University of Padua
Department of Information Engineering

M. Sc. Degree in Computer Engineering

Conformational Analysis of Protein Structural Ensembles

Course: Structural Bioinformatics

Candidates:
Mikele Milia, mikele.milia@studenti.unipd.it
ID Number 1218949

Giulia Pezzutti, giulia.pezzutti@studenti.unipd.it
ID Number 1234012

Federica Vettor, federica.vettor.1@studenti.unipd.it
ID Number 2019160

Project supervisor:
Prof. Damiano Piovesan

Accademic Year 2020 - 2021

Contents

List of Figures

List of Tables

In this report, the main project choices and the results for five structural ensemble PED00020 of measles virus nucleoprotein are reported.

1 Task 1

The goal of the first task is to implement a tool to identify conformational relationships thanks to models structural features within an ensemble. The input is a set of files, each containing the PDB structure of one ensemble, which will be analyzed independently.

1.1 Single conformation features

For each conformation, the following structural features are extracted and subsequently saved:

- *Radius of gyration* (RG) is computed from the coordinates of each atom and their barycenter.
- *Relative accessible surface area* (ASA) is computed for each residue thanks to DSSP[1].
- *Secondary structure* (SS) is determined for each residue thanks to the Ramachandran regions[2] associated to Phi and Psi angles. For an easiest subsequent analysis, each value is converted into an integer as reported in table 1.
- *Distance matrix* contains the pairwise distances between each pair of residues. For an easiest analysis, since it is symmetric, only the linearized version of its upper triangular sub-matrix is considered.

For the presented computations, all protein chains have been considered in order to provide a complete analysis of the structure.

Secondary Structure	**Int**
-	0
Beta-sheet	1
Polyproline I-II	2
Alpha-helix	3
Left-handed Helix	4

Table 1: Conversion for Secondary Structure values

1.2 Representative conformations

In order to extract the representative conformations, the models are clustered through a *K-Medoids* approach and a customized metric function, selecting

[1]**DSSP**: the values obtained for the conformations within an ensemble are equal. It may happen that the DSSP does not return the ASA value for some residues: for this limitation, a check on the array dimension and, if needed, a zero-padding are performed.

[2]**SS**: for the first and last residue, for which the value is not computable, '-' is inserted.

the optimal number of clusters thanks to *silhouette* score. *K-Medoids* has been chosen since at the end the representative conformations will be medoids selected between the input set points, without generating unseen conformations.

The implemented metric[3] is tailored for each feature set, in order to measure the dissimilarity between models. The following distance metrics have been chosen according to the meaning and the behaviour of single features:

- Absolute difference between radius of gyration.
- Euclidean distance of ASA vectors.
- Normalized hamming distance between SS vectors using the scoring matrix reported in table 2 and defined accordingly to their definition and properties.
- Cosine distance between distance matrix.

	0	1	2	3	4
0	0	1	1	1	1
1	1	0	1	1	1
2	1	1	0	1	1
3	1	1	1	0	0.5
4	1	1	1	0.5	0

Table 2: Scoring matrix Secondary Structure values

In table 3 we report both the optimal number of clusters and the corresponding silhouette scores obtained with the described clustering procedure for the analyzed ensembles.

Ensemble	**Cluster**	**Silhouette**
001	4	0.538269
002	3	0.511878
003	4	0.499020
004	3	0.521289
005	3	0.545507

Table 3: Silhouette values obtained during clustering

Afterwards, a weighted graph is built, inside which each node is a representative model and edges length between pairs of nodes is proportional to their distance according to the built metric. In figures 1, 2, 3, 4 and 5, built graphs are reported.

[3]**Metric**: take in input features vector of two conformations and compute their distance as sum of partial features distances.

