# Detection of Hydrophobic Clusters in Molecular Dynamics Protein Unfolding Simulations Using Association Rules

Paulo J. Azevedo[1], Cândida G. Silva[2], J. Rui Rodrigues[2], Nuno Loureiro-Ferreira[2], and Rui M. M. Brito[2,3,*]

[1] Departamento de Informática, Universidade do Minho, 4710-057 Braga, Portugal
pja@di.uminho.pt
[2] Centro de Neurociências de Coimbra, Universidade de Coimbra, 3004-517 Coimbra, Portugal
[3] Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, 3004-535 Coimbra, Portugal
brito@ci.uc.pt

**Abstract.** One way of exploring protein unfolding events associated with the development of Amyloid diseases is through the use of multiple Molecular Dynamics Protein Unfolding Simulations. The analysis of the huge amount of data generated in these simulations is not a trivial task. In the present report, we demonstrate the use of Association Rules applied to the analysis of the variation profiles of the Solvent Accessible Surface Area of the 127 amino-acid residues of the protein Transthyretin, along multiple simulations. This allowed us to identify a set of 28 hydrophobic residues forming a hydrophobic cluster that might be essential in the unfolding and folding processes of Transthyretin.

## 1 Introduction

One of the most challenging problems in molecular biology today 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. This sequence of amino-acids is encoded by the linear sequence of nucleotides in a gene, but protein function is mediated by its exquisite three-dimensional structure. Predicting the 3D structure of a protein from the linear sequence of amino-acids is as yet an unsolved problem today, and a challenge for those eager to harness the information content of the genomes.

In recent years, the issues of protein folding became also pivotal in the understanding of a series of human and animal diseases, generally know as conformational dis-

* The authors acknowledge the support of the "Fundação para a Ciência e Tecnologia" and the program FEDER, Portugal, through projects POSI/SRI/39630/2001/CLASS (to PJA), POCTI/BME/49583/2002 (to RMMB) and the Fellowships SFRH/BD/1354/2000 (to NLF), SFRH/BD/16888/2004 (to CGS) and AI/06/02 (to JRR). We thank the Center for Computational Physics, Departamento de Física, Universidade de Coimbra, for the computer resources provided for the MD simulations.

orders or amyloid disorders, and ranging from Alzheimer´s to bovine spongiform encephalopathies (BSE). 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), a human plasma protein responsible for such amyloid diseases as Familial Amyloid Polyneuropathy (FAP), Familial Amyloid Cardiomyopathy (FAC) and Senil Systemic Amyloidosys (SSA), using both experimental and computational methodologies [1,2,3]. One way of exploring the unfolding events that may be responsible for TTR aggregation is through the use of Molecular Dynamics Protein Unfolding Simulations (MDPUS). However, we know today that in an ensemble of protein molecules not all of them follow the same folding or unfolding route. Thus, multiple simulations are required in order to have some idea of the conformational space available to a protein molecule in its unfolding process. These simulations are computationally expensive and generate a huge amount of data. In order to contrast, compare and characterize the molecular properties associated with each simulation, Data Mining techniques are required.

In the present paper, we report the use of Association Rules, a specific Data Mining technique to identify relations between atomic elements of the data, in order to detect potential coordinated movements of different amino-acid residues in unfolding simulations of the protein Transthyretin. In particular, through the analysis of the Solvent Accessible Surface Area (SASA) of each amino-acid residue along multiple unfolding simulations of TTR, we identify 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.

## 2 Molecular Dynamics Protein Unfolding Simulations

Molecular Dynamics (MD) simulations have recently become an important tool to explore folding and unfolding processes in proteins [4,5,6] and we have put forward some of the challenges facing the researcher when comparing and contrasting the results of multiple MD simulations in different proteins [7].

In Molecular Dynamics simulations, molecules are treated as spheres connected by springs and classical mechanics are used to calculate forces and velocities. Although this treatment is highly approximated and does not take into account quantum effects, the realism of the simulation depends on the ability of the potential energy function to reproduce the inter-atomic interactions characteristic of the molecular system under study. In fact, several decades of research in small molecules and biological macromolecules allowed the definition of a generally accepted set of empirical functions, and today MD is a well established method for studying equilibrium protein dynamics and non-equilibrium processes such as protein folding and unfolding.

