with spectral, geometry-based shape comparison methods and protein structure comparison methods. For all the shape comparison methods, blind all-to-all dense comparisons have



(a) ubiquitin (b) thioredoxin (c) MIP

Fig. 4. Illustration of the 50th value of the WKS on the SESs (upper panel) and their corresponding SWIMs (lower panel) for (a) ubiquitin 5xbo_A_1, (b) thioredoxin 1trv_A_1 and (c) Macrophage Inflammatory Protein 1hun_A_1.

been performed on the whole dataset. Dense comparison is performed when all-to-all descriptor points are compared during the computation of the surface shape similarity. The L_1 distance is used for the HKS and WKS spectrum-based methods, as described in [2,50]. The metrics used for the protein structures comparison methods are those proposed by the authors, which are RMSD, TM-score and DeepScore, respectively, for CE, TM-Align and DeepAlign.

3.1. Illustration of PLO3S performance with selected examples

Three proteins were selected from the SHREC19 dataset, herein used as the benchmarking dataset, to illustrate the PLO3S method functioning and outputs. The selection includes two proteins presenting similar SWIMs, ubiquitin and thioredoxin, and one protein presenting a dissimilar SWIM compared to the two former proteins, Macrophage Inflammatory Protein (MIP, Fig. 4). Ubiquitin and thioredoxin display a globally spherical shape whereas MIP is more elongated. The SWIMs of MIP are less dense, as illustrated by the large blue area representing values of the WKS close to 0. The red area is located at the bottom of the SWIM (close to the pole) and then distorted due to the mapping of the unit sphere to the 2D plane. In the SWIM descriptor, the most singular values are the most qualitatively discriminative values for the eigenvalues of the WKS. In this example, the red area contains the most qualitatively discriminative values.

The dissimilarity scores between different conformations of ubiquitin, thioredoxin and MIP are presented in Fig. 5. The ubiquitin and thioredoxin SWIMs are very similar, as evidenced by dissimilarity scores ranging from 0 to 40. On the contrary, the dissimilarity scores of MIP, compared to ubiquitin or thioredoxin, vary from 30 to 255.

Within the MIP class, the dissimilarity scores are separated into two groups of respectively 10 and 20 conformations. Dissimilarity scores within these two groups range from 0 to 50 and are similar to intra-class scores of ubiquitin and thioredoxin (see the lower right corner of Fig. 5,

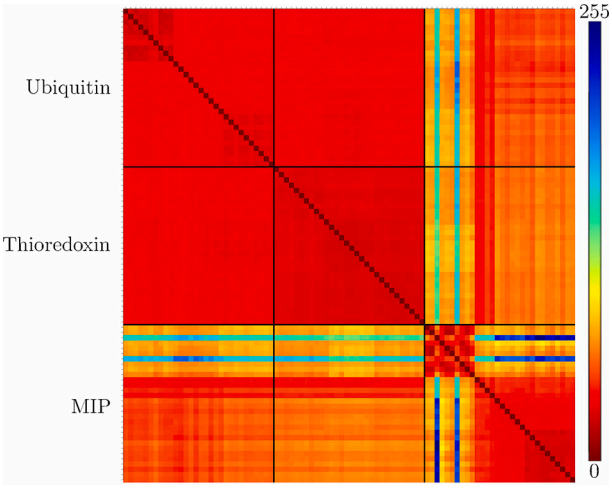


Fig. 5. Matrix of dissimilarity scores of PLO3S on three selected examples: ubiquitin, thioredoxin and Macrophage Inflammatory Protein (MIP).

which exhibit two red squares corresponding to the two MIP groups). The scores between these two groups are significantly different, ranging from 60 to 255. The scores of the third and seventh conformations (1hun_A_3, 1hun_A_7) are even more distinct with scores varying from 120 to 255.

The first ten MIP conformations display a more cylinder-like shape (Fig. 6a and 6b) than the last twenty conformations (Fig. 6c).

Similar to intra-class scores, the first ten MIP conformations have a significantly high dissimilarity score with all ubiquitin and thioredoxin SWIMs, varying from 60 to 255, and the scores of the third and seventh MIP conformations (1hun_A_3, 1hun_A_7) ranging from 120 to 255.

