A Fully Automatic and Efficient Methodology for Peptide Activity Identification Using Their 3D Conformations

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

▶ **To cite this version:**

Azzam Alwan, Rémi Cogranne, Pierre Beauseroy, Edith Grall-Maës, Nicolas Belloy, et al.. A Fully Automatic and Efficient Methodology for Peptide Activity Identification Using Their 3D Conformations. IEEE Access, 2021, 9, pp.92143 - 92156. 10.1109/access.2021.3091939 . hal-03311364

HAL Id: hal-03311364
https://hal.science/hal-03311364v1

Submitted on 31 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 21, 2021, accepted June 7, 2021, date of publication June 23, 2021, date of current version July 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3091939

A Fully Automatic and Efficient Methodology for Peptide Activity Identification Using Their 3D Conformations

AZZAM ALWAN[1,2], RÉMI COGRANNE[1], (Member, IEEE), PIERRE BEAUSEROY[1], (Member, IEEE), EDITH GRALL-MAËS[1], NICOLAS BELLOY[3], LAURENT DEBELLE[3,4], STÉPHANIE BAUD[3], MANUEL DAUCHEZ[3], AND SÉBASTIEN ALMAGRO[3]

[1]Research Unit LIST3N, M2S laboratory, Troyes University of Technology, 10004 Troyes, France
[2]Relyfe Europe, 51100 Reims, France
[3]UMR CNRS 7369 MEDyC, Université of Reims Champagne-Ardenne, 51100 Reims, France
[4]The University of Manchester, School of Biological Sciences, Division of Cell Matrix Biology and Regenerative Medicine, Manchester M13 9PL, U.K.

Corresponding author: Rémi Cogranne (remi.cogranne@utt.fr)

ABSTRACT Over the past decades, the understanding of peptides and proteins biological functions has been an active research topic. Latest research works in this field have suggested that protein conformations may be a key feature for gaining insights into protein biological functions. However, analyzing small and highly flexible protein chunks, namely oligopeptides made of a handful of amino acids, remains challenging because of their dynamics and wide range of conformations. In this paper, a statistical methodology based on unsupervised statistical learning is proposed for analyzing 3D conformations small and highly flexible elastin-derived peptides. The goal of this study is twofold: first, is it aimed at identifying the most frequent conformations of each peptide and to study their stability. Second, and most important, it is aimed at comparing main conformations of different elastin-derived peptides to identify the ''signature'' than can be linked to a biological activity. The main strength of the present work is to propose a method for confirmation recognition that is not affected by peptide rotations or translations and, hence, avoids the use of the complex superposition methods. In addition, the proposed approach uses Kernel PCA to eliminate atypical peptide conformations. Due to the instability of those peptides, removing outliers is crucial since they may dramatically impact clustering results. To extract the most frequent conformations, we propose to use a hierarchical clustering method. Eventually, a peptide activity detector is defined based on comparison of main conformation found in different peptides. The main interests of the proposed method are twofold: first, it is fully automatic method, second, it does not require any additional information or expertise and, third, it can identify conformations accurately that make peptides enabling a given biological activity. Experimental results on a large dataset of peptides conformations highlight the relevance and efficiency of the proposed method.

INDEX TERMS Automatic conformation identification, hierarchical classification, flexible peptide conformation, structure classification of protein, outliers detection.

I. INTRODUCTION

In the 1970s, Anfinsen *& al.* demonstrated in [1] that the function of a protein is encoded in its sole amino-acid sequence. Therefore, it came clear that the 3D structure of a protein is fully determined by the sequence of amino-acid is it made of. This sequence encodes the 3D shape of the protein as well as its biological function. Since then, considerable research efforts have been made to understand the relationships between such an amino-acid sequence, the ensuing 3D conformations of a protein, its dynamics and its possible biological functions. Despite significant advances [2], this task still remains extremely challenging for proteins. This is even more for highly dynamic structures such as peptides. In the field of biology, peptides (very small protein sequences) have been found to regulate key biological functions by behaving, for instance, as cardiac hormones [3] or antimicrobial molecules [4]. Further, antigenic peptides are

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

used to design vaccines. As a consequence, they deserve considerable attention. Because, peptides are small molecules, their structure is more dynamic than that of proteins which are much larger. Consequently, the analysis of molecular simulations aiming at understanding the variation of their 3D structure over time (trajectories) is difficult. This holds particularly for peptides derived from elastin.

Elastin is a polymeric macromolecule from the extracellular matrix responsible for the elasticity and proper functioning of tissues such as lung, skin or arteries [5]. More than 82% of tropoelastin (monomer of elastin) chains are built from 5 amino acids, out of 20, with numerous permutations. Elastin is synthetized during infancy and it is extremely stable. Nevertheless, it is degrading fatefully as we age and in age-related disorders such as type 2 diabetes, atherosclerosis or aneurysm formation [6].

Elastin degradation releases small peptides, so-called elastin-derived peptides (EDP), which are characterized by a wide range of biological activities [7], [8]. The activity of these peptides relies on their interaction with dedicated receptors. Only certain very specific peptides ''active'' conformations match with those receptor and, hence, are able to trigger and sustain a biological function [9]. A serious limit in our understanding of EDP functions is their considerable conformational variability because it hampers the analysis of their 3D trajectories when they are in water. Indeed, water is a plasticizer for elastin so that the elastin-water system is characterized by a very high entropy [10]. Extracting structures from chaos is no easy task.

In this work, we design a statistical method that aims at analyzing the trajectories of peptides in order to identify and extract recurrent conformations along their trajectory. Our method is computationally efficient, without alignment or superposition, while being fully automated. Further, we apply this procedure to EDP trajectories in view of identifying their bioactive conformers.

A. PRIOR WORKS AND CONTRIBUTION OF THE PRESENT PAPER

A vast majority of prior methods for identifying peptides main conformations can be divided into two categories;

The first category is based on protein or peptide composition in terms of amino acid sequence [11]–[15] which bring important physicochemical property of the whole chain. The amino-acid sequence is thus used to extract features that are assumed to capture all information about peptide key functions. Therefore, such features can be used as an input of any machine learning method with the purpose to identify what in the composition gives birth to a given biological function of interest. When peptide functionalities are known one can use to so-called supervised method while, on the opposite, peptides whose natures are unknown are analyzed with non-supervised methods.

As examples, the method proposed in [11] used peptides sequence to extract 20 features representing each amino acid occurrence frequency. A fuzzy non-supervised method has been applied to cluster the protein sequences and their main conformations. A different approach developed [14] proposes to use the amino acid sequence to extract physico-chemical properties such as its molecular weight, isoelectric point, length of amino acids, atomic composition, etc.... A labeled dataset is used to compare the accuracy of several supervised neural network with respect to identification of main conformations.

Interestingly, those works show that the amino-acid sequence is loosely related to biological function. However, such approaches are limited because they can hardly take into account all information related to dynamics and 3D conformation of peptides while it seems crucial to match a target receptor [16].

Methods from the second category try to relate 3D conformations of peptides with their biological functionalities [17]–[23]. In this category, the main approach consists in comparing 3D conformations using alignment such as DALI [17] or SSAP [19]. These two methods consist in representing each a peptide through the so-called distance matrix that contains all distances between all pairs of atoms. The DALI method [17] consists in chopping the distance matrices into parts that correspond to hexapeptide in order to find similar chunks in different 3D conformations. Although DALI is accurate, it comes at a very expensive computational cost since it will search similarities because of chunks of conformations: the number of combinations grows with the square of amino acid sequence length. In addition, DALI has mostly being evaluated over macro proteins whose large size makes them stable over molecular dynamic simulations: all conformations are gathered closely around several main conformations, with little or almost no spread. On the opposite, this paper focuses on elastin peptides which are arranged in a large number of very similar sequences (82% of them are composed of the same 5 amino acids). In addition their small size makes them very flexible; they all share a wide range of confirmations, many being identical.

In recent years, various alternative approaches have been proposed. A superposition of pairwise distance matrices has been proposed in [23], to determine the similarity between conformations. After reducing a conformation to the position of 7 most important atoms from the backbone, the root-mean-squared deviation (RMSD) is computed to cluster together similar conformations. This approach is simple but its efficiency has been assessed over a rather simple peptide with very little degree of freedom. A different approach has been proposed in [15], [18] modeling peptide backbone conformations with a parameterized curve which capture its global shape. Again, a measure of distance between curves can be used to compare them and assign those as being from the same conformation. While interesting, a vast majority of prior works suffer from the two main limitations. First of all, non-supervised approaches require prior information about the correct number of clusters. Second, peptide 3D conformation is reduced to the sole backbone. This prevents taking into account information from side chains, in general

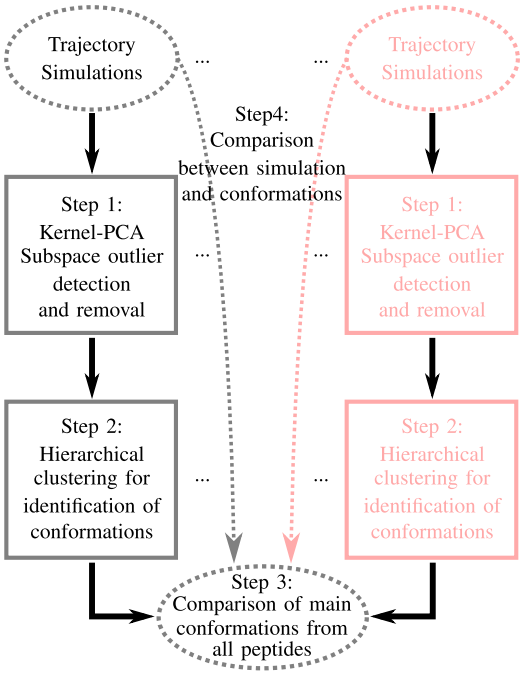


FIGURE 1. Overview of the proposed methodology. From dynamic trajectory simulation the first consists in KPCA-based outlier removal followed by main conformations classification using hierarchical clustering. The main conformations from all simulations are compared to get insights with respect to biological functionality.

to lower computational complexity, while their impact on the biological function remains mostly unstudied.

