An efficient statistical methodology for peptide 3D structure clustering

Azzam Alwan, Rémi Cogranne, Pierre Beauseroy, Edith Grall-Maës, Laurent Debelle, Nicolas Belloy, Stéphanie Baud, Manuel Dauchez, Sébastien Almagro

An efficient statistical methodology for peptide 3D structure clustering

Azzam Alwan*, Rémi Cogranne*, Pierre Beauseroy* , Edith Grall-Maës*,
Laurent Debelle†, Nicolas Belloy†, Stéphanie Baud†, Manuel Dauchez†, and Sébastien Almagro†,
*Troyes University of Technology, UMR CNRS 6281 ICD/ROSAS/LM2S,10004 Troyes France.
Email adresses: {firstname.last}@utt.fr
†Univ. of Reims Champagne-Ardenne, UMR CNRS 7369 MEDyC,51100 Reims France.
Email adresses: {firstname.last}@univ-reims.fr

***Abstract*—The analysis of proteins and peptides conformations is of crucial interest to gain insights on their biological functions; it has therefore been an active research topic over the past decades. However, analyzing conformations of small size and highly flexible peptides remains a challenge due to their instability and their large number of different shapes. In this paper, an efficient methodology is proposed to analyze 3D structures of highly flexible elastin-derived peptides and to find out their principal conformations using a clustering algorithm. This methodology is based on a special representation of peptide structures, which has the great advantage to be neither affected by peptides' translations nor rotations, hence, avoiding the use of a complex superposition method. In addition, the proposed approach uses for the first time Kernel PCA to remove outlier structures that are not frequently present and do not resemble any other peptide structures. Outlier removal is very important in this context because, due to the instability of those peptides, a small portion of very different conformations, that seldom occur, can heavily affect the ensuing clustering results. Finally, the proposed approach latest step consists in hierarchical clustering, used as a non-supervised classification method to gather together similar structures. Experimental results, obtained using an existing database, show the relevance and the efficiency of the proposed method.**

***Index Terms*—Clustering, flexible peptide structure, Hierarchical classification, Outliers detection, Kernel PCA**

I. Introduction

According to the World Health Association [1], cardiovascular diseases are the first cause of death: more people die every year due to cardiovascular problems than to any other health issue. One of the main actors of the vascular system integrity is Elastin, a polymer of a protein called tropoelastin. It confers resilience to the artery wall and its degradation, generated by aging and/or diseases (such as diabetes or atherosclerosis), is one of the symptoms/causes of such diseases. This degradation produces small fragments of protein called peptides. These peptides could be considered as circulating "molecular signals" that potentially activate response from organs and cells. The activity of a peptide is related to its three-dimensional structure and consequently to its physico-chemical properties. Therefore, for a better understanding of cardiovascular aging, it is crucial to identify among elastin peptides, the repetitive conformations, that may be related with their biological functions. These repetitive sequences, and their three-dimensional (3D) structures, are frequently observed in trajectories of molecular dynamics simulations. From examination of trajectories, it appears that these repetitive shapes may lead to detect a specific signature of a "key element" in molecular signaling. The aim of this first approach is to automatically identify, in this very large set of potential molecules, redundant molecular structures that will allow us to discriminate some specific 3D structures which could be related to the "active peptides" produced by the degradation of elastin.

Based on peptides' characteristics, many methods have been proposed to identify repetitive conformations of peptides with the goal to understand their functionality. These studies could be divided into two main categories. The first one relies on the peptide sequence [2]–[4]. Most of the methods from this category extract biological properties from peptide sequences and then use them in machine learning to determine some key functions. Actually, methods from this category are not sufficient to deduce peptide main functions because the functionality is related to both structures and thermodynamics. This especially explains why two peptides with different sequences may, or may not, have similar function depending on whether they may present the same conformation during a dynamic behavior. On the other hand, the second category, more closely related with the present paper, relies on the three-dimensional geometric structure of peptides [5]–[12]. The majority of methods in this category aim at solving the problem of comparing several protein structures by improving the methods of structures alignment e.g DALI and SSAP [6], [8]. These two methods use the distance matrix to represent each protein structure. They have the major advantage of comparing different proteins with different sizes. Despite its accuracy and sensitivity, this method is still very computationally expensive because it explores all the possibilities of comparisons. Thus, as we will treat structures having the same size, these methods are not convenient for us as we cannot profit from their main advantages. We also have to keep in mind that elastin peptides are very specific (82% the tropoelastin sequences are composed of 5 amino-acids, with numerous repetitive patterns) and lead to the elasticity of the tissues, faraway from globular domains or transmembrane structures found in proteins with