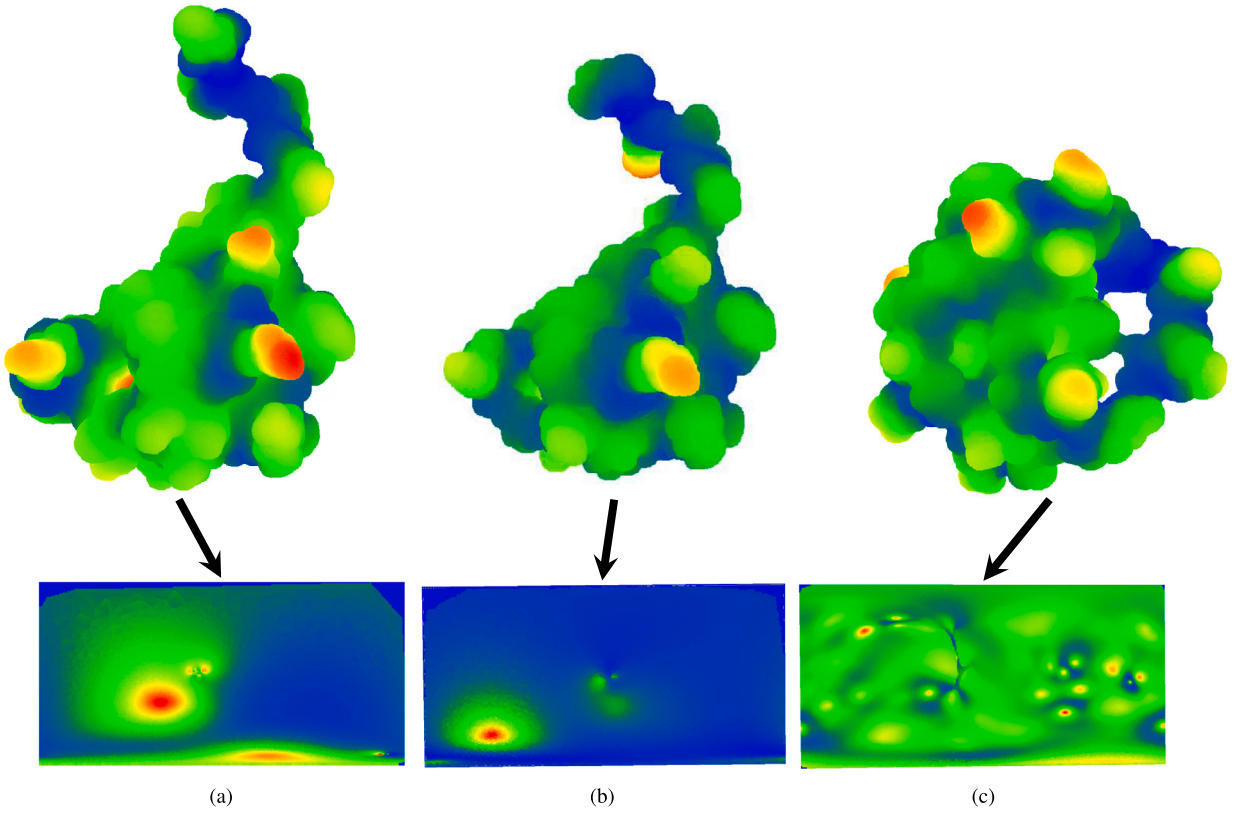


Fig. 6. Illustration of the 50th value of the WKS on the SES (upper panel) and their corresponding SWIMs (lower panel) for three structures of the Macrophage Inflammatory Protein (MIP): (a) 1hun_A_1, (b) 1hun_A_3 and (c) 1b50_B_1.

Table 1
Average computation time (ACT) for one descriptor and comparison for CE, TM-Align, DeepAlgin, PLO3S, HKS and WKS in seconds.

	Descriptor	Comparison
CE	na	0.8 s
TM-Align	na	< 0.1 s
DeepAlign	na	0.1 s
PLO3S	23 min 35 s	0.7 s
HKS	1 min 10 s	4 min 1 s
WKS	9 min 36 s	3 min 11 s

3.2. Evaluation of PLO3S in enrichment

The overall results for PLO3S are shown in Fig. 7. Precision decreases steadily from 0.9 to 0.2 for the PLO3S method (Fig. 7a) as the recall increases from 0.05 to 1. The recall increases up to 0.9 when considering the first 62% of the results, and then increases steadily up to 1. This indicates a compromise between downsizing the dataset and removing a maximum of true negatives. Negative Predictive Value (NPV) ranges from 0.93 to 0.99 with a peak at 37% of the dataset size (Fig. 7c). This peak corroborates the recall value stabilizing at 62% of the dataset size as the NPV is determined by the negatives while the recall is based on the positives.

The computation time of the descriptor SWIM and of the comparison of two SWIM descriptors are shown in Table 1. On average, 23 minutes and 35 seconds are necessary to generate a SWIM. The average computational time associated with the comparison of two SWIM descriptors is 0.7 seconds.