The present paper proposes a novel method for analyzing simulated molecular dynamics of different peptides with the following two main goals. First of all, it is aimed at identifying in a simple and efficient manner the main or most frequent 3D conformations from those simulation. This first step must be fully automatic while it must also allow the practitioner to visualize the results and tune the cluster method accordingly. Second, the ultimate goal is to be able to identify which peptide can be associated with a biological activity and, more precisely, what conformation may be key that enable such activity. An overview of the main steps involved in the proposed methodology is presented in Figure 1.

It should be noted that the present work focus on elastin-derived oligopeptides of small size, hence, whose conformations are extremely flexible and very unstable over time. We especially show that in this context side chains are essential and relevant to capture fine differences between numerous and similar conformations.

Thus, the main contributions of this paper are the following:

1) We study very similar yet very small and, hence, extremely flexible peptides. On a practical point of view, their conformations keep fluttering, they can take a wide range of forms which change abruptly and suddenly over time. This behavior acts like a strong noise that make the identification of main conformations very difficult.
2) In order to extract main conformations in a fully automatic manner we propose a novel method based on two statistical tools. First a sharp filtering method is applied to remove outlier conformations. Then, an unsupervised-learning method is proposed for identification of main conformations such that it can be tuned by practitioners.
3) We leverage a specific representation of conformation as well as a relevant similarity measure between two of them; this reduces the computational complexity dramatically.
4) Eventually, we propose a detection rule to determine whether a peptide is active or not.

The present paper is organized as follows: Section II formally states the problem of peptide main conformations identification. Then, Section III presents the proposed methodology for determining peptides most representative conformations from the molecular dynamic simulation. The validation of this method as well as the setting of its parameters is detailed in Section IV. Application on real data is presented in Section V. Section VI addresses the problem of comparing main conformations extracted from several different peptides. Finally, the ultimate goal of identifying active peptides and main underlying conformations is addressed in Section VI-A. Section VII concludes the paper.

II. IDENTIFYING MAIN CONFORMATIONS: PROBLEM STATEMENT

In order to determine the peptide conformation that triggers a biological function, the first step of the proposed approach is to determine each peptide main conformations. This section states the problem of identifying main conformations as well as properties one can expect from a suitable solution.

A. DEFINITION AND PROBLEM STATEMENT

A peptide essentially consists of a set of atoms lying in the 3 dimensional space and linked to each other. The coordinates of these atoms are referred to as the structure of the peptide denoted $\mathbf{S} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)^\top$ where N is the total number of atoms and $\mathbf{a}_n \in \mathbb{R}^3$ is the 3D coordinate of n-th atom. In a real situation, atoms of a structure all keep moving rapidly and randomly under thermodynamic influences. In the present paper, we use dynamic molecular simulations [24] to reproduce accurately thermodynamic motions of atoms via a sequence of structure denoted $\mathbb{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_T)$.

Formally speaking, two structures $\mathbf{S}_t$ and $\mathbf{S}_u$ belong to the same conformation if, and only if, there is a rotation, characterized by matrix $\mathbf{R}$, and a translation, characterized by vector $\boldsymbol{\zeta} \in \mathbb{R}^3$, such that:

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

with $\mathbf{1}_N \in \mathbb{R}^N$ containing only ones.

Therefore the conformation $\mathbb{F}_u$ can be defined as the set of all structures that are symmetric to $\mathbf{S}_u$:

$$\mathbb{F}_u = \left\{ \mathbf{S}_t \,\middle|\, \exists \mathbf{R}_t \,,\; \exists \boldsymbol{\zeta}_t \,,\; \mathbf{S}_u = \mathbf{S}_t \mathbf{R}_t + \mathbf{1}_N \boldsymbol{\zeta}_t^\top \right\}. \tag{2}$$

In general, two structures are not exactly symmetric. It is therefore more realistic to define a conformation as the set of all structures that are symmetric to $\mathbf{S}_u$ up to epsilon:

$$\mathbb{F}_u^\epsilon = \left\{ \mathbf{S}_t \,\middle|\, \min_{\mathbf{R}_t \,, \boldsymbol{\zeta}_t} \mathrm{d}\left(\mathbf{S}_u \,,\; \mathbf{S}_t \mathbf{R}_t + \mathbf{1}_N \boldsymbol{\zeta}_t^\top \right) < \varepsilon \right\}, \tag{3}$$

where $\mathrm{d}(\cdot, \cdot)$ can be any distance, from a mathematical point of view, and ϵ is the maximal distance between two structures from the same conformation.

B. PRACTICAL DIFFICULTIES AND CONSIDERATIONS

Before moving into practical considerations, let us point out that from the definition of main conformations, Eq. (3), a given structure $\mathbf{S}_t$ may either belong to one single main conformation, to two or more main conformations, or does not belong to any of the main conformations. The latter case represents a structure that is referred to as an outlier: it is quite far away from all cluster of conformations and, hence such it has no interest to understand how conformations are related to biological activities. The former case, in which a structure $\mathbf{S}_t$ belongs to two, or more, main conformations is more troublesome. The most likely explanation is that this constitutes a transitional structure between the two conformations and, as such, also does not constitute key elements for understanding activity.

As a consequence, we will assign a structure $\mathbf{S}_t$ to the conformation C_t using following assignment rules:

$$\begin{cases} L_t = 0 \;\text{ if }\; \forall k \in 1, \ldots, K,\; \mathbf{S}_t \notin \mathbb{F}_k^\epsilon \\ L_t = 0 \;\text{ if }\; \exists\, (k, l) \text{ such that } \mathbf{S}_t \in \mathbb{F}_k^\epsilon \text{ and } \mathbf{S}_t \in \mathbb{F}_l^\epsilon \\ L_t = k \;\text{ if }\; \mathbf{S}_t \in \mathbb{F}_k^\epsilon \text{ and } \forall l \neq k,\; \mathbf{S}_t \notin \mathbb{F}_l^\epsilon \end{cases} \tag{4}$$

Equations (1) - (4) formally state the first fundamental problems of present work and enable us to point out the main difficulties we have to address.

First of all, in the definition (3), it is very difficult to define ϵ *in practice* since this ''tolerance'' essentially depends on the activation of a target biological function which is hardly measurable.

Second, it is important to note that definition (3) fits well with the most general case in which, due to electrical properties of atoms, most structures are close to one of the few so-called ''metastable structure'' [25] between which transitions are very rare. On the opposite, elastin peptides are highly flexible and elastic and, hence, they have many widely different conformations between which transitions occur very frequently.

Additionally, Equation (1)-(3) highlight the superposition problem that must be solved: finding for each and every structure the rotation and the translation that minimizes a distance with a given main conformations. Solving this problem is computationally demanding and does not always lead to an optimal solution. Hence, we suggest finding a manner to bypass this issue.

In practice, one cannot know the number of main conformations in sequence $\mathbb{S}$. In addition, the method proposed in the present paper will be used by practitioners without knowledge on data processing and that would rather visualize, inspect and interpret the similarities between the obtained main conformations. Therefore, the proposed clustering method must automatically determine the main conformations and their number while it must be flexible enough to allow practitioners to visualize the conformation and adjust the results accordingly.

Last, but not least, the reference structure, denoted $\mathbf{S}_u$ in Eq.(1)-(3) is not known and must be automatically extracted from the sequence itself. To this end, we must carefully take into account the outliers, as defined in Eq. (4), because even though it they have no interest, they may have a dramatic impact especially when it comes to automatic identification of main conformations. One must note that, generally speaking, automatic detection and removal of an outlier is a difficult task.

All those difficulties may be summarized as follows: it is wished to design a method that inspects a sequence of very unstable structures, that very often moves randomly and abruptly, and without any prior information or even a clear definition of main conformations, it must automatically define the main conformations and, on the opposite, must eliminate the outliers structures that do not belong to a single main conformation.

III. METHODOLOGY FOR FINDING THE MAIN CONFORMATIONS

Once the main difficulties and objectives have been clearly pointed out, this section describes the method for identifying main conformation from peptides dynamic trajectory.

First of all, let us state that in order to tackle to superposition problem, it is proposed to represent a peptide by its distance matrix; formally, for a structure $\mathbf{S}_t$ the distance matrix $\mathbf{M}_{\mathbf{S}_t}$ is defined by its components:

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

where $\|\mathbf{v}\|_2$ stands for the Euclidean norm of vector $\mathbf{v}$, k and l are the atoms index.

