Identification of structure clusters in molecular dynamics trajectories from Residue Interaction Networks

Edoardo Spadetto[1], Riccardo Carangelo[2], and Alessandro Benetti[3]

[1]University of Padua, Department of Physics, M.Sc. in Physics of Data , mail: edoardo.spadetto@studenti.unipd.it, ID: 1236682
[2]University of Padua, Department of Biology, M.Sc. Industrial in Biotechnologies, mail: riccardo.carangelo@studenti.unipd.it, ID: 1190173
[3]University of Padua, Department of Mathematics, M.Sc. in Computer Science , mail: alessandro.benetti.1@studenti.unipd.it , ID: 1210974

Keywords: *Protein Clustering, Molecular Dynamics Trajectories, Weak Residue Interactions, Structural Bioinformatics, Bioinformatics Algorithms, Network Analysis, Stability Analysis*

Abstract

The aim of this work is to provide a simple, yet effective, algorithm usable as a bioinformatic prediction tool for a fast and efficient *in silico* identification of the most relevant clusters through the analysis of the "trajectories" data derived from a generic molecular dynamics experiment. In this article we try the algorithm on sample proteins, performing a PCA of 60 dimensions , which is susequently clustered using an Agglomerative Clustering. Further results are then calculated, like "the contact maps and"¡¡ the mean values of the RMSD for the clusters. We provide also several structural comparisons of the representative models for the clusters with the wild type variant of the same proteins.

1 Introduction

A molecular dynamics simulation outputs several information and structures about a specific protein in question. This kind of *in silico* analysis explore the behaviour of a specific protein in a given simulation box environment, with specific chemical and physical boundary conditions and specific timeframe settings. For each frame of the simulation it is possible to get a snapshot, that could be a .pdb file which contains a static set of 3D coordinates of every atom in the protein in that certain frame. These files can be visualized or used to conduct various tests[1]. In this case, from each .pdb file it was computed an interaction network file containing information related to bonds occurring between the residues of a protein. The purpose of this project is to compute a clustering of the interaction network files, in order to identify snapshot's subgroups displaying similar conformations. This way we can find out how many, and which, persistent states are present during the timeframe of any possible molecular dynamics simulation. Such clusters may potentially contain all sorts of conformations, from stable to less stable states of the molten globule, until the totally denatured states which could possibly occur. An efficient individuation of these states allow us to infer about the behaviour of the proteins in a given set of simulation conditions and also to use the results for further biochemical and medical analysis.

2 Methods

All the interaction network files (RING output files) were produced by submitting all the .pdb files derived from molecular dynamics simulations to the Residue Interaction Network Generator tool (RING)[2]. To build the vector representing each snapshot we used several attributes that can be divided in two main parts. The first part consists in $2N^2$ dimensions, where N represent the dimension of the set of residues appearing in all the RING output files. These store for each couple of amino acids the energy (E) and the inverse distance ($1/d$) of the bond occurring between the two residues. The second part of the vector is obtained from representing each snapshot as a graph[6]. The graphs are obtained by using as nodes the amino acids appearing in the corresponding RING output file, and the ones not appearing in the file, but nec-

essary to get a consistent structure of the protein chains. Moreover, this choice led the protein chains represented through the graph to have as terminal residues the ones with the minimum and maximum indexes appearing in the corresponding RING output file for that respective peptidic chain. As edges we used the peptidic bonds and the ones appearing in the RING output files with weight equal to the inverse of energetic linear density of the bond ($^d/_E$). From the graphs then were computed the betweenness centrality, the degree centrality and closeness centrality[6] of each node obtained from the residues appearing in the RING output file. These values then were stored in a data-set with the number of rows equal to the number of snapshots and number of columns equal to $3N$.

The complete data-set, obtained by merging the data from the two parts, had a number of rows equal to the number of snapshots and a number of columns equal to $2N^2+3N$. To reduce the dimensionality we performed at first a scaling of all the variables to a (0,1) range and then a PCA. The result of the 60-dimensional PCA has been clustered trough an Agglomerative Clustering algorithm, with the prescription of a ward linkage. The number of groups chosen is the one that maximized the silhouette score of the clusters. With the single purpose of visualization we also computed a 3D PCA on the 60-dimensional data-set, and its output is the one used in the 3D graphs. After the clustering was computed we decided to compute two scores, one is based on the scaled data-set before applying the PCA and the other one is based on the .pdb files and it is computed using the RMSD, a mathematical tool used to evaluate the distance travelled by each atom[8] in two different snapshots.

The first score, that will be called M is:

$$M_{\alpha,\beta} = \frac{\sum_{i\in\alpha}\sum_{j\in\alpha}\sqrt{(\vec{X_i}-\vec{X_j})^2}}{\#\alpha\cdot\#\beta} \qquad (1)$$

where X_i and X_j represent two vectors of dimensions $2N^2+3N$ from the original data-set, the one obtained after performing the scaling, but before the 60-dimensional PCA. These vectors are associated with two different snapshots, in fact α and β identify the clusters that these snapshots belong to.