regular secondary structures (i.e. helices, strands or sheets and turns). Recently, several state-of-the-art works proposed to model the protein structure as a continuous parameterized curves rather than a sequence of discrete points [5], [7]. Although the efficiency of these methods, there is a risk of losing information related to the amino acid side chain due to the function used to interpolate between the discrete points to obtain the curves. Beside that, another approach was suggested and consists in using Self-Organizing Maps (SOM) method that tends to replace protein structures by neurons vectors [9], then to submit it to a hierarchical clustering to extract the representative conformations. The only inconvenient of this method is that the SOM is very sensitive to the intuitive of its parameters which can significantly affect the results. Furthermore, a new approach is proposed in [12]. It is based on the data density function. The authors use the root-mean-squared deviation (RMSD) as a distance metric for clustering. Additionally, the suggested method was applied on a protein with relatively well-understood dynamics, but even so, did not return the expected result for the number of clusters. Thus, as our objective is to observe the relation between the different clusters in the peptide dynamics and to avoid the alignment between structures as it is computationally expansive, we decided not to take this method into account. Indeed, all the clustering methods, that are mentioned previously, are only applied on the C^{α} of the protein structures. Implicitly it leads to the conclusion that the side chains have no effect on the conformation variation, and that was proven wrong in the scope of our work. For those reasons, prior state-of-the-art methods are not adapted to tackle our problem, that is, to find the main patterns of peptides which are so flexible and are not regular as the protein treated in the previews methods.Thus the main contributions of the present paper are the following:

1. The proposed method includes a highly efficient algorithm to detect outliers conformations (that is, relies on a first step that aims at detecting outliers conformation) in order not to take into account the non-repetitive structures which may heavily impact the results of clustering.
2. An accurate description of the similarity between the main conformations can be easily obtained and illustrated after the automatic classification.
3. A novel approach combining two statistical methods is proposed to enhance the performance of the clustering method and to make it simpler and easier.

The rest of the paper is organized as follows. The problem formulation is presented in Section 2. In Section 3, the proposed Methodology is described. Experimental results evaluating the efficiency of the proposed method are presented in Section 4. Finally, conclusions are drawn in Section 5.

II. Problem formulation

This section formally states the problem of identifying patterns that frequently occur among peptide structures before presenting the proposed method to address this problem.

A peptide is essentially a set of related atoms in space. The three-dimensional position of the N atoms that compose a peptide will be referred to as the structure of the peptide and will be noted as $\mathcal{S}$. Of course, such structure is a set of N three positions, or vectors : $\mathcal{S} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)^T$ with $\mathbf{a}_n \in \mathbb{R}^3$ the position of n-th atom. This paper studies sequence over time of peptide structures since, as for any molecule, the atom move relatively to each other under thermodynamical influence. Let us denote $\mathbb{S}$ the inspected sequence of T structures, that is $\mathbb{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_T)$.