The main advantage of this representation (5) is that, thanks the properties of Euclidean distance between vectors, it remains unchanged under translation and rotation. Therefore, the comparison of two structures $\mathbf{S}_t$ and $\mathbf{S}_j$ can be simply carried out by the difference between their distance matrices $\mathbf{M}_{\mathbf{S}_t} - \mathbf{M}_{\mathbf{S}_j}$ [26]. The distance matrix $\mathbf{M}_{\mathbf{S}_t}$ will thus be considered as an observation representing the structure $\mathbf{S}_t$ in the so-called ''*conformational space*''.

A. OUTLIERS DETECTION

As previously noted, elastin peptides are highly flexible and dynamic and hence give birth to peculiar or atypical structures. In order to achieve a higher robustness these outliers

must be removed or discarded before identifying main conformations. In our context, small elastin-derived peptides are made of about $N = 80$ atoms which makes the distance matrix be made of $Q = \frac{N(N-1)}{2} = 3160$ different distances. In such high-dimensional space, several methods have been proposed for outliers remove, see for instance [27]. In this context, angle-based methods or those based on nearest neighbors are among the most popular. However it can become quite computationally demanding especially when sequences are made of numerous structure.

In our case, it seems obvious that all the distance matrix $\mathbf{M}_{\mathbf{S}_t}$ does contain some redundancy. Methods based on dimensionality reduction and subspace outlier detection are extremely relevant and among those, Principal Component Analysis (PCA) is certainly a fundamental tool. In brief, for a sequence of structures $\mathbf{m}_{\mathbf{S}_t}$, $t = \{1, \ldots, T\}$ each put into a vector of $Q = \frac{N(N-1)}{2}$ components, the PCA seeks the linear projector $\mathbb{R}^Q \rightarrow \mathbb{R}^q$, represented by the orthonormal matrix $\mathbf{P}_q$ of size $q \times Q$, which reduces observations to Q components while minimizing the mean square error:

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{m}_{\mathbf{S}_t} - \mathbf{P}_q\mathbf{P}_q^{\top}\mathbf{M}_{\mathbf{S}_t}\right\|_2^2. \tag{6}$$

A few algebra show that, for a single observation $\mathbf{m}_{\mathbf{S}_t}$, PCA reconstruction error is given by:

$$Err_{\text{Rec}}\left(\mathbf{m}_{\mathbf{S}_t}\right) = \sum_{t=1}^{T}\left\|\mathbf{m}_{\mathbf{S}_t} - \mathbf{P}_q\mathbf{P}_q^{\top}\mathbf{M}_{\mathbf{S}_t}\right\|_2^2 \tag{7}$$

$$= \left\|\mathbf{m}_{\mathbf{S}_t}\right\|_2^2 - \left\|\mathbf{P}_q^{\top}\mathbf{M}_{\mathbf{S}_t}\right\|_2^2 \tag{8}$$

The key idea of PCA for subspace outlier identification is illustrated in Fig. 2. Since the goal is to minimize the mean square reconstruction error (6) outliers will be poorly represented and hence will feature a much higher reconstruction error (7).

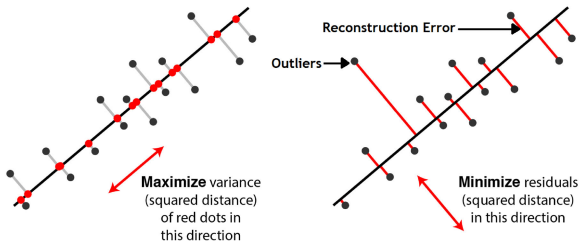


FIGURE 2. Illustration of PCA: The black line represents the first principal axis along which the variance is maximal (left) while reconstruction error is minimal (right).

While PCA is efficient for this purpose [27], [28], it can be noted that it is limited to the case where observation can be represented in a linear subspace. In our case, due to the nature of distance matrix, we have noted that this assumption does not hold true. Instead of using a high-dimensional subspace we have to leverage a non-linear subspace for representation of observation, using the so-called Kernel PCA (or K-PCA) [29].

To turn PCA into a non-linear representation, a mapping function Φ is applied to transform the observation $\mathbf{m}_{\mathbf{S}_t}$ from the "*conformational space*" $\mathbb{R}^d$ to the so-called feature space $\mathcal{F}$. Kernel PCA simply consists in applying the same PCA method over the transformed data, in the feature space $\mathcal{F}$.

In order to apply the transformation of data Φ while keeping a low computational complexity one must note that the application of PCA exclusively uses scalar products. Therefore, one does not need to explicitly defines the mapping function Φ but rather to know on to compute a scalar product in the feature space:

$$\Phi(\mathbf{m}_{\mathbf{S}_t})^{\top}\Phi(\mathbf{m}_{\mathbf{S}_u}) = k\left(\mathbf{m}_{\mathbf{S}_t}, \mathbf{m}_{\mathbf{S}_u}\right).$$

This so-called "Kernel trick" consists in defining the kernel function $k(\cdot, \cdot)$ that corresponds to the scalar product in a feature space without explicitly defining the mapping Φ [30].

Following the method explained in the linear case of the PCA, see Fig. 2, it is proposed in the present paper to classify as an outlier the observations with the highest reconstruction.

In the present paper, we have used the most universal kernel [29] which is the Gaussian Kernel defined by:

$$k\left(\mathbf{m}_{\mathbf{S}_t}, \mathbf{m}_{\mathbf{S}_u}\right) = \exp\left(-\frac{\left\|\mathbf{m}_{\mathbf{S}_t} - \mathbf{m}_{\mathbf{S}_u}\right\|_2^2}{2\sigma^2}\right). \tag{9}$$

KPCA has shown a great efficiency especially for outlier removal [29], which is the one of interest in this paper. The reader is referred to [29], [31] for assessment and details.

For the application of this methodology, two parameters have to be determined: the number of eigenvectors q used for K-PCA and the width σ of the Gaussian kernel (9). The selection of these parameters as well as the number of observations that shall be classified as "outliers" will be discussed in Section IV.

B. CLUSTERING METHOD

Regarding the recognition of the main conformations, we propose to use a hierarchical clustering method due to its flexibility and accuracy. This method is one of the most widely used for unsupervised classification problems [32]. It consists in considering each observation as a cluster and to merge iteratively the two clusters which are the closest to each other, according to some distance. Obviously such a methodology works without any prior information. Moreover, allows representing the relation between conformation by showing how clusters are merged, see for instance Figure 7. In addition, it also helpful for practitioners to inspect each conformation at each level in order to be able to adjust the result accordingly.

We will briefly explain the proposed hierarchical method; to this end, let us consider I clusters denoted $\mathcal{C}_i$ with n_i observations for clusters $i \in \{1, \ldots, I\}$. Let us also denote $\mathbf{D}$

the matrix whose element d_{ij} is the distance between clusters i and j. The Hierarchical Agglomerative Clustering algorithm can be described as follows:

Step1: Find the smallest distance $d_{i,j}$ remaining in $\mathbf{D}$.
Step2: Merge clusters i and j into a new cluster noted k
Step3: Update the set of distances in $\mathbf{D}$.

The three steps above are carried out iteratively until one single cluster remains.

Obviously, the definition of the distance between clusters is fundamental is such a clustering method. While many different distances have been proposed, see for instance [33], in the present paper we have used Ward which is defined as:

$$d_{i,j} = \frac{n_i n_j}{(n_i + n_j)} \left\| \bar{\mathbf{m}}_i - \bar{\mathbf{m}}_j \right\|_2^2, \tag{10}$$

where $\bar{\mathbf{m}}_i$ stands for the expectation (geometrical mean) of cluster i which is simply given by the average of all data that belongs this cluster:

$$\bar{\mathbf{m}}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} \mathbf{m}_{\mathbf{S}_t}$$

We have selected this distance because it is relatively simple, it offers ensuing good performance and because it is generally considered as the most suitable distance when data have unknown distributions [34], which fit well with our case.

To help the practitioner visualize the main conformations resulting from this method, we have chosen to represent each cluster using a single conformation, that is the average.; the clustering will hence results in also defining the dictionary of main representative conformations for a peptide dynamic trajectory simulation.

IV. VALIDATION AND PARAMETERS SELECTION

The methodology proposed in the present paper involves several parameters whose setting must be validated, namely the number q of principal components in the KPCA, the σ width of Gaussian kernel and the default number of main conformations from a sequence. The main difficulty is that ''real data'', which results from dynamic trajectory simulations are not provided with a ''ground truth'' that could be used to assess the relevance of the method. A review of methods used to overcome this problem has been presented in [35]; following the suggestions from the prior work, it is proposed in this paper to use an artificial dataset of structures such that we can investigate in order to design a methodology for parameter validation.

A. ASSESSMENT ON ARTIFICIAL DATASET

The virtual dataset we will use for this purpose has been generated randomly in a controlled manner; this dataset is illustrated in Fig. 3 and is made of:

- an artificial 9 atoms planar molecule that is each atom is defined by a two-valued (x, y) coordinates;
- 4 main conformations to be found;
- 1300 structures derived from the main conformations by adding noise;

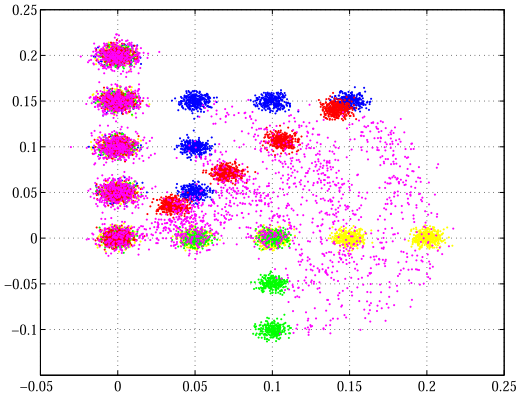


FIGURE 3. Illustration of structures from computer-generated data; note the 9 different atoms position which are equidistant from each other gathered around four main conformations (shown, for readability, in blue, red, yellow and green). The transitional structures are shown in purple.