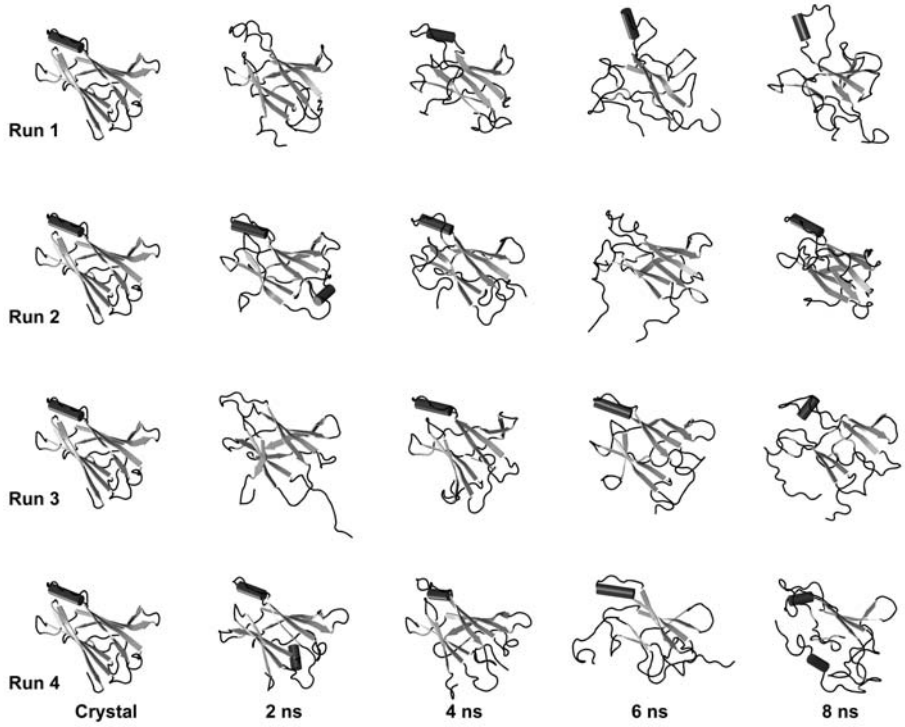


**Fig. 1.** Secondary structure ribbon representations of the monomer of WT-TTR along four different Molecular Dynamics unfolding simulations. Beta-strands, alpha-helices and turns and coil are represented by arrows, cylinders and tubes, respectively. The difference in the setup of each of the four runs resides in the assignment of initial atomic velocities

### 2.1 Simulation Details

Initial coordinates for Transthyretin were obtained from the crystal structure (PDB entry 1tta) [8] and hydrogen atoms were added. All minimization and MD procedures were performed with the program NAMD [9], using version 27 of the CHARMM force field [10]. All atoms were explicitly represented. Internal waters were placed with the program Dowser [11] and the program Solvate (http://www.mpibpc.mpg.de/abteilungen/071/solvate/node8.html) was used to add solvent water molecules and $Na^{+}Cl^{-}$ ions around the protein. The complete system was comprised of 45,256 atoms.

The system was minimized, equilibrated and heated to the target temperature. Several simulations were performed using *Centopeia*, a Linux computer cluster at UC. Control simulations, at 310 K, and several unfolding simulations, at 500 K, were performed for up to 10 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. Figure 1 shows a set of representative trajectories for the thermally-induced unfolding of a TTR monomer.

## 2.2 Trajectory Analysis

Several global molecular properties, such as radius of gyration ($R_g$), root mean square deviation (RMSD), secondary structure and native contacts, among others, may be calculated along each trajectory in order to characterize and map the unfolding events. Here we calculated the Solvent Accessible Surface Area (SASA) of each individual amino-acid residue along the MD unfolding trajectories in order to study potentially correlated behavior among different residues.

