A Closer Look on Protein Unfolding Simulations through Hierarchical Clustering

Pedro Gabriel Ferreira*[1], Cândida G. Silva†[1], Rui M. M. Brito† and Paulo J. Azevedo*
*Department of Informatics, University of Minho
4700-057 Braga, Portugal
Email: pja@di.uminho.pt
†Chemistry Department, Faculty of Science and Technology, and Center for Neuroscience and Cell Biology
University of Coimbra, 3004-535 Coimbra, Portugal
Email: brito@ci.uc.pt

***Abstract*— Understanding protein folding and unfolding mechanisms are a central problem in molecular biology. Data obtained from molecular dynamics unfolding simulations may provide valuable insights for a better understanding of these mechanisms. Here, we propose the application of an augmented version of hierarchical clustering analysis to detect clusters of amino-acid residues with similar bahavior. These clusters hold similar global pattern behavior of solvent accessible surface area (SASA) variation in unfolding simulations of the protein Transthyretin (TTR). Classical hierarchical clustering was applied to build a dendrogram based on the SASA variation of each amino-acid residue. The dendrogram was enriched with background information on the amino-acid residues, enabling the extraction of sub-clusters with well differentiated characteristics.**

I. Introduction

One of the unsolved paradigms in molecular biology is the protein folding problem, *i.e.* the acquisition of the functional three-dimensional structure of a protein from its linear sequence of amino acids, ultimately determined by the sequence of bases in a gene. With the identification of several diseases as protein folding disorders and the advent of many genomic projects, protein folding has become a central issue in molecular sciences research.

In recent years, protein misfolding has been associated with a large number of human diseases such as Alzheimer's, Parkinson's and Huntington's, all sharing the appearance of insoluble fibrillar deposits of proteins. Although the proteins involved differ in sequence, structure and function, the amyloid pathologies share common molecular mechanisms. In particular, it seems that in all studied cases, due to proteolysis, mutation or unfolding events, the normally soluble proteins are converted into molecular forms prone to aggregation, leading to cytotoxic oligomeric species and amyloid fibrils. We have been particularly interested in the structural characterization of the molecular species present in the aggregation pathway of Transthyretin (TTR), using both experimental [1] and computational [2], [3] methodologies.

Human TTR is a homotetrameric protein with 127 amino acids per subunit. Several single point mutations in TTR enhance the amyloidogenicity of the protein and lead to diseases such as Familial Amyloid Polyneuropathy (FAP) and Familial Amyloid Cardiomyopathy (FAC). Among the numerous pathogenic variants, Leu55→Pro (a proline replacing a leucine in position 55) is one of the most amyloidogenic and Val30→Met (a methionine replacing a valine in position 30) is one of the most prevalent [4]. It is today generally believed, that partial or even extensive unfolding of the native protein is required for aggregation and amyloid formation. One way of exploring the unfolding events that may be responsible for TTR aggregation is through the use of molecular dynamics (MD) protein unfolding simulations. In MD simulations, one tracks the atoms in only one protein molecule as a function of time. However, it is known today that in an ensemble of protein molecules not all of them follow the same folding or unfolding route, requiring multiple simulations to be carried out in order to obtain an approximate idea of the conformational space available to a protein molecule in its folding or unfolding processes.

To explore the unfolding routes of monomeric species of TTR, several molecular dynamics simulations of the TTR subunits of wild-type (WT-TTR) and the variant Leu55→Pro (L55P-TTR) were performed, at high temperatures. These MD simulations are computationally expensive and generate a huge amount of data, making the comparison of different trajectories a difficult task [5]. Data mining techniques might provide the technology required to contrast, compare and characterize the variation of molecular properties associated with each simulation. Azevedo *et al.* [6] reported the use of association rules in the identification of a group of hydrophobic residues moving in a coordinated fashion and most likely forming a hydrophobic cluster essential in the folding and unfolding processes of Transthyretin. In another work, Ferreira *et al.* [7] proposed an algorithm that extracts approximate motifs, *i.e.* motifs that capture portions of time series with a similar and eventually symmetric behavior.

In the present paper, we report the use of clustering analysis techniques to detect clusters of different amino-acid residues showing similar global solvent exposure pattern in unfolding simulations of the protein Transthyretin. We applied hierarchical clustering to build a dendrogram based on the variation of the solvent accessible surface area (SASA) of

[1]Both authors contributed equally to the present work.

each amino-acid residue along each MD simulation.. Next, we enriched the dendrogram information with background knowledge reflecting the amino-acid residue position in the protein linear sequence, the variation of their coordinates during the unfolding simulations and hydrophobicity values. Finally, sub-clusters with well differentiated characteristics were extracted.

II. Molecular Dynamics Simulations

Molecular dynamics (MD) simulations of temperature-induced unfolding have been recently reported as a tool to explore the early molecular events triggering amyloid fibril formation in globular proteins such as lysozyme [8], human prion protein [9], immunoglobulin kappa light chain [10], and transthyretin [2], [11], [12].

In MD simulations, successive configurations of the system are generated by integrating Newton's laws of motion. The result is trajectory that specifies how the positions and velocities of the atoms in the system vary with time.