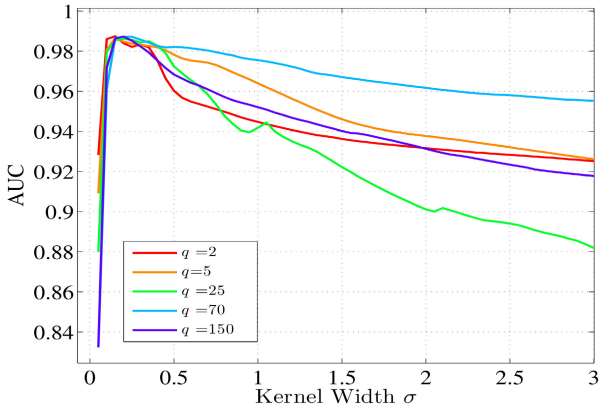


FIGURE 4. Outlier detection efficiency, using AUC over computer generated data, see Fig. 3, as a function of σ and for 5 different number q of KPCA.

- 350 transitional structures between every two pairs of conformations.

The value of additive Gaussian noise as well as the probability of transition is set to mimic data from dynamic trajectories simulations.

First of all, Figure 4 shows the efficiency of the proposed method for outlier detection. Note that while the Receiver Operational Characteristic (ROC) Curve allows presenting the empirical probabilities of outlier detection as a function of the false alarm probability (that is the classify a noisy conformation observation as an outlier); in the present paper, however, the results are shown using the Area Under the ROC Curve (AUC) because it summarizes with one single real value the detection accuracy: the higher the AUC the more accurate the detection, see [36] for more details.

Figure 4 shows outlier detection accuracy through, with the AUC metric, as a function of kernel width σ and for a few different number q of KPCA. This figure shows three important things: first of all, the proposed KPCA subspace outlier detection method is very efficient since its AUC reaches 0.99 which means that all outliers can be removed with almost

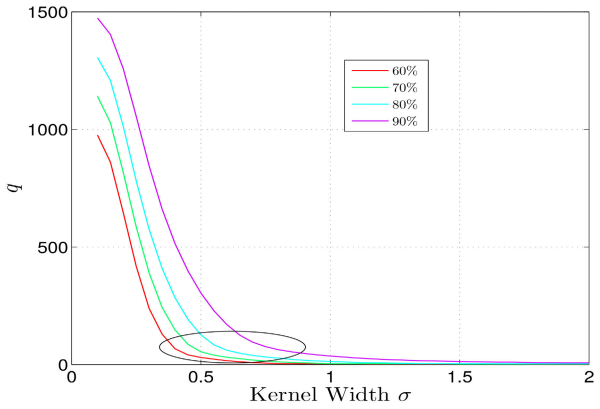


FIGURE 5. Number of eigenvectors q versus σ for 4 different inertia rates (11) over the simulated *DB*2. The circled zone represents the most relevant set of values.

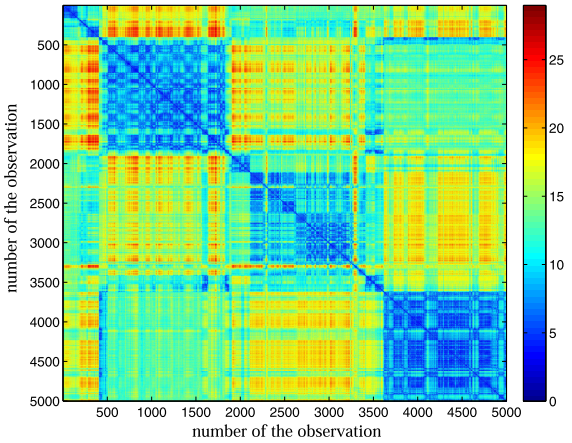


FIGURE 6. Matrix representing the euclidean distance between pairs of observations coming from the 5 000 successive structures of the peptide EGFEPG. It is used to present the similarity between them.

no error. Second, this figure shows that a relatively small kernel width is more relevant; this is not surprising because conformations are ''rather close'', the kernel should therefore be set to distinguish those clusters which are nearby each other. Last and not least, this figure shows that the setting of parameters from results of this simulation is difficult since the same accuracy for outlier detection is reached regardless of the number of KPCA. Intuitively, it seems desirable to keep the number of KPCA eigenvectors q as low as possible but this statement must be supported.

In order to be able to extend the proposed KPCA outlier detection to real trajectories simulation, and especially the underlying method for parameters assessment, we proposed to measure the ''inertia rate'' which is defined by:

$$I = \frac{\sum_{d=1}^{q} \lambda_d}{\sum_{d=1}^{Q} \lambda_d} \tag{11}$$

where λ_i is the i^{th} largest eigenvalue associated with the i^{th} eigenvector from KPCA and Q is the total dimension of observations.

Roughly speaking, the so-called ''inertia rate'' express of the average fraction of observations that are preserved within KPCA subspace and.

Figure 5 shows the required number of PCA eigenvectors q requires achieve specific inertia rate values, as a function of Gaussian kernel width σ. Interestingly, the curves presenting the couple of parameters (σ, q) clearly highlights an inflexion point in Figure 5. Intuitively, this point corresponds to an optimal setting as a larger Gaussian kernel width seems quite inefficient for representing the observations while, on the opposite, the number of principal components must increase significantly when kernel width is only slightly reduced.

This method can be justified observing that a small enough value for σ can push the Gaussian function $k\left(\mathbf{m}_{\mathbf{S}_t}, \mathbf{m}_{\mathbf{S}_u}\right)$ to be close to zeros for any two observations t and u with $t \neq u$. Which means that data is spread all apart and, hence, more eigenvector is required to represent them all. On the opposite, for very large values of σ, $k\left(\mathbf{m}_{\mathbf{S}_t}, \mathbf{m}_{\mathbf{S}_u}\right) \approx cst$, hence almost all samples are gathered and assumed similar, hence can be represented with low number of principal components, and the use of kernel PCA becomes meaningless as explained in [29].

This method, which essentially aims at finding the optimal values of parameters (σ, q) can be applied blindly, without information on the dataset, hence its application over ''real data'' from dynamic trajectories simulations.

V. CLUSTERING RESULTS ON REAL DATA

A. REALISTIC DATASETS

In the present work, the dataset of ''real'' peptide structures actually comes from a dynamic trajectories simulations from 12 considered different peptides. Each simulation has been acquired with a sampling period of 200ns, is it made of a total 40 000 structures whose size, in terms of the number of atoms, ranges from $N = 79$ to 87. Prior studies have shown 4 peptides (namely APGVGV, GVGVAP, PGVGVA, VAPGVG) are not associated with any activity while, on the opposite, 5 peptides are active (GVAPGV, PGAIPG, VGVAPG,EGFEPG, LGTIPG); in addition 3 are of unknown activity (PGAYPG, VGLAPG, VVGPGA).

B. VISUALIZATION AND CONFORMATION IDENTIFICATION ASSESSMENT

As already explained, blindly identification of main conformations, without any knowledge on the expected outcomes can hardly be assessed and presented.

For the sake of clarity and presentation, we have proposed to focus on a sequence of 5 000 structures from peptides EGFEPG. On the one hand, this sequence is rather representative of the whole dataset while, on the other hand, seems particularly stable, with a handful of main confirmations, among which the structures are not extremely similar, and transitions that generated outliers.

This small structure sample is represented in Figure 6 via the distance between each and every structure. This figure highlights large blue squares along the diagonal among which distance between structures is much smaller, hence

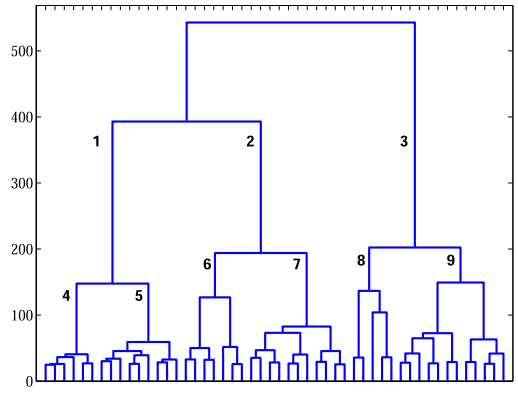


FIGURE 7. Illustrative example of the dendogram obtained from the 5 000 peptides structures EGFEPG 6 showing how clusters are gathered (*x*-axis) when increasing the minimal distance between centers (*y*-axis).

likely made of the same conformation. On the opposite the figure shows large yellow/red area apart from the diagonal which indicates that structures are much more different. We will use this small dataset to study the impact of parameters (σ, q) on the proposed method for outlier detection as well as to visualize ensuing main identified conformations to get insights from practitioners.

First of all, we assumed that 20% of the data are outliers; this rate may seem very high but since we are only interested in the most representative conformations, we would rather drop too many structures than having clusters affected with those outliers. We used the same results presented in Figure 5 to identify the relevant values for parameters (σ, q) which results in $\sigma = 5$ and $q = 12$; those values are naturally larger as compared to those found for artificial data because of their large size (in terms of atoms) and the higher number of less stable of main conformations.