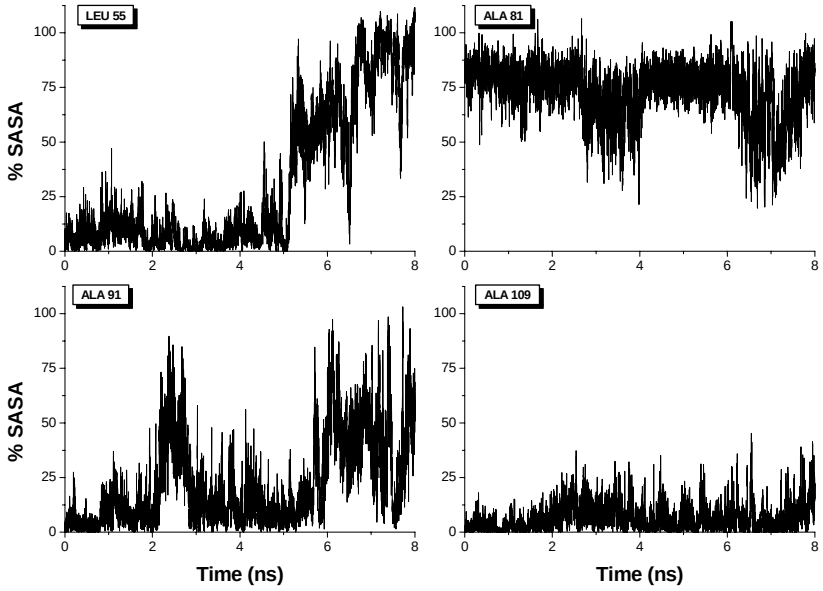


**Fig. 2.** Variation of the Solvent Accessible Surface Area (SASA) for individual amino-acid residues along a Molecular Dynamics unfolding simulation of the protein Transthyretin, at 500 K. 0% indicates an accessible surface of 0 $Å^2$ and 100% indicates a SASA for X equal to what is determined in the tripeptide Ala-X-Ala

The solvent accessible surface area (SASA) is the surface of the protein available to a spherical probe of 1.4 Å diameter, and was calculated using the program *naccess* [12]. The monomer of TTR has 127 amino-acid residues and each simulation trajectory analyzed here is constituted by 8,000 frames (one frame saved per ps simulated).

Thus, for each simulation we have 127 plots (one per residue) of SASA *vs* time with 8,000 points (one point per frame). Figure 2 shows four examples of these plots. Leu55 is unexposed in the native structure and in the first half of the simulation, but it becomes highly exposed to the solvent late in the simulation. Ala109 is always unexposed to the solvent, even late in the simulation when the protein is already denatured. In general terms, some of the residues roughly follow the SASA patterns shown in Figure 2, but several other patterns are also observed. In order to find groups of residues that change solvent exposure in a coordinated fashion during the unfolding simulations, we have searched for Association Rules as detailed below.

## 3 Searching for Association Rules

Association Rules [13] represent a pattern language to describe relations among atomic elements (items) of the data. They hold simple and clear semantics and are of the form:

$$A_1 \,\&\, A_2 \,\&\, A_3 \,\&\, \ldots \,\&\, A_n \rightarrow C$$

A rule is derived from a co-occurring set of items (itemset). In the present case, the itemset could be $C,A_1,A_2,A_3,...,A_n$. For the specific problem of SASA data analysis, items correspond to attribute/value pairs (residue / SASA value). The consequent C may be a set of items but here we only consider rules with a single item as consequent.

Quality and usability of the rules are measured through two types of metrics - predictability and incidence. Traditional metrics are *Support* (for incidence) and *Confidence*. *Support* is calculated by itemset counting among the transactions (records) contained in the data. *Confidence* corresponds to the strength of the rule and is obtained from the conditional probability of the consequent knowing the antecedent.

The aim of a rule generator algorithm is to derive high strength and interesting (surprising) rules. The user provides a minimal incidence (*Support*) value to avoid considering rare phenomena in the data. Thus, a minimal *Support* value filters out items (and itemsets) that occur in a low number of records. A minimal *Confidence* threshold is also supplied to select only high strength rules. The number of frequent items (which occurrence satisfies the minimal *Support*) is an important parameter for the computational complexity of the data analysis.

### 3.1 The Data

The studied data is composed of four different unfolding simulations of WT-TTR and L55P-TTR (with a Proline replacing a Leucine in position 55). The data describes SASA variations along 8000 frames (records) corresponding to the 8 ns of each run (simulation). 127 attributes (amino-acid residues that constitute the protein) are present. We removed the temporal label present in each frame so that only intra-transactional relations would be extracted. These datasets turn out to be filled with very dense data, which makes rule generation into a computational hard task.