A. Simulation Details

Initial coordinates for WT-TTR were obtained from the crystal structure (PDB entry 1tta [13]) and hydrogen atoms were added. All minimization and MD procedures were performed with the program NAMD [14], using version 27 of the CHARMM force field [15]. All atoms were explicitly represented. The complete system was comprised of 44,556 atoms.

The system was minimized, equilibrated and heated to the target temperature. Control simulations, at 310 K, and several unfolding simulations, at 500 K, were performed for up to 8 ns. The simulations were carried out using periodic boundary conditions and a time step of 2 fs, with distances between hydrogen and heavy atoms constrained. Short range non-bonded interactions were calculated with a 12 Å cut-off, and long range electrostatic interactions were treated using the particle mesh Ewald summation (PME) algorithm. More details on the simulations can be found elsewhere ([2], [5]). In Fig. 1, it is depicted a set of trajectories for control and unfolding molecular dynamics simulations of the WT-TTR monomers.

B. Trajectory Analysis

In order to have a more quantitative description of the unfolding pathways of a protein, several molecular properties might be calculated along the MD trajectories. Radius of gyration (R_g), root mean square deviation (RMSD), secondary structure, native contacts and solvent accessible surface area, among others, may be calculated along each trajectory in order to characterize and map the unfolding events.

The solvent accessible surface area (SASA) of a protein is defined [16] as the locus of the center of a probe sphere (representing a solvent molecule) as it rolls over the van der Waals surface of the protein. The solvent accessible surface area was computed using the program NACCESS [17], using a spherical probe of 1.4 Å diameter, mimicking a water molecule. Global or partial SASA may be calculated. In order to study the individual behavior of the 127 amino acid residues constituting the TTR subunit, we calculated SASA for each one of the residues along the unfolding trajectories. Thus, for each simulation we have 127 plots of SASA vs time, with 8,000 points. Several SASA variation patterns are observed, with some residues, even far apart in the protein sequence, sharing similar behavior.

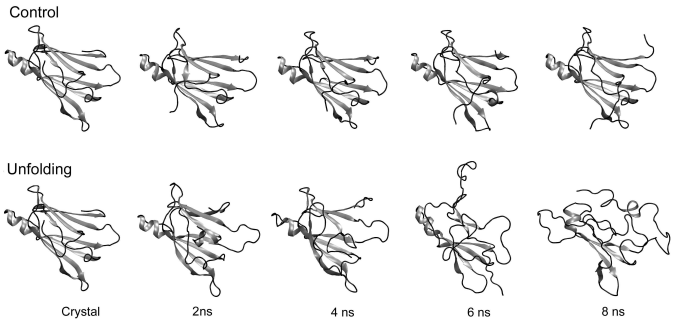

Fig. 1. Secondary structure ribbon representations of the monomer of WT-TTR along control and unfolding molecular dynamics simulations. β-strands, α-helices, and turns and coil are represented by arrows, cylinders and tubes, respectively.

III. Method

Clustering is the arrangement of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often proximity according to some defined distance measure. Clustering has been applied with success in many domains. It has been shown to be particularly useful in life sciences, namely in microarray gene expression data [18].

For the particular problem of protein unfolding, it is important to understand how amino-acids residues relate to each other during this process. In fact, finding clusters of residues that change solvent exposure in a coordinated fashion during one unfolding simulation or across several unfolding simulations might be of great importance to define the folding nuclei for a protein [5]. The solvent accessible surface area (SASA) is the molecular property that quantifies the exposure of the residues to the solvent surrounding the protein. Through the application of classical hierarchical clustering to the SASA variation of each residue of WT-TTR along MD unfolding simulations, an hierarchical tree –dendrogram– can be built expressing how residues cluster together, based on their global SASA trend. Although dendrograms contain useful information, their interpretation may prove to be a difficult task due to the amount and complexity of the data under study.

However, additional information about the residues chemical properties and their behavior along the unfolding trajectory is available which can be used to enrich the dendrogram and help in the identification of sub-clusters with well differentiated characteristics. These sub-clusters will be defined in a way that intra-cluster similarity variation is minimized

and inter-cluster variation is maximized. This helps in a better understanding of the sub-clusters differences, which is typically difficult in a classical dendrogram. The additional information on the residues considered to define the sub-clusters consists in the following properties: (P1) the distance of the residues in the protein linear sequence, (P2) the spatial distance of the residues along the MD unfolding trajectory and (P3) the hydrophobic propensity. The first property is the trivial information obtained from the residues positions in the linear sequence of the protein. Although, adjacent residues tend to show similar surface exposure, finding residues far apart in the linear sequence of the protein with similar behavior may provide unexpected insights in the residues associations. The second property quantifies the overall spatial variation of the sub-cluster. This can be achieved by means of Eq. (1), which measures the deviation of the residues in relation to a central point of the sub-cluster. The third property is used to provide an average value of hydrophobicity of a set of residues. The initial values of hydrophobicity for each residue were defined according to the Radzicka and Wolfenden [19]. The positive values indicate that the residues are hydrophobic, and the negative values indicate that the residues are hydrophilic.

$$\frac{\sum_{i=1}^{n} ||X_i - Y_i||}{N} \tag{1}$$

The overall approach of the method presented here may be summarized in the following four steps:

1) Build a dendrogram through classical hierarchical clustering of the amino-acids for the SASA property;
2) Into each node of the dendrogram insert the background information; calculate the information for each parent node through the values of the child nodes (Bottom-up traversal);
3) Perform a top-down traversal of the extended dendrogram; split into two sub-clusters when significant inter-cluster variation is detected. Apply recursively the procedure in the sub-clusters and stop when sub-clusters achieve a user defined minimum number of amino-acids;
4) Output sub-clusters and the respective information.