The second score, S, is:

$$S_{\alpha,\beta} = \frac{\sum_{i\in\alpha}\sum_{j\in\alpha} RMSD(PDB_i, PDB_j)}{\#\alpha\cdot\#\beta} \qquad (2)$$

where PDB_i and PDB_j are .pdb files associated to the snapshots i and j. The score S is a pseudo-metric that quantifies on average how much a cluster is dissimilar to another one and it will be represented with matrices. The M computes an analogous quantity based on the information contained in the RING output files, not in the .pdb.

Additionally, it has been also possible to select the most representative snapshot for each cluster and then perform a structural comparison analysis. The selection of the representative snapshots has been achieved by finding the ones which minimized their 60-dimensional euclidean distance from the center of their corresponding cluster.

The results of the subsequent structural comparison analysis have been obtained using UCSF Chimera[3]. Through the Chimera's MatchMaker tool we can obtain a visual structural overlap of two conformations and also evaluate the RMSD. The MatchMaker tool is able to superimpose two structures by creating at first sequence alignments, after which the aligned residues pairs are fit (one point per residue). The RMSD calculation is performed on the C_α atoms of all the amino acids and using the Needleman-Wunsch algorithm on a BLOSUM-62 matrix and through pruning iteratively all atom pairs until no pair exceeds a distance of 2.0 Å. The two snapshots represented by the two red dots in **??** have been compared at first with the experimental structure, in order to highlight the conformational differences, and secondly each other.

3 Results

Here we present and discuss the output of the algorithm on several protein structures, the chain A of the STIM1, the P16, and the Frataxin. Such outputs of the algorithm allowed us to discover how many clusters are formed during the molecular dynamics processes and highlighted also which are the representative .pdb files for each cluster.

3.1 STIM1 analysis

The first result proposed is the one obtained from the chain A of the STIM1 protein. From Figure 1 we can appreciate a sort of trajectory in the 3D PCA space. This suggests that the conditions applied in the molecular dynamic simulation effectively modified the protein structure. Performing the clustering on the 60-dimensional data-set we obtained 3 clusters, whose 3D representation can be seen in Figure 13. Looking at the M and S scores (Figure 3 and Figure 2) we noticed that they report similar results, suggesting in both cases that each cluster is more similar to itself than to the other ones.

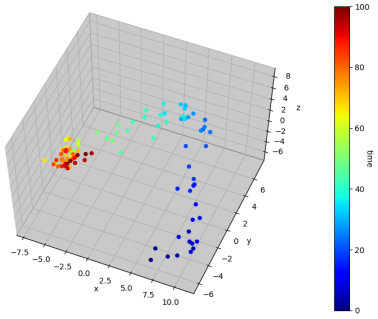

Figure 1: Output of the 3D PCA operated on the chain A of the STIM1 protein data-set. Already with 3 dimension it is possible to appreciate a significant trajectory in the PCA output space.

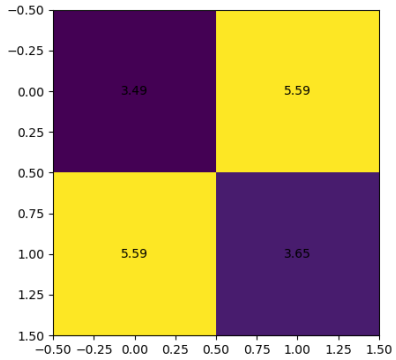

Figure 2: Matrix representation of the S score obtained from the .pdb files of the STIM1's chain A clusters. The darker color of the diagonal shows that the clusters are more similar between themselves than to the others, hence validating the clustering.

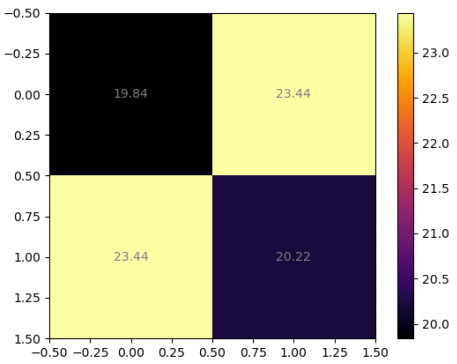

Figure 3: Matrix representation of the M score of the STIM1 clustering. The darker color of the diagonal shows that the clusters are more similar between themselves than to the others.

3.2 P16 analysis

The analysis on the P16 protein produced again 2 clusters, but the results were poorer with respect to the ones obtained for the STIM1 protein (Figure 9). In fact, comparing this outcome with the previous case we can observe that the trajectory in the 3D-PCA space is less defined, but anyway still recognizable (Figure 8). Looking at the scores we observe that the M matrix (Figure 7) confirms the results but, the S score (Figure 10) does not, in fact the dissimilarity of the first cluster with itself is slightly greater than the one with the other cluster. The discrepancy between the M and S scores could mean that is not present a consistent subdivision in clusters for this molecular dynamic simulation. This idea is strengthened by the lower difference between the diagonal and off-diagonal elements in the M matrix respect to the one showed for the M score of the chain A of the STIM1. In general, a discrepancy between the M and S matrices is not necessarily related to the absence of a consistent clustering, but it could rather mean that S score is not representative to appreciate a structural variation in the protein.