As described in the introduction, the goal of the present paper is to propose a method for automatically finding, without manual intervention, conformations that occur the most frequently within $\mathbb{S}$ a sequence of peptide structures. Formally speaking, two structures $\mathcal{S}_t$ and $\mathcal{S}_u$ have exactly the same conformation if and only if there exist a rotation, characterized by matrix $\mathcal{R}$, and translation, characterized by vector $\boldsymbol{\zeta} \in \mathbb{R}^3$, such that

$$\mathcal{S}_u = \mathcal{S}_t\mathcal{R} + \mathbf{1}_N\boldsymbol{\zeta}^T \tag{1}$$

with $\mathbf{1}_N$ a vector of N row containing only ones. And consequently the conformation $\mathbb{F}$ can be considered as the set of all structures that have the same conformation:

$$\mathbb{F} = \left\{\mathcal{S} | \mathcal{S}_u = \mathcal{S}_t\mathcal{R} + \mathbf{1}_N\boldsymbol{\zeta}^T\right\} \tag{2}$$

with $\mathcal{S}_u$ the structure from which the conformation is defined, or any instance of structure in the set $\mathbb{F}$. The main conformations of a peptide sequence $\mathbb{S}$ are the ones in which falls the higher number of structures from sequence $\mathbb{S}$. However, it is obvious that because the moves of each atom is somewhat random, finding two structures with exactly the same conformation, as defined in Eq. (2), is very unlikely. Because we are interested in finding structures that are associated with the same function as the peptides, one can define a conformation in a wider sense as:

$$\mathbb{F} = \left\{\mathcal{S} | \mathrm{distance}(\mathcal{S}_u, \mathcal{S}_t\mathcal{R} + \mathbf{1}_N\boldsymbol{\zeta}^T) < \varepsilon\right\} \tag{3}$$

where ε defines the upper bound on distance from structure $\mathcal{S}_u$ that is allowed to consider a structure $\mathcal{S}$ as having the same conformation; ε is thus the maximal distance that affects the function of peptides.

Based on the "practical" definition (3) of the conformation, it is obvious that any clustering algorithm should assign all the observations, or structures, from the inspected sequence $\mathbb{S}$, with the conformation they fall within, if any. Structures that do not fall within any main conformation can be considered as outliers, those can typically occur when changing from one main conformation to another, also referred to as "transitional conformations". Let us denote $L(t)$ the label of structure $\mathcal{S}_t$ and $\{\mathbb{F}_1, \mathbb{F}_2, \ldots, \mathbb{F}_K\}$ the set of all K main conformations; the described assignment rule can be defined formally as follows:

$$\begin{cases} L(t) = k \;\text{ if }\; \mathcal{S}_t \in \mathbb{F}_k \\ L(t) = 0 \;\text{ if }\; \forall k \in 1, \ldots, K\,,\, \mathcal{S}_t \notin \mathbb{F}_k \\ L(t) = 0 \;\text{ if }\; \mathcal{S}_t \in \mathbb{F}_k \; and \; \mathcal{S}_t \in \mathbb{F}_l \end{cases} \tag{4}$$

The Equations (1) - (4) state the core problems of this paper, that is defining the K main conformations from a sequence of peptide structures $\mathbb{S}$, and also allow to formalize the main

difficulties addressed in this work. The first one is due to the alignment between the peptide structures, see Equations (1) - (3), which is required to defined conformations independently from the rotation and translation of peptide structures. However, this is computationally expensive and is not always reliable in our purpose. Hence, the proposed method must avoid using structure alignments but must carefully take into account rotation and translation properly. The second difficulty is the unknown number of outliers within the sequence of structures $\mathbb{S}$. Moreover, the definition of outliers (3) is itself fuzzy, this is due to the fact that the distance that defines a conformation is not known and may changes for various classes. The proposed approach must be able to avoid the effects of the structures that do not resemble to all conformations without knowing how to set the tolerance of similarity. The third difficulty is the unknown number of the main conformations in the sequence $\mathbb{S}$. Indeed, several methods exist to perform clustering without any prior knowledge of the number of clusters [13], [14]. However, it is not enough for us as we are interested in studying the similarities between the different main conformations. Thus, the proposed method must be flexible allowing thus the user to visualize the different clusters found, without additional computational costs, and to adjust the results manually when user needs to divide a conformation into two or more sub-classes