To confirm these results, we propose to visualize the impact of those parameters on main conformation identification by measuring the evolution of their centers through the inter-center distance defined as:

$$\|\bar{\mathbf{m}}_r - \bar{\mathbf{m}}_s\|_2 \,. \tag{12}$$

Figure 8 shows the evolution of distances between centers (12) when increasing the numbers of eigenvalues used with proposed method of subspace outlier detection based on KPCA. Note that for the sake of readability and clarity we used the same scale for all the sub-figures. Figure 8 clearly shows that, from approximately 12 eigenvectors, the three main conformation centers remain very close when increasing the number of principal components. Roughly speaking, this emphasizes that, when the kernel width σ is kept unchanged, increasing numbers of principal components in the KPCA do not bring much information about main conformations. This *post-hoc* justify our choice to set $q = 12$ with $\sigma = 5$.

Another manner to visualize the results of main conformation identification is via the so-called dendogram presented in the Figure 7. In this figure, the y-axis represents the distance between clusters centers, see Eq. (10); on the opposite, the x-axis shows structures. Together the dendogram show the main conformations, also referred to as the clusters, and how they are merged together for different distances.

Figure 7 clearly shows that the 5 000 samples, illustrated in Figure 6, are made of three main conformations (numbered from 1 to 3) but also that each of those can be split into two (numbered from 4 to 9).

C. ANALYSIS AND COMPARISON WITH PRIOR ARTS

We have stated that one the main contribution of the present method is to able practitioners to visualize the outcome and manually adjust the settings according to their needs. As an illustrative example for this feature, Figure 9 shows the most representative structures for each main of the three conformations. A practitioner would certainly be interested in comparing the two sub-clusters that give birth to one cluster, *e.g.* subclusters 4 and 5, in order to decide what is the most relevant outcome from expertise point of view. Such example is illustrated in Figure 10 which shows clearly that subcluster 4 and 5 share the same backbones while differences are on the orientation of the ultimate side-chain orientation. With those tools of hierarchical clustering along with the proposal for visualizing most representatives conformations and performing comparison, the practitioner can use of its expertise to guide the process.

Current art methods for protein main conformations identification usually focus only on protein backbones. By doing so, it is assumed that side chains are not relevant for characterizing the main conformations, hence on the ensuing functionality. While this allows simplification, by dramatically reducing the number of atoms, this strong assumption has, up to our knowledge, never been proved and in fact seldom studied. The result illustrated in Figure 10 deeply question this assumption.

Though the present paper does not aim at answering this question definitively our results seems to indicate that side chains may be of high interest for main conformations identification. To provide further evidence along this direction, we have contrasted the results obtained with and without side chains using the proposed method over the 40 000 structures, from sequence ''VGVAPG''. We have chosen this specific peptide for simplicity and clarity of the presentation because since it is one of the least flexible from our dataset it has a limited number of main conformations, thus making the comparison easier.

The result shown in Table 1 compares the fraction of structures that are classified into the same main conformations with or without side chains. While almost 94% of the data from the third main conformation are classified into the same conformation either side chains are used or not; this fraction is down to approx. 77.5%. With more than 23% and 15% of data classified into different main conformations (for the first and second ones) it is obviously questionable to assume that side chains is not relevant. We would claim that such results, along with those obtained with other peptides,

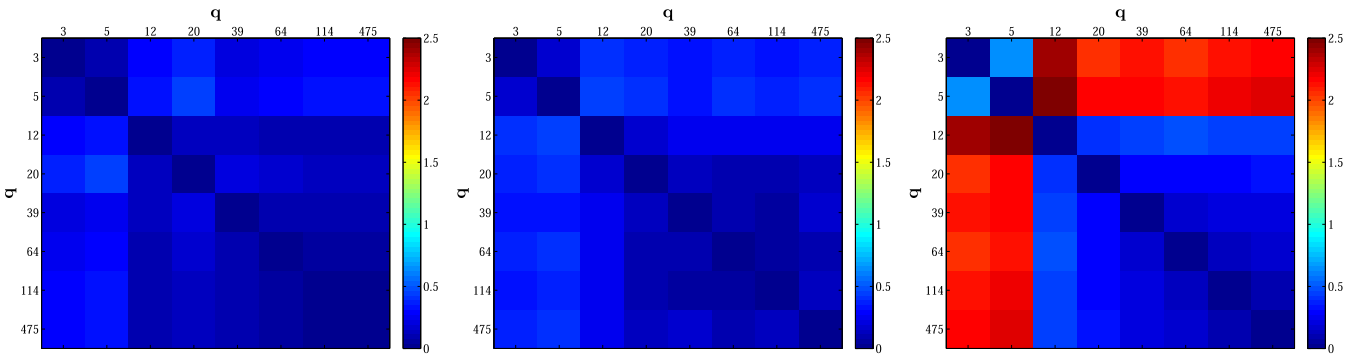
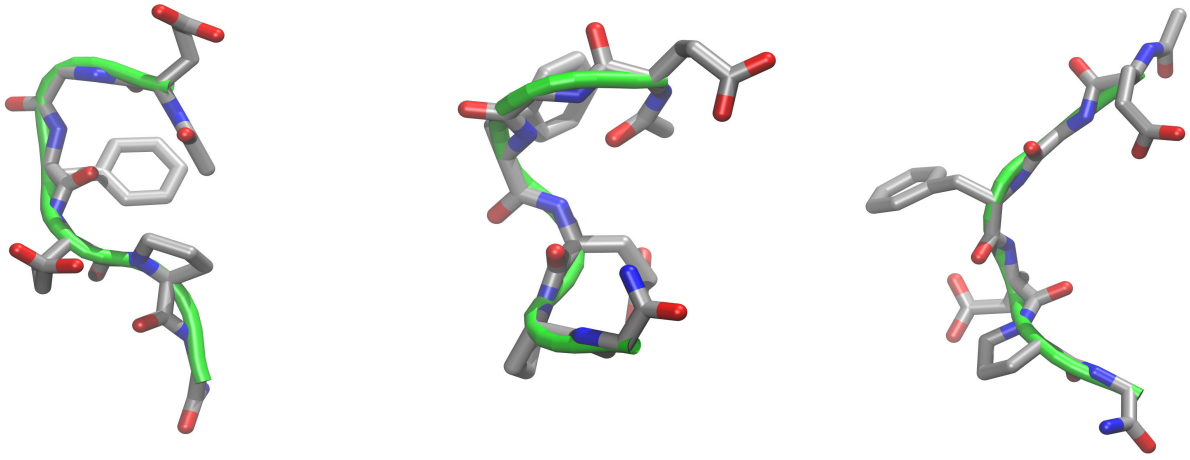


FIGURE 8. Evolution of distances between class centers, see Eq. (12) as a function of the number of eigenvectors used in kernel PCA.



(a) Main conformation form Cluster #1 (b) Main conformation form Cluster #2 (c) Main conformation form Cluster #3

FIGURE 9. Visual outcome of the hierarchical clustering method: the three most representative structures of all three main conformations.

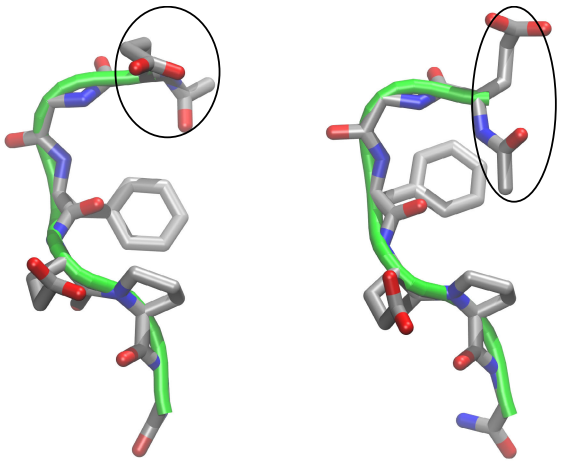
TABLE 1. Percentage of the structures classified into possible main conformations with and without side chains.

classes	w1	w2	w3
wo1	77.46%	22.53%	0
wo2	15.09%	84.52%	0.39%
wo3	0	6.03%	93.9%

that side chains shall be considered for clustering small and flexible elastin-derived peptides. Their impact on the ensuing biological activity should also be considered.

Last but not least, we would like to emphasize the relevance of the proposed method via a comparison with prior arts and especially the most relevant and popular one, namely DBSCAN [37]. This alternative approach is also interesting because it is one of the very few approach that performs both outlier removal and clustering. Roughly speaking, the principle of DBSCAN is to gather iteratively data that are within a certain range ϵ. Clusters with less than *MinPts* are assumed to be made of outliers. It is thus simple, efficient, and only require the user to set those two parameters (ϵ, ϵ).

