STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXII**, Number 1, 2017
DOI: 10.24193/subbi.2017.1.08

USING COMPUTATIONAL INTELLIGENCE MODELS FOR ADDITIONAL INSIGHT INTO PROTEIN STRUCTURE

MARIA-IULIANA BOCICOR[1], ALESSANDRO PANDINI[2], GABRIELA CZIBULA[1], SILVANA ALBERT[1], AND MIHAI TELETIN[1]

Abstract. Proteins are large, complex molecules with crucial roles in the functioning of living organisms. Understanding the underlying mechanisms by which proteins achieve their structures and substructures, as well as those involved in the conformational transitions may contribute to a deeper comprehension of the involved biological processes. This paper investigates a new machine learning perspective upon analyzing protein conformational transitions and introduces a new formalization for the problem, with the more general goal of uncovering interesting patterns in protein conformational transitions. This study represents the starting point of a research which is being conducted in order to obtain a better comprehension of proteins' structures and, implicitly, functions, by investigating *computational intelligence* methods for analyzing and deducing proteins conformational transitions.

1. Introduction

Proteins are large, complex molecules with crucial roles in the functioning of living organisms: they can be building blocks in the body (structural proteins), they catalyze biochemical reactions in metabolism (enzymes) or they may execute key tasks in maintaining the cellular environment. Moments after a protein is synthesized it folds, forming a stable three-dimensional (3D) structure, which is known to define the protein's function and which is entirely dictated by the linear sequence of amino acids composing the protein [26]. According to various external factors from the protein's environment (e.g. temperature, interaction with other molecules), modifications in the protein structures occur during their biological functions. Thus, a protein

Received by the editors: May 14, 2017.

2010 *Mathematics Subject Classification.* 68T05, 62H30.

1998 *CR Categories and Descriptors.* I.2.6 [**Computing Methodologies**]: Artificial Intelligence – *Learning*; I.5.3 [**Computing Methodologies**]: Pattern Recognition – *Clustering*.

Key words and phrases. Protein conformations, Computational Intelligence, Machine learning, Self-organizing maps.

will acquire a limited number of alternative conformations (belonging to the same fold), having the ability to transition between them [25]. Understanding protein conformational transitions and protein dynamics is essential for the comprehension of biomolecular interactions. This is of paramount importance in the process of developing new drugs that can inhibit proteins' uncontrolled behaviour, which can arise in pathological cases (such as protein incorrect folding or mutations) [15].

The contribution of the paper is summarized as follows. Our first goal is to explore a new *machine learning* perspective upon studying protein conformational transitions. Starting from the current state-of-the-art which refers to the analysis of conformational changes in proteins, we propose a new computational model for the problem of predicting protein conformational transitions. Secondly, we aim to provide an intuition upon the applicability of *machine learning* techniques for uncovering interesting patterns in the structure of proteins. The study performed in this paper represents the starting point of a research which is being conducted in order to obtain a better comprehension of proteins' structures and, implicitly, functions, by investigating *computational intelligence* methods for analysing and deducing proteins conformational transitions. The long-term goal of our research is to contribute to a better understanding and to offer additional insight into the construction and functioning of proteins.

The rest of the paper is organized as follows. Section 2 presents the motivation of our approach, highlighting the importance and relevance of understanding *protein conformational transitions*, but the difficulty of the problem as well. The biological background related to our approach is given in Section 3. The current state-of-the-art, as well as the limitations of existing approaches related to the analysis of protein structure are presented in Section 4. Section 5 introduces our *machine learning* perspective on the problem, together with an incipient computational model. A case study which highlights the applicability of *machine learning* methods for analyzing protein conformational transitions is described in Section 6. The conclusions of the paper, as well as directions for continuing our research are pointed out in Section 7.

2. Motivation

Although the stable 3D structure of a protein is defined by a unique topology (i.e. fold), this structure is not static and it is now widely accepted that proteins are dynamic objects [25]. According to various external factors from the protein's environment (e.g. temperature, interaction with other molecules), modifications in proteins' structures occur during their biological functions. A protein will thus acquire a limited number of conformations and will have

the ability to transition between alternative conformations. Understanding protein dynamics and how these conformational transitions occur is essential for the comprehension of biomolecular interactions, which is of paramount importance in the process of developing new drugs that can inhibit proteins' uncontrolled behaviour [15].

