Optimal Identification of Semi-Rigid Domains in Macromolecules from Molecular Dynamics Simulation

Stefan Bernhard*, Frank Noé

Free University Berlin, DFG Research Center MATHEON, Berlin, Germany

Abstract

Biological function relies on the fact that biomolecules can switch between different conformations and aggregation states. Such transitions involve a rearrangement of parts of the biomolecules involved that act as dynamic domains. The reliable identification of such domains is thus a key problem in biophysics. In this work we present a method to identify semi-rigid domains based on dynamical data that can be obtained from molecular dynamics simulations or experiments. To this end the average inter-atomic distance-deviations are computed. The resulting matrix is then clustered by a constrained quadratic optimization problem. The reliability and performance of the method are demonstrated for two artificial peptides. Furthermore we correlate the mechanical properties with biological malfunction in three variants of amyloidogenic transthyretin protein, where the method reveals that a pathological mutation destabilizes the natural dimer structure of the protein. Finally the method is used to identify functional domains of the GroEL-GroES chaperone, thus illustrating the efficiency of the method for large biomolecular machines.

Citation: Bernhard S, Noé F (2010) Optimal Identification of Semi-Rigid Domains in Macromolecules from Molecular Dynamics Simulation. PLoS ONE 5(5): e10491. doi:10.1371/journal.pone.0010491

Editor: Markus J. Buehler, Massachusetts Institute of Technology, United States of America

Received October 5, 2009; **Accepted** April 14, 2010; **Published** May 13, 2010

Funding: This work was funded by the DFG Research Center MATHEON (http://www.matheon.de). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bernhard@math.fu-berlin.de

Introduction

The mechanical properties of biomolecules and their complexes are essential to molecular function, because many molecular processes are accompanied by conformational changes, in which domains of the molecule must be able to move with respect to each other [1–5]. For example the mechanical properties of actin are strongly coupled to polymer formation and degradation [6]. Such a coupling between different functional states and aggregation states of molecules and their mechanical properties are ubiquitous in biology. Understanding the nanomechanics of the biomolecules, i.e. the semi-rigid domains and their relative mobility for each given conformational or aggregation state, is thus one of the key questions in molecular biophysics allowing for both (i) the understanding/analysis of the molecular nanomechanics and (ii) paving the ground for efficient large-scale coarse-grained simulations [7–9].

The first step to analysis and simulation of molecular nanomechanics is the identification of the rigid and flexible parts of biomolecules in different chemical, conformational or aggregate states considered. Conventional experimental techniques, like for example nuclear magnetic resonance (NMR), provide limited information about these processes.

One approach to identify the rigid and flexible parts in biomolecules is to partition the system into domains (also called "groups" or "clusters" in other works) that are nearly rigid. In the coarse-grained model, these domains can only move as a rigid body with six degrees of freedom (3 translation + 3 rotation). Such a low dimensional model of the original high-dimensional dynamics yields itself easily to the understanding of essential mechanical properties of the molecule and how they change between conformations. Clearly, such a model only approximates the real mobility and the approximation error will depend on the number of domains considered and on the flexibility/rigidity of the molecule in the conformation considered. Consequently, such a model is better suited for describing functional transitions or aggregation than for processes involving much flexibility, such as folding.

Several methods for the identification of nearly rigid domains in biomolecules have been proposed that produce similar but not identical results. They can be categorized into model-based methods, where structural aspects such as hydrophobicity, topology, structural homology or for e.g. identical sequence motifs serve to identify the smallest building blocks [10–13]. In this category there are also a variety of methods that try to optimize certain structural properties of protein domains, such as the distance-mapping [14], interface area [15], specific volume [16] and compactness of the domain [17]. In [18] a cluster method is proposed that uses contact measures and fuzzy logic to define protein domains.

Data-based approaches in contrast define domains based on data of the flexibility of the biomolecule, such as MD simulations [19,20]. One approach to obtain correlated motion of atoms within the molecule is (quasi) harmonic analysis, namely Principal Component Analysis (PCA) and Normal Mode Analysis (NMA) [21,22]. Here, the motions that contribute most to the variation between the molecular configurations are described by the dominant eigenmodes of the covariance matrix or the Hessian of the potential, respectively. The subspace of the first few eigenmodes contains most of the flexibility and a number of methods have been developed to use this information in order to identify domains [23–25]. Other data-based approaches are based on dynamical clustering [26], hierarchical clustering of correlation patterns (HCCP) [27], and the hinge detection algorithm [19,28]. The latter algorithm assumes that collections of atoms move as rigid bodies connected by hinges or axes of rotation. Recently, [29] has proposed a optimal method to decompose proteins into rigid domains using equilibrium fluctuations of inter-residue distances.

Normal-mode-based techniques are limited by the fact that they only use local information of the energy landscape. PCA-based clustering methods do not suffer from this limitation, but still require all structures to be fitted to a mean or reference structure before calculating the covariance matrix. Such a fitting procedure works well as long as the structures are very similar, but if very large conformational changes are involved, then structures which are very similar to each other but very different from the reference structure may become very different after the fitting and thus produce a misleading covariance matrix. Thus, it is desirable to use a method that works with internal coordinates only. Moreover, there is a lack of the domain identification techniques that avoid ad-hoc assumptions and parameter choices that indirectly influence the number of clusters. It would be rather desirable to have an explicit control of the clustering error by adjusting the number of domains, or to have the method select the number of domains such that the clustering error is below a certain threshold.

The proposed method works by defining (i) a distance-deviation matrix between atoms based on dynamical data, (ii) formulating the clustering problem as a quadratic optimization problem that is based on this matrix and (iii) solving this clustering problem to optimality and obtaining an assignment of atoms to clusters. To illustrate the strengths and limitations of this approach a number of example systems are considered: two artificial peptides Ala$_5$ and $MR121-GSGSW$ and the two biomolecules transthyretin and the chaperone complex GroEL-GroES.