IV. Classical Hierarchical Clustering

Data clustering algorithms can be *hierarchical* or *partitional*. Hierarchical algorithms find successive clusters using previously established smaller clusters, whereas partitional algorithms determine all clusters at once. Hierarchical clustering can be divided in two types: *agglomerative* and *divisive* [20]. Agglomerative clustering starts by attributing a cluster to each object. Next, it merges one cluster with other successively into larger clusters. The procedure stops when all objects are covered by a cluster or when only k clusters are left. The divisive methods use a top-down approach. They start by considering all the objects in a single cluster, and successively subdividing the clusters into small sub-clusters. It stops when all k clusters are achieved or when the distance between closest clusters is above a defined threshold. The parameter k is defined by the user. The procedure stops when the complete tree is built.

An important question that arises during the clustering preparation is how to assess the similarity between objects. Several measures can be applied. The cosine function, Pearson's correlation, Jaccard's coefficient or Euclidean distance are some of the commonly used metrics. Both Euclidean distance and correlation measures have a clear biological meaning. Euclidean distances are applied when the idea is looking for identical patterns, while correlation measures are used in the cases where trends of the patterns are the subject of the analysis [21], [22]. Cosine and correlation based measures are well suited for clustering of both low and high dimensional datasets [18] and are data scale independent, which is not the case of the Euclidean distance.

A. The data

The studied data was obtained from one MD unfolding simulation of WT-TTR. The data is composed of 127 time series with 8000 points each, describing SASA variation of each amino-acid residue along the unfolding simulation. The dataset was prepared so that each amino-acid residue is considered as an object and each time frame as a dimension.

Data normalization is applied to scale all values into a defined range. The dataset was normalized according to Eq. (2), where S_{min} and S_{max} correspond to the minimum and maximum values found in each time series, respectively, and X is the value being normalized. After normalization, all the values in a time series are within the range $[0, 1]$ and the series ready to be directly compared.

$$\frac{X - S_{min}}{S_{max} - S_{min}} \tag{2}$$

B. Similarity Definition

To determine the similarity among the SASA time series, a similarity measure needs to be defined. Since we are looking for time series with similar trend or tendency, the Pearson correlation coefficient [23] was chosen as the similarity measure, as defined in Eq. 3.

$$P(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

The values of P, between series X and Y, are within -1 and 1. The values -1 and 1 indicate a strong negative or positive linear relation between X and Y, respectively. A P value of 0, indicates no linear relation between two time series.

C. Hierarchical Clustering Computation

The CLUTO [24] clustering toolkit was used to compute the clustering solutions. CLUTO provides access to its various clustering and analysis algorithms via the *vcluster* and *scluster* programs.

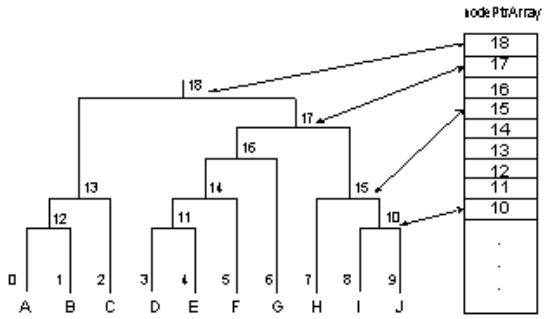

Fig. 2. Extended Tree (*exTree*) for the example objects A to J and *nodePtrArray*; some connections examples are represented.

The hierarchical agglomerative clustering, combined with the Pearson correlation coefficient as similarity measure, was computed using the *vcluster* program, which treats each object as a vector in a high-dimensional space. Different parameter settings were tested to find the best clustering solution, *i.e.* the solution with smaller entropy and larger purity values [24]. This corresponds to a solution with ten clusters ($k = 10$), using the complete linkage function ($i2$), *i.e.* the distance between clusters is determined by the most distant pair of objects of each cluster. As a stop criterion the minimum number of amino-acid residues per sub-cluster was set to 4. The *plottree* and *plotmatrix* options were used to plot the dendrogram and the similarity matrix.

V. Augmented Hierarchical Clustering

This section describes in detail the process used to build the extended dendrogram –*exTree*, shown in Fig. 2. The cluster and the tree files outputted by CLUTO are scanned and used to recreate the dendrogram. An auxiliary data structure *nodePtrArray*, which contains pointers to all nodes of *exTree*, is also built to help performing a bottom-up traversal of the dendrogram. This traversal, from the leaves –the residues corresponding nodes– to the root, is performed so that the information of each parent node is calculated based on the information of the child nodes. Next, the dendrogram is traversed in a top-down manner: for each node where a maximization of the inter-cluster dissimilarity is found, the tree is splitted and a cluster is formed and outputted. The procedure is applied recursively until the stop criterion is met. In the following two sub-sections, we describe in detail each of these steps.