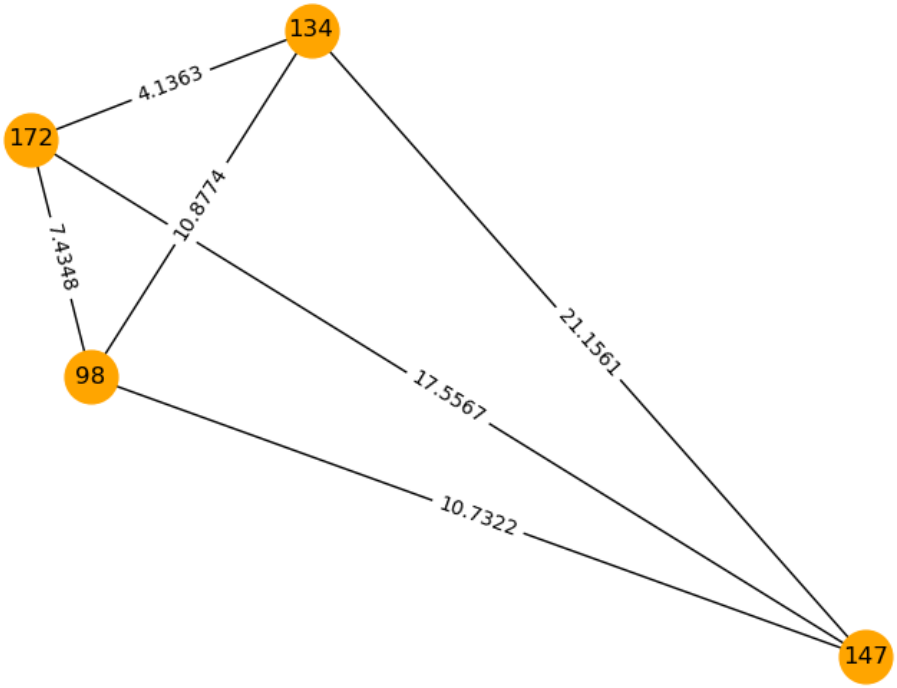

Figure 1: Graph of ensemble 001.

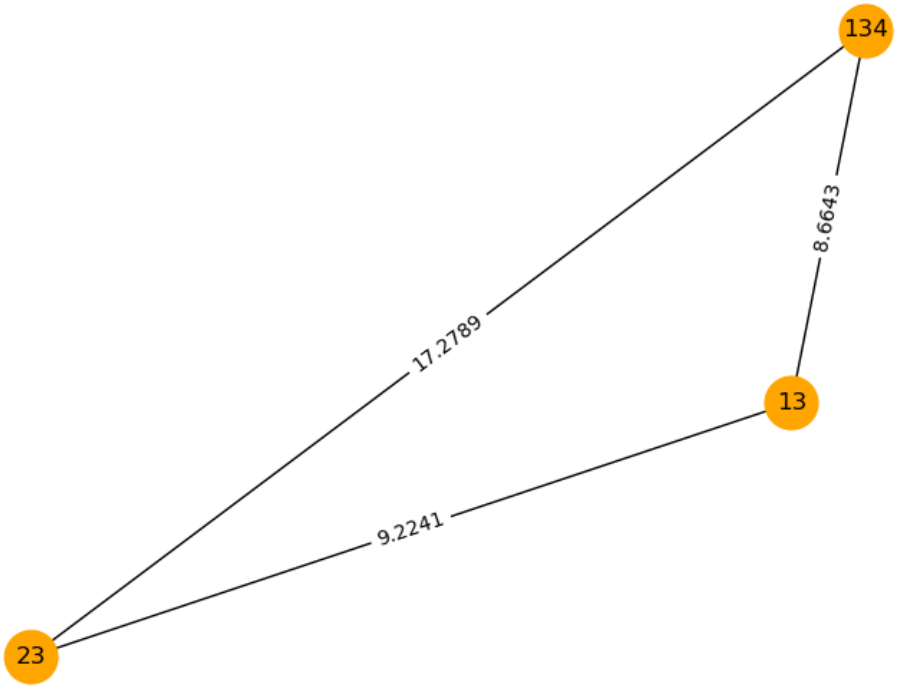

Figure 2: Graph of ensemble 002.