III. METHODOLOGY

This section presents the proposed methodology for peptides clustering that achieves the goals laid out in the previous section. In the first place, regarding the difficulty that is related to the rotation and translation effects of structures in space, it is proposed to represent each peptide structure as a distance matrix $\mathcal{M}$, which characterizes the distance between atoms of structure [15]. Consequently, the rotation and translation effects are automatically overcome. Taking the same set of peptide structures defined in the previous part which is composed of N atoms and has *T* structures. The "distance matrix" $\mathcal{M}_{\mathcal{S}_i}$ of a structure $\mathcal{S}_i$ is defined as follows:

$$\mathcal{M}_{\mathcal{S}_i}(k,l) = \|\mathbf{a}_k - \mathbf{a}_l\|_2 , \tag{5}$$

where $\|\mathbf{v}\|_2$ stands for the Euclidean norm of vector $\mathbf{v}$.
The matrix $\mathcal{M}_{\mathcal{S}_i}$ is considered as an observation in the so-called "conformational space". The dimension of this space changes according to the number of atoms in structure. Thus, for the current peptide, the dimension of its space is equal to $N(N-1)/2$, due to the symmetry of $\mathcal{M}$. In the second place, as it is noted previously, it exists structures that can be considered as outliers because of their unique conformations and can have effects on the clustering results. Hence an automatic outlier detection algorithm is proposed to detect these outliers. Moreover, this algorithm gives the possibility to the user to adjust the detection accuracy in order to ensure his clustering results. Finally, the next step is to use a clustering method on the remaining data to detect the principal conformations and to study the relations between them. Thus, an algorithm that does not need a prior knowledge of the clusters' number has been proposed. It is able to show the relations between the different structures and it is intelligent enough to adjust easily the number of clusters according to the data and the requirements.
In the following part, the method that is used to ignore the transitional conformations (outliers) will be described in Section III-A. Then in Section III-B, the clustering method that is used to classify the data will be explained.

A. Outliers detection

To detect properly the main conformations, the transitional conformations which can be found between them must be eliminated from the data. Recently, Kernel PCA has been investigated for outliers detection, and it has shown a high performance. It will be briefly described in the following subsection.

1) Kernel PCA: Kernel PCA is an extension of PCA to deal with the data non-linearity. Recently this method has been used for outliers detection. It demonstrates a competitive performance as compared to other outliers detection methods [16]–[18]. Its concept is to map the data from the original space to a higher dimensional feature space $\mathcal{F}$ through a nonlinear transformation $\mathbf{\Phi}$, then the PCA will be applied in this space and a reconstruction error $\mathbf{Recc}$ will be calculated for each observation as an index of outliers detection. In the linear case, for a point $\mathbf{x}_i \in \mathbb{R}^z$, $\mathbf{Recc}$ is performed as follows :

$$\mathbf{Recc} = \|\mathbf{x}_i - \mathcal{V}_q \mathcal{V}_q^T \mathbf{x}_i\|_2 \tag{6}$$

where $\mathcal{V}_q$ is the matrix containing the q principal components, $\mathcal{V}_q \mathcal{V}_q^T \mathbf{x}_i$ is the estimated point after reconstruction, and q represents the number of the principal components. In $\mathcal{F}$, Equation (6) is achieved as following

$$\mathbf{Recc} = (\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_i)) - (\mathcal{W}\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathcal{W}\mathbf{\Phi}(\mathbf{x}_i)) \tag{7}$$

where $\mathcal{W}$ is a matrix contains the q eigenvectors that are obtained from Kernel PCA. For more details about this treatment see [18].
Indeed, the principal components are affected directly by the main clusters that constitute the majority of the data. In other words, the observations that belong to these clusters are closer to the principal components than those which are outside of them. By consequence, the main observations have a low reconstruction error values and the outliers have the greatest reconstruction error values. To employ this method, two parameters have to be optimized; the number of eigenvectors q used for PCA and the kernel parameter (σ in the case of Gaussian kernel), $\mathcal{K}(x_i, x_j) = exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)$. The choice of these parameters will be discussed in Section IV-B