A. Node Annotation

Each non-leaf node of the dendrogram is considered to be the root of a possible sub-cluster and it is annotated with a tuple containing information on properties P1, P2 and P3 as described in Section III (see Fig. 3).

For each non-leaf node other auxiliary information is kept:

- Average value of the positions in the linear sequence of the residues in the cluster
- Average vector of the tri-dimensional coordinates of the residues in the cluster along the simulation.

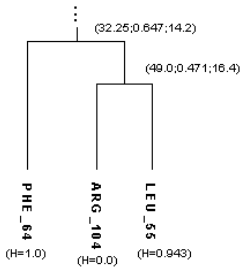

Fig. 3. Example of a sub-cluster in the extended tree and the respective node information. Amino-acids correspond to sub-clusters 127 and 128 in Fig. 4.

Input : tp $(tree\ pointer)$
$list\ of\ tuples\ (P1, P2, P3),\ one\ for\ each\ residue$
foreach $non-leaf\ node$ **do**
  if $ClusterSize = 2$
  then the information tuple is calculated between the two leaf nodes (residues);
  if $ClusterSize > 2$
  then the information tuple is calculated between a leaf node and a centroid of the other child sub-cluster or between the centroids of the two sub-clusters;
end

Algorithm 1: Pseudo-algorithm for the node annotation process.

B. Tree Traversal

After the dendrogram is extended with background information, the next step is to find those sub-clusters showing well differentiated characteristics. The tree is splitted in a certain node if there is a maximization of the inter-variation of the respective child nodes. To measure the dissimilarity between tuples of information of two different nodes the weighted Euclidean distance was applied (Eq. 4). The weights ω_i, used in the equation, give the ponderation factor to the different properties used to discriminate the sub-clusters. Since the parameters of the tuple have different scales, normalization is required. This can easily be done by Eq. 2, where the minimum and maximum values of each parameter are required.

$$|\overrightarrow{X} - \overrightarrow{Y}| = \sqrt{\sum_{i=1}^{n} \omega_i \cdot (X_i - Y_i)^2} \ where \ \sum_{i=1}^{n} \omega_i = 1 \quad (4)$$

Algorithm 2 shows the pseudo-code of the function *findSubClusters* that summarizes step 3 and 4 of the method (Section III). It describes the details of the top-down traversal of the tree and the respective output of the clusters. The algorithm is initially called passing a pointer to the root of the tree, the splitting threshold and as the stop criterion the minimum size of a cluster, *findSubClusters(tree→root, splitThreshold, minClusterSize)*.

VI. Results

Potentially, hierarchical clustering allows us to subdivide the collection of patterns of SASA variation into discrete classes. Figure 4 shows the dendrogram obtained when the Pearson's

input : tp $(tree\ pointer)$, δ $(Split\ Threshold)$, $minSize$
$ed = eucliDist(tp.Left, tp.Right)$;
if $((tp.Left.size < minSize)\ OR\ (tp.Right.size < minSize))$ **then**
    $outputCluster(tp)$;
end
if $(ed > \delta)$ **then**
    $outputCluster(tp.Left)$;
    $outputCluster(tp.Right)$;
else
    $findSubClusters(tp.Left, \delta, minSize)$;
    $findSubClusters(tp.Right, \delta, minSize)$;
end

Algorithm 2: Pseudo-algorithm of the function *findSubClusters*.

TABLE I
DIFFERENT VALUES TESTED FOR THE SPLIT THRESHOLD (δ) AND PROPERTIES PONDERATION FACTORS (ω_i).

	Threshold	Ponderation Factor			#clusters
		P1	P2	P3	
HC 1	0.05	1.0	0.0	0.0	4
HC 2	0.05	0.0	1.0	0.0	6
HC 3	0.05	0.0	0.0	1.0	4
HC 4	0.05	0.5	0.5	0.0	4
HC 5	0.05	0.5	0.0	0.5	4
HC 6	0.05	0.0	0.5	0.5	4
HC 7	0.05	0.33	0.33	0.33	18
HC 8	0.1	1.0	0.0	0.0	7
HC 9	0.1	0.0	1.0	0.0	6
HC 10	0.1	0.0	0.0	1.0	8
HC 11	0.1	0.5	0.5	0.0	6
HC 12	0.1	0.5	0.0	0.5	13
HC 13	0.1	0.0	0.5	0.5	6
HC 14	0.1	0.33	0.33	0.33	6
HC 15	0.8	1.0	0.0	0.0	20
HC 16	0.8	0.0	1.0	0.0	18
HC 17	0.8	0.0	0.0	1.0	20
HC 18	0.8	0.5	0.5	0.0	20
HC 19	0.8	0.5	0.0	0.5	20
HC 20	0.8	0.0	0.5	0.5	20
HC 21	0.8	0.33	0.33	0.33	20

correlation coefficient is used to study the variation of SASA values of each one of the 127 amino-acid residues along a MD unfolding simulation of WT-TTR. Pearson correlation determines the extent to which the SASA values of the two residues are linearly related to each other. By applying this metric we were looking for clusters of residues with similar global SASA behavior along the MD simulation.