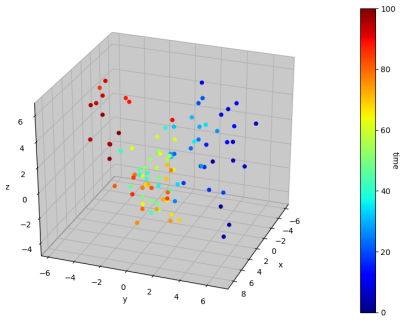

Figure 4: Output of the 3D PCA operated on the P16 protein. The graph figures the 60-dimensional data-set condensed in 3 dimensions trough a PCA. The color gradient represents the temporal index of each snapshot.

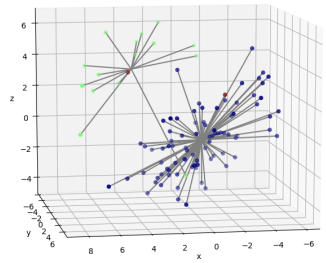

Figure 5: Clustering outcome for the P16 protein. The 3D space represents the output of a PCA applied to the 60 dimensional data-set.

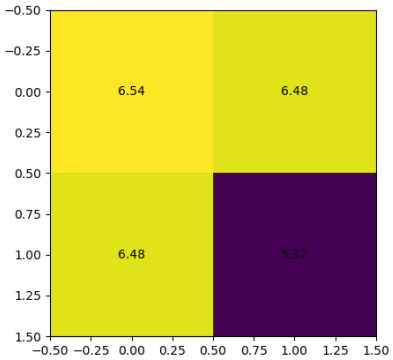

Figure 6: Matrix representation of the S score obtained from the .pdb files of the P16 protein. The fact that an off-diagonal element has a lower score with respect to the one in the diagonal suggest a poor quality clustering.

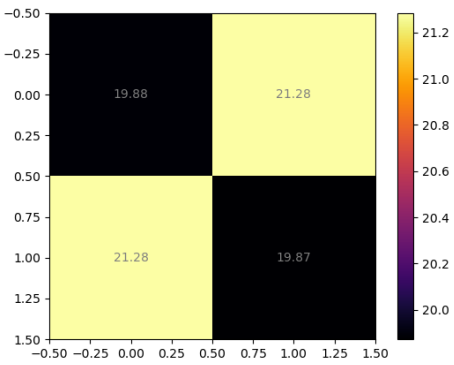

Figure 7: Matrix representation of the M score obtained from the P16's data-set.

3.3 Frataxin analysis

The data-set regarding Frataxin's snapshots, condensed in 3D trough the PCA, pictures an evident trajectory Figure 8. The number of clusters that maximized the silhouette score was 3. Looking at the dissimilarity scores M and S (Figure 11 and Figure 10) we observe that these are consistent with each other.

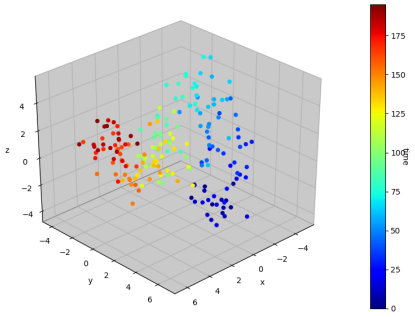

Figure 8: Output of the 3D PCA operated on the Frataxin protein. The graph figures the 60 dimensional data-set condensed in 3D trough a PCA. The color gradient represent the temporal index of each snapshot.

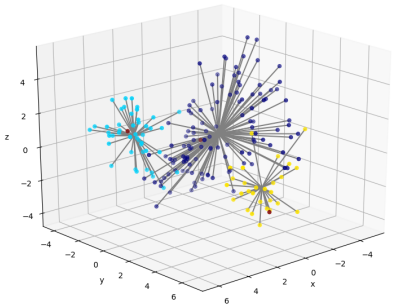

Figure 9: Clustering outcome for the Frataxin protein. The 3D space represents the output of a PCA applied to the 60 dimensional data-set.

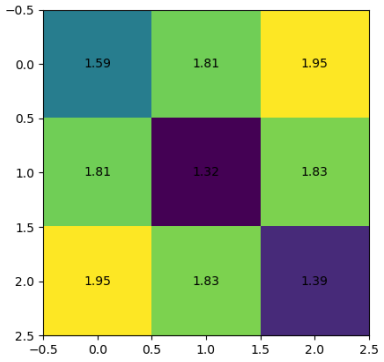

Figure 10: Matrix representation of the S score obtained from the .pdb files of the Frataxin protein.