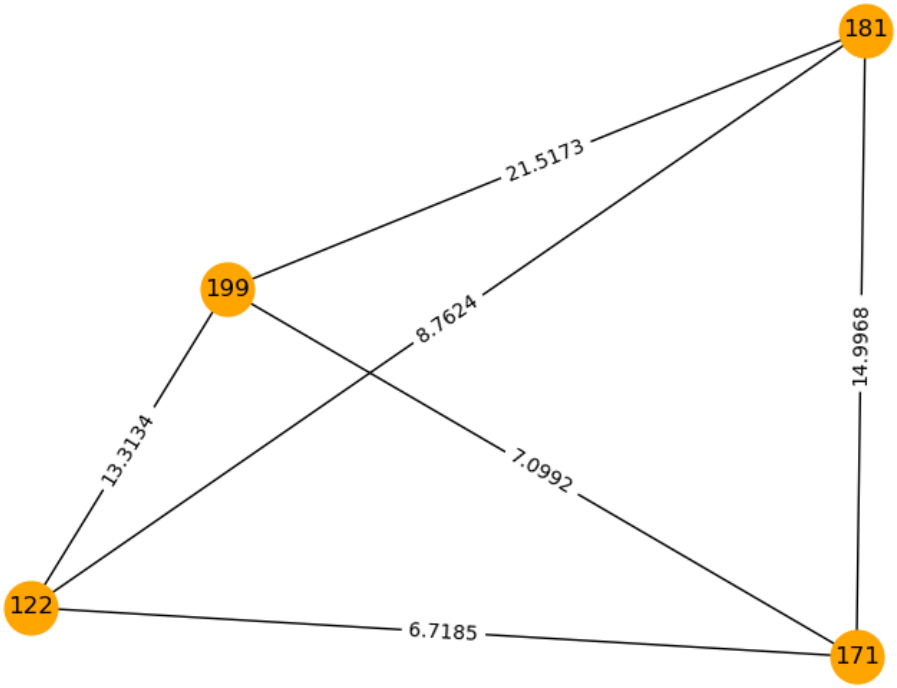

Figure 3: Graph of ensemble 003.

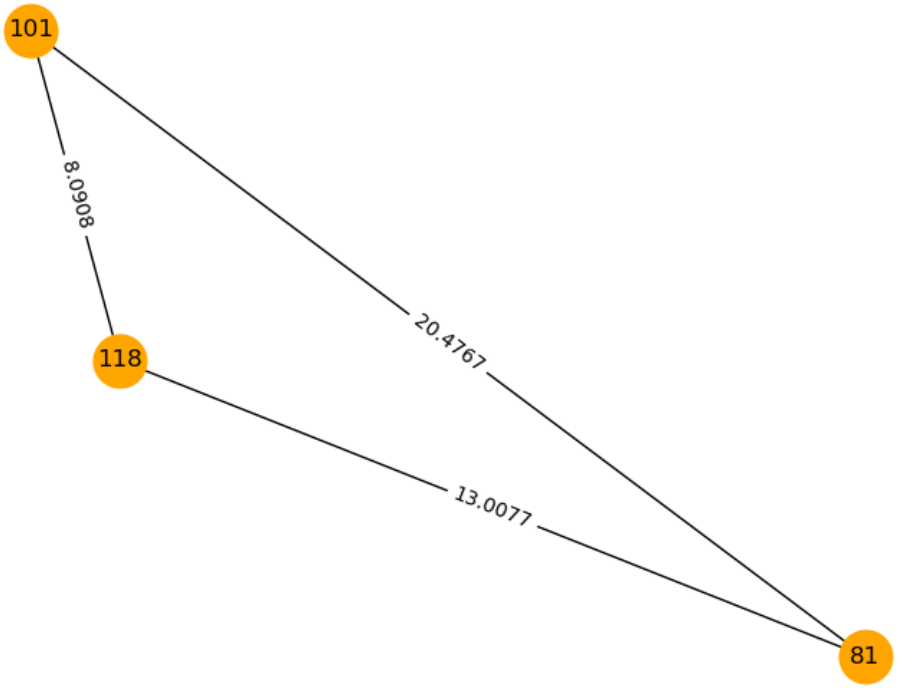

Figure 4: Graph of ensemble 004.