B. Clustering Analysis

Hierarchical clustering is one of the popular clustering algorithms [19]. This method has been selected due to its flexibility and its accuracy. In this method it is not necessary to define the number of cluster beforehand. Since our goal is to study the data dispersion as well as to find out the main conformation, hierarchical clustering provides an accurate description of the

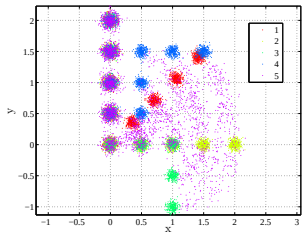


Fig. 1: Illustration of the 4 main conformations in blue, red, green and yellow. The transitions are shown in purple.

data structures by using a dendrogram which represents similarity within the data on different levels. Moreover, by cutting the dendrogram at different levels, different clustering could be found with different numbers of clusters. To understand how it works, and for more details, one can refer to [20].

IV. Experiments and results

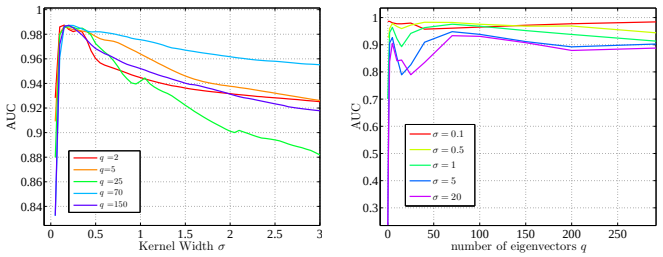
This section describes the data that are used in the experiments, discusses the selection of the proposed method's parameters (q, σ), evaluates the proposed method and finally shows the results obtained with the real data set of peptides.

A. Data sets

In this paper, two data sets have been used. The first, *DB1*, is a real data set. It consists of sequences of 3D related points. They represent the spatial configuration of the atoms composing the various peptides from elastin. Each peptide is composed of N atoms and $40\,000$ samples have been considered. Each sample correspond to a specific 3D position of all the atoms at sampling time with fixed to 200ns in molecular dynamics trajectory simulations.
The second data set, *DB2*, has the same features of peptides in order to evaluate our approach. It is a simulated dataset that is formed of 9 atoms with each 1300 positions for 4 original conformations, and 350 observations for the transitional conformations. These 350 samples should be removed by the outlier detection step. The dynamics of atoms is assumed to be Gaussian noise as illustrated in Fig. 1.

B. Common core for all experiments

In this section, the selection of the proposed method's parameters is discussed. As mentioned previously, two main parameters have to be optimized, the number of eigenvectors, q, used in kernel PCA, and the Gaussian Kernel standard deviation σ. In fact, we are in unsupervised case. Hence, in order to evaluate the proposed approach and to optimize the parameters, one has to use methods that must be independent of the number of outliers and clusters. A list of these methods has been presented in [21]. Since there is no information regarding the real data set, it is proposed to follow the method that consists to apply our algorithm on a simulated data (*DB2*), where the number of outliers and clusters is known, and study its capacity to detect the outliers. To evaluate detection performance the ROC (Receiver operating characteristic) curve has been used; this ROC curve shows the pobability of correct outlier detection against the probability of misclassifying one



(a) The effects of σ (b) The effects of q

Fig. 2: AUC criterion in respect to : σ in (a) and q in (b) with various values of q and σ respectively.