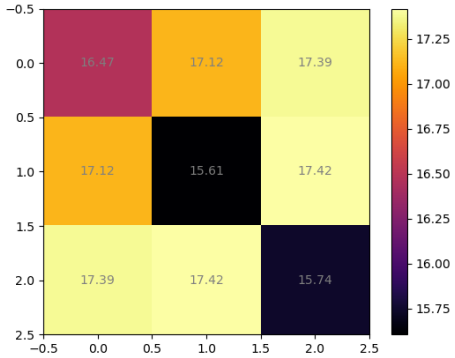

Figure 11: Matrix representation of the M score obtained from the Frataxin's data-set.

4 Applications

In this section we use the results outputted from the algorithm for making biochemical observations about two analysed proteins, in order to show some applications which uses the output information. In particular, we discuss below about two proteins, the chain A of the STIM1 and the P16, which are both subunits of a bigger structure (the STIM1 and the cdk6_P16INK4A complex, respectively).

4.1 Loss of function in STIM1's chain A

The STromal Interaction Molecule 1 (STIM1) is a fundamental membrane protein which has an important role in the flow of Ca^{2+} in and out of *H. sapiens* cells. The structure of this protein is shown in Figure 12. Below we show the results of the algorithm operated on the chain A of the STIM1 protein (UniProtKB entry Q13586 and PDB entry 2K60)[4][5], which includes the amino acids between positions 58 and 201. Performing the 60-dimensional PCA on the data-set, we kept the 84.1% of the variance ratio of the original data-set. The clustering applied to the 60D data-sets, obtained the maximum silhouette score for 2 clusters(**??**). The Figure 1 shows instead a 3D visualization of the protein dynamics in the PCA space.

The clustering operated by the algorithm is validated using S score. In Figure 2 there is the matrix generated to evaluate the conformational similarity intra- and inter-clusters. In the figure we see how the darker diagonal shows the fact that the quantitative similarity among elements of the same clusters is higher with respect to the one measured between the two clusters.

The first comparison is made between the native structure and the snapshot STIM1_14.pdb, representative for the first cluster. In this case we found a RMSD value of 1.184 Å across 39 kept atom pairs(with 73% of the C_α atoms excluded from the calculation). Deleting our threshold, the RMSD reached a value of 4.871 Å across all the 144 atom pairs. As we can see in Figure 13, the regular secondary structures (α-helices and β-sheets) seem preserved, but relocated, while the random coils seem more affected. Despite this fact, we can notice in the figure, in which we compare just the amino acids of the Ca^{2+}-binding site, how the two conformations seem very preserved (we obtained indeed a RMSD of 0.508, which is a quantitative clue of how low is the variation between the two sites).

The second comparison is made between the experimental structure and the snapshot STIM1_64.pdb, representative for the second cluster. In this case we found a RMSD value of 1.035 Å across 39 kept atom pairs (with 73% of $C-\alpha$ atoms excluded from the calculation once again).

As happened for the previous comparison, the RMSD reached the very high value of 5.984 Å across all the 144 pairs of amino acids. As we can see in Figure 14 and as we saw in Figure 13, the regular secondary structures (α-helices and β-sheets) seem again preserved, but relocated, while the random coils present more conformational variation. With respect to the previous comparison, this one shows a greater variation between the snapshot and the native structure (also the RMSD on all the pairs is higher). Moreover, we can notice how the binding site is extremely altered, with some side chains completely moved away and probably no more able to coordinate the calcium. This fact is confirmed by the higher RMSD between the amino acids of the binding site, which has a value of 0.882.

It is interesting to notice from our previous observation that the random coils seem more affected than α-helices and β-sheets and this is a signal of how the weak interactions contribute to stabilize a structure (helices and sheets form a

greater amount of these interactions than random coils, which are more dynamic and fragile). Moreover, despite every conformational conservation in the binding site, the overall highlighted differences in the two representative snapshots are remarkable and could result in a significant alteration of the catalytic efficiency (or even a complete loss of function) of the two models against the native structure. Between the two snapshots we found a RMSD of 1.255 Å across 59 remained amino acid pairs (which reaches the value of 5.313 Å across all the 144 pairs of amino acids), this suggests a clear conformational variation between the two snapshots representing the discovered clusters. Indeed, we can identify the presence of various non-overlapping zones, despite the secondary structures seem basically conserved. We can observe how the Ca^{2+}-binding sites result distant as well, although a comparison operated just between all the amino acids of the two Ca^{2+}-binding sites gives back a RMSD of 0.763 (Figure 15). All these observations must be contextualized, because this protein we analyzed is just a chain evolutionary projected to be with other subunits for a cohoperative stabilization. Despite we don't know the conditions of the molecular dynamics, we can suppose that, in general, this subunit tends to loss is function once separated from the other parts of the complete protein.