We exposed all datasets to a discretization process that, according to the analysis, reduced the number of values per attribute to 2 or at most 3. The values correspond to low ([0,25[ ), high ([75,100[ ) and medium SASA values. The latter was interpreted as a null value, which the system was programmed to ignore. Thus, the data was discretized to mainly consider unexposed or highly exposed amino-acid residues, along the MD simulations. Hence, we managed to reduce the number of frequent items being considered and consequently the complexity of the computational problem. In general, discretization reduces the complexity of the problem and leads to the derivation of higher quality rules.

### 3.2 Rule Generation

The standard algorithm to derive association rules is *Apriori* [13,16]. This algorithm is divided in two main steps: i) mining of frequent patterns (the extraction of itemsets that satisfy minimal support); and ii) rule generation (to derive rules using the frequent itemsets). The first step is the computational hard task and it has received considerable attention from the Data Mining community. Several proposals exist in the literature. We use CAREN (developed in [14]) which includes an algorithm for mining frequent patterns based on depth first expansion with bitwise representation. CAREN also implements several features for rule derivation and selection, namely antecedent and consequent filtering (item or attribute specification), max/min number of items in a rule, different metrics, $\chi^2$ test during itemset mining (which significantly reduces the number of relevant itemsets), improvement filtering on rules (to eliminate redundant rules), etc [15,16]. There are several metrics to evaluate association rules. We used the standard confidence metric. However, other metrics are available in the CAREN system.

Several extractions (queries) were performed on each discretized simulation. Each query was designed to answer a relevant biochemical question. For example, we were particularly interested in verifying which chemical classes of amino-acid residues behaved in a similar fashion and, among these classes, which particular residues behaved similarly. We designed queries to relate the following groups of amino-acid residues:

i) hydrophobics *vs* hydrophobics
ii) aromatics *vs* aromatics
iii) hydrophobics *vs* hydrophilics
iv) positively charged *vs* negatively charged

For instance, to derive rules that represent the interaction between hydrophobic residues (i) we designed a script to invoke the CAREN system with specific parameters: consequent filter, list of hydrophobic residues as possible antecedents, specific minimal number of items as 4, among other less relevant parameters. The minimal number of items varied, depending on the query being addressed. For example, in query i) we used 4 and in query iv) we used 2 as the minimal number of items.