Table 2 shows the results, in terms of the number of clusters and number of outliers, obtained when using DBSCAN over



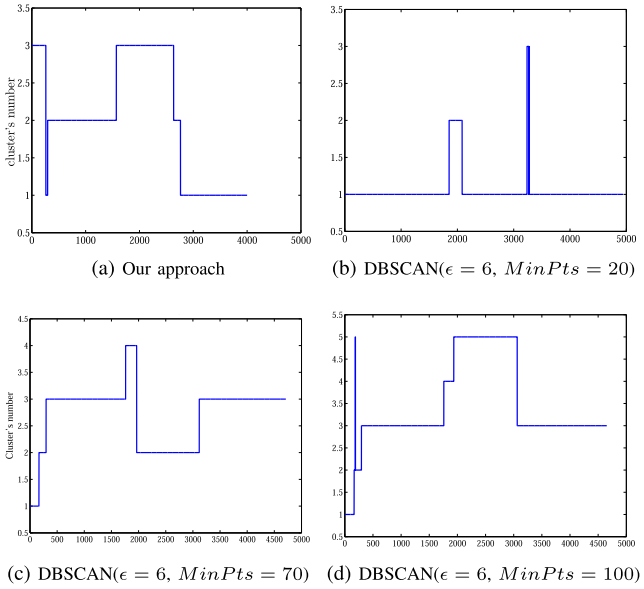
(a) Center of Subcluster #4 (b) Center of Subcluster #5

FIGURE 10. Representation of the most representatives conformations for two sub-clusters 4 and 5, as labeled in Fig. 7.

the same set of 5 000 with different values for parameters ϵ (in columns) and *MinPts* (in rows). We have carried out a grid-search approach and refined the grids in order to focus on results that are sound, *i.e.* that does not put all data into

TABLE 2. Number of outliers and clusters obtained with the DBSCAN method. The results are shown in this form: (number of outliers, number of clusters). The columns correspond to the minimal number of points ***MinPts***, and the lines correspond to the radius of the circles ϵ.

ϵ \ $MinPts$	18	22	25	28	32	36	38	43	49
6.2	(34 , 3)	(39 , 3)	(61 , 3)	(91 , 3)	(99 , 4)	(102 , 4)	(109 , 4)	(123 , 4)	(141 , 4)
6.4	(27 , 2)	(32 , 2)	(42 , 3)	(54 , 3)	(87 , 3)	(92 , 3)	(95 , 4)	(98 , 4)	(114 , 4)
6.6	(7 , 2)	(12 , 2)	(27 , 2)	(38 , 2)	(72 , 2)	(76 , 2)	(81 , 2)	(85 , 3)	(88 , 3)
6.8	(2 , 2)	(7 , 2)	(8 , 2)	(24 , 2)	(65 , 1)	(67 , 1)	(68 , 1)	(73 , 1)	(75 , 2)



(a) Our approach (b) DBSCAN($\epsilon = 6$, $MinPts = 20$)

(c) DBSCAN($\epsilon = 6$, $MinPts = 70$) (d) DBSCAN($\epsilon = 6$, $MinPts = 100$)

FIGURE 11. Comparison between the classification of peptide series over time using the proposed method (a) and the well-known DBSCAN approach with various settings (b-d).

only one cluster or split all data and assumed that the dataset is made only of outliers.

First of all, Table 2 shows that this approach is hardly usable in the context of elastin-derived peptide structures because it is extremely sensitivity. Parameters (ϵ, ϵ) must be finely tuned and do not generalize for various peptides. More important, a brief comparison of the results obtained with the proposed approach and those obtained with DBSCAN, presented in Figure 11 show that the low relevance of the latter. Indeed when obtaining the same number of clusters, one can notice that DBSCAN always put a vast majority of data into the same cluster and only a very few number of samples into the others which do not match with ground truth, see Figure 6.

VI. ANALYSIS OF PEPTIDE ACTIVITY

The ultimate goal of the present paper is to study whether the biological activity can be due to its spatial conformations. To this end, this section focuses on finding potential "*conformational signatures*" that would distinguish active peptides from non-active ones. To address this problem we must compare the main conformations from each and every peptides.

From the collaboration between scientists from data analysis and molecular biological sciences we propose to cluster the molecular dynamic simulations using a dissimilarity threshold equal to 200 for all peptides, see Figure 7.

Let us state first that we have experimentally found that methods that directly compare structures of different sizes [38] are highly inefficient in the present context; it indeed stated as similar conformations that were considered as very different from the point of view of molecular biology. As a consequence, we keep with the approach described in Section III that essentially consists in calculating the Euclidean distance the distance matrix representing each conformation. However this is only possible for peptides with the same size. To this end, we have to restrain the comparison to the atoms of the backbone.

The method we proposed is the following. For each and every main conformations extracted from all peptides, we extracted the most representative structure (the observation with the least distance from the conformation center). Then, we measure if the same structure can be found in other peptides by finding the conformation which minimizes the distance with the given structure. Eventually, we determine a threshold such that active peptide share similar conformation while non-active peptide does not.

Fig. 12 show such a binary result; each row corresponds to a single conformation while each column corresponds to the set of conformation extract for a given peptides. The first four peptides are those without biological activity while the last five are the active ones. A blue color indicates a similarity between a specific conformation (in row) and all those extracted from a given peptide (in column) higher than the determined threshold; on the opposite, red color indicates that a given conformation cannot be found in a given peptides. The long blue lines on the bottom right corner of Fig. 12 clearly points out that active peptides share several similar conformations. On the opposite, the top left corner show very little similarity between conformations from inactive peptides.

Interestingly, one can note that only 6 different conformations that can be found in active peptide. Even more interesting, among those conformations two are shared among all active peptides. This confirms the assumption it is wished to verify, activity of a peptide could be due to the presence of

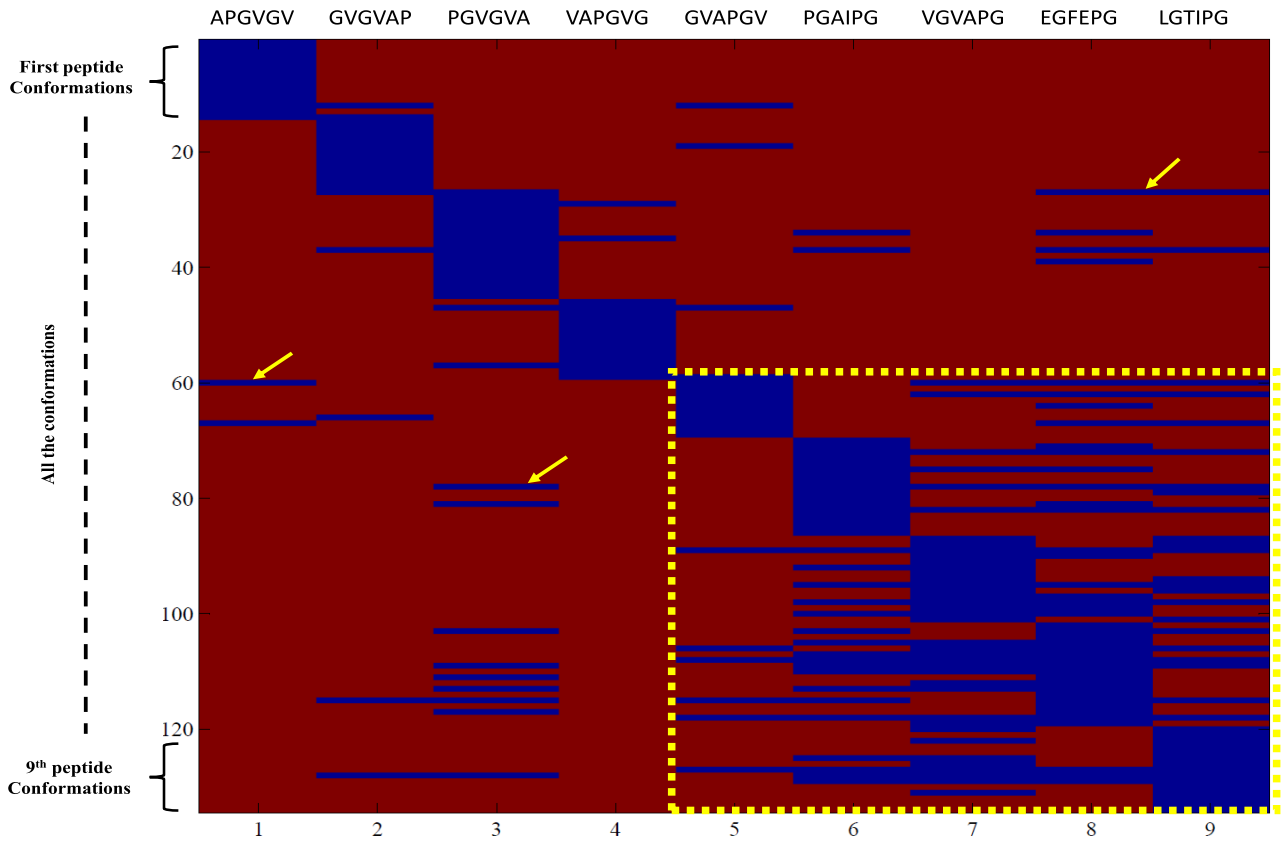
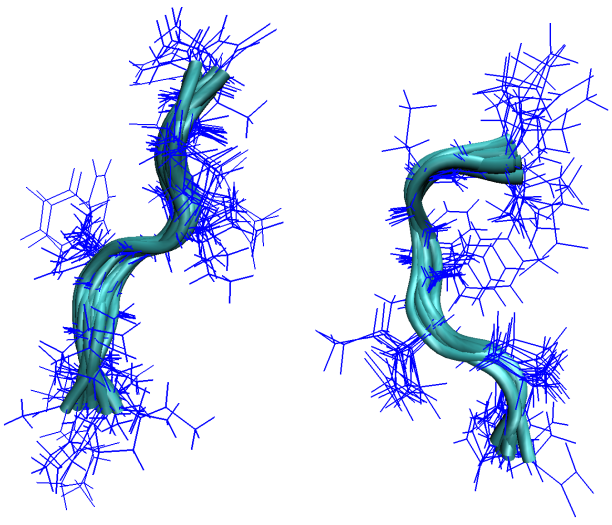


FIGURE 12. Matrix that represents the identical conformations in each peptide. Each column represents a peptide. Each line corresponds to a conformation.The blue color indicates that they are identical, and red indicates the opposite.