When investigating the role of conformational transitions in biological function from a computational perspective, the first stage is devising a formalisation of the problem in question, which involves specific domain knowledge and thus a collaboration between biologists, chemists, physicists and computer scientists. Various formal abstractions of problems related to protein structure, or their equivalent transformations have been proven to be NP-hard or NP-complete [7, 8], which means that there are no algorithms which can solve these problems in realistic time. The complexity of the *protein conformational transitions* problem is further increased by the high dimensionality of the space to be explored. For such classes of problems, heuristic techniques inspired from artificial intelligence and mathematical optimisation are certainly suitable candidates. In addition to the difficulties mentioned above, obtaining sufficient relevant experimental biological data for thorough analyses and understanding is time-consuming and financially expensive.

Both the importance and the complexity of the problem motivate us to investigate the usefulness of *machine learning* models and methods for the analyzing and detecting the conformational changes in proteins. Our perspective on the problem is new, to the best of our knowledge it has not been investigated in the literature, yet. We are confident that *machine learning* based solutions are applicable and may lead to interesting and valuable information, due to these models' ability to discover hidden patterns in data.

3. Background

Proteins are large molecules, having significant roles in the structure, development and functioning of living organisms. They are composed of basic building blocks - *amino acids* - small molecules which chain together in order to create proteins. The amino acids sequence forms the primary structure of the protein, which can be represented as a string of symbols representing the 20 amino acids (they are encoded by the letters of the alphabet). Although the sequence of amino acids is linear, the protein does not have an extended conformation, as intramolecular forces between the amino acids lead to a folding of the protein. As soon as it is synthesized as a linear sequence of amino acids, a protein folds in a matter of seconds to a stable three dimensional structure called the protein's native state. This structure of the protein is very important, as it defines the protein's function. However, proteins are

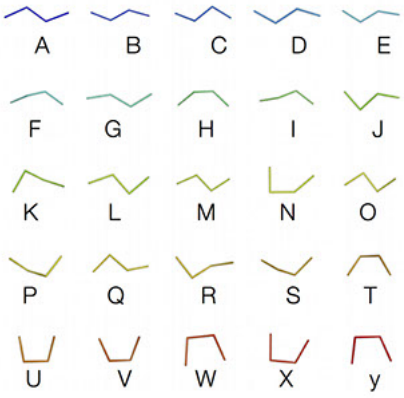


FIGURE 1. Structural elements and their associated symbols of the SA. Figure source: Alessandro Pandini [19].

dynamic molecules and undergo slight changes in their structures, according to the function they are fulfilling and depending on environmental conditions. Understanding and tracing these conformational changes (or transitions) could help us gain new insight into the way proteins function.

When studying protein conformations, it can be noticed that there are several frequently occurring conformations for small fragments. These so-called *states* have been determined with various methods and encoded in Structural Alphabets (SA) [18], which contain codes for the re-occurring short conformations. There are several types of SA, derived using various methods [16]. These are particularly useful in computational applications, as they allow representing a three dimensional structure via a one dimensional array (a sequence of characters of the alphabet), thus facilitating analysis of protein structure.

In our study we employ the structural alphabet derived by Pandini et al. in [18]. This is composed of 25 codes, represented by 25 letters of the (conventional) alphabet, each letter representing the short structural (three dimensional) element composed of four amino acids in the linear sequence of the protein. The structural elements and their associated letters of the SA are depicted in Figure 1. The structural element is characterized by the two angles between consecutive amino acids (more specifically, between the alpha carbon atoms of these amino acids) and by the torsion angle formed by all four atoms [18].

Let us consider a protein Pr, whose primary sequence is composed of n amino acids: $Pr = p_1p_2 \cdots p_n$. Then, a structural conformation of protein Pr can be represented as a sequence of letters of length $n - 3$, where each letter encodes the structure formed by four amino acids in the primary sequence:

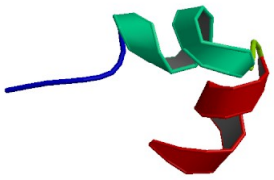


FIGURE 2. 3D view of protein 1HP9. Image from the RCSB PDB (`www.rcsb.org`) [2] of PDB ID 1HP9 [24] [2].