3.3. Comparative evaluation of the performance of PLO3S in enrichment with spectral geometry based shape comparison methods

We compared our PLO3S method with two other spectral geometry based methods for shape comparison, WKS and HKS, which have already shown good performance in 3D shapes comparison [26,25].

The PLO3S method is the top performer for the precision-recall metric (Fig. 7a). The precision for WKS and PLO3S is about 0.9 for a 0.05 recall, while the precision of HKS is 0.37. The precision of WKS and HKS methods is lower than 0.05 for a recall of 1 while the PLO3S precision is superior, around 0.2. Both HKS and WKS descriptors are computed and compared on 3D surfaces. However, PLO3S descriptor, called Surface Wave Interpolated Maps (SWIM), is based on a 2D space. Despite having one less dimensional space, SWIM shows superior performance to the HKS and WKS descriptors.

The PLO3S method has the highest recall compared to HKS and WKS for all sizes of the positives set (Fig. 7b). The recall curve of PLO3S increases rapidly at the beginning and stabilizes around 62% of the dataset size, while the recall of HKS and WKS increases steadily.

The PLO3S method outperforms both HKS and WKS methods for Negative Predictive Value (NPV) (Fig. 7c). NPV ranges from 1 to 0.9, as the size of the true negatives set is approximately 13 times larger than the size of the true positives set. The NPV for PLO3S ranges from 0.93 to 0.99 and its maximum is at 37% of the dataset size. On average, HKS and WKS show stable NPV, around 0.93 for WKS and 0.92 for HKS which is globally inferior to PLO3S.

We determined the average descriptor computation time and comparison time in Table 1 for the three methods. The descriptor computation time includes the processing time from the input mesh to the final descriptor. PLO3S has the slowest descriptor computation time with 23 minutes and 35 seconds, while the fastest descriptor computation time is 1 minute and 34 seconds (HKS). On the contrary, the fastest com-