Several clustering solutions were computed testing different combinations of properties ponderation factors and the split threshold. The description of experiments conducted is given in Table I. Many of the experiments led to the same clustering solutions. Others produced less interesting results as the residues were more or less equally divided in the clusters without any specific characteristic of interest distinguishing them. The most common solution was the one obtained with the parameters described for experiment HC 21 in Table I. This was considered the most interesting solution for analysis for two main reasons: (i) split threshold was set to a high value ($\delta = 0.80$) and consequently the tree will only be splitted when the dissimilarity value between two tuples of information is really significant; (ii) all properties contribute equally to the characterization of the clusters obtained. Figure 4 shows the similarity matrix and the dendrogram produced by CLUTO. The final sub-clusters produced by the method proposed here are shown in Table II.

TABLE II
CLUSTERS RESULTING FROM THE TRAVERSAL OF THE EXTENDED DENDROGRAM FOR A SPLIT THRESHOLD $\delta = 0.8$ AND $\omega_i = 0.33$.

Cluster [a]		Sub-cluster [b]	P1 [c]	P2 [c]	P3 [c]
I	1	{35, 68, 37, 8}	31.00	9.37	-0.56
	2	{84, 90, 80, 31, 72}	23.75	14.92	-3.31
II	3	{92, 94, 93, 85, 89, 22, 24}	2.00	14.08	-1.22
	4	{124, 46, 120, 45}	76.50	5.34	-1.61
III	5	{41, 70, 115, 87}	28.50	19.62	-0.91
	6	{3, 4, 2, 5}	2.00	2.65	-1.18
IV	7	{101, 39, 34, 62, 43, 36, 42, 100}	43.50	6.73	-4.80
	8	{44, 99, 57, 67}	8.75	11.94	-0.97
V	9	{117, 38, 96, 98, 83, 52, 95, 122, 76, 81}	50.87	12.87	-2.05
VI	10	{69, 97, 65, 6, 7, 121, 21, 66}	24.86	11.19	-2.23
VII	11	{103, 63, 59, 18, 20, 82, 77, 78, 102, 26}	33.75	25.26	-2.32
VIII	12	{30, 32, 29, 47, 48, 27, 9, 33}	14.13	6.48	-0.49
	13	{11, 10, 58, 60, 12, 104, 55, 64, 54}	10.48	6.41	-0.64
	14	{28, 53, 116, 61}	40.00	14.74	-0.49
	15	{13, 49, 14, 16, 15, 23}	13.50	8.86	-0.18
	16	{127, 74, 75, 86, 25, 106, 118}	37.81	8.75	-3.14
IX	17	{126, 123, 125, 50, 51, 1, 56}	15.00	25.11	-3.21
	18	{113, 114, 111, 73, 79, 91, 88}	11.88	8.78	1.25
X	19	{112, 110, 19, 119, 108, 17}	71.50	8.38	1.25
	20	{109, 71, 105, 107, 40}	36.00	9.06	1.61

[a] Clusters obtained by the classical hierarchical clustering.
[b] Sub-clusters obtained from the application of our method.
[c] For all properties, the absolute values are shown.

The 127 residues that constitute the WT-TTR protein were clustered in 20 clusters. Some clusters are constituted by residues that are close in the protein sequence (clusters 3 and 6). Other clusters (12, 13, 14, 15 and 16) show that high correlation can be found between residues that are far apart in the protein sequence (Fig. 3). In this case, the method shows that the residues are spatially close during the the simulation.

Further analysis will be given to sub-clusters 3 to 4, 12 to 16 and 19 to 20, originated from clusters II, VIII and X, respectively, after the background information has been introduced on the nodes of the dendrogram (Figures 3 and 4).

The identification of sub-clusters 3 and 4 from cluster II is a good example of the use of the background information. The residues in sub-cluster 4 are far apart in the protein sequence (high value of P1) but are close in the spatial arrangement of the protein (low value of P2). On the other hand, the residues in sub-cluster 3 show exactly the opposite behavior: they are close in the protein sequence (low value of P1) but are far apart in the spatial arrangement of the protein (high value of P2). The distinction between the residues in these two sub-clusters would not be clarified if additional background information had not been introduced in the analysis.