The immediate use of the method is to understand dynamic processes in large macromolecules and their complexes which involve changes of molecular rigidity. This includes processes like conformational changes, ligand binding and protein aggregation [30–32]. Besides this, the outcome of the method can be used in a number of other biophysical problems, including the coarse-grained simulation of macromolecular encounters and association.

Materials and Methods

The principal objective of this work is to develop a new coarse-graining technique to partition large molecular systems *optimally* into semi-rigid domains, thus providing a simple model of molecular nanomechanics. The proposed method is data-based and meets the following requirements:

1. Optimal and unique molecular partitioning for given data and number of domains
2. Works with internal coordinates only and is thus independent of a reference structure
3. Can be applied to characterize models with multiple conformations without “overlooking” rarely populated conformations
4. Error measure for coarse-grain quality and ability to adjust the accuracy by the number of domains or the maximum acceptable clustering error
5. Simple applicability and robustness - no parameters other than number of domains
6. Model independent, so that experimental findings are easily incorporated
7. Efficient and simple implementation

Molecular rigidity and distance deviation

Inter atomic distance-deviation is a common metric used for the identification of rigid domains in proteins [33,34]: Within a rigid domain, the euclidean distance between pairs of atoms remains constant, while it fluctuates for atom pairs that lie in different rigid domains moving relative to each other.

The analysis of local molecular rigidity is based on the distance deviation matrix $\mathbf{S}$, whose elements S_{ij} are the Euclidean distance deviations, between the atoms i and j in the molecule, defined as:

$$S_{ij}=\sqrt{\langle(d_{ij}-\langle d_{ij}\rangle)^2\rangle}, \tag{1}$$

where $\langle\rangle$ indicates the ensemble average and $d_{ij}=||\mathbf{x}_i-\mathbf{x}_j||$ is the Euclidean distance between the atomic positions $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{S}$ is symmetric ($\mathbf{S}=\mathbf{S}^T$), but not necessarily positive definite, it has dimensions $N\times N$ for an N atomic molecule. In practice, the ensemble average Eq. 1 may be estimated via a time expectation value from a molecular dynamics simulation. Of course, the reliability of the estimate and thus the result of our method will depend on the length of the simulation: Only if all relevant conformations of the molecule have been visited with a probability according to the Boltzmann distribution, will Eq. 1 converge. The distance-deviation can be computed for all solute atom pairs, or for a reduced set of representative atoms, such as α-carbon atoms in order to reduce memory consumption when analyzing large macromolecules. We note that it is possible to use the matrix of squared distance deviations instead of using distance deviations. Alternatively to using simulations, Eq. 1 can be computed from realizations of an NMR ensemble or several x-ray structures of the same molecule. The mean row value of $\mathbf{S}$ is a measure for the flexibility of individual atoms.

Cluster membership probability

Most methods in the literature [19,27,28] assume that each atom is uniquely assigned to one domain. This results in a so called integer optimization problem, which is very hard to solve [35]. Reference [36] has suggested using a fuzzy membership, where formally each atom $i\in\{1,\ldots,N\}$ may participate in different domains $m\in\{1,\ldots,M\}$ with a certain membership probability $X_{mi}\in[0,1]$. $X_{mi}=0$ means that the motion of the atom is independent of the motion of the domain, and $X_{mi}=1$ means they are perfectly synchronized. A natural normalization condition for $\mathbf{X}\in\mathbb{R}^{M\times N}$ is that the total membership probability sums up to one,

$$\sum_{m=1}^{M}X_{mi}=1 \quad \forall\ i\in 1,...,N \tag{2}$$

As a direct consequence we can write the probability $P_{m_{ij}}$ of finding the atoms i and j within the same domain m as

$$P_{m_{ij}}=X_{mi}X_{mj} \tag{3}$$

As it will turn out, the optimal grouping into domains is always unique in practice ($X_{mj}\in\{0,1\}\ \forall\ m,i$). Nevertheless, the introduction of the fuzzy memberships is essential as it allows the clustering problem to be formulated as continuous quadratic optimization problem, which, in contrast to integer optimization problems can be solved efficiently for very large systems.

Optimization problem for identifying semi-rigid domains

We define the optimal partition of the molecule into domains as the one that minimized deviations within the domains:

$$\text{minimize} \quad q(\mathbf{X}) = \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N} X_{mi}X_{mj}S_{ij} = trace(\mathbf{XSX}^T) \qquad (4)$$

This objective function measures the error describing the amount of distance deviations neglected by confining the motion of the atoms within their domains. Since a partitioning using M domains can always realize a $M-1$ domain partitioning as a special case, increasing the number of domains relaxes the optimization problem and the optimal error is thus monotonically decreasing (see also section "MR121-GSGSW peptide"), for $M=N$, the solution $\mathbf{X}=\mathbf{I}$ and $q(x)=0$ is obtained.

In contrast to heuristic coarse-graining methods the minimization problem in Eq. 4 leads to an optimal partitioning of the molecule according to the number of domains chosen. Furthermore the partitioning has no bias towards equally-sized domains, i.e. it allows for domains of very different sizes if this is requested by the structure of $\mathbf{S}$.