5 Supersecondary structures preservation in the P16

The Cyclin-dependent kinase inhibitor 2A (P16) (UniProtKB entry P42771 and PDB entry)[4][5] negatively regulates the proliferation of normal cells and strongly interacts with CDK6 protein, forming with the latter a complex named CDK6_P16INK4A. In Figure 16 we can see the native structure of this protein, which presents a series of four helix hairpins connected by long loops. The N-terminal and C-terminal portions are characterized by random coil regions. In this section we study the variations of this supersecondary structures between the two discovered clusters and in relation to the native structure. In doing this kind of comparison, we don't consider N- and C-terminal coils, since they could affect our results. Indeed, how we can observe in Figure 16 they are in total contact with the solvent and they do not seem to make relevant interactions with the rest of the protein. Therefore, we expect them to be very free-to-move, without assuming any kind of specific conformation. We can appreciate this in Figure 17, which demonstrates how the lack of conformation in the terminal loops could affect the results of our comparison.

What we can infer at first sight in Figure 18 and Figure 19 is that α-helices are not subjected to significant deformations, although they tend to deviate from the native position, because of the conformational changes of the loops which interconnects them. In fact, the loops are in general more dynamic in relation to regular secondary structures. In the figures we can observe a clear distancing of the two snapshots in comparison with respect to the native form of the protein. Such deviation could destroy the hydrophobic packing interaction that two helices keep when found in a hairpin supersecondary structures, leading to a loss of conformational stability and a possible consequent loss of function.

In particular, we obtained a RMSD value of 1.284 Å across 54 kept atom pairs (with 55% of the C_α atoms excluded from the calculation and a value of 3.101 for all the 121 atom pairs) for the native P16-snapshot_30 comparison, and a RMSD value of 1.217 Å across 53 kept atom pairs (with 56% of the C_α atoms excluded from the calculation and a value of 2.750 for all the 121 atom pairs) for the P16-snapshot_94 comparison. Also, the RMSD value of the two snapshots without the terminal coils is 0.752 Å across 112 kept atom pairs (with 7% of the C_α atoms excluded from the calculation and a value of 1.476 Å for all the 121 atom pairs). We can observe that in both cases the RMSDs are not enough large to affirm that the three structures are structurally different from each other. This is further evident when we compare the two snapshots alone, how it is shown in Figure 20. The figure shows that the two snapshots have a very similar structure, despite they should belong to two distinct clusters. We can certainly affirm that the helices move mildly during the time of the dynamics, but it doesn't seem to be a large conformational variation. These results can justify what we obtained from the algorithm. In fact, the clustering seems to be not so exhaustive (how it is shown in the respective matrices) and all the structures analyzed until now, for the P16, are coherent with all the outputs and with the problems of the algorithm to find significant clusters.

6 Conclusions

The algorithm constitutes an efficient method to extrapolate data from molecular dynamics snapshots but surely it needs further tests on other molecular dynamics simulations to verify its reliability. The M and S scores are two built-in powerful tools to validate the results and we also noticed that a well-defined 3D trajectory (obtained from the PCA) is symptom of a succeeded clustering procedure. Although performing a structural comparison seems the only way to dispel any doubts about the effective success of the procedure. Additionally, among the snapshots of a molecular dynamics simulation, there are some of them which identifies metastable states and others which identifies the transitions between these. A good way to improve the algorithm could be to make it able to recognize such types of states, in order to appreciate which are the intermediate metastable states and what characterize the temporal transition from one to another. Our examples provide a simple application of what it is possible to realize exploiting the potential of this algorithm. Indeed, it is possible to apply the same procedures to study the most relevant conformations occurring in a molecular dynamics simulation also for large proteins, using them for further *in silico* studies.

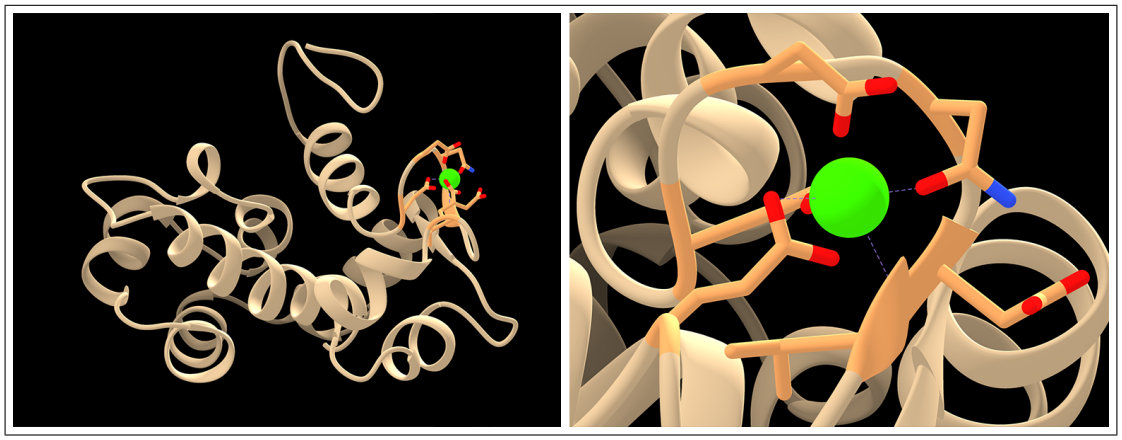

Figure 12: *On the left*: 3D low energy structure of the chain A of the STIM1 (which presents the same conformation of the first .pdb snapshot) obtained through NMR analysis applied to a stable solution of the protein; the residues involved in the binding of the Ca^{2+} (the green sphere) are highlighted in green orange and their sides chain are shown. *On the right*: detail of the Ca^{2+}-binding site, which shows the atoms directly involved in the calcium coordination complex (the interaction is represented by a dashed line), which sees as primarily involved amino acids D76, D78, N80, D82, V83 and E87.