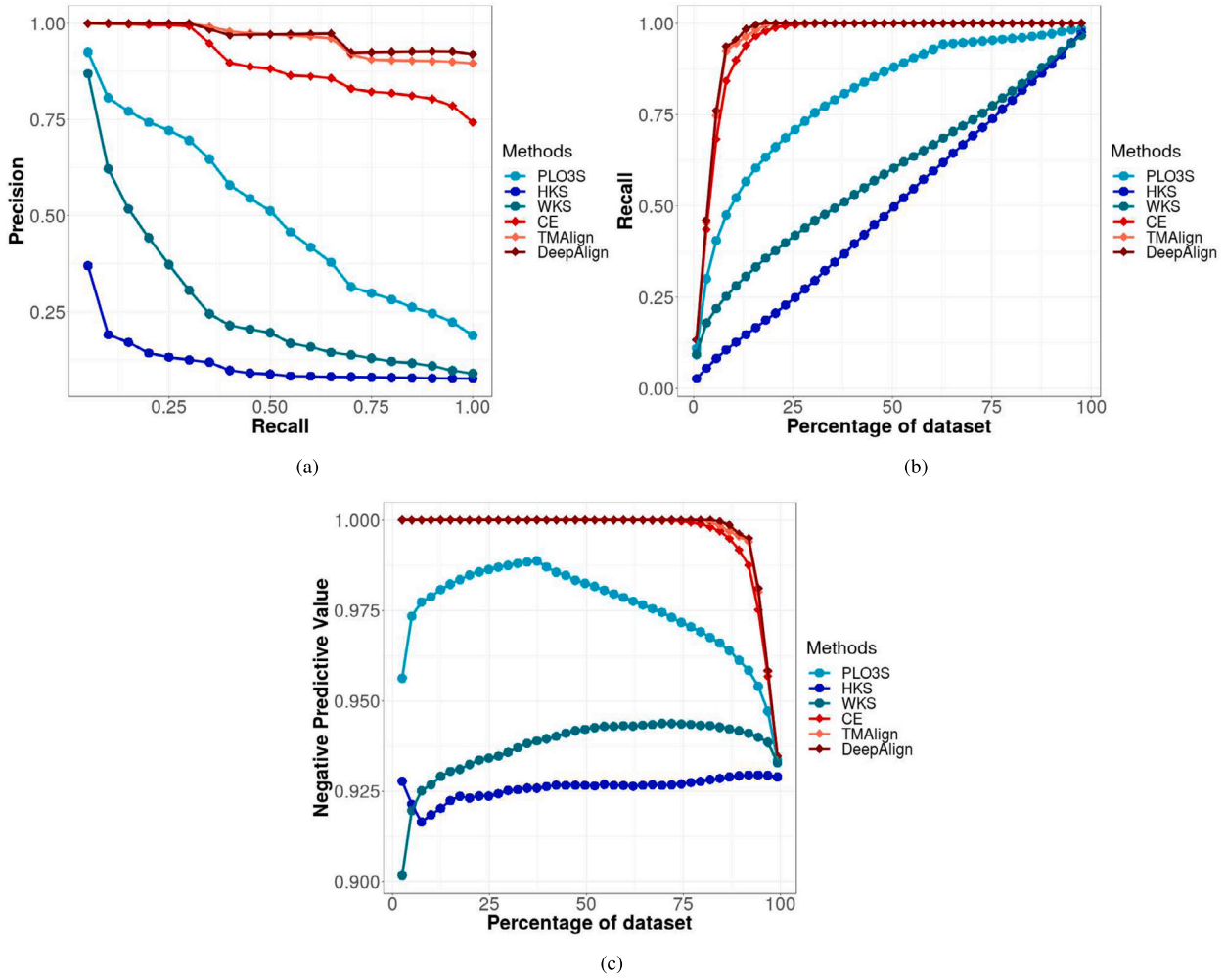


Fig. 7. a Precision-Recall curves for PLO3S, HKS, WKS, CE, TM-Align and DeepAlign. b Recall values for different thresholds for positives. c Negative Predictive Values for different thresholds for negatives.

parison time is 0.7 seconds for PLO3S, while the slowest is HKS with 4 minutes and 1 second on average.

3.4. Comparative evaluation of the performance of PLO3S in enrichment with protein structure comparison methods

The three standard protein structures comparison methods, CE, TM-Align and DeepAlign outperformed the PLO3S method in all three measures, precision-recall, recall and negative predictive value (Fig. 7). While the precision of PLO3S is about 0.9 for low recall and 0.2 for high recall, DeepAlign and TM-Align have higher overall precision, ranging from 0.89 to 1 for different associated recall.

The PLO3S recall stabilizes around 0.9 for a dataset size of 62%. The recall of protein structure comparison methods is 1 after the first 20% of the results.

The PLO3S NPV is inferior to the protein structure comparison methods, with its highest NPV of 0.99 at 37% of the dataset size. The NPV of the protein structure comparison methods equals 1 for a 2% to 75% dataset size and reaches 0.93 for 100% of the dataset.

The protein structures comparison methods do not require the computation of a descriptor because the comparison and scoring parts are directly based on the structure (Table 1). The comparison time is 0.8 second, 0.7 second, 0.1 second and inferior to 0.1 second for CE, PLO3S, DeepAlign and TM-Align, respectively. The comparison time of PLO3S is thus similar to the comparison times of the protein structure comparison methods.

4. Discussion

4.1. Illustration of PLO3S method with selected examples

In order to illustrate the PLO3S method functioning and outputs, three proteins, ubiquitin, thioredoxin and MIP, were selected and further analyzed. The SWIMs of ubiquitin and thioredoxin display similar patterns, which is corroborated by low values of their corresponding dissimilarity scores. Conversely, the high values in dissimilarity scores obtained by comparing the SWIMs of MIP and ubiquitin and thioredoxin illustrate their difference in shape since MIPs display an elongated shape.

An elongated shape, not easily projectable into a unit sphere, may be prone to more noise brought by the projection. This is emphasized by the differences between the scores of the group of 10 conformations of MIP (that present more elongated shapes) and the scores of the other 20 conformations of MIP (Fig. 6a and Fig. 6b) compared to the second group of 20 MIP conformations (Fig. 6c). All these conformations were gathered under the same MIP protein label by the SCOPe classification, but when investigating further we noticed that these two groups of conformations belong to similar yet distinct proteins. The 10 first MIP conformations are from the Macrophage Inflammatory Protein 1 alpha (MIP-1alpha), and the 20 last MIP conformations are from the Macrophage Inflammatory Protein 1 beta (MIP-1beta). The conformations available in the dataset for these two proteins present different shapes and thus different surface topologies. Even if MIP-1alpha and

MIP-1beta display 68% sequence identity, PLO3S allowed to highlight their surficial dissimilarity.

All the dissimilarity scores of the first group of 10 MIP conformations are high compared to all other scores, indicating the noise sensitivity of an elongated shape projected onto a unit sphere. In particular, the third and seventh conformations of MIP (1hun_A_3 and 1hun_A_7) show significantly higher scores than the other conformations of the 10 MIP conformations. Because the scores for the third and seventh MIP conformations are higher in all respects than the others, they indicate outliers probably caused by the elongated shape not fitting well on the unit sphere.