The minimization problem in Eq. 4 together with the normalisation condition in Eq. 2 can be written into a standard quadratic optimization problem with linear constraints that is solved here in order to identify the optimal partitioning into domains.

$$\text{minimize } q(\mathbf{x}) = \mathbf{x}^T\mathbf{H}\ \mathbf{x} \qquad (5)$$

$$\text{such that}$$

$$\mathbf{Ax}=\mathbf{b}$$

$$0 \leq \mathbf{x}$$

Here $\mathbf{H} \in \mathbb{R}^{MN \times MN}$ is the symmetric Hessian matrix,

$$\mathbf{H} = \begin{bmatrix} \mathbf{S} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{S} \end{bmatrix} \qquad (6)$$

and $\mathbf{x} \in \mathbb{R}^{MN}$ is a column vector containing the membership probabilities

$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)^T = (X_{1,1}, \ldots, X_{1,N}, X_{2,1}, \ldots, X_{M,N})^T, \qquad (7)$$

with X_i being the i-th row of $\mathbf{X}$. The constraint matrix $\mathbf{A} \in \mathbb{R}^{MN \times MN}$ and the column vector $\mathbf{b} \in \mathbb{R}^{MN}$ represent the equality constraints on $\mathbf{x}$. According to Eq. 2

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{I}_2 & \cdots & \mathbf{I}_{M-1} & \mathbf{I}_M \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \qquad (8)$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, and

$$\mathbf{b}_k = \begin{cases} 1 & \text{for} \quad k \leq N \\ 0 & \text{for} \quad k > N \end{cases}. \qquad (9)$$

Because the Hessian matrix $\mathbf{H}$ is just a composition of submatrix $\mathbf{S}$, one may reduce problem size by introducing the subvectors, $\mathbf{x}_m = (X_{m,1}, \ldots, X_{m,N})^T$ for each domain and reassemble $\mathbf{Hx}$ from the products $\mathbf{Sx}_m$. The numerical implementation is described in section "Numerical implementation".

Numerical implementation

The present quadratic optimization problem is solved using an active set method similar to that of Gill et al., described in [37]. The solution procedure involves two phases: the first phase involves the calculation of a feasible point $\mathbf{x}$ (if one exists), the second phase involves the generation of an iterative sequence of feasible points that converge to the solution.

Besides the sparse definition of $\mathbf{A}$ and $\mathbf{b}$ one may reduce the size of the problem from one $MN \times MN$-dimensional problem to M $N \times N$-dimensional problems. Because $\mathbf{H}$ is block diagonal, one may compute $\mathbf{Hx}$ as the piecewise product $\mathbf{Sx}_m$ and reconstruct the vector

$$\mathbf{Hx} = (\mathbf{Sx}_1, \mathbf{Sx}_2, \ldots, \mathbf{Sx}_{M-1}, \mathbf{Sx}_M)^T \qquad (10)$$

in a subsequent computation. This modification reduces the memory consumption significantly (by a factor of M), because instead of $\mathbf{H} \in \mathbb{R}^{MN \times MN}$ only $\mathbf{S} \in \mathbb{R}^{N \times N}$ has to be held in the memory. The involved increase of computation time is insignificant. With this modification the problem size solvable on desktop computers is up to 65,000 particles. We note that in large molecular systems these particles may be chosen to be backbone or α-carbon atoms, so that the number of atoms of the molecule can be much larger.

Initial condition and "successive restart"

Even though the method is robust for low M (see section "MR121-GSGSW peptide") it was found that for larger M the solution depends on the initial condition (IC) that is provided to the solver. A permutation of the domains only modifies the labels and not the grouping, thus there exist at least 2^M equivalent solutions. Unfortunately, there are also multiple non equivalent local minima where the solver may get trapped. In order to avoid being trapped in a bad local minimum it is advisable to choose a good initial condition $\mathbf{x}_{IC}$.

One approach to escape from local minima is applying stochastic methods such as Monte Carlo sampling, simulated annealing or genetic algorithms. Another simple approach that has shown to work well in practice is to use the solution obtained for $(M-1)$ domains to construct $\mathbf{X}$ for M domains. This heuristic approach may done by identifying the cluster membership subvector, $\mathbf{x}_{max}$, that has the maximum average contribution to $q(\mathbf{x})$ per member. Formally this is expressed by

$$\mathbf{x}_{max} = argmax_m \frac{\mathbf{x}_\mathbf{m}^\mathbf{T}\mathbf{S}\mathbf{x}_\mathbf{m}}{\sum_{i=1}^{N} X_{mi}}. \qquad (11)$$

The memberships within the subvector $\mathbf{x}_{max}$ are distributed over two domains by substituting $\mathbf{x}_{max}$ by $\mathbf{x}^*$, with elements $x_i^* = c_i\ x_{max_i}$ and appending $\mathbf{x}^{**}$ with elements $x_i^{**} = (1-c_i)\ x_{max_i}$ as $\mathbf{x}_M$. Here

$c_i \in [0,1]$ is either deterministically or randomly chosen for every atom in $\mathbf{x}_{max}$. To assure that the number of atoms is conserved the sum of the membership probabilities over $\mathbf{x}^*$ and $\mathbf{x}^{**}$ is one for all member atoms.

$$\mathbf{x}_{IC} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}^*, \ldots, \mathbf{x}_{M-1}, \mathbf{x}^{**}) \qquad (12)$$

This procedure assures that the clustering error is monotonically decreasing with increasing number of domains (see section "MR121-GSGSW peptide").

Clustering quality and number of domains

The error can be used as tool to choose the number of domains M, either by prescribing a desired q^* and asking for the smallest number of domains with $q(x_{opt}) \leq q^*$, or by looking for gaps in the error series and choosing M such that $q_M(x_{opt}) \ll q_{M-1}(x_{opt})$. In some applications it may be desirable to have the number of domains selected automatically rather than by the user. One possible method is to select the number of domains such that the clustering error stays below a user-imposed bound, q_{tol}. For this, it is useful to define the normalized clustering error as:

$$\overline{q}(x) = \frac{1}{N}\sum_i \overline{q_i}(x) = \frac{1}{N}\sum_i \frac{\sum_m \sum_j P_{m_{ij}} S_{ij}}{\sum_m \sum_j P_{m_{ij}}}. \qquad (13)$$

Here, $\overline{q_i}(x)$ is the normalized error for atom i. Eq. 13 is not what is optimized here, but a measure for the mean distance deviations of pairs within domains for a given number of clusters. It therefore has a direct physical interpretation and measures the quality of a clustering of the molecule into semi-rigid domains. Alternatively, Eq. 13 can be modified to use the matrix of squared distance deviations, leading to the RMSD as error measure. $\overline{q}$ is identically zero when the molecule consists of M^* perfectly rigid domains and $M > M^*$ is used. $\overline{q}$ is also useful in order to make an automated choice of M: It can be set to a value the user considers as small enough, such as $0.05\,nm$. Based on this rationale, the optimal clustering is chosen by the following algorithm:

1. Compute distance-deviation matrix, $\mathbf{S}$
2. Set $M = 2$
3. Compute optimal clustering X_M of based on X_{M-1}.
4. If $\overline{q}(X_M) < q_{tol}$ return X_M
4. Else $M := M+1$, Go to 3.