sample from main conformations as an outlier. However, to avoid having a curve for each setting, we reduce the evaluation to the Area Under the Curve (AUC) [22] which summarize the performance of a binary-detector. The effects of σ and q on the outliers detection are shown in Fig. 2a and Fig. 2b respectively. It can be seen in Fig. 2a that for a small value of σ , Kernel PCA has a good performance for almost all the q values used. When the value of σ increases, the AUC value is decreased. That is why a large value of σ is not recommended. Despite that, the optimal value of σ is still unknown. In addition, by observing the yellow and red curves in Fig. 2b, that correspond to small σ values, it can be seen that these curves are rather stable and q are very similar in terms of AUC values. Otherwise, when σ is far from its reasonable value which is estimated to be small, the results are not conclusive. Accordingly, it is not needed to assign a large value for q to have a good performance as kernel PCA could achieve it with a small value. Furthermore, the optimal value of q also depends on the data inertia rate, which represents the rate of information in the data after performing Kernel PCA. Fig. 3 presents q versus σ for four different inertia rates. In order to observe the behavior of Kernel PCA for different inertia rates, the coordinates of points of each curve in Fig. 3 are considered as parameters for Kernel PCA in order to calculate the detection and false alarm probabilities. Looking at the illustrations in Fig. 4, it is clear that the circled zone (where β has its highest values and υ has its lowest values) has a better performance than the other regions. By comparing that to the curves in Fig. 3, we see that this zone corresponds to the inflection points. Consequently, these illustrations help us to estimate the two optimal parameters. σ has a range of optimal values illustrated in Fig. 3. If it has a small value, we discretize the data, in the sense that each point becomes an eigenvector, resulting in a poor representation of the data variance. Furthermore, if the value of σ is large, the kernel PCA becomes meaningless, as explained in [18]. Meanwhile, increasing the value of q has no influence on the outliers detection. Hence, it is preferable to choose a low value for q, because with large values of q, the Kernel PCA treatment becomes time consuming.

C. Real data

After performing the proposed algorithm on the simulated data $(DB2)$ and finding the procedure to determine the optimal

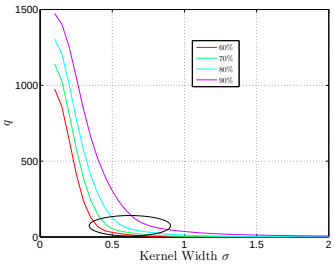
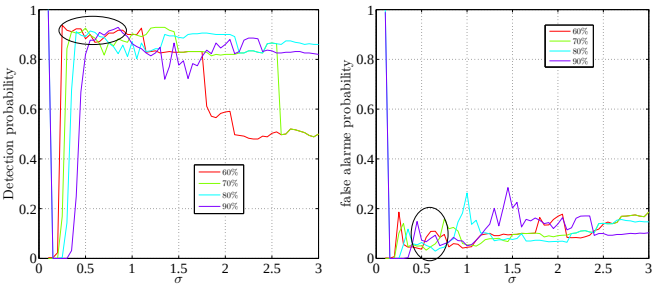


Fig. 3: Number of eigenvectors q versus σ for four different inertia rates.



(a) Detection probability (β) (b) False alarm probability (υ)

Fig. 4: Probability of detection and false alarm for various inertia rates. Taking the point coordinates of each curve in Fig. 3 as parameters for the kernel PCA algorithm. Each step in the abscissa corresponds to a σ value and its corresponding q value on the curves of Fig. 3.

parameters, we will apply our proposed approach on the target data (*DB1*) to see if it fulfills our expectations. Since the data are not labeled, it is not possible to recognize outliers from clusters. However, by looking at the available data of one of the peptides, we find 5 000 successive observations that can be classified into 3 clusters (We will use this amount of peptide structures only for clarity, otherwise we can apply our method on all the 320 000 structures). That can be concluded from the representation of the similarities between them. Knowing that these 5 000 observations can be classified into 3 clusters, it is suggested to apply the proposed classification method on them and to find out the optimal parameters. However, in this case, we cannot use the ROC curve to evaluate the results because the number of outliers is unknown. Hence, it is proposed to use a method that consists in exploiting the knowledge of the class number in these 5 000 observations, and to use the stability of the class center positions, after clustering, as an evaluation criterion. Then, to determine the optimal parameters (σ, q), we apply the following steps: first, we illustrate the curve that represents the number of eigenvectors q versus σ to have a range of reasonable values for σ, as shown in Fig. 3. Next, we take a value for σ from this range and then we use our method with it and with various values of q. In order to ignore the transitional forms, it is proposed to consider 20% of the data as outliers. Indeed, our main objective is to find the main forms and not to know how many outliers there are (it will be studied in a extension of this paper). So, even if we consider such a number of observations as outliers, it is not a problem