(a) The first conformation (b) The second conformation

FIGURE 13. The two main conformations extracted from the active peptides only.

this specific conformation. The most representative structures that features those conformations are presented in Figure 13.

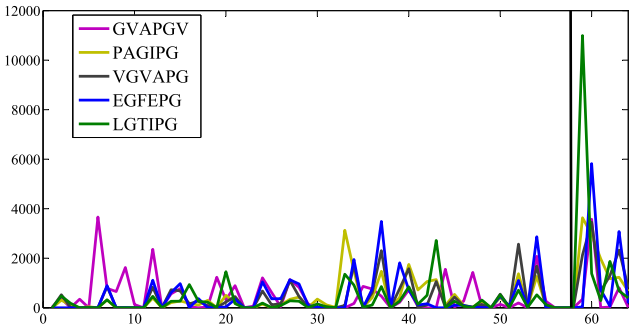
On the opposite, we noted that among inactive peptides several main conformations are shared with active peptides. However, the two conformations that are shared by all active peptide are never found among inactive one. This is also an important clue that support the assumption that a given biological activity may be due to a specific conformation that matches a target receptor.

A. IDENTIFICATION OF ACTIVE PEPTIDES

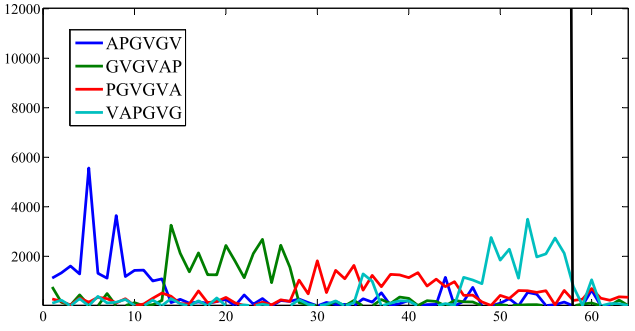
We propose to verify with the three peptides whose activity is unknown, namely PGAYPG, VGLAPG, VVGPGA, the two main results that results from analysis of peptides activity. Those results are, first, that active peptides share the same two conformations and, second, that non-active peptides almost do not share similar conformations at all.

To this end, we applied the same method described in Sections III and IV for outliers removal. Then we propose to assign each and every observation, from a given sequence, to the closest one among the 64 different reference conformations extracted from the 9 whose activities are known.

Figure 14 represents the application of this method to the 9 whose activities are known; obviously, it confirmed the previous mentioned point, while active structure are very often similar to the main conformation of active peptides, the non-active peptide do not share many similar structures. Table 3 shows the same results under a different form; as shown in Figure 14, we have divided the 64 main different



(a) For active peptides



(b) For the non-active peptides

FIGURE 14. The number of elements close to each conformation. The x-axis corresponds to the conformations of all the peptides. The first 58 conformations belong to the inactive peptides, and the last 6 conformations belong to the active peptides.

TABLE 3. The average effective structure of each peptide within the conformations which belong to the non-active peptides (first column) and active peptides (second column).

Conformations index / Peptide	$1 \rightarrow 58$ (non active)	$59 \rightarrow 64$ (active)
APGVGV (non active)	**539.70**	116.16
GVGVAP(non active)	**544.12**	73.5
PGVGVA(non active)	**513.60**	368.5
VAPGVG(non active)	**529.89**	211
GVGVAP(active)	469.27	**797**
PGAIPG(active)	351.18	**1938.5**
VGVAPG(active)	367.36	**1782.2**
EGFEPG(active)	380.9	**1650.7**
LGTIPG(active)	282.5	**2602.5**
PGAYPG(unknown)	435.2	**1126.3**
VGLAPG(unknown)	372.41	**1733.3**
VVGPGA(unknown)	**516.5**	340

conformations into two groups: those extracted from inactive peptides (from 1 to 58) and those extracted from active peptides (from 59 to 64). Table 3 shows the average number of structures whose closest conformation is in each group. From those results, it is obvious that the average of structures closest to the first group is much higher for inactive while, on the opposite, the number of structures closest to the second group is much higher to active peptide. Note that Table 3 includes the results to the three peptides whose activity is unknown, see latest rows. We can conclude from these results that peptides PGAYPG and VGLAPG should allow biological activity while, on the opposite, the peptide VVGPGA shall not trigger the target biological function. Those conclusions have then been supported empirically.

VII. CONCLUSION

The present paper proposes a novel statistical methodology, based on the clustering of 3D structures, with the ultimate goal of analyzing the biological activity of highly flexible elastin-derived peptides. This novel method combines different statistical algorithms to detect the conformation of peptides that trigger a given biological functionality. The proposed method is based on the following two main steps: (1) Kernel-PCA is used for outlier removal (2) Hierarchical clustering allows identification of main conformation into a flexible manner that can be adjusted by practitioners. We have presented a method for parameters selection and assessed its relevance of a set of dynamics trajectories simulations. This allows us to identify two conformation which is always found among active peptides while, on the opposite, structures from non-active peptides are rarely similar.

It is expected to extend the proposed method for blind identification of peptide activities. Besides, our future work consists in quantifying the biological activity of peptides in order to be able to relate this amount with the frequency at which a given conformation occurs in our simulations.

REFERENCES

[1] C. Anfinsen and H. Scheraga, "Experimental and theoretical aspects of protein folding," *Adv. Protein Chem.*, vol. 29, pp. 205–300, Jan. 1975.
[2] H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling, "Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent," *J. Amer. Chem. Soc.*, vol. 136, no. 40, pp. 13959–13962, Oct. 2014.
[3] Y. Nakagawa, T. Nishikimi, and K. Kuwahara, "Atrial and brain natriuretic peptides: Hormones secreted from the heart," *Peptides*, vol. 111, pp. 18–25, Jan. 2019.
[4] M. Zasloff, "Antimicrobial peptides of multicellular organisms: MY perspective," *Antimicrobial Peptides*, vol. 4, pp. 3–6, May 2019.
[5] S. M. Mithieux and A. S. Weiss, "Elastin," *Adv. Protein Chem.*, vol. 70, pp. 437–461, Jun. 2005.
[6] J. Uitto, L. Ryhänen, P. A. Abraham, and A. J. Perejda, "Elastin in diseases," *J. Investigative Dermatol.*, vol. 79, no. 1, pp. 160–168, 1982.
[7] A. Wahart, T. Hocine, C. Albrecht, A. Henry, T. Sarazin, L. Martiny, H. El Btaouri, P. Maurice, A. Bennasroune, B. Romier-Crouzet, S. Blaise, and L. Duca, "Role of elastin peptides and elastin receptor complex in metabolic and cardiovascular diseases," *FEBS J.*, vol. 286, no. 15, pp. 2980–2993, Aug. 2019.
[8] A. Le Page, A. Khalil, P. Vermette, E. H. Frost, A. Larbi, J. M. Witkowski, and T. Fulop, "The role of elastin-derived peptides in human physiology and diseases," *Matrix Biol.*, vol. 84, pp. 81–96, Nov. 2019.
[9] C. Kawecki, L. Duca, S. Blaise, B. Romier, H. el Btaouri, P. Gillery, L. Debelle, and P. Maurice, "Vieillissement matriciel et impacts vasculaires," *Hématologie*, vol. 21, no. 4, pp. 221–229, 2015.