Computational performance

The computational performance of the method was demonstrated for the examples discussed in the "Results" section. From Table 1 is seen that the method is very efficient even for a large number of particles/domains.

Molecular models and simulation setup

To demonstrate the performance and usefulness of the method we have applied it to a series of molecular systems:

1. A $1\,\mu s$ MD trajectory of Ala$_5$, containing 36 solute atoms.
2. A $2\,\mu s$ MD trajectory of the artificial peptide MR121-GSGSW [38] (i.e. a chromophore MR121 is connected with GLY-SER-GLY-SER-TRP), containing 81 solute atoms.
3. 500 ns MD trajectories of the wild type of transthyretin (PDB ID code, 1DVQ) [39], containing 2,257 solute atoms, and two point variants 58Arg, 58His, containing 2,265 and 2,260 solute atoms respectively. The point mutants were generated by Modeller Release 9v5 [40].
4. A $2\,ns$ MD trajectory of the chaperone GroEL-GroES (PDB ID code, 1GRU), containing 72,716 solute atoms.

Table 1. Computation time for selected molecular systems.

System	no. Atoms	time in seconds		
		M = 2	M = 5	M = 81
MR121-GSGSW	81	0.015	0.12	14.46
Transthyretin	229 C_α	0.048	0.94	94.72
Transthyretin	2257	1.32	92.34	$4.45 \cdot 10^3$
GroEL-GroES	8015 C_α	181.77	1537.01	$>4.2 \cdot 10^4$

Computation time for selected molecular systems with $M = \{2, \ldots, 81\}$ domains. Computations were done on a usual desktop computer with CPU@2.5 GHz and 6.5 GB Ram, time is given in seconds.

doi:10.1371/journal.pone.0010491.t001

All molecular dynamics trajectories were generated by the molecular dynamics package Gromacs 3.3 [41] using the standard distribution force field GROMOS96 43a2. The solutes were solvated in SPC216 water in a cubic box with at least $1\,nm$ of water on each side of the solute. The structures were equilibrated with a $10\,ps$ molecular dynamics simulation constraints on all bonds of the protein. A subsequent energy minimisation without position restraints was performed with a steepest descent minimization. The production runs were done with LINCS constraints [42] on the hydrogen bond length and a $2\,fs$ time step, the trajectory was written every $2\,ps$. The electrostatic interactions were computed using the smooth Particle Mesh Ewald algorithm (PME), where the full direct and reciprocal space parts were calculated each step with a lattice spacing of $0.12\,nm$. The Van der Waals interactions were computed with a cut-off at $1\,nm$. All simulations were performed with Berendsen temperature coupling and isotropic pressure coupling to $1\,atm$. The temperatures used were $293\,K$ for systems $1+2$ and $300\,K$ for systems $3+4$.

Results

We illustrate our approach on a number of test systems. In all cases the distance deviation matrix $\mathbf{S}$ was computed from the data and the optimization problem in Eq. 5 was solved for a series of consecutive domains numbers M using the successive restart approach described in the "Materials and Methods" section.

Application to Small Model Systems

Numerical example. Interestingly, although the method formally allows for fuzzy memberships, the optimal assignment of atoms to domains is always unique in practice, thus obtaining an exact partitioning of atoms into domains. For example, consider a hypothetical 3-atom system with the distance-deviation matrix

$$\mathbf{S} = \begin{bmatrix} 0 & 0.7 & 0 \\ 0.7 & 0 & 0.3 \\ 0 & 0.3 & 0 \end{bmatrix}, \qquad (14)$$

which has the optimal solution for M = 2 subunits:

$$\mathbf{X}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 1\end{bmatrix} \text{ or by permutation } \mathbf{X}=\begin{bmatrix}0 & 1 & 1\\ 1 & 0 & 0\end{bmatrix}. \qquad (15)$$

Here atom 1 is placed in the first and atoms 2 and 3 are placed in the second domain. As described in Section "Cluster membership probability" the elements of the membership matrix either converge to one or to zero, i.e. the atoms tilt over to the domain that produces the smallest clustering error when including this atom. In other words, the clustering error is minimal when each atom is fully assigned to the subunit it belongs to most.

Now consider the case,

$$\mathbf{S}=\begin{bmatrix}0 & 0.5 & 0\\ 0.5 & 0 & 0.5\\ 0 & 0.5 & 0\end{bmatrix}, \qquad (16)$$

that has the solution for $M=2$

$$\mathbf{X}=\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 0\end{bmatrix} \text{ or by permutation } \mathbf{X}=\begin{bmatrix}0 & 1 & 0\\ 1 & 0 & 1\end{bmatrix}, \qquad (17)$$

with total error $q=0.025$. In contrast the fuzzy solution

$$\mathbf{X}=\begin{bmatrix}1 & 0.5 & 0\\ 0 & 0.5 & 1\end{bmatrix} \text{ or by permutation } \mathbf{X}=\begin{bmatrix}0 & 0.5 & 1\\ 1 & 0.5 & 0\end{bmatrix} \qquad (18)$$

has a higher total error of $q=0.028$. Finally, consider the pathological case of an off-diagonally uniform distance matrix which represents entirely uncorrelated motion

$$\mathbf{S}=\begin{bmatrix}0 & 0.5 & 0.5\\ 0.5 & 0 & 0.5\\ 0.5 & 0.5 & 0\end{bmatrix}, \qquad (19)$$

which may be found for gas particles. In this case the solution for $M=2$ is degenerated:

$$\mathbf{X}=\begin{bmatrix}1 & a & 0\\ 0 & 1-a & 1\end{bmatrix} \qquad (20)$$

In this case the total error of $q=1$ was found for all $a\in[0,1]$. According to the uniformity of $\mathbf{S}$ this degenerate case is never found in practice for macromolecules. Even nearly unstructured proteins will have some structure in $\mathbf{S}$ because of their bonding topology and minor deviations in $\mathbf{S}$ from the uniform case will cause the atoms to be uniquely assigned to one domain such that the error is minimum.