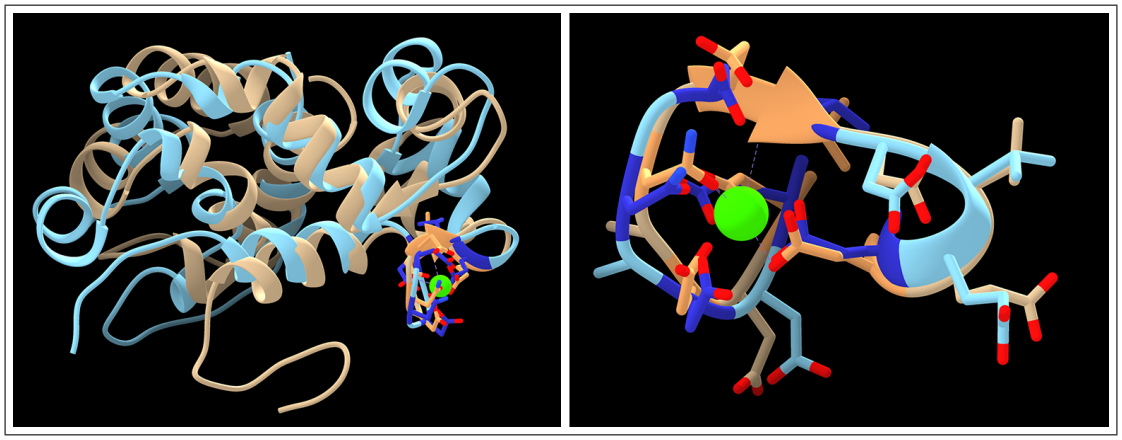

Figure 13: *On the left*: Comparison between the experimental structure of the chain A of the STIM1 (in light brown) and the snapshot STIM1_14.pdb (in cyan); the Ca^{2+}-binding sites are highlighted. *On the right*: Comparison of the binding sites which shows the similarity of the two substructures, in light orange is shown the binding site of the native chain A of the STIM1 and in medium blue is shown the binding site of the snapshot.

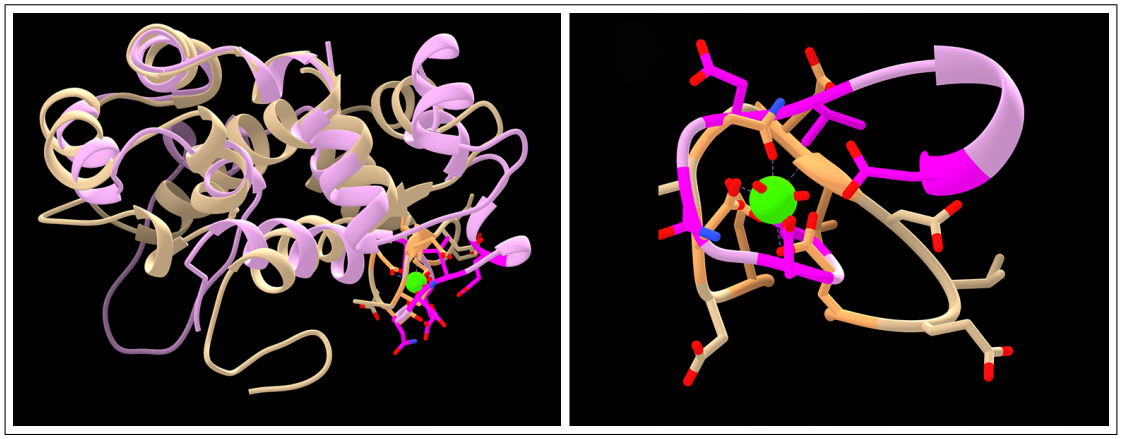

Figure 14: *On the left*: Comparison between the experimental structure of the chain A of the STIM1 (in light brown) and the snapshot STIM1_14.pdb (in plum); the Ca^{2+}-binding sites are highlighted. *On the righ*: Comparison of the binding site which shows the conformational difference between the two substructures; in light orange is shown the binding site of the native chain A of the STIM1 and in magenta is shown the binding site of the snapshot.