$\mathcal{C} = s_1 s_2 \cdots s_{n-3}$. Slight changes in a protein's conformation lead to different conformations (thus different representations). One could imagine a sliding window of length four, passing over the protein's structure and each group of four amino acids is represented by a symbol of the SA. An as example, we present protein $1HP9$ [1] (a toxin from scorpion venom), which has a short primary sequence (22 amino acids): GHACYRNCWREGNDEETCKERC. A three dimensional view of this protein is illustrated in Figure 2. Five possible SA representations of this protein from the analysis of conformations available in a database of molecular simulations [13] are shown below (the symbols in these representations are symbols of the structural alphabet [18]):

- `QSUWNSVVVPRIJUUVVUV`
- `QSUWNSVVVPRIJUUVUUV`
- `RSUWNSVVVPRIKUUVVUV`
- `QSUWNSVVVPRGKUUVVUV`
- `QSUWNSVVVPRGKUUVVUV`

It can be noticed that all these conformations have the same length of 19 symbols ($= 22 - 3$) and there are very slight differences among them. The amount of changes is consistent with the timescale of the original simulation: conformations were recorded at intervals of 1 picosecond.

4. Literature review

Several theoretical models have been proposed for modelling conformational transitions, among which we mention those introduced by Miyashita et al. [14], Whtford et al. in [27], Skjaerven et al. [22]. These were employed by physics-based computational methods, such as molecular dynamics [17] or Monte Carlo [12] to simulate the movement of atoms. However, although having the potential to offer valuable information about protein structure, these simulations

[1] http://www.rcsb.org/pdb/explore/explore.do?structureId=1hp9

[2] This image is used according to RCSB PDB Policies & References: `http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html`.

are extremely computationally expensive and thus their time intervals are considerably shorter than those of real biological conformational changes. Normal Mode Analysis [22] and simplifications of it have been used in several cases for modelling protein conformational transitions: Schuyler et al. present in [21] a tool which is able to generate a transition pathway from a source to a destination conformation and Al-Bluwi et al. use in [1] robotics inspired methods (motion planning algorithms) to model conformational transitions.

Another technique presented by Haspel et al. in [9], who propose to trace conformational changes from a start to a goal conformational state by mapping the protein to a reduced representation, capturing low-energy conformations with the help of a coarse-grained physics based energy function and applying a sampling-based motion planning algorithm (again, inspired from robotics). The limitations of these later solutions are that they either use relatively simple energy functions (which thus only consider a small number of energy parameters), or they provide approximations of the paths, which require further refinement.

Raveh et al. introduce in [20] an approach called *PathRover* that, based on initial external constraints can generate motion pathways. The motion planning algorithm takes into account any available prior information and incorporates it into the algorithm of rapidly exploring random trees (RTT). This solution's main advantage is that by using initial constraints, it narrows down the search in high-dimensional spaces thus being significantly faster. They managed to do that by integrating their solution into *Rosetta* - modelling framework that aggregates algorithms for computational modeling and analysis of protein data. In order to successfully integrate it, they had to provide energy functions, optimising protocols and techniques for sampling.

The generated pathways are the result of partial data assimilation in sampling-based motion planning of molecules. As a result, each pathway has to form a sequence that satisfies all the initial restrictions while consisting of clash-free low-energy conformations. The challenge still remains in extracting physical features from simulated motion and being able to bridge experimental and computational observations. Significantly less options are explored in [20] because of the use of partial input but there is no learning involved based on existing findings.

Cortés et al. propose in [4] a computational approach based on path planning. The technique is intended to predict the motions of the molecules of the proteins. It is mentioned that motion planning techniques have lots of applications in computational biology and that they can be successfully applied on protein study. The proposed approach is split in two main stages, a *geometric filtering phase* and an *energy based computation* applied only on the

solutions extracted from the first stage. One of the advantages of this split is the increase in computational speed. The approach analysis shown that the filtering stage is very effective and that it is capable to present very important knowledge to biologists. However, the second stage still has some limitations, since it cannot exploit all the provided knowledge.