We conclude that no fuzzy memberships are found in macromolecules. Note, however, that the introduction of a fuzzy membership was still essential, because using this formulation we could express the optimization problem as a continuous quadratic optimization problem. The solution to this kind of problem is much easier than the solution to the integer optimization problem emerging by the priori assumption that the memberships must be integer values.

Polyalanine. As a first example the optimization method was applied to Ala$_5$ in order to demonstrate that the method can identify meaningful domains. Some of the resulting coarse-grain structures and the clustering error for $M=\{2,...,N\}$ are shown in Figure 1. The sub-structures are approximately equally sized and represent the optimal partitioning for a given number of domains. As the number of domains is increased the size of the domains diminishes. For $M=6$, the method successfully identifies the domains that are nearly rigid due to bond angle, angle, improper dihedral and ω-angle constraints: There are 4 domains containing the 4 peptide planes including the first but excluding the second C_α plus the CH_3 side chain. The remaining two domains contain the N-terminal and the C-terminal (see Figure 1). The small remaining clustering error reflects the vibrations still allowed within the domains, mainly due to flexibility in the improper and ω-dihedrals. For $M=19$ the method clusters the system into domains containing one backbone atom each along with the one side chain atom connected to it. Finally the method is shown to be consistent in the limit, because for $M=N$ every atom is placed into a single domain ($\mathbf{X}=\mathbf{I}$), and the error is zero $\bar{q}_{M=N}=0\ nm$ (not shown as structure).

MR121-GSGSW peptide. In order to study a more complex system, the method was applied to the MR121-GSGSW peptide. Figure 2 shows a series of molecular coarse-grain structures for selected numbers of domains $M=\{2,3,4,6,7\}$. This series shows clearly how the flexible parts of the molecule subdivide into finer domains. $M=3$ separates the GSGS chain and the chromophores, $M=4$ splits also the GS domains. Using more domains accounts for smaller decrease of the error until $M=6$ the system is split into individual residues.

The corresponding distance-deviation matrix is shown in Figure 3 (top). It is structured into blocks along the diagonal (values are close to zero), that represent the almost rigid regions of the molecule. The values on the diagonal are zero ($S_{ij}=0\ \forall\ i=j$) (blue), while values far from the diagonal are large (red). To identify rigid domains within the peptide we have employed the quadratic optimization method for $M=6$. The convergence of the method depends on the size of the molecular system, the number of domains chosen and the initial conditions. For six domains in

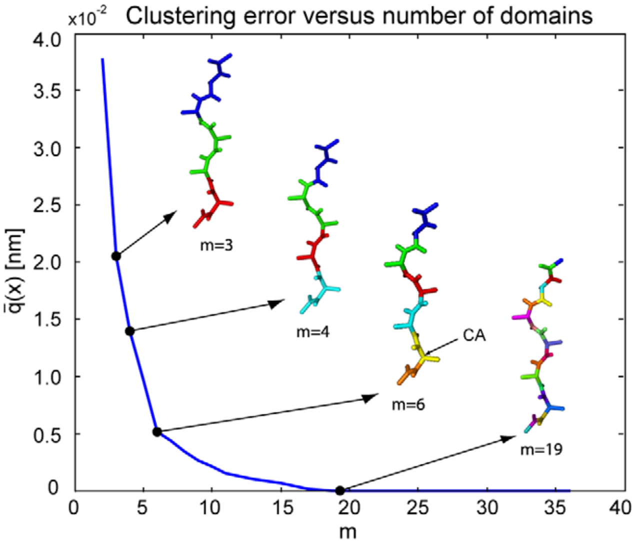

Figure 1. Clustering error of Ala$_5$ for $m\in\{2,\ldots,36\}$ and corresponding coarse-grain structures for $m=\{3,4,6,7,19\}$ domains. The decrement of the clustering error is very steep for $m\leq 5$ and relatively flat afterwards, suggesting that $m=6$ is a good choice for the number of domains ($q_{tol}\approx 0.005\ nm$). The molecule is partitioned into its four peptide planes and two end groups containing the C- and N-terminus respectively.
doi:10.1371/journal.pone.0010491.g001

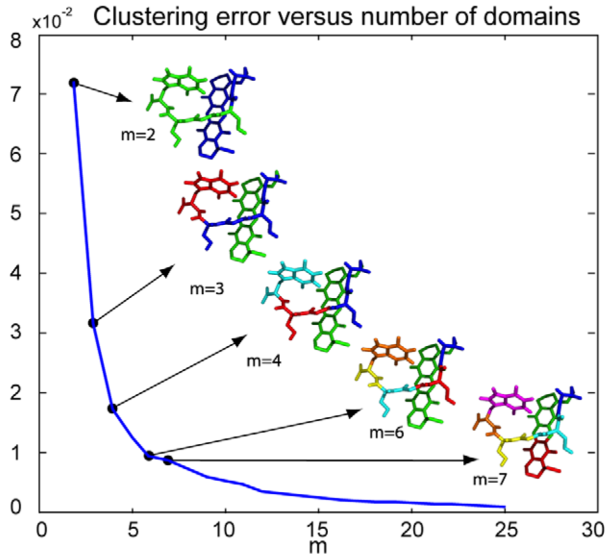

Figure 2. Clustering error of MR121-GSGSW for $\mathrm{m}\in\{2,\ldots,25\}$ and corresponding coarse-grain structures for $\mathrm{m}=\{2,3,4,6,7\}$ domains. The decrement of the clustering error is very steep for $m\leq 5$ and relatively flat afterwards. For $m=6$ the number of domains is well balanced with the expected error ($q_{tol}\approx 0.01\ nm$).
doi:10.1371/journal.pone.0010491.g002

the artificial peptide (81 atoms) it converged within ~ 100 iterations, and a few *ms* on a standard desktop computer.