Through classical hierarchical clustering, 20% of the

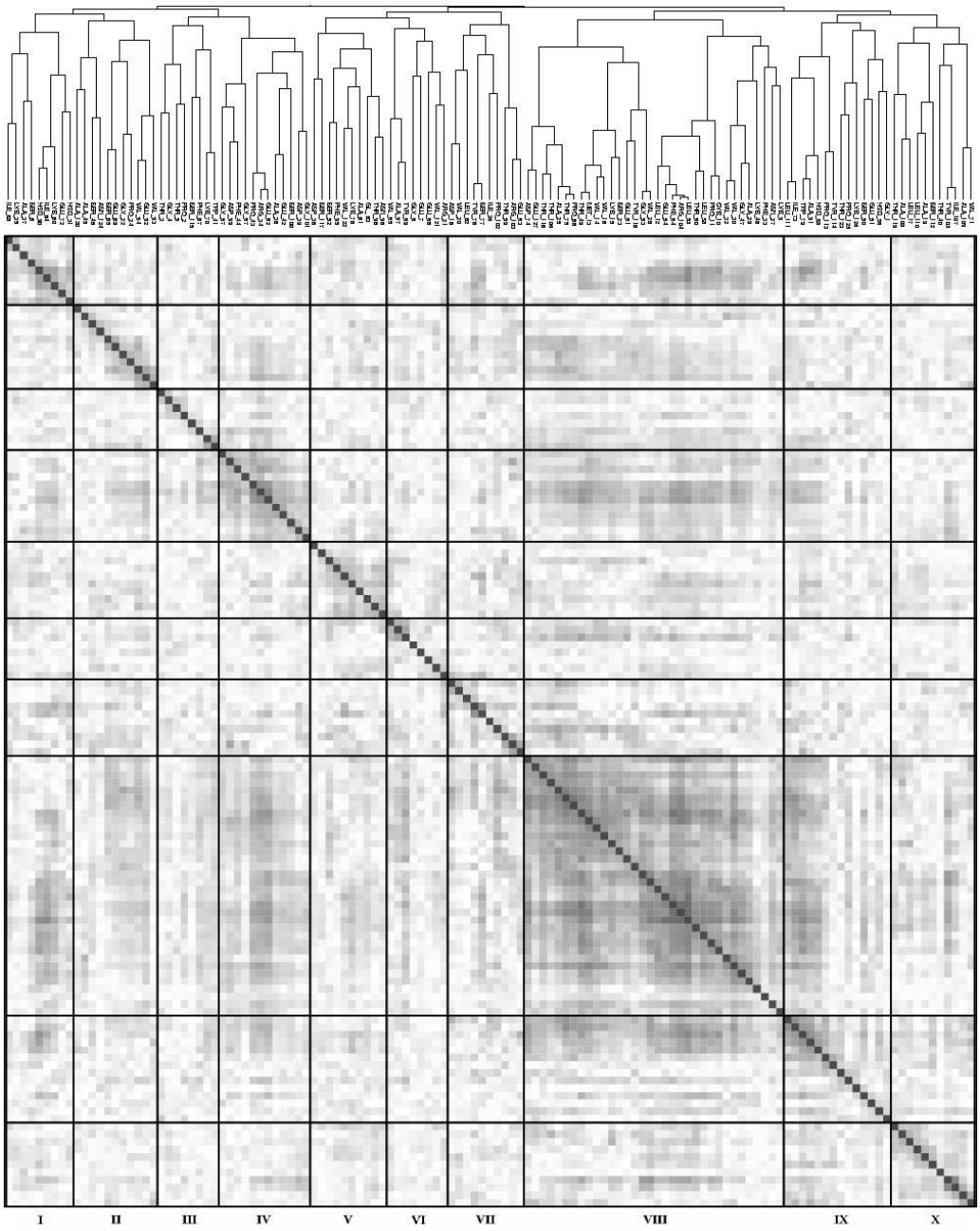

Fig. 4. The dendrogram and the respective similarity matrix, obtained through the application of classical hierarchical clustering to the dataset.

residues of the protein (34 residues) were placed in Cluster VIII . However, our method allowed a better discrimination among the residues constituting this cluster. In fact, the method made possible to identify, at least, four different patterns of SASA behavior (Fig. 5, Panels A, B, C and D) based solely on the residues background information. The overall trends found may be described as follows. Sub-cluster 13 is mainly composed of hydrophobic residues behaving in the same fashion: the residues remain unexposed in the first half of the simulation, and from that point on go progressively to regions of high exposure to the solvent (Fig. 5, Panel A). Although, the residues in sub-cluster 14 have similar characteristics to the residues in sub-cluster 13, the SASA pattern is sightly different. The residues stay unexposed in the first part of the trajectory, after which they become highly exposed to the solvent. Around the middle of the simulation the residues tend to become hidden, and get progressively exposed until the end of the simulation (Fig. 5, Panel B). Sub-cluster 15 is mainly composed of residues that are hydrophobic and that are involved in the formation of well defined protein topologies, like β-sheets. The SASA values of these residues seem to be varying in a certain range of values (Fig. 5, Panel