The study we conducted on the current state-of-the-art on the problem of identifying proteins conformational transitions revealed that a *machine learning* based computational model has not been investigated in the literature, yet.

5. Theoretical model. Our proposal

As opposed to other approaches in the literature (Section 4), we tackle the problem of determining conformational transitions in proteins from a different angle and we derive a different formalization for it, starting from a data set of more than 300 proteins and their associated conformations. As described in Section 3, a protein Pr or length n can be viewed as a word over the alphabet of 20 letters representing amino acids $\mathcal{A} = \{G, P, A, V, L, I, M, C, F, Y, W, H, K, R, Q, N, E, D, S, T\}$: $Pr = p_1 p_2 \dots p_n$, where $p_i \in \mathcal{A}, \forall i \in \{1, 2, \dots, n\}$.

For each protein we are given thousands of different conformations, obtained by molecular dynamics simulations. Each conformation is converted into its SA representation. The structural alphabet is composed of the 25 letters shown in Figure 1: $\mathcal{SA} = \{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y\}$. It is important to remark that although same symbols are being used both for amino acids and for structural elements, these are actually completely different concepts (this is important to be remembered when processing and experimenting on the data).

For each protein $Pr = p_1 p_2 \dots p_n$ in the data set, we are given a large number m of experimentally determined conformations (for the data set we use, $m = 10000$). Therefore, for each protein we have a set $\mathcal{S} = \{c_j \mid c_j = c_j^1 c_j^2 \dots c_j^{n-3},\ j \in \{1, 2, \dots, m\},\ c_j^k \in \mathcal{SA},\ k \in \{1, 2, \dots n-3\}\}$ of conformations. Considering all these conformations, a distribution matrix is computed for each protein, which holds information about the SA elements' distribution, for each position k, $\forall k \in \{1, 2, \dots, n-3\}$. This frequency matrix can be interpreted as a "profile" of the protein dynamics where for each fragment position we have a probabilistic measure of the occurrence of each letter in the alphabet. An example of such a matrix, for the 5 conformations of the protein $1HP9$ presented in Section 3, is given in Table 1. For each position in the SA representation we compute the probability of occurrence of each symbol of

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0.6	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0.6	0	0	0	0	0	0
N	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Q	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0.2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
S	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
U	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
V	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	1
W	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 1. Distributions of SA symbols for the example presented in Section 3.

the SA on that specific position. For simplicity, in Table 1 we only show the distributions of the symbols occurring in the 5 conformations, but on a real world example, all symbols of the SA are considered.

Furthermore, other useful biological information about the protein and the composing amino acids can be considered (structure related information). For instance, a property an amino acid is characterized by is the relative solvent accessibility (RSA), which measures the solvent exposure of the amino acid. This property is numerical and it can have different values for the same amino acid, belonging to different structural environments. Another example would be the amino acid's hydrophobicity, a physical property which measures how much the amino acid is repelled by water. This is important, as hydrophobic forces are decisive factors in the protein folding process.

Considering the input information described above, we formulate the problem of determining protein conformational transitions as follows:

- *Given:*
  - A protein, as a string of amino acids.
  - Other structurally significant, biologic characteristics of amino acids (e.g. RSA values, hydrophobicity).
  - A *small* number of conformations (e.g. 10 conformations, determined using molecular dynamic methods).
- *The requirement is to solve any or both problems below:*

- Generate a matrix of probability distributions similar to the one presented in Table 1, corresponding to all m possible conformations (even though these are not known).
- Generate all m possible conformations for the protein.

Our aim is to further formalize the problem, considering various combinations of possible input data in order to be able to approach it from a machine learning perspective. Nonetheless, both requirements are difficult and conventional machine learning techniques are very probably not sufficient for satisfying results, therefore a more thorough investigation, as well as new or hybrid techniques are demanded in order to solve any of the two formulations of the problem.

6. Experiments

In this section we aim to give an empirical confirmation of our hypothesis that *machine learning* methods are applicable for analyzing proteins conformational transitions. More specifically, our focus is to highlight that *unsupervised learning* methods are able to capture patterns among the conformations of the same protein, as well as relationships between related proteins, relations which are confirmed from a biological perspective.