The resulting membership matrix, **X**, and the corresponding coarse-grain structure are shown in Figure 3 (bottom). The colors show the assignment of atoms to domains. The elements of the membership matrix either converge to one or to zero, i.e. the atoms tilt over to the domain that produces the smallest clustering error when including this atom.

In order to test the optimality of the results, we have repeatedly solved the clustering problem for the MR121-GSGSW peptide using different initial conditions: (i) for given M, each atom is assigned a random membership to each domain, $X_{mi}\sim \mathrm{uniform}\,[0,1]$ and then normalized so that $\sum_m X_{mi}=1$; (ii) only $M=2$ is using a random initial condition while the solutions for $M>2$ are found by successive restart from the previous solution with the atoms of the largest-error domain split into the two new domains by an initial assignment of $X_{mi}=0.5$. Figure 4 shows a comparison for the clustering error for both cases, with ten realizations for the random initial condition plotted. The results are identical independent of the initial condition for small M, which suggests that these solutions are likely to be globally optimal. For large M the solution based on random initial condition gets trapped in different but only slightly suboptimal local minima, while the successive restart solution is monotonically improving for increasing M. In all cases studied, the heuristic successive restart scheme possesses a useful monotonicity property, and performs better than optimization of random guesses.

Application to Biological Complexes

Transthyretin. The transport protein transthyretin (TTR) is primarily synthesized in liver, choroid plexus, and the retina. The primary function is the transport of thyroxine and retinol binding protein (RBP). Both molecules can bind to the homo-tetrameric structure of TTR, which is found at a physiological pH of $7-7.4$. In contrast the 28 kDa dimer structure is observed at pH >7 and titration of 2% sodium dextransulfate (SDS). It has two identical 127-amino-acid monomers (A - blue) and (B - green) (see Figure 5) with an extensive β-sheet structure that form β-sandwiches [43]. The interactions between the two monomers involve electrostatic and hydrophobic forces.

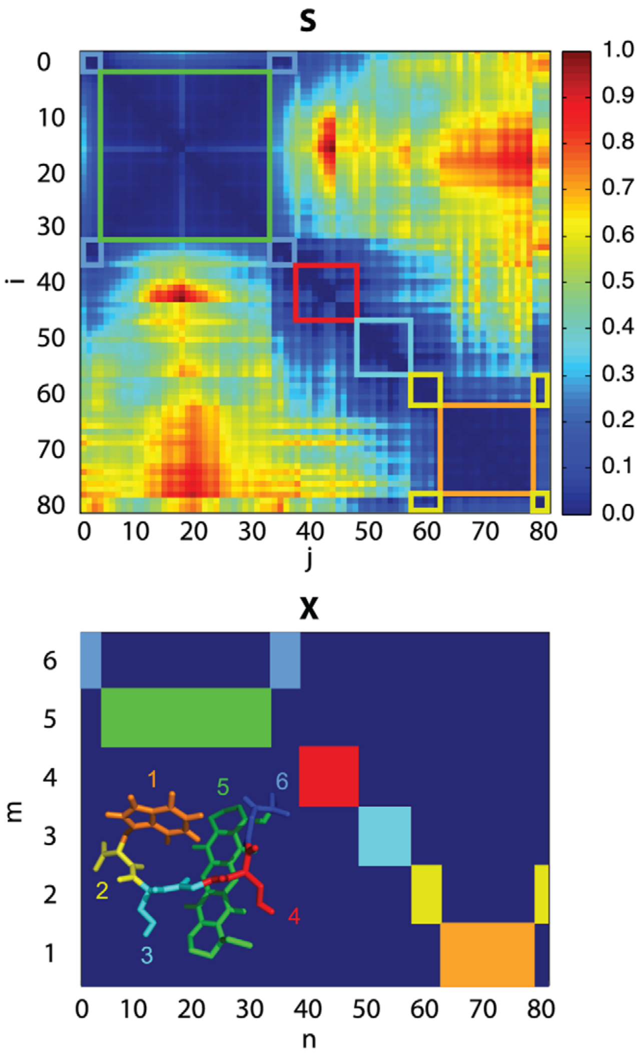

Figure 3. Distance deviation matrix S for MR121-GSGSW (top) and membership matrix, X, for m = 6 clusters (bottom). The colors relate the semi-rigid regions in the distance deviation matrix to the molecular coarse-grain structure and the membership matrix.
doi:10.1371/journal.pone.0010491.g003

Transthyretin is one of the human proteins known to be associated with local amyloidosis. Amyloid fibrils are the polymerized form of the protein, their internal structure mainly consists of cross β-sheets, arranged perpendicular to the long axis of the fibrils [44]. Both point variants of TTR and the native protein are known to deposit as amyloid fibrils in the extra-cellular region, where they cause neurodegeneration and organ failure (for reviews on amyloidosis see [45,46]). Transthyretin is known to be associated with the amyloid diseases senile systemic amyloidosis (SSA), familial amyloid polyneuropathy (FAP), and familial amyloid cardiomyopathy (FAC). Other known amyloidogenic diseases are for e.g. Alzheimer's disease, type 2 diabetes and the transmissible spongiform encephalopathies which are characterized by proteinaceous deposits in the affected relevant organs.