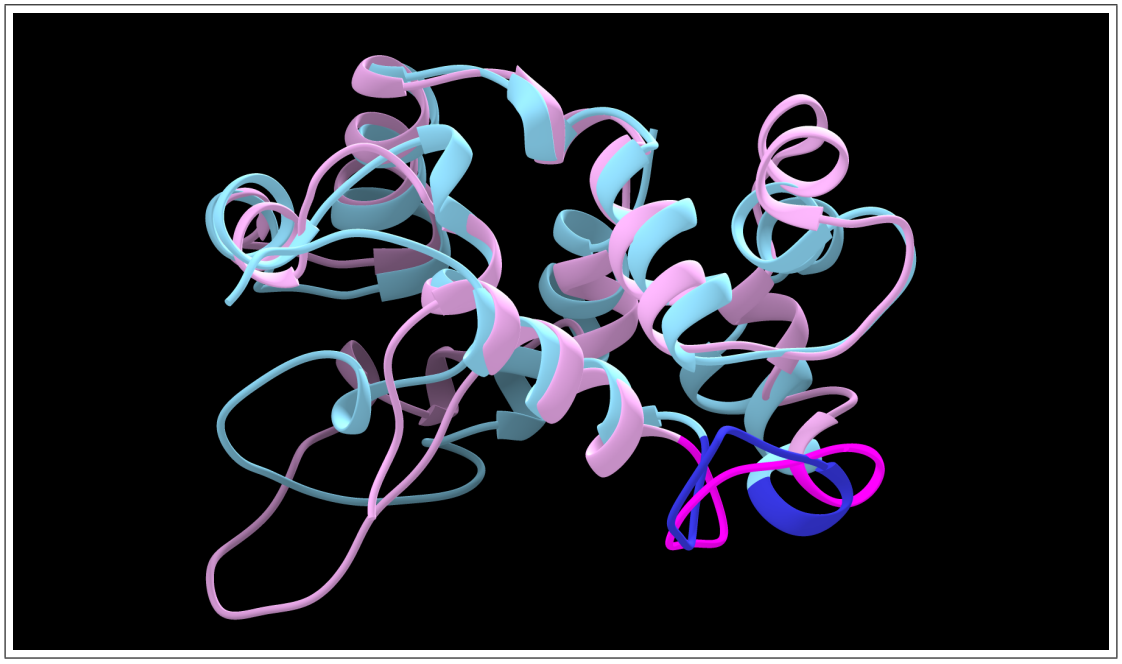

Figure 15: Structural comparison of the snapshots 14 (cyan) and 64 (light brown); the 12 amino acids long Ca^{2+}-binding sites (positions 76-87)[4] are highlighted.

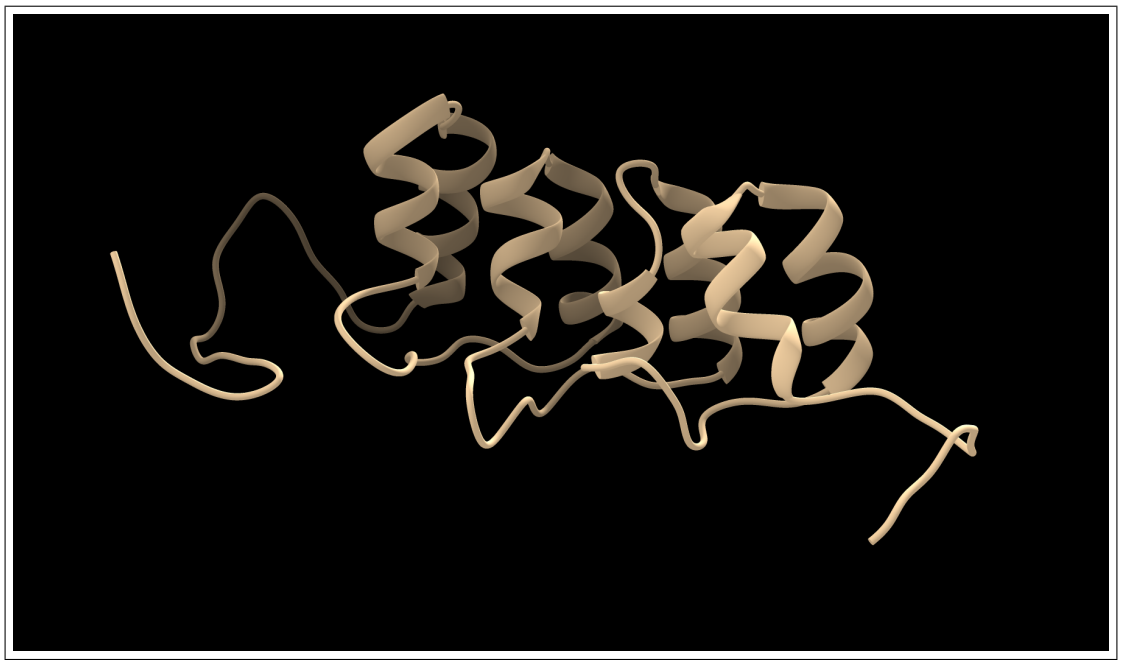

Figure 16: 3D native conformation of the P16 experimentally derived through NMR analysis applied to a stable solution of the protein. We can easily appreciate the four hairpins structure linked by loops and the N- and C-terminal coils.