As mentioned in the section 2, we noticed this point in a preliminary study with non-globular proteins and we decided to exclude them from the benchmarking dataset for this first study. However, the outcomes obtained with the MIP protein showed that as the PLO3S is based on the use of the local SWIM descriptor, the sharpest features of the proteins can still be recognized with high performances.

Finally, these results show that the impact of the side-chain flexibility of the surface residues on the SWIM dissimilarity scores is limited: the intra-class dissimilarity scores resulting from the side-chain flexibility remain low, while large conformational motions (the MIP case) greatly increase the dissimilarity scores.

4.2. Evaluation of PLO3S in enrichment

In order to evaluate the PLO3S performance in shape retrieval, we used a benchmarking dataset with 403 protein conformations for 14 protein shape classes. The performance of the PLOS3S method was measured using Precision-Recall and negative predicted value curves. Around a threshold of 62% of the dataset size, the recall of the PLO3S method reaches a value of 92%. The highest value for NPV is around 38% which means that further analysis of the last 38% of the dataset is not relevant. Thus, the dataset size can be reduced by 38% (while discarding less than 10% of the real positive objects, *i.e.* proteins that are indeed surficial homologs to the query). This allows to dramatically reduce the effort required to screen a large dataset in a hierarchical protocol.

The performance of PLO3S in terms of recall and NPV indicates that our method meets its main objective: selecting most of the positive objects, *i.e.* proteins that are indeed surficial homologs to the query, while discarding a large number of negative objects, *i.e.* proteins that are not surficial homologs to the query, with high confidence. This allows to safely decrease the size of very large datasets in a context of protein surface shapes screening.

The average computational time required to compute the SWIMs, the PLO3S descriptors, accounts for most of the time of the method. This preprocessing step is performed only once for a given object and the resulting SWIMs can be stored for later comparisons/screens in a SWIM database describing protein surface shapes. Here, the critical aspect for a large database screening is the computational time required to compute the shape dissimilarity between two protein shapes, which is satisfactory (0.7 seconds in average).

4.3. Comparative evaluation of the performance of PLO3S in enrichment

Spectral geometry through the spectra of the Laplace-Beltrami operator is a common approach in the field of computer vision to compare surfaces [36,50,2,47,8,4,5]. Through spectral geometry, the geometry and topology of a shape is represented by its spectrum, which are the eigenvalues of the Laplace-Beltrami operator. The HKS, WKS and PLO3S descriptors are constructed with the eigenvalues and eigenfunctions of the Laplace-Beltrami operator.

These spectra have multiple properties, the central one being the invariance to isometry [36,50,2]. Invariance to isometry prevents high-amplitude, non-rigid transformations to modify the values of the descriptor, which is an important feature to take into account for protein shape comparison since proteins are dynamic objects that display different conformations.

To our knowledge, no method has been designed to find local surficial similarities of proteins independently of the sequence using the spectra of the Laplace-Beltrami operator. The Surface Wave Interpolated Maps (SWIM) descriptor has been developed on a protein surfaces dataset and is based on the spectra of the Laplace-Beltrami operator through the Wave Kernel Signature (WKS) descriptor. The SWIM descriptor also reduces the dimensions to a 2D space, which reduces the computation time compared to classic spectra-based methods used for 3D objects comparison such as HKS and WKS.

Although SWIM is a 2D descriptor and some information is lost when the dimensions are reduced, the precision-recall curves indicate a better performance of PLO3S compared to the other computer vision methods. Protein surfaces are rough, displaying many variations over their surfaces. Since computer vision methods are designed to be applied on objects that often display a flat surface, a rough protein surface can be considered as a noisy signal which decreases their performance in retrieval. In PLO3S, this problem is overcome by smoothing the surface shape (an average of the points of the 3D surface is projected on the same point on the 2D unit sphere).

For a set of n objects, an all-against-all comparison requires $n \times n$ comparisons, while only n descriptor computations are required (once per object). For this reason, the speed of the descriptors comparison affects the method time to a higher degree than the descriptor computation time. The slow computation of SWIM compared to the computation of HKS and WKS descriptors is not an issue in this context since it is a one-time operation that can be pre-processed. On the contrary, the computation of dissimilarity in PLO3S is 272 times faster than in WKS and 344 times faster than in HKS. This is due to (1) the definition of SWIM in the 2D space, and (2) the creation of a specific data structure that can be manipulated by the GPU with GPGPU. To the best of our knowledge, there is no GPGPU implementation of HKS and WKS.