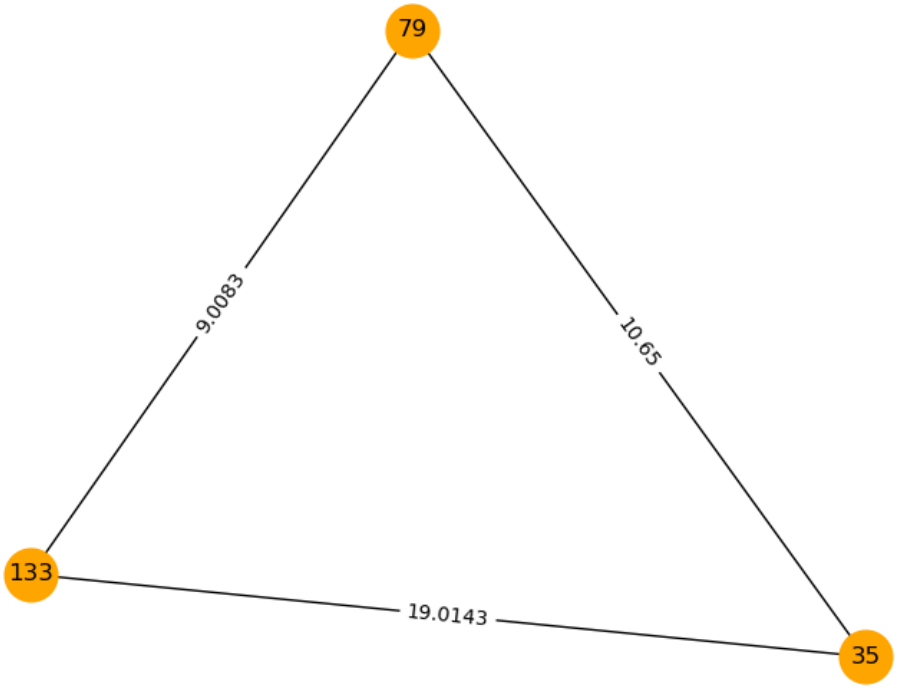

Figure 5: Graph of ensemble 005.

A high results variability can be observed, in terms of selected medoids but in general not in the optimal number of clusters. This is probably caused by two main factors:

- Proximity of the points in the features space.
- Random initialization of medoids during clustering.

1.3 Pymol image

A Pymol image is generated to visualize the 3D structure of one representative conformation and the variability of each residue. We decided to generate one image for each ensemble, instead of the superimposition of all the conformations, because latter would give a too messy result without any important information. In order to reduce the computing time needed to compare the features[4], we have decided to take into consideration only features of representative conformations, since each one of them is a good approximation of its own cluster.

The i-th residue variability is calculated using a customized metric, which considers a window of 19 (9 before and 9 after) residues around the current one. After extracting the features vectors of those residues for each considered conformation, the metric calculates the residue's variability as the sum of the following partial distances:

- Standard deviations of ASA values for each residue are calculated and their mean is computed to obtain an average window value.

[4]**Pymol**: the time needed for the Pymol image generation considering the features over all the models is about one hour.

- Normalized hamming distance between each pair of SS vectors is calculated using the scoring matrix reported in table 1.

- Cosine dissimilarities are computed and summed for each pair of analyzed conformations between distance vectors of corresponding residues.

Note that the radius of gyration is not used since it is not residue-dependent.

A BWR color-map[5] is used to highlight which residues own the highest or lowest variability inside each Pymol image. Looking at figures 6, 7, 8, 9 and 10, it is possible to notice that structures are in general very disordered since contains few well-defined secondary structures, that in these cases are only alpha-helices. This low complexity may be caused by the probably required high energy level of folding. Further considerations about the protein structure are provided in the conclusions of this report.

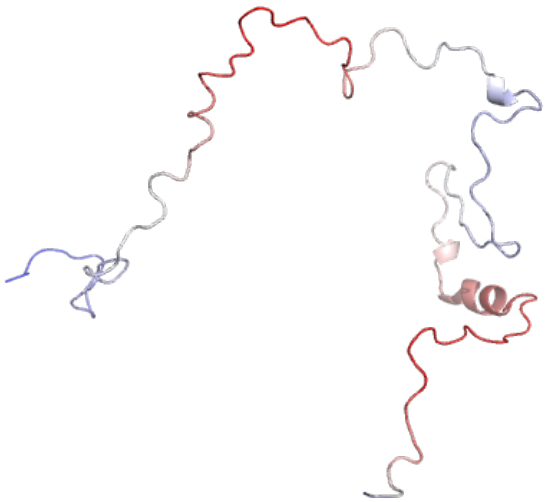

Figure 6: Pymol image of ensemble 001 (model 098).