Figure 17: Superposition of 18 native NMR experimentally-determined conformations of the P16, taken from the same sample in different times; in magenta are highlighted the α-helices, in order to distinguish them from the rest of the protein, which is all constituted by random coils (there are not β-sheets); it is possible to see the high conformational variation of the N-terminal and C-terminal regions in solution, against the stability of the internal supersecondary structures and the loop that link them.

Figure 18: Conformational comparison between the native P16 and the snapshot 30 from the molecular dynamics .pdb.

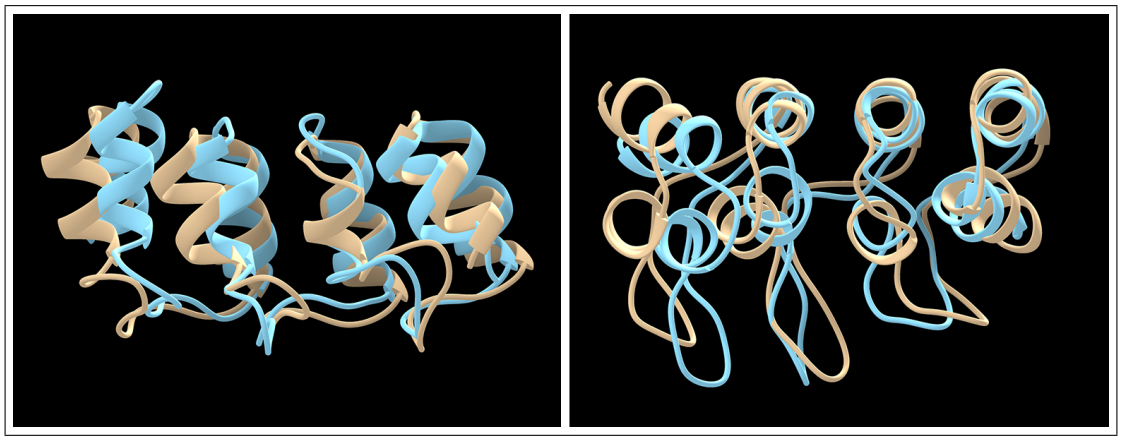

Figure 19: Conformational comparison between the native P16 and the snapshot 94 from the molecular dynamics .pdb.

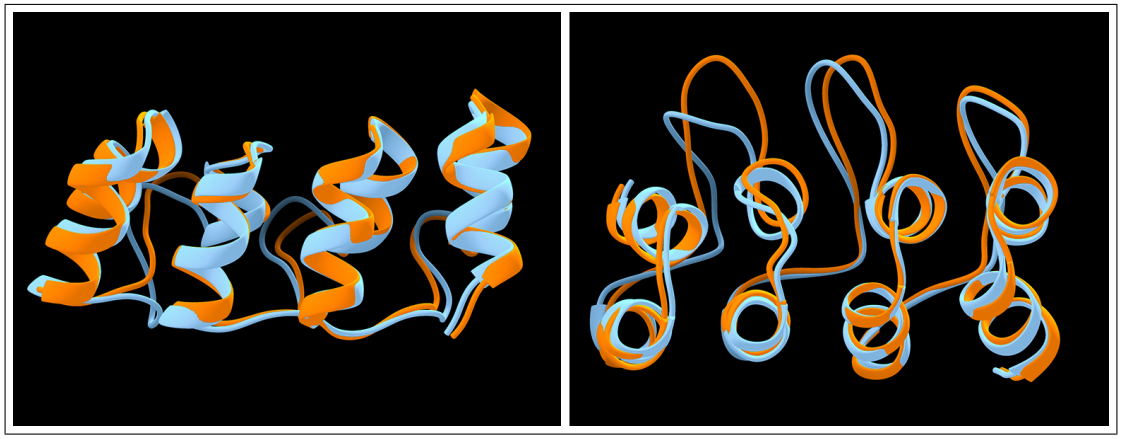

Figure 20: Conformational comparison between the native P16 and the snapshot 94 from the molecular dynamics .pdb.

References

[1] Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. Nature structural biology, 9(9), 646-652.

[2] Piovesan, D., Minervini, G., and Tosatto, S. C. (2016). The RING 2.0 web server for high quality residue interaction networks. Nucleic acids research, 44(W1), W367-W374. Site: http://old.protein.bio.unipd.it/ring/

[3] UCSF Chimera–a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12.

[4] UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1), D506-D515. Site: https://www.uniprot.org/

[5] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... and Bourne, P. E. (2000). The protein data bank. Nucleic acids research, 28(1), 235-242. Site: https://www.rcsb.org/

[6] Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States). Site: https://networkx.github.io/documentation/stable/index.html

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410. Site: https://blast.ncbi.nlm.nih.gov/Blast.cgi

[8] Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation, GitHub, http://github.com/charnley/rmsd