The performances of PLO3S in enrichment are similar or higher than the reference computer vision methods evaluated in the present work with a faster comparison time to compute the dissimilarity. PLO3S can be used to quickly and efficiently decrease the size of a very large dataset while retaining the proteins that are surficial homologs to the query. Our method can be used for a screening in a big data environment to reduce a protein dataset that can be refined with a finer-grained method in a hierarchical protocol.

PLO3S is a local protein surface comparison method based on a surface descriptor that relies only on the surface shape; therefore, it is independent of the sequence, structure or fold of the protein. It differs from most of the structure-based comparison methods that rely on the atomic 3D coordinates of the main-chain atoms. The lower performance of PLO3S when compared to structure-based comparison method was expected as the ground truth essentially relies on the structural classification of proteins. However, PLO3S displays limited decrease of performance when compared to structure-based comparison methods and the best performance among the other evaluated computer vision methods. Thus, PLO3S represents an alternative approach for cases where the surficial properties are the key point of a study. Therefore, our method complements the structure-based comparison methods in a variety of tasks, such as the search for potential surficial homologs in a drug discovery project or in a classification or annotation process.

4.4. Comparison of 3D surfaces projected on 2D maps

Protein surfaces have been studied for decades but attempts to compare protein surfaces are relatively recent. PLO3S can be assimilated to some methods previously described for protein surfaces comparison that rely on 3D-to-2D projection. However, these methods differ from PLO3S by the choice of the descriptors (such as the Zernike polynomial formalism [30,16]), the methods used for the 3D-to-2D projection [22,46], or by the decomposition of the surface into local patches [12].

Moreover, unlike PLO3S, none of these methods allow for a global, quantitative comparison of the entire surface of proteins. Most focus on the search of local similarities using patches (defined as a group of points of the surface that are close to each other in the 3D space) extracted from the protein surface [12,30,16]. A few methods may compare the whole surface (and any physico-chemical property projected on it), but they use a single-view projection and are thus limited to the comparison of homologous proteins 3D-aligned prior to any computation to allow for a correct comparison [22,46].

While not described here, PLO3S may be adapted to use any physico-chemical property (or any combination of multiple physico-chemical properties) to help the comparison of protein surfaces. One may add a charge value (or any other property of interest) to each point of the 3D mesh prior to the computing the WKS descriptor, in order to easily build a hybrid shape-property descriptor that may increase the PLO3S output. In this work, we focus on a first proof-of-concept that protein surfaces can be compared globally in high throughput manner, *i.e.* they can be used for screening purpose, as PLO3S does not suffer from the limitations of the other methods (patch comparison only or homologous proteins comparison only). Despite the high throughput of PLO3S, when compared to the protein comparison methods from the bio-informatics field, it suffers from a lower performance when considering the true positives only. The design of PLO3S, which renders the comparison step of 3D surfaces ultra-fast, is therefore well-suited to the application we envisioned. It is designed to filter out true negatives from a pre-compiled surface database before applying a more precise and computationally-intensive method.

5. Conclusions

In the present work, we introduced PLO3S that relies on the SWIM descriptor, a 2D representation of the surface topology based on a conformal projection of the protein surface. The SWIM descriptor allies the advantages of being spectrum-based, *i.e.* invariant to isometry and accounting for local surficial features of the shape, and of being in 2D, allowing very fast computation of the shape dissimilarity for the screening of large protein surface datasets. In addition, SWIM is a local descriptor that allows for a partial comparison and can therefore be used to find similarities between protein regions.

The performance of PLO3S in enrichment has been evaluated in a blind comparison of protein surfaces using a subset of the SHREC19 protein shapes dataset. The PLO3S method can be used as a fast, coarse grained protein surface shape screening method that efficiently eliminates proteins displaying dissimilar surfaces to downsize large datasets of protein surface shapes.

Since proteins with a related function often share a similar surface while potentially displaying a low sequence, structure or fold similarity [44,18,27], PLO3S was developed to be used in complement to protein structure comparison methods in different applications where the identification of protein surficial homologs is important such as target fishing for adverse interaction screening or poly-pharmacology in a drug discovery pipeline and protein-protein interactions annotation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Léa Sirugue reports financial support was provided by European Research Council (grant agreement n° 640283). Matthieu Montes reports financial support was provided by European Research Council (grant agreement n° 640283).

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 640283).