We considered an experiment consisting of *seven* proteins (codes: 1ASH, 1DLW, 1ECA, 1C52, 1CCR, 1APQ, 1COU in the Protein Data Bank [2]), taken from three different superfamilies (1.10.490.10, 1.10.760.10, 2.10.25.10). The superfamilies for the proteins were determined using **CATH Protein Structure Classification** database [3] which is a publicly available online resource that provides information on the evolutionary relationships of protein domains [5]. In this database, two proteins are considered in the same superfamily if there is a similarity between their three-dimensional structure [11].

Table 2 illustrates the superfamilies for the seven proteins considered in our experiment, as well as the similarity index between the proteins belonging to the same superfamily, as provided by the FATCAT algorithm (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists) [28].

From Table 2 we observe that the proteins from the first two families have a similarity index about 20%, while the proteins from the third family have the lowest similarity index of about only 5%.

In order to test our hypothesis that *unsupervised learning* models are able to capture the biological relationships between proteins data, we performed the following experiment.

#	Superfamily	Proteins	Similarity index
1	**1.10.490.10**	{1ASH, 1DLW, 1ECA}	1ASH - 1DLW: 20.57% 1ASH - 1ECA: 25.85% 1ECA - 1DLW: 19.08%
2	**1.10.760.10**	{1C52, 1CCR}	1C52 - 1CCR: 27.10%
3	**2.10.25.10**	{1APQ, 1COU}	1APQ - 1COU: 4.92%

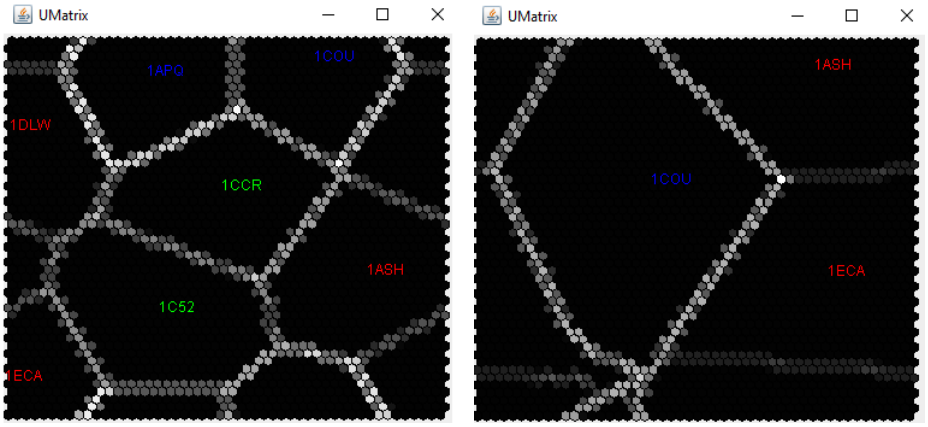
TABLE 2. Sample proteins

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
0	0	0	0	0	0	0.021	0	0.031	0.021	0.031	0	0	0.052	0	0.052	0.042	0.063	0.105	0	0.221	0.305	0.052	0	0

TABLE 3. Probabilities of occurrence of SA symbols for the example presented in Section 3.

We are further considering the theoretical model introduced in Section 5. For each protein, 10000 conformational transitions are known. The specific data we use is retrieved from MoDEL, a database which includes representatives from different protein families and fold arrangements [13]. Our current experiment's goal is to investigate whether biologically relevant correlations could be found within the given numerical data and to mine this given data in order to discover significant signals than can later be used by machine learning strategies to solve the problem described and defined in Sections 3 and 5. For this purpose, we use a further simplified representation of a protein: instead of the frequency matrix, we use a frequency vector, constructed as follows. For each of the 25 letters l_i ($1 \leq i \leq 25$) from the structural alphabet and each protein Pr, we compute the probability $p_{l_i}^{Pr}$ of occurrence of each letter l_i in the conformational transitions of protein Pr. Thus, a protein Pr may be visualized as a 25-dimensional vector containing the probabilities of occurrence of the symbols from the structural alphabet in the given protein, $Pr = (p_{l_1}^{Pr}, p_{l_2}^{Pr}, \ldots, p_{l_{25}}^{Pr})$. For the protein example presented in Section 3 (1HP9), including the 5 presented conformations, the frequency vector is presented in Table 3.