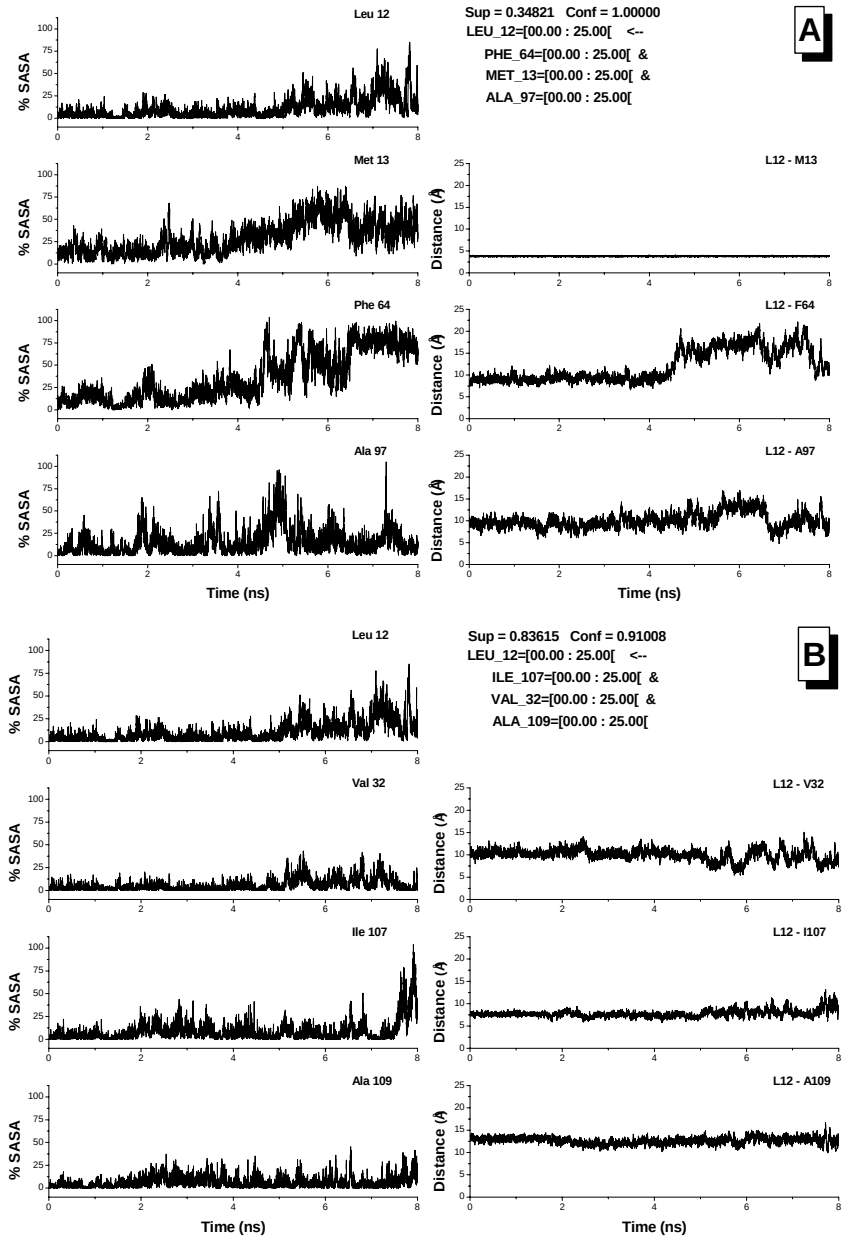


**Fig. 3.** Example of two Association Rules (at the top of each panel) generated by CAREN and involving hydrophobic residues. Panel **A** shows a low *Support* rule and Panel **B** a high *Support* rule. The plots on the left show the change in relative SASA along one MD unfolding simulation of WT-TTR for the residues involved in each derived rule. The plots on the right show inter-residue distances for each pair of residues involved in each derived rule

This difference is justifiable by chemical arguments. In the first case (i), we were looking for interactions between 4 or more residues - a hydrophobic cluster. In the last case (iv) we were looking for interactions between pairs of positively and negatively charged residues - a salt bridge.

### 3.3 Rule Selection

At this stage, we were mostly interested in finding amino-acid residues that had similar patterns of solvent exposure. Rules with high support are more significant since they represent a phenomenon that persists along a larger interval of data frames. As a first approach, we considered rules with *Support* above 30% and *Confidence* above *90%*. It turns out that the only rules having very high *Support* are those relating hydrophobic residues with hydrophobic residues. All other queries generate rules with low *Support*. Figure 3 shows two examples of hydrophobic rules, one with relatively low *Support* (Fig. 3A) and another with very high *Support* (Fig. 3B). In both cases, the rules have 3 antecedents and 1 consequent, but there are rules involving 5, 6, 7 and more antecedents.

In order to have a deeper understanding of the processes being revealed by the Association Rules, we determined the inter-residue distances among all the residue pairs involved in each rule (Fig. 3, Panels on the right). While for the lower *Support* rule (Fig. 3A) the inter-residue distances vary more widely, for the high *Support* rule (Fig. 3B) inter-residue distances vary much less along the unfolding simulation. This metric clearly reveals amino-acid residues that not only have similar SASA behaviors but also maintain the same spatial relation among them along the unfolding simulation. Using a similar treatment for all the rules obtained, we were able to define a set of 28 hydrophobic amino-acid residues that share two main characteristics: i) they remain unexposed to the solvent; and ii) they display a constant spatial relation during most of the unfolding simulation. As may be seen in Figure 4, the 28 identified hydrophobic residues are concentrated in the interior of the protein. Thus, we may conclude that this set of 28 residues out of 127 may constitute a hydrophobic cluster essential in TTR unfolding and folding.

## 4 Conclusions

In this report we show that Data Mining techniques, such as searching of Association Rules, applied to the analysis of a massive quantity of data generated in Molecular Dynamics Protein Unfolding Simulations, are in fact very useful in the detection of hidden relations among the constituents of the molecular system under study. Association rules appear as a pattern language with high potential to express common behavior among amino-acid residues during the protein unfolding process. Rules are simple objects with clear reading. They are easy to interpret and useful for future prediction tasks.