References

[1] Angenent S, Haker S, Tannenbaum A, Kikinis R. On the Laplace-Beltrami operator and brain surface flattening. IEEE Trans Med Imaging 1999;18:700–11.
[2] 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); 2011. p. 1626–33.
[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[4] Bronstein AM, Bronstein MM, Kimmel R. Numerical geometry of non-rigid shapes. Springer Science & Business Media; 2008.
[5] Bronstein MM, Kokkinos I. Scale-invariant heat kernel signatures for non-rigid shape recognition. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2010. p. 1704–11.
[6] Canterakis N. 3d Zernike moments and Zernike affine invariants for 3d image analysis and recognition. In: 11th Scandinavian conf. on image analysis. Citeseer; 1999.
[7] Cole R, Vishkin U. Faster optimal parallel prefix sums and list ranking. Inf Comput 1989;81:334–52.
[8] Craciun D, Levieux G, Montes M. Shape similarity system driven by digital elevation models for non-rigid shape retrieval. In: Pratikakis I, Dupont F, Ovsjanikov M, editors. Eurographics workshop on 3D object retrieval, the eurographics association; 2017.
[9] Craciun D, Sirugue J, Montes M. Global-to-local protein shape similarity system driven by digital elevation models. In: IEEE BioSmart; 2017.
[10] Fortune S, Wyllie J. Parallelism in random access machines. In: Proceedings of the tenth annual ACM symposium on theory of computing; 1978. p. 114–8.
[11] Fox NK, Brenner SE, Chandonia JM. Scope: structural classification of proteins—extended, integrating scop and astral data and classification of new structures. Nucleic Acids Res 2014;42:D304–9.
[12] Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein M, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods 2020;17:184–92.
[13] Gao Y, Melonakos J, Melonakos J, Tannenbaum A. Conformal flattening itk filter. Available from: https://doi.org/10.54294/msr7a5, 2006.
[14] Gao Z, Rostami R, Pang X, Fu Z, Yu Z. Mesh generation and flexible shape comparisons for bio-molecules. Comput. Math. Biophys. 2016;4.
[15] Gramada A, Bourne PE. Multipolar representation of protein structure. BMC Bioinform 2006;7:1–13.
[16] Grassmann G, Miotto M, Di Rienzo L, Gosti G, Ruocco G, Milanetti E. A novel computational strategy for defining the minimal protein molecular surface representation. PLoS ONE 2022;17:1–17.
[17] Guzenko D, Burley SK, Duarte JM. Real time structural search of the protein data bank. PLoS Comput Biol 2020;16:e1007970.
[18] Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. PLoS Comput Biol 2019;15:e1006969.
[19] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 1992;89:10915–9.
[20] Johnson AE, Hebert M. Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Trans Pattern Anal Mach Intell 1999;21:433–49. https://doi.org/10.1109/34.765655.
[21] Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J. Molecular surface representation using 3d Zernike descriptors for protein shape comparison and docking. Curr Protein Pept Sci 2011;12:520–30.
[22] Kontopoulos DG, Vlachakis D, Tsiliki G, Kossida S. Structuprint: a scalable and extensible tool for two-dimensional representation of protein surfaces. BMC Struct Biol 2016;16.
[23] La D, Esquivel-Rodríguez J, Venkatraman V, Li B, Sael L, Ueng S, et al. 3d-surfer: software for high-throughput protein surface comparison and analysis. Bioinformatics 2009;25:2843–4.
[24] Langenfeld F, Axenopoulos A, Benhabiles H, Daras P, Giachetti A, Han X, et al. Protein shape retrieval contest. In: Biasotti S, Lavoué G, Veltkamp R, editors. Eurographics workshop on 3D object retrieval, the eurographics association; 2019.
[25] Langenfeld F, Peng Y, Lai YK, Rosin PL, Aderinwale T, Terashi G, et al. SHREC 2020: multi-domain protein shape retrieval challenge. Comput Graph 2020;91:189–98. https://doi.org/10.1016/j.cag.2020.07.013.
[26] Li C, Hamza AB. Spatially aggregating spectral descriptors for nonrigid 3d shape retrieval: a comparative survey. Multimed Syst 2014;20:253–81.
[27] Machat M, Langenfeld F, Craciun D, Sirugue L, Labib T, Lagarde N, et al. Comparative evaluation of shape retrieval methods on macromolecular surfaces: an application of computer vision methods in structural bioinformatics. Bioinformatics 2021.
[28] Mak L, Grandison S, Morris RJ. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. J Mol Graph Model 2008;26:1035–45.
[29] Mavridis L, Ritchie DW. 3d-blast: 3d protein structure alignment, comparison, and classification using spherical polar Fourier correlations. In: Biocomputing 2010. World Scientific; 2010. p. 281–92.
[30] Milanetti E, Miotto M, Di Rienzo L, Monti M. 2d Zernike polynomial expansion: finding the protein-protein binding regions. Comput Struct Biotechnol J 2021;19:29–36.