Considering the above modelling, each of the seven proteins considered in our case study is represented as a multi-dimensional vector. Our focus is to test if the conformational transitions of the proteins provide useful information regarding their three-dimensional structure and if an *unsupervised learning* model is able to capture this type of biological relationships between the proteins.



(A) Proteins 1ASH, 1DLW, 1ECA, 1C52, 1CCR, 1APQ and 1COU.

(B) Proteins 1ASH, 1ECA and 1COU.

FIGURE 3. U-Matrix visualization.

We will use a *self-organizing map* (SOM) as an *unsupervised learning* model. SOMs are known to be powerful *data mining* tools for visualizing high-dimensional data. A *self-organizing map* [23] is a type of artificial neural network that is trained using *unsupervised learning* to provide a low-dimensional representation of the high-dimensional input space, called a *map* [6]. The *topological mapping* is the main characteristic of the unsupervised mapping provided by a SOM, more exactly the input samples which are close to each other in the input space will be mapped into neighboring neurons on the output map.

6.1. **Results and discussion.** We mapped the *seven* proteins described above (considering their 25-dimensional representations) on a SOM having a *torus* topology. For the SOM visualization, we use the U-Matrix method [10] with the following interpretation: the lighter regions express data that are dissimilar while darker regions contain data that are similar.

Figure 3a depicts the U-Matrix visualization of the SOM trained on the *seven* proteins. Visualizing the U-Matrix for the resulting map, we clearly observe three regions corresponding to the three protein families described in Table 2.

Figure 3b illustrates the U-Matrix visualization of the SOM trained on only *three* proteins: 1ASH, 1ECA and 1COU. From these, only the first two belong to the same superfamily. This can be visualized on the U-Matrix, since there is

a clear separating boundary between the protein 1COU and the class formed by the other two proteins.

The results previously described and depicted in Figures 3a and 3b indicate the potential of unsupervised *machine learning* models (the *self-organizing map*, in our case) to uncover patterns encoded in the conformational transitions of proteins.

7. Conclusions and future work

In this paper we have investigated the problem of analyzing the conformational transitions of proteins, with the more general goal of contributing to a comprehensive understanding of the problem. We presented the current state-of-the-art approaches and we proposed a new computational perspective on the problem, based on *machine learning*. Our proposal represents the starting point of a research initiated on the topic approached in this paper, our long-term goal being to offer additional insight into the construction and functioning of proteins.

We also highlighted, through a *data mining* experiment, that the information obtained through analyzing proteins conformational transitions capture the relationships between related proteins, relations which are confirmed from a biological perspective.

Starting from the computational model proposed in Section 5, future work will be done in order to apply concrete supervised *machine learning* methods (e.g. *artificial neural networks*, *support vector machines*) for predicting the conformational transitions of proteins, as well as the matrix of probability distributions associated to protein conformations.

References

[1] Ibrahim Al-Bluwi, Marc Vaisset, Thierry Siméon, and Juan Cortés. Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC Structural Biology*, 13(1):S2, 2013.

[2] H.M Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242.

[3] CATH: Protein Structure Classification Database at UCL. CATH - Gene3D. `http://www.cathdb.info`.

[4] J. Cortés, T. Siméon, V. Ruiz De Angulo, D. Guieysse, M. Remaud-Siméon, and V. Tran. A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, 21(1):116–125, January 2005.

[5] Natalie L. Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45(D1):D289, 2017.

[6] N. Elfelly, J.-Y. Dieulot, and P. Borne. A neural approach of multimodel representation of complex processes. *International Journal of Computers, Communications & Control*, III(2):149–160, 2008.

[7] Aviezri S. Fraenkel. Complexity of protein folding. *Bulletin of Mathematical Biology*, 55(6):1199 – 1210, 1993.

[8] Christophe Guyeux, Nathalie M.-L. Cote, Jacques M. Bahi, and Wojciech Bienia. Is protein folding problem really a NP-complete one? First investigations. *Journal of Bioinformatics and Computational Biology*, 12(01):1350017–1350041, 2014.