[5]**BWR** : is a color-map that use blue, white and red color to characterize respectively low, mid and high values inside a plot.

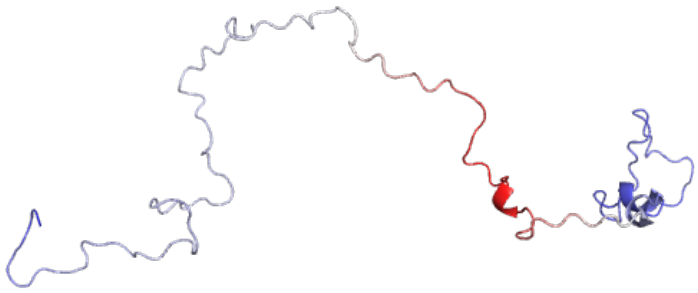

Figure 7: Pymol image of ensemble 002 (model 023).

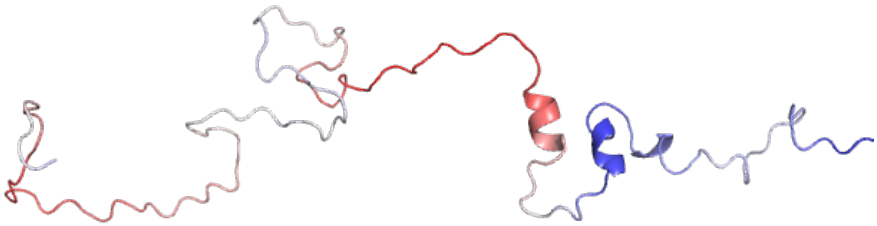

Figure 8: Pymol image of ensemble 003 (model 122).

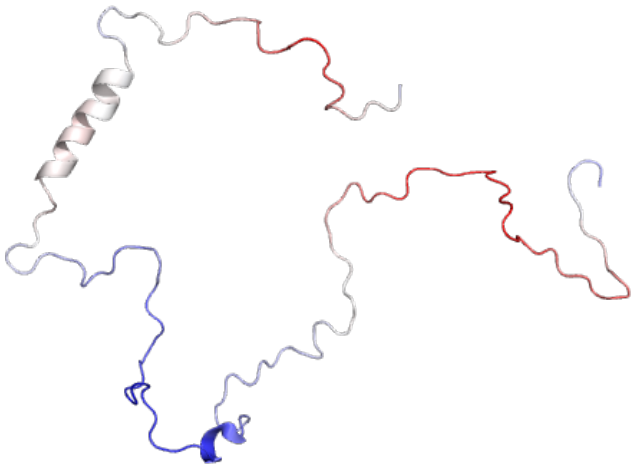

Figure 9: Pymol image of ensemble 004 (model 118).

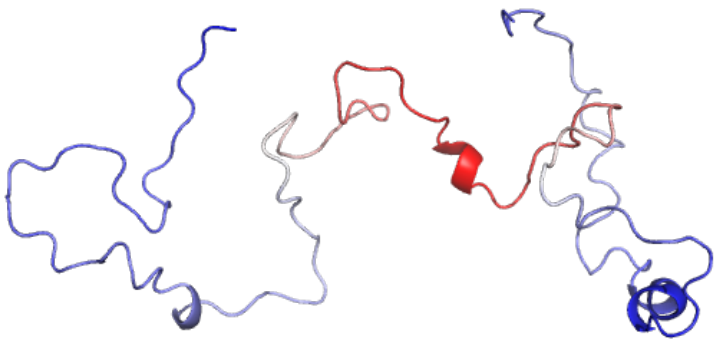

Figure 10: Pymol image of ensemble 005 (model 133).

2 Task 2

The goal of the second task is to implement a tool able to describe the relationship between ensembles of a PED, from global and local points of view. Relationships between different ensembles will be evaluated thanks to PED features. The input are files, each containing ensemble features, computed during Task 1.