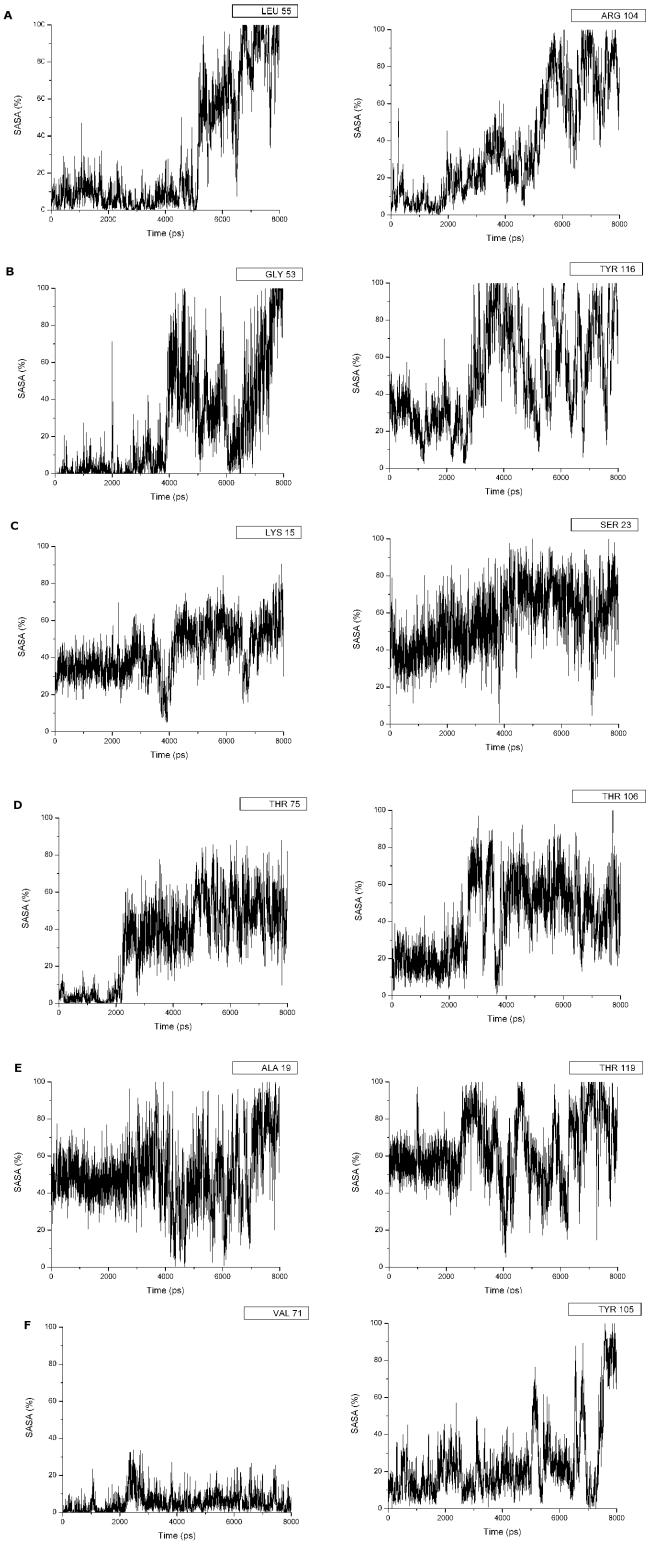

Fig. 5. Examples of SASA pattern behaviors obtained with the augmented hierarchical clustering methodology.

C). The residues in sub-cluster 16 are involved in less well defined topologies, like turns and loops. In the beginning of the simulation, these residues are buried (SASA $\leq$ 25%), but they rapidly become exposed (Fig. 5, Panel D). Sub-cluster 12 shows the worst correlation values and it was not possible to identify a pattern of SASA behavior for the residues that constitute this sub-cluster.

The values of correlation found in sub-clusters 19 and 20 are similar (Fig. 4). Thus, splitting cluster X would not be intuitively trivial. From Table II, we can see that the difference between the two sub-clusters is mainly due to property P1. Although in the two sub-clusters, residues that are far apart in the protein sequence seem to be better correlated than residues that are close in the protein sequence, different SASA variation trend are observed (Fig. 5, Panels E and F). While in sub-cluster 20 residues show low exposure along the simulation(Fig. 5, Panel F), in sub-cluster 19 residues have higher exposure to the solvent (Fig. 5, Panel E).

VII. Conclusion

In this paper, we show that hierarchical clustering may be a useful tool to identify and compare patterns of solvent accessible surface area variations of individual amino-acid residues in protein unfolding simulations. However, an intuitive interpretation of the dendrogram obtained by classical hierarchical clustering computation may prove to be a difficult task. Since additional information is available for the characterization of the MD unfolding simulations (see Sect. II-B), it has been used to enhance the cluster analysis process. The idea is to use the supplementary information to enrich the dendrogram, allowing to find sub-clusters with well differentiated characteristics, and providing the means for a better understanding of the cluster differences. This demanded a top-down traversal of the hierarchical tree, where sub-clusters are isolated when the children of a node show an overall significant difference.

The results here reported show that the proposed approach proved to be useful since it allows the identification of well characterized sub-clusters. Moreover, it uncovers new relations between the residues, in particular between those far apart in the sequence. This may provide new and helpful insights to the comprehension of the folding and unfolding processes of the protein.

Acknowledgment

The authors acknowledge the support of the ”Fundação para a Ciência e Tecnologia” and the program FEDER, Portugal, through grant POCTI/BME/49583/2002 (to RMMB) and the Fellowships SFRH/BD/13462/2003 (to PGF) and SFRH/BD/16888/2004 (to CGS). We thank the Center for Computational Physics, Departamento de Física, Universidade de Coimbra, for the computer resources provided for the MD simulations.