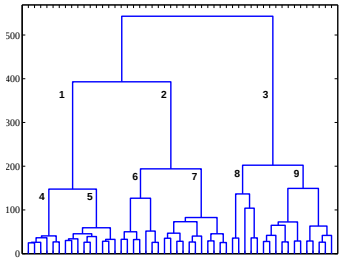
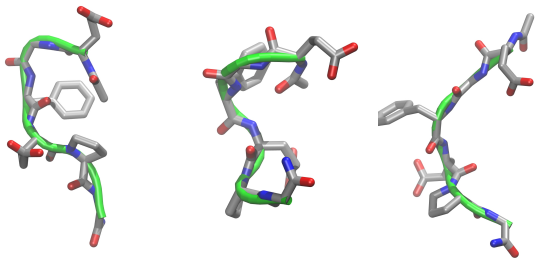


Fig. 5: Dendrogram obtained after applying the hierarchical clustering on the 4 000 observations without Outliers.



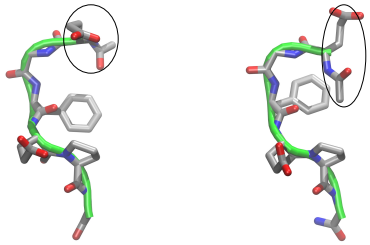
(a) center 1 (b) center 2 (c) center 3

Fig. 6: The three main peptide's structures that result from cutting the dendrogram in (Fig. 5) into three clusters.

for us in this approach, and it can be modified according to the requirements and the domain expert's demands. In our case, this percentage can be justified from a biological point of view. Due to space limitation, we will not illustrate the variation of the three centers positions of the three clusters in function of q. Briefly, after $q = 12$ the positions of the three centers become stable which means that we do not need more than 12 eigenvectors to perform our method. We applied this simulation for different values of σ from the values' range that is noted above and the same results are obtained. That is why, we adopt σ equals 5, and q equals 12 for our algorithm.

D. Results

In this paper, only a restricted number of cases have been presented. To show the performance of the proposed method, the Kernel PCA is applied on the 5 000 observations (From *DB1*) that are used to determine the two parameters, with $\sigma = 5$ and $q = 12$. Then the 1 000 observations that have the greatest values of **Recc** are considered as outliers, and they are eliminated from the data. Next, hierarchical clustering is employed on the 4 000 remaining observations to find out their main conformations. The dendrogram in Fig. 5 shows clearly that these latter observations can be divided into 3 clusters. We used CH (Calinski and Harabasz) as index for clustering validity [23]. Fig. 6 shows the 3 main conformations present in these observations. These three structures are chosen to be the closest 3D structures to the three resulting clusters' centers. To ensure that this dendrogram reflects accurately the similarity between structures, we divided each main cluster to two sub-clusters, then we compared the difference between them in 3D dimension and in theoretical measure. Fig. 7 shows the



(a) subcenter 4 (b) subcenter 5

Fig. 7: The structures which correspond to the centers of sub-clusters 4 and 5 in Fig. 5.

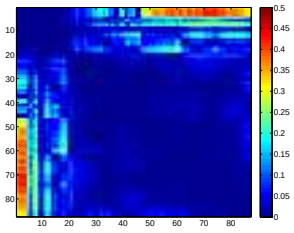


Fig. 8: Difference between the distance matrices $\mathcal{M}$ of the centers of th sub-clusters 4 and 5. Blue color means low values and red color high values

difference in 3D between the centers of classes that have the number 4 and 5 in Fig. 5. In other hand, Fig. 8 illustrates the difference between the distance matrices $\mathcal{M}$ which belong to them. We can see clearly that these two figures lead to the same results that the only first 20 atoms at the top, which are side chains, have changed their directions and the others atoms have been conserved their forms.