2.1 Ensembles features

For each ensemble, PED features are extracted from the previously computed models features. In details, the obtained features are the following:

- *Radius of gyration*[6] (RG) is retrieved for each model.
- *SS entropy* is calculated for each position across ensemble's conformations, exploiting its probabilistic definition.
- *Median ASA* is computed for each position across ensemble's conformations.
- *Median RMSD* is computed using roto-translation matrices across ensemble's conformations.
- *Median distance matrix* is computed by calculating median distance for each pair of equivalent positions across ensemble's conformations.
- *Standard deviation* (STD) *distance matrix* is computed calculating standard deviation distance for each pair of equivalent positions across ensemble's conformations.

2.2 Global metric

The global metric aims to provide an accurate measure of dissimilarity between ensembles pairs represented with their features, building an evaluation that takes into account the different nature of the features evaluated. Given two PED features vectors, the sum of the following partial distances is returned:

- Absolute difference of RG mean values extracted from each features vector, after zero-padding removal.
- Chebyshev distance[7] calculated between entropies vectors.
- Euclidean distance calculated between median ASA vectors.
- Euclidean distance calculated between median RMSD vectors.
- Cosine distance calculated between median distance vectors.

[6] **RG**: It may happen that an ensemble contains a lower number of conformations with respect to the others. To avoid alignment problems, zero-padding is performed.

[7] **Chebyshev distance** : a.k.a maximum metric, is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension.

Standard deviation distance matrix has not been taken into consideration for the global evaluation, since it is a kind of variability measure for each residue and that does not provide useful information from a global point of view.

Results of global distances application inside PEDs are provided through an *heatmap* which shows distances between each pair of ensembles thanks to a color-map and a *dendrogram* which exploits a linkage matrix to apply a *complete* linkage approach. Figures 11 and 12 show respectively distances and similarity relationships between analyzed ensembles. From the heatmap analysis, it is possible to notice that ensembles 002 and 003 are the closest ones, whereas 001 and 002 are the furthest ones. In general, all ensembles seem to be evenly spaced from each other. In addition, a more precise understanding of the spatial distribution can be inferred from the dendrogram, in which two clearly visible sub-clusters are revealed, in detail the first containing 002, 003 and 004 and the second 001 and 005.

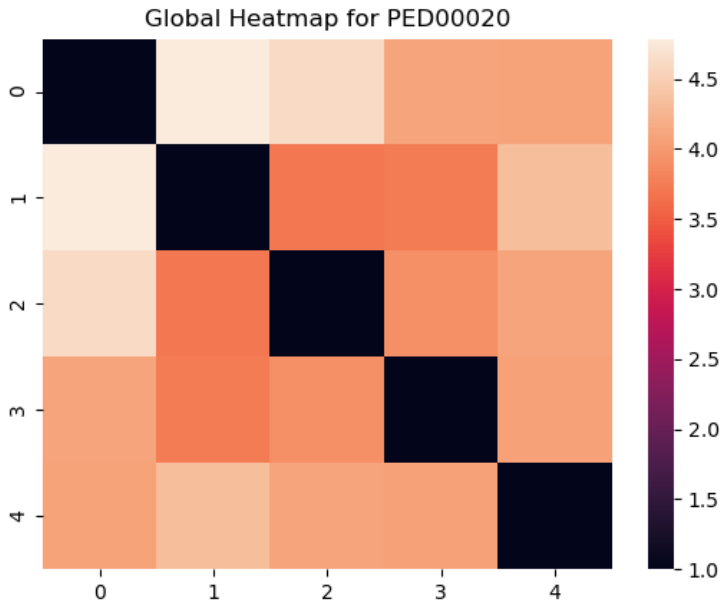

Figure 11: Heatmap of all the ensembles.