[31] Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 5115–24.
[32] Mylonas SK, Axenopoulos A, Daras P. Deepsurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. Bioinformatics 2021;37:1681–90.
[33] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
[34] Papadakis P, Pratikakis I, Theoharis T, Perantonis S. Panorama: a 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. Int J Comput Vis 2010;89:177–92.
[35] Pawłowski K, Godzik A. Surface map comparison: studying function diversity of homologous proteins. J Mol Biol 2001;309:793–806.
[36] Reuter M, Wolter FE, Peinecke N. Laplace-spectra as fingerprints for shape matching. In: Proceedings of the 2005 ACM symposium on solid and physical modeling. ACM; 2005. p. 101–6.
[37] Reuter M, Wolter FE, Peinecke N. Laplace–Beltrami spectra as 'shape-DNA' of surfaces and solids. Comput Aided Des 2006;38:342–66. https://doi.org/10.1016/j.cad.2005.10.011.
[38] Rubner Y, Tomasi C, Guibas LJ. A metric for distributions with applications to image databases. In: Sixth international conference on computer vision (IEEE cat. no. 98CH36271). IEEE; 1998. p. 59–66.
[39] Ruiz Echartea ME, Chauvot de Beauchêne I, Ritchie DW. Eros-dock: protein–protein docking using exhaustive branch-and-bound rotational search. Bioinformatics 2019;35:5003–10.
[40] Rusu RB, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3d registration. In: 2009 IEEE international conference on robotics and automation; 2009. p. 3212–7.
[41] Rusu RB, Blodow N, Marton ZC, Beetz M. Aligning point cloud views using persistent feature histograms. In: IEEE/RSJ international conference on intelligent robots and systems, 2008. IEEE; 2008. p. 3384–91.
[42] Rusu RB, Bradski G, Thibaux R, Hsu J. Fast 3d recognition and pose using the viewpoint feature histogram. In: 2010 IEEE/RSJ international conference on intelligent robots and systems; 2010. p. 2155–62.
[43] Sael L, La D, Li B, Rustamov R, Kihara D. Rapid comparison of properties on protein surface. Proteins, Struct Funct Bioinform 2008;73:1–10.
[44] Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, et al. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins, Struct Funct Bioinform 2008;72:1259–73.
[45] Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. Patchdock and symmdock: servers for rigid and symmetric docking. Nucleic Acids Res 2005;33:W363–7.
[46] Schweke H, Mucchielli MH, Chevrollier N, Gosset S, Lopes A. SURFMAP: a software for mapping in two dimensions protein surface features. J Chem Inf Model 2022;62:1595–601.
[47] Shi Y, Lai R, Wang DJ, Pelletier D, Mohr D, Sicotte N, et al. Metric optimization for surface analysis in the Laplace-Beltrami embedding space. IEEE Trans Med Imaging 2014;33:1447–63.
[48] Shilane P, Min P, Kazhdan M, Funkhouser T. The Princeton shape benchmark. In: Shape modeling applications, 2004, proceedings. IEEE; 2004. p. 167–78.
[49] Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Eng 1998;11:739–47.
[50] Sun J, Ovsjanikov M, Guibas L. A concise and provably informative multi-scale signature based on heat diffusion. In: Computer graphics forum. Wiley Online Library; 2009. p. 1383–92.
[51] Tombari F, Salti S, Di Stefano L. Unique shape context for 3d data description. In: Proceedings of the ACM workshop on 3D object retrieval. ACM; 2010. p. 57–62.
[52] Wang S, Ma J, Peng J, Xu J. Protein structure alignment beyond spatial proximity. Sci Rep 2013;3:1–7.
[53] Xu D, Zhang Y. Generating triangulated macromolecular surfaces by Euclidean distance transform. PLoS ONE 2009;4:e8140.
[54] Yin S, Dokholyan NV. Fingerprint-based structure retrieval using electron density. Proteins, Struct Funct Bioinform 2011;79:1002–9. https://doi.org/10.1002/prot.22941.
[55] Yin S, Proctor EA, Lugovskoy AA, Dokholyan NV. Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci 2009;106:16622–6.
[56] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins, Struct Funct Bioinform 2004;57:702–10.
[57] Zhang Y, Skolnick J. Tm-align: a protein structure alignment algorithm based on the tm-score. Nucleic Acids Res 2005;33:2302–9.
[58] Zheng WM, Liu X. A protein structural alphabet and its substitution matrix clesum. In: Transactions on computational systems biology II. Springer; 2005. p. 59–67.