V. Conclusions

This paper studies the clustering of 3D structures of highly flexible elastin-derived peptides computed from numerical molecular dynamics simulations. It introduces a new strategy that combines different statistical methodologies to detect the main conformations of a peptide during a period of time. It proposes to use the distance matrix as a representation for the peptide's structure to avoid the effects of both translations and rotations over time. Moreover, it employs for the first time Kernel PCA to detect outliers conformations that do not resemble to any others. The removal of such outliers has been shown to be very efficient as it ensures a classification that is not impacted by those points that essentially lie between most frequently found conformations. Using Hierarchical Agglomerative Clustering method on the remaining data is eventually proposed to ensure a large flexibility in the classification of different conformations. Indeed by setting the threshold, one is free to choose between a coarse and a fine grain classification and is also able to visualize the difference between cluster centers to help tune the method according to the data and the requirements. Experimental results, performed using an existing database and simulated data, have shown that the proposed method is valid and efficient.

References

[1] W. H. Organization, *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization, 2009.

[2] S. Bandyopadhyay, "An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection," *Fuzzy Sets and Systems*, vol. 152, no. 1, pp. 5–16, 2005.

[3] J. cheol Jeong, X. Lin, and X.-W. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 8, no. 2, pp. 308–315, 2011.

[4] S. Vipsita, B. K. Shee, and S. K. Rath, "An efficient technique for protein classification using feature extraction by artificial neural networks," in *2010 Annual IEEE India Conference (INDICON)*. IEEE, 2010, pp. 1–5.

[5] W. Wu, A. Srivastava, J. Laborde, and J. Zhang, "An efficient multiple protein structure comparison method and its application to structure clustering and outlier detection," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. IEEE, 2013, pp. 69–73.

[6] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of molecular biology*, vol. 233, no. 1, pp. 123–138, 1993.

[7] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[8] C. A. Orengo and W. R. Taylor, "Ssap: sequential structure alignment program for protein structure comparison," *Methods in enzymology*, vol. 266, pp. 617–635, 1996.

[9] D. Fraccalvieri, A. Pandini, F. Stella, and L. Bonati, "Conformational and functional analysis of molecular dynamics trajectories by self-organising maps," *BMC bioinformatics*, vol. 12, no. 1, p. 158, 2011.

[10] G. Bottegoni, W. Rocchia, M. Recanatini, and A. Cavalli, "Aclap, autonomous hierarchical agglomerative cluster analysis based protocol to partition conformational datasets," *Bioinformatics*, vol. 22, no. 14, pp. e58–e65, 2006.

[11] L. V. Nedialkova, M. A. Amat, I. G. Kevrekidis, and G. Hummer, "Diffusion maps, clustering and fuzzy markov modeling in peptide folding transitions," *The Journal of chemical physics*, vol. 141, no. 11, p. 09B611_1, 2014.

[12] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, "Persistence-based clustering in riemannian manifolds," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 41, 2013.

[13] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[14] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[15] A. E. Torda and W. F. van Gunsteren, "Algorithms for clustering molecular dynamics configurations," *Journal of computational chemistry*, vol. 15, no. 12, pp. 1331–1340, 1994.

[16] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. C. Platt *et al.*, "Support vector method for novelty detection." in *NIPS*, vol. 12, 1999, pp. 582–588.

[17] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern recognition letters*, vol. 20, no. 11, pp. 1191–1199, 1999.

[18] H. Hoffmann, "Kernel pca for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863–874, 2007.

[19] A. D. Gordon, "A review of hierarchical classification," *Journal of the Royal Statistical Society. Series A (General)*, pp. 119–137, 1987.

[20] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.

[21] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[22] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[23] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, and M. M. Charrad, "Nbclust: An r package for determining the relevant number of clusters in a data set," *J. Stat. Soft*, vol. 61, pp. 1–36, 2014.