References

[1] Quintas, A., Vaz, D. C., Cardoso, I., Saraiva, M. J. and Brito, Rui M. M., ”Tetramer dissociation and monomer partial unfolding precedes protofibril formation in amyloidogenic transthyretin variants,” *J. Biol. Chem.*, vol. 276, pp. 27207-27213, 2001.

[2] Rodrigues, J. R. and Brito, R. M. M., "How important is the role of compact denatured states on amyloid formation by transthyretin?," *Amyloid and Amyloidosis*, CRC Press, pp. 323-325, 2004.
[3] Correia, B. E., Loureiro-Ferreira, N., Rodrigues, J. R. and Brito, Rui M. M., "A structural model of an amyloid protofilament of Transthyretin,", *Protein Sci.*, vol. 15, pp. 28-32, 2006.
[4] Brito, R. M. M., Damas, A. M. and Saraiva, M. J., "Amyloid Formation by Transthyretin: From Protein Stability to Protein Aggregation," *Curr. Med. Chem. - Imun., Endoc. & Metab. Agents*, vol. 3, pp. 349-360, 2003.
[5] Brito, R. M. M., Dubitzky, W. and Rodrigues, J. R., "Protein folding and unfolding simulations: A new challenge for data mining," *OMICS: A Journal of Integrative Biology*, vol. 8, pp. 153-166, 2004.
[6] Azevedo, P. J., Silva, C. G., Rodrigues, J. R., Loureiro-Ferreira, N. and R. M. M. Brito, "Detection of Hydrophobic Clusters in Molecular Dynamics Protein Unfolding Simulations Using Association Rules," *Proc. 6th International Symposium ISBMDA 2005, Lect. Notes Comput. Sc.*, vol. 3745, pp. 329-337, 2005.
[7] Ferreira, P. G., Azevedo, P. J., Silva, C. G. and Brito, R. M. M., "Mining Approximate Motifs in Time Series," *Lect. Notes Art. Intl.*, vol. 4265, pp. 77-89, 2006.
[8] Moraitakis, G. and Goodfellow, J. M., "Simulations of Human Lysozyme: Probing the Conformations Triggering Amyloidosis," *Biophys. J.*, vol. 84, pp. 2149-2158", 2004.
[9] Shamsir, M. S. and Dalby, A. R., "One Gene, Two Diseases and Three Conformations: Molecular Dynamics Simulations of Mutants of Human Prion Protein at Room Temperature and Elevated Temperatures," *Proteins*, vol. 59, pp. 275-290, 2005.
[10] Nowak, M., "Immunoglobulin Kappa Light Chain and Its Amyloidogenic Mutants: A Molecular Dynamics Study," *Proteins*, vol. 55, pp. 11-21, 2004.
[11] Armen, R. S., Alonso, D. O. and Daggett, V.", "Anatomy of an amyloidogenic intermediate: conversion of beta-sheet to alpha-sheet structure in transthyretin at acidic pH," *Structure*, vol. 12, pp. 1847-1863, 2004.
[12] Lei, M. and Yanga, M. and Huo, S., "Intrinsic versus mutation dependent instability/flexibility: a comparative analysis of the structure and dynamics of wild-type transthyretin and its pathogenic variants," *J. Struct. Biol.*, vol. 148, pp. 153-168, 2005.
[13] Hamilton, W., Steinrauf, L. K., Liepnieks, J., Braden, B. C., Benson, M. D., Holmgren, G., Sandgren, O. and Steen, L., "The X-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val30Met variant to 1.7 Å resolution," *J. Biol. Chem.*, vol. 268, pp. 2416-2424, 1993.
[14] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, "NAMD2: Greater scalability for parallel molecular dynamics," *J Comp.Physics*, vol. 151, pp. 283-312, 1999.
[15] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J. and et al., "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, pp. 3586-3616,1998.
[16] Lee, B. and Richards, F. M., "The interpretation of protein structures: estimation of static accessibility," *J. Mol. Biol.*, vol. 55, pp. 379-400, 1971.
[17] Hubbard, S. J. and Thornton, J. M., "NACCESS, Computer Program," Department of Biochemistry and Molecular Biology, University College London, 1993.
[18] Zhao, Y. and Karypis, G., "Clustering in the life sciences," *Methods Mol. Biol.*, vol. 224, pp. 183-218, 2003.
[19] Radzicka, A. and Wolfenden, R., "Comparing the Polarities of the Amino Acids: Side-Chain Distribution Coefficients between the Vapor Phase, Cyclohexane, 1-Octanol, and Neutral Aqueous Solution," *Biochemistry*, vol. 27, pp. 1664-1 670, 1988.
[20] Han, J. and Kambler, M., "Data Mining, Concepts and Techniques," *Morgan Kaufmann*, 2001.
[21] Heyer, L. J., Kruglyak, S. and Yooseph, S., "Exploring expression data: Identification and Analysis of coexpressed genes," *Genome Research*, vol. 9, pp. 1106-1115, 1999.
[22] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 14863-14868, 1998.
[23] Zar, J. H., "Biostatistical Analysis," *Prentice Hall*, 1998.
[24] Karypis, G., "CLUTO - A clustering toolkit," University of Minnesota, 2003.