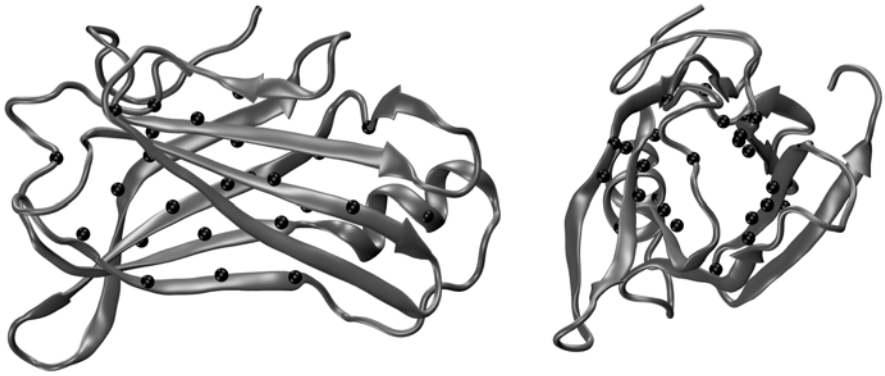

**Fig. 4.** Schematic representations of the backbone structure of the monomer of Transthyretin. Black spheres indicate the positions of the $C_\alpha$ atoms of the 28 hydrophobic residues identified by the Association Rules. The two views are related by a 90º rotation

In the case of the molecular system studied here - the unfolding behavior of the protein Transthyretin -, we searched for Association Rules among the variation profiles of Solvent Accessible Surface Area (SASA) of each one of the 127 residues of TTR in several unfolding simulations. This allowed us to define a group of 28 hydrophobic residues which appear to form a hydrophobic cluster essential in the unfolding and folding processes of TTR. Thus, the application of specific data mining techniques to an extremely large set of data generated from protein unfolding simulations helped us uncover new biochemically relevant knowledge.

## References


1. Quintas, A., Vaz, D.C., Cardoso, I., Saraiva, M.J.M., Brito, R.M.M.: Tetramer Dissociation and Monomer Partial Unfolding Precedes Protofibril Formation in Amyloidogenic Transthyretin Variants. J. Biol. Chem. 276 (2001) 27207-27213
2. Brito, R.M.M., Damas, A.M., Saraiva, M.J.S.: Amyloid Formation by Transthyretin: From Protein Stability to Protein Aggregation. Current Medicinal Chemistry - Immun. Endoc. & Metab. Agents 3 (2003) 349-360
3. Rodrigues, J.R., Brito, R.M.M.: How important is the role of compact denatured states on amyloid formation by transthyretin? In: Gilles Grateau, Robert A. Kyle and Martha Skinner (eds.): Amyloid and Amyloidosis. CRC Press (2004) 323-325
4. Daggett, V.: Molecular Dynamics Simulations of the Protein Unfolding/Folding Reaction. Acc. Chem. Res. 35 (2002) 422 - 429
5. Pande V.S., Baker I., Chapman J., Elmer S.P., Khaliq S., Larson S.M., Rhee Y.M., Shirts M.R., Snow C.D., Sorin E.J., Zagrovic B.: Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers 68 (2003) 91-109

6. Beck, D.A.C., Daggett, V.: Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods 34 (2004) 112-120
7. Brito, R.M.M., Dubitzky, W., Rodrigues, J.R.: Protein Folding and Unfolding Simulations: A New Challenge for Data Mining. OMICS A Journal of Integrative Biology 8 (2004) 153-166
8. Hamilton, J.A., Steinrauf, L.K., Braden, B.C., Liepnieks, J., Benson, M. D., Holmgren, G., Sandgren, O., Steen, L.: The X-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30-Met variant to 1.7 Å resolution. J. Biol. Chem. 268 (2003) 2416-2424
9. Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., Schulten, K.: NAMD2: Greater scalability for parallel molecular dynamics. J. Comp. Physics 151 (1999) 283-312
10. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., et al.: All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102 (1998) 3586-3616
11. Zhang, L., Hermans, J.: Hydrophilicity of cavities in proteins. Proteins: Struct. Func. Gen. 24 (1996) 433-438
12. Hubbard, S.J., Thornton, J.M.: NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London (1993)
13. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Databases, Chile (1994)
14. Azevedo, P.J., Jorge, A.M.: The CLASS project: http://www.niaad.liacc.up.pt/~amjorge/Projectos/Class/
15. Azevedo, P.J.: The Caren System: http://www.di.uminho.pt/~pja/class/caren.html
16. Azevedo, P.J.: CAREN – A java based Apriori implementation for classification purposes. Research Report – Departamento de Informática, Universidade do Minho (2003)