[10] M. A. Lillie and J. M. Gosline, "The effects of hydration on the dynamic mechanical properties of elastin," *Biopolymers*, vol. 29, nos. 8–9, pp. 1147–1160, Jul. 1990.
[11] S. Bandyopadhyay, "An efficient technique for superfamily classification of amino acid sequences: Feature extraction, fuzzy clustering and prototype selection," *Fuzzy Sets Syst.*, vol. 152, no. 1, pp. 5–16, May 2005.
[12] R. Daras, D. Zarpalas, D. Tzovaras, and M. G. Strintzis, "3D shape-based techniques for protein classification," in *Proc. IEEE Int. Conf. Image Process.*, Feb. 2005, p. 1130.
[13] J. Cheol Jeong, X. Lin, and X.-W. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 308–315, Mar. 2011.
[14] S. Vipsita, B. K. Shee, and S. K. Rath, "An efficient technique for protein classification using feature extraction by artificial neural networks," in *Proc. Annu. IEEE India Conf.*, Dec. 2010, pp. 1–5.
[15] 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 *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2013, pp. 69–73.
[16] S. Leavitt and E. Freire, "Direct measurement of protein binding energetics by isothermal titration calorimetry," *Current Opinion Struct. Biol.*, vol. 11, no. 5, pp. 560–566, Sep. 2001.
[17] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *J. Mol. Biol.*, vol. 233, no. 1, pp. 123–138, Sep. 1993.
[18] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in Euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, Jul. 2011.
[19] C. A. Orengo and W. R. Taylor, "SSAP: Sequential structure alignment program for protein structure comparison," *Methods enzymology*, vol. 266, pp. 617–635, Oct. 1996.
[20] D. Fraccalvieri, A. Pandini, F. Stella, and L. Bonati, "Conformational and functional analysis of molecular dynamics trajectories by self-organising maps," *BMC Bioinf.*, vol. 12, no. 1, p. 158, 2011.
[21] 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, Jul. 2006.
[22] L. V. Nedialkova, M. A. Amat, I. G. Kevrekidis, and G. Hummer, "Diffusion maps, clustering and fuzzy Markov modeling in peptide folding transitions," *The J. Chem. Phys.*, vol. 141, no. 11, 2014, Art. no. 09B611.
[23] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, "Persistence-based clustering in Riemannian manifolds," *J. ACM*, vol. 60, no. 6, p. 41, 2013.
[24] S. Baud, L. Duca, B. Bochicchio, B. Brassart, N. Belloy, A. Pepe, M. Dauchez, L. Martiny, and L. Debelle, "Elastin peptides in aging and pathological conditions," *Biomol. Concepts*, vol. 4, no. 1, pp. 65–76, Feb. 2013.
[25] W. Huisinga and B. Schmidt, *Metastability Dominant Eigenvalues Transfer Operators*, vol. 49. Berlin, Germany: Springer, Jan. 2006, pp. 167–182, doi: 10.1007/3-540-31618-3_11.
[26] T. F. Havel, I. D. Kuntz, and G. M. Crippen, "The theory and practice of distance geometry," *Bull. Math. Biol.*, vol. 45, no. 5, pp. 665–720, 1983.
[27] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Stat. Anal. Data Mining*, vol. 5, no. 5, pp. 363–387, Oct. 2012.
[28] F. Keller, E. Muller, and K. Bohm, "HiCS: High contrast subspaces for density-based outlier ranking," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 1037–1048.
[29] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognit.*, vol. 40, no. 3, pp. 863–874, Mar. 2007.
[30] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
[31] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 582–588, 1999.
[32] A. D. Gordon, "A review of hierarchical classification," *J. Roy. Stat. Soc. Ser. A*, vol. 150, pp. 119–137, Mar. 1987.
[33] G. N. Lance, "A general theory of classificatory sorting strategies: II. clustering systems," *Comput. J.*, vol. 10, no. 3, pp. 271–277, Mar. 1967.
[34] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classification*, vol. 1, no. 1, pp. 7–24, Dec. 1984.
[35] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, Jun. 2014.
[36] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[37] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, Jan. 2007.
[38] J. Lu, G. Xu, S. Zhang, and B. Lu, "An effective sequence-alignment-free superpositioning of pairwise or multiple structures with missing data," *Algorithms Mol. Biol.*, vol. 11, no. 1, p. 18, Dec. 2016.

AZZAM ALWAN received the Diploma and M.S. degrees from the Grenoble Polytechnic School and the Ph.D. degree in systems optimization and safety from the University of Technology of Troyes, in 2019. He is currently a Senior Research Engineer with Relyfe.com. In 2020, he was considered among the top eight innovators in the Middle East and Africa region due to his innovation of a new deep learning architecture to recognize the behavior of smartphone users. His research interests include chronic diseases prediction and medical document analysis.

RÉMI COGRANNE (Member, IEEE) received the Ph.D. degree in systems safety and optimization from the University of Technology of Troyes (UTT), France, in 2011. He was regularly invited as a Visiting Scholar with Binghamton University, from 2014 to 2020. He has been an Associate Professor with UTT, since 2013. His research interests include hypothesis testing applied to image forensics, steganalysis, steganography, and computer networks. He has published more than 75 articles, among which two received best paper awards, and three International patents. He is a member of the Editorial Board of IEEE TRANSACTIONS ON INFORMATION FORENSICS (T-IFS) and *Security* and an Elected Member of IEEE Technical Committee on IFS. He has been a TPC Chair of IEEE WIFS 2021, the General Chair of ACM IH&MMSec 2019, and the Main Organizer of ALASKA Steganalysis Challenge (https://alaska.utt.fr).

PIERRE BEAUSEROY (Member, IEEE) received the B.E. degree in computer science and the master's degree in control of complex systems, in 1988, and the Ph.D. degree in control of complex systems from the University of Technology of Compiègne, France, in 1992. Since 1993, he has been a Researcher with the University of Technology of Troyes and a member of the Modeling and System Dependability Laboratory. From 2011 to 2016, he directed a laboratory before being appointed as a Head of the ROSAS Department (operational research, applied statistics, and simulation). He was appointed as a Professor, in 2010. He has been a partner in over 25 projects with national and industrial research groups. His research interests include detection, pattern recognition, and machine learning. He has been a recipient of the National Doctoral Supervision and Research Award, since 2006.

EDITH GRALL-MAËS received the Ingenieur degree from the Compiègne University of Technology, France, in 1989, and the Ph.D. degree (Doctorat ès Sciences Techniques) from the Federal Polytechnic School of Lausanne, Switzerland, in 1993. In 1995, she joined the Troyes University of Technology, France, where she is currently an Associate Professor. She is a partner of industrial projects and program projects. Her research interests include supervised and unsupervised learning decision methods mainly based on data for diagnosis and prognostic for various applications, such as medical, aeronautics, and energy.

NICOLAS BELLOY received the master's degree in structural biology and nanobiology and the Ph.D. degree in molecular modeling from the Université de Reims Champagne Ardenne, in 2005 and 2009, respectively. He is currently a Research Engineer with the Université de Reims Champagne Ardenne, where he has been a member, since 2009. His main research interests are molecular dynamics and molecular modeling of extracellular matrix components to describe their conformational dynamics and interactions. He is also interested in developing new approaches based on rigid bodies simulation to address these questions. His research interests include biology, physics, and computer science.

LAURENT DEBELLE received the Ph.D. degree in biochemistry and biophysics from the University of Reims Champagne Ardenne (URCA), France, in 1995. He served as a Postdoctoral Researcher with the Laboratory of Organic Chemistry, the University of Potenza, Italy, for several years. In 1998, he was recruited as an Associate Professor of biochemistry with URCA. In 2009, he was appointed as a Full Professor. Since 2019, he has been an Honorary Senior Researcher with the Division of Cell Matrix Biology and Regenerative Medicine, School of Biological Sciences, Faculty of Biology Medicine and Health, University of Manchester, U.K. He has been invited to present his work in major matrix-related congresses. The approaches he uses are at the interface of biology, biophysics, and numerical simulations. He has authored more than 40 articles and one patent. His research interests include elastin structure-function relationships and their significance in pathophysiological states and notably during aging. He is a member of the French Society for Matrix Biology, the International Society for Matrix Biology, and the French Biophysics Society.

STÉPHANIE BAUD received the Ph.D. degree in physics from the University of Franche-Comté, in October 2004. From February 2005 to August 2007, she carried out a thematic conversion in biophysics. During her Ph.D. training at the Molecular Structure and Function Department, The Hospital for Sick Children, Toronto, she developed numerical simulations on elastic proteins. In September 2007, she was recruited at the Extracellular Matrix and Cellular Dynamics (MEDyC) Research Unit, University of Reims Champagne-Ardenne. In 2020, she was appointed as a Full Professor. For the last years, she has been working in close collaboration with a team of experimental biologists, trying both to respond to their structural issues/challenges and to bring them new lines of work such as new molecules to be tested from molecular modeling work. She was involved in the research projects like structural characterization and structure/function relationship of matrikines, study of the interactions between matrikines and proteins constituting the extracellular matrix, and methodological developments dedicated to the study of biological systems.

MANUEL DAUCHEZ received the Ph.D. degree in physics and computational biophysics from the University of Lille, in 1990. He is working in the field of molecular modeling of glycoconjugates and developments of corresponding force fields. After a Postdoctoral Research position with the University of Oregon, USA, working on the junction between canonical forms of DNA, he came back to France to develop a potential energy function in order to reproduce vibrational spectra. In September 1993, he was recruited as an Assistant Professor with the University of Reims Champagne-Ardenne (URCA) to develop molecular modeling methodologies in a health laboratory working on lung diseases. In 2000, he was appointed as a Full Professor and joined the Extracellular Matrix Unit, French National Center CNRS, URCA. Since 2000, he has been working to promote molecular modeling, molecular dynamics, and corresponding developments and applying these methodologies to study the structures/functions/dynamics of extracellular matrix (ECM) components. Since last five years, he has been in charge of the chair of excellence in research in molecular modeling (MAgICS) installed at the Reims Champagne-Ardenne University. He has promoting all these numerical simulations in the complexity of the ECM. He was involved in numerous projects to characterize structures and dynamics of peptides, proteins studied alone or in interactions and to decipher the interactions between matrikines and matricellular proteins. During the last decade, he proposed numerous methodological developments dedicated to the study of the ECM biological systems.

SÉBASTIEN ALMAGRO received the Ph.D. degree in biophysics from Joseph Fourier University, Grenoble I, in 2003. He has been an Associate Professor with the University of Reims Champagne Ardenne (URCA), France, since 2010. He has published around 20 articles in a broad range of biological domains from chromosome structure, molecular motors and cell adhesion, protein traffic and recycling to membrane, and molecular dynamics. His research interests include structural biology, cell dynamics, and image analysis.