[9] N. Haspel, M. Moll, M. L. Baker, W. Chiu, and L. E. Kavraki. Tracing conformational changes in proteins. In *2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop*, pages 120–127, Nov 2009.

[10] S. Kaski and T. Kohonen. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. In *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, pages 498–507. World Scientific, 1996.

[11] Michael Knudsen and Carsten Wiuf. The CATH database. *Human Genomics*, 4:207–212, 2010.

[12] I. Lotan, F. Schwarzer, and J.C. Latombe. Efficient energy computation for monte carlo simulation of proteins. *Lecture Notes in Computer Science*, 2812:354–373,, 2003.

[13] Tim Meyer, Marco D'Abramo, Adam Hospital, Manuel Rueda, Carles Ferrer-Costa, Alberto Prez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluis Gelp, and Modesto Orozco. MoDEL (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure*, 18(11):1399 – 1409, 2010.

[14] Osamu Miyashita, Peter G. Wolynes, and Jos N. Onuchic. Simple energy landscape model for the kinetics of functional transitions in proteins. *The Journal of Physical Chemistry B*, 109(5):1959–1969, 2005.

[15] G. Morra, M. Meli, and G. Colombo. Molecular dynamics simulations of proteins and peptides: from folding to drug design. *Current Protein and Peptide Science*, 9:2181–2196, 2008.

[16] B. Offmann, M. Tyagi, and A.G. de Brevern. Local protein structures. *Current Bioinformatics*, 2(3):165–202, 2007.

[17] Kei-ichi Okazaki, Nobuyasu Koga, Shoji Takada, Jose N. Onuchic, and Peter G. Wolynes. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 103(32):11844–11849, 2006.

[18] A. Pandini, A. Fornili, and J. Kleinjung. Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics*, 11(97):1–18, 2010.

[19] Alessandro Pandini. Structural alphabet tools for molecular simulations. `http://people.brunel.ac.uk/~csstaap2/software.html`. [Online; accessed 12-May-2017].

[20] B. Raveh, A. Enosh, O. Schueler-Furman, and D. Halperin. Rapid sampling of molecular motion with prior information constraints. *PLoS Computational Biology*, 5(2), February 2009.

[21] Adam D. Schuyler, Robert L. Jernigan, Pradman K. Qasba, Boopathy Ramakrishnan, and Gregory S. Chirikjian. Iterative cluster-nma: A tool for generating conformational transitions in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):760–776, 2009.

[22] Lars Skjaerven, Siv M. Hollup, and Nathalie Reuter. Normal mode analysis for proteins. *Journal of Molecular Structure: {THEOCHEM}*, 898(13):42 – 48, 2009.
[23] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10:151–159, 1999.
[24] K.N. Srinivasan, V. Sivaraja, I. Huys, T. Sasaki, B. Cheng, T.K. Kumar, K. Sato, J. Tytgat, C. Yu, B.C. San, S. Ranganathan, H.J. Bowie, R.M. Kini, and P. Gopalakrishnakone. kappa-hefutoxin1, a novel toxin from the scorpion heterometrus fulvipes with unique structure and function. importance of the functional diad in potassium channel selectivity. *J.Biol.Chem*, 277:30040–30047, 2002. PDB ID: 1HP9.
[25] Nobuhiko Tokuriki and Dan S. Tawfik. Protein dynamism and evolvability. *Science*, 324(9524):203–207, 2009.
[26] D. Voet and J. Voet. *Biochemistry*. Wiley, 4 edition, 2011.
[27] Paul C. Whitford, Osamu Miyashita, Yaakov Levy, and Jos N. Onuchic. Conformational transitions of adenylate kinase: Switching by cracking. *Journal of Molecular Biology*, 366(5):1661 – 1671, 2007.
[28] Yuzhen Ye and Adam Godzik. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, 32:582–585, 2004.

[1] Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania
E-mail address: {iuliana, gabis, albert.silvana}@cs.ubbcluj.ro, tmic1334@scs.ubbcluj.ro

[2] Department of Computer Science, Brunel University, London, England
E-mail address: alessandro.pandini@brunel.ac.uk