Transthyretin aggregation to amyloid fibers has been the subject of many studies [43–48], however the molecular mechanisms are

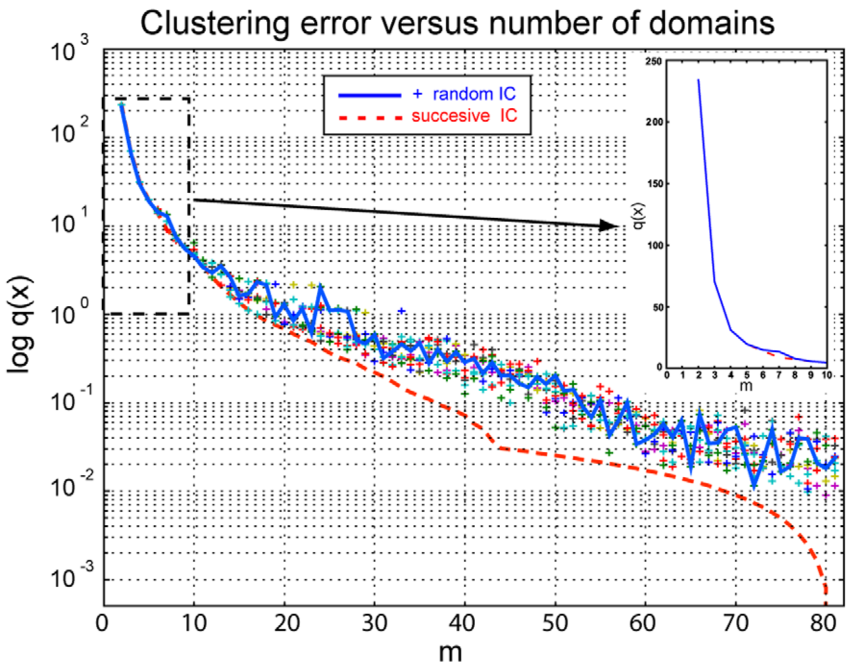

Figure 4. Dependence of clustering error on the choice of the initial condition. When using a random assignment to clusters for the first step $M=2$ followed by successive restart (dashed red line) the error is monotonically decreasing. Choosing random initial conditions for all M (one realization highlighted as blue solid line, 9 more realizations indicated by "+"), the optimization gets trapped in slightly different local minima for large M. For small M the method robustly identifies the same minimum independent of the initial condition, indicating that global optimality is achieved in this case.

doi:10.1371/journal.pone.0010491.g004

still not completely understood. Structural modifications and their effect on conformational stability were studied by structural and computational analyses [49] and experimentally by urea and temperature induced unfolding [50,51]. It is proposed that amyloidogenicity of TTR is associated with anomalous structures that favour oligomer and fibril formation. The structures are assumed to be the product of complex dissociation via destabilisation [52] and subsequent unfolding and folding of the protein [53]. It could be verified that prior fibril formation the homo-tetramer dissociates into two dimers [54]. Whether the dimers need to dissociate into monomers before fibrillation can occur is still unclear. However it is assumed that dimer dissociation is the result of a mechanism called "edge exposure", where the displacement of residues $115-123$ (inner β-strand) and residues $22-41$ (outer β-strand) flattens the dimer structure [55,56].

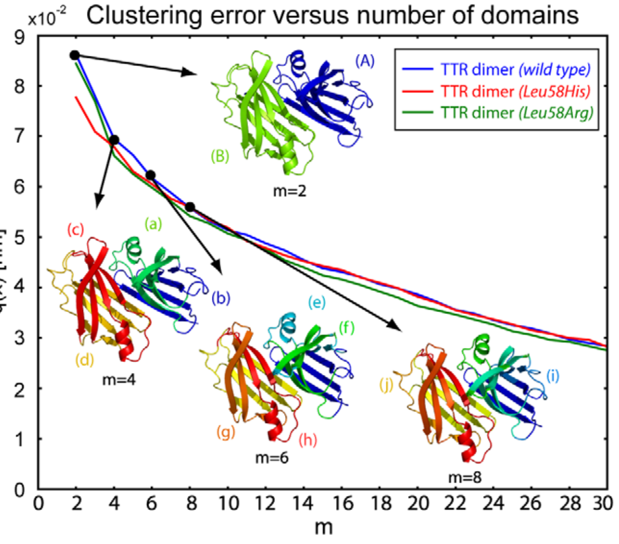

Figure 5. Clustering error of native Transthyretin for m$\in\{2,\ldots,30\}$ and corresponding coarse-grain structures for m$=\{2,4,6,8\}$ domains. The obtained coarse-grain structures separate the dimer into two monomers (A) and (B) for $M=2$ and identify the β-sandwich structure (a)+(b) and (c)+(d) in the two monomers for $M=4$. For $M=6$ the method additionally identifies the α-helical structure (e) and (h), for $M=8$ two flexible loops (i) and (j) are found.

doi:10.1371/journal.pone.0010491.g005

To date, a large number of TTR variants have been associated with amyloid formation [57]. Here we study the structural rigidity of the wild-type and two variants commonly found, where the leucine of residue 58 in the dimers is replaced by arginine or histidine (TTR-58Arg and TTR-58His) and investigate their possible role in destabilisation and dissociation of the dimer structure. Both variants are known to be amyloidogenic, however the phenotypic difference of FAP between the 58His and 58Arg mutations suggest differences in the secretion efficiency or aggregation characteristics of the TTR variants [58].

In Figure 5 we show the clustering error for increasing number of domains and the coarse-grain structures for $M=\{2,4,6,8\}$ domains. As expected for $M=2$ the atoms from each monomer are placed in separate but symmetric domains. This separation is maintained for larger values of M. For $M=4,6,8,\ldots$ the domains found in the two monomers are nearly, but not perfectly symmetrical, as a result of limited statistical accuracy of the molecular dynamics trajectory. For $M=4$ the algorithm identifies two β-sheets (b) and (d) in the dimer and two structures (a) and (c) including β-sheets and the α-helices. At $M=6$ the structures (a) and (c) are split into one block containing two β-strands (f) and (g) and one block containing two β-strands and the α-helix (e) and (h).

For $M = 8$ the two loops in the outer region containing two short β-strands are found to be separate domains (i) and (j).