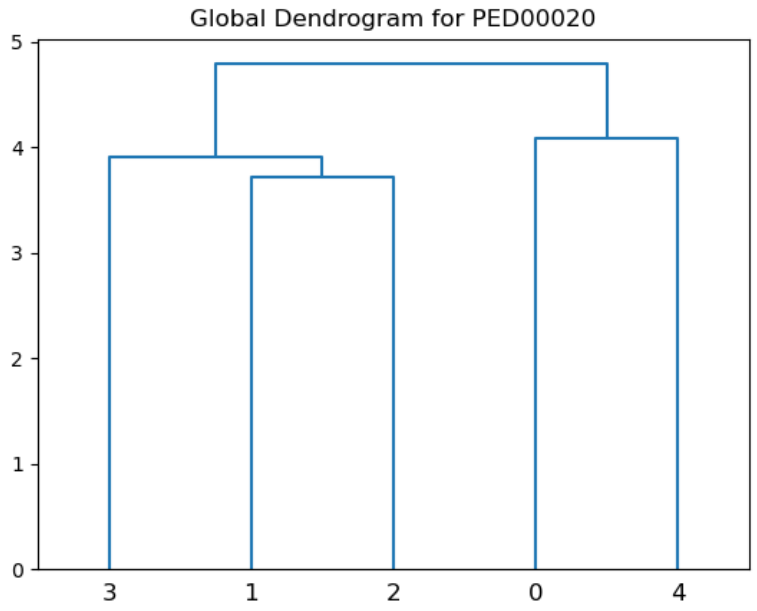

Figure 12: Dendrogram of all the ensembles.

2.3 Local metric

The local metric evaluates the ensembles' features variability for each residue. The i-th residue variability is calculated considering a window of 19 (9 before and 9 after) residues around the current one. After extracting the features vectors of those residues, a variability score is calculated independently for each type of feature as follows:

- Mean window value of the standard deviations calculated among conformational entropies for each residue.
- Mean window value of the standard deviations calculated among conformational ASA for each residue.
- Mean window value of the standard deviations calculated among conformational RMSD values for each residue.
- Mean window value of standard deviation's trimmed mean[8] of STD distance vectors for each residue.

Each variability score is then used to calculate the residue mean variability value. In figure 13 each residue variability in terms of ensembles' features is shown.

[8]**Trimmed mean**: used to to avoid outliers influence.

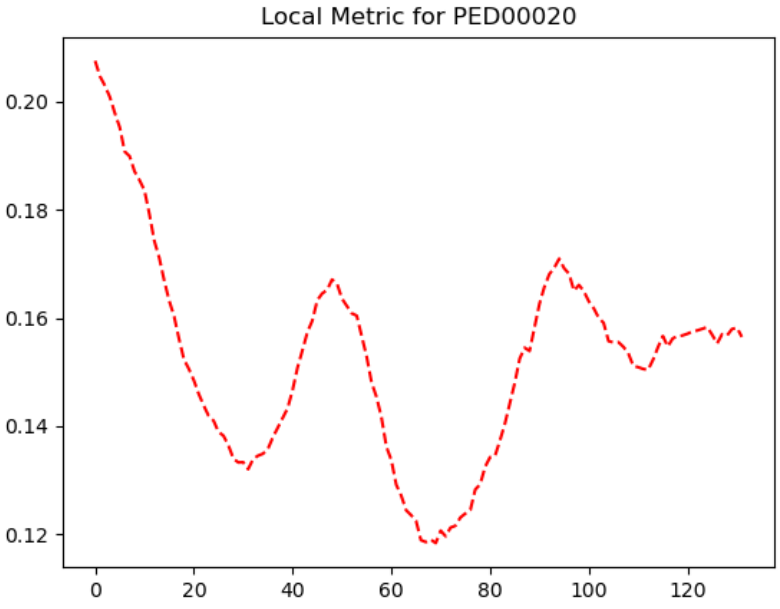

Figure 13: Plot of local score.

3 Conclusions

Since the graph reported in figure 13 shows an average value among all the ensembles, it is not possible to make an exhaustive comparison between its behaviour and the structure variability in Pymol images (figures 6, 7, 8, 9 and 10), even if it allows to infer general considerations about disordered content.

The measles virus nucleoprotein seems to be quite disordered: in particular, all the ensembles share a central high variability region (that could be associated with one of the local maxima), whereas some of them have another variable region in one of the extremities. Due to the smaller available window on the trailing parts, correspondent scores could be less trusted with respect to the central parts of the protein in which, except ensemble 005, is located the most stable region.

The previous analysis can lead to these inferences but the final protein study should be performed with additional information. Since the protein has the stable region in its center, these residues might be in charge of some functional properties. On the contrary, the most variable regions could not be useful from a functional point of view or could allow the protein to respond to stimuli in different ways.