To demonstrate the applicability to experimental data we used the method to partition molecular structures of transthyretin obtained by x-ray crystallography. Besides the wild type structure (PDB code 1DVQ), which was also used in the molecular dynamics simulation, five related structures of transthyretin complexed with resveratrol, diclofenac, flurbiprofen, DDBF, oFLU, and PHENOX (PDB codes 1DVS, 1DVT, 1DVU, 1DVX, 1DVY, 1DVZ) [39] have been used to generate the distance deviation matrix. Due to in sequence mutations in the structures 1DVX and 1DVZ we cleaned up the structure files to leave only comparable α-carbon atoms in all six pdb-files. In the 1DVX file we removed residue 9+127 BLEU, 110+228 BSER and 113+231 BTHR, while in 1DVZ 7+124 BLYS and 10+127 BLEU where removed. We note that these six crystallographic x-ray structures correspond to different chemical or crystallographic states, so that the structural differences between them are not expected to be idendical to the structural differences within the Boltzmann-weighted ensemble of a solvated TTR in a single chemical state. Nevertheless, it is expected that the differences in the crystallographic realizations are sensitive to the molecule's instrinsic flexibility, so that a comparison between the simulation-based and X-ray-based results is interesting. The clustering result obtained from the x-ray structures and clustering error with coarse-grain structures for $M = 2{,}4{,}6{,}8$ are shown in Figure 6. The clustering of x-ray structures yield similar coarse-graining as obtained by the clustering of molecular dynamics data. The dimer is separated into two monomers (A) and (B) for $M = 2$, while for $M = 6$ the β-sandwich (b) + (c) + (g) + (f) and α-helical (e) + (h) structures within the dimers are identified (compare top right structure generated from MD data). However, for increasing number of domains ($M = 8$) the clusterings are different. Note that the clustering error found for the X-ray structures is much smaller, indicating that the crystallographic realizations are much more similar to each other than the structures accessible to the dynamics of TTR in a solvent simulation, likely owing to crystal lattice constraints.

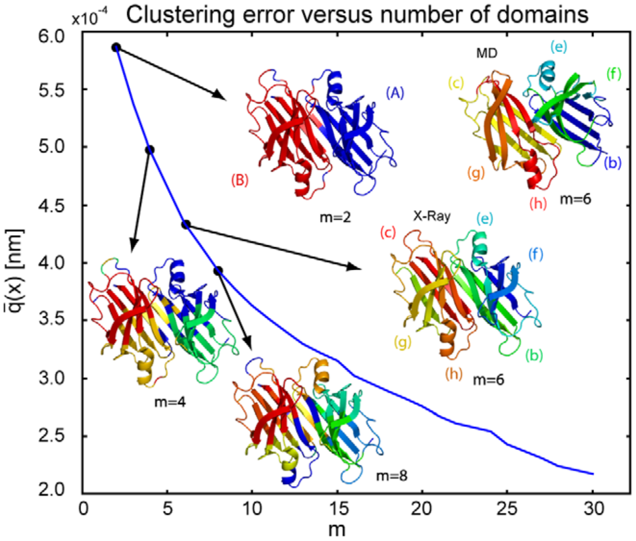

Figure 6. Clustering error obtained from six x-ray structures of Transthyretin for $m \in \{2, \ldots, 30\}$ and corresponding coarse-grain structures for $m = \{2,4,6,8\}$ domains. As for molecular dynamics data the obtained coarse-grain structures separate the dimer into two monomers (A) and (B) for $M = 2$ and identify the β-sandwich (b) + (c) + (f) + (g) and α-helical (e) + (h) structures in both monomers for $M = 6$.
doi:10.1371/journal.pone.0010491.g006

The distance-deviation matrices in Figure 7 show two large blocks along the diagonal (blue) that indicate that the internal rigidity within each of the two associated monomers is much larger than the rigidity between the monomers. The off-diagonal regions in the matrices (yellow-green-red) represent the inter-monomeric rigidity and are related to the stability or binding strength between the monomers. Large values in the matrix (red) indicate low stability, while small values (blue) are related to high stability of the dimer.

The structural modification induced by the amyloidogenic variants (TTR-58Arg and TTR-58His) contribute to an de/increase in rigidity in some regions of the structure (see increasing red regions in the variants compared to the native TTR in Figure 7), which lead to local de/stabilisation of the dimer. The overall stability of the dimer is directly related to the difference $\Delta q_{12} = q_{M=1} - q_{M=2}$. Here, the binding strength $\Delta q_{12_{Arg}} = 4.0 \cdot 10^4$ for TTR-58Arg is increased and $\Delta q_{12_{His}} = 5.84 \cdot 10^4$ for TTR-58His is decreased compared to the binding strength of the wild-type $\Delta q_{12_{wt}} = 5.12 \cdot 10^4$. The decreased stability for TTR-58His variant compared to the wild type protein is supported by urea and thermal induced unfolding experiments [51] and computational studies that are based on an energy functions derived from non-redundant x-ray structures [49].

However in addition to the overall stability, increased atomic motion of specific regions in the dimer may influence the stability of the dimer and favor transient dissociation. The local flexibility/rigidity of atoms is reflected by $\overline{q_i}(x)$, i.e. the mean row value of $\mathbf{S}$. The method is thus able to determine the relevant substructures that may cause destabilisation by taking the row average of the distance deviation matrix (see Figure 7). The peaks indicate residues that have increased distance deviation with respect to all other residues, i.e. the most flexible regions in the dimer. In Figure 7 (right) the structures are color coded according to the row average of $\mathbf{S}$, the mutated residue 58 is colored purple. In agreement with [51] the results indicate that compared to the wild-type protein the 58His and 58Arg variants are mainly destabilized at the monomer-monomer interface. In comparison to the wild-type TTR (Figure 7 top), it is clearly seen that the 58His variant increases the total distance deviation between the monomers (see average value of the row mean) and the peak values at residues 11,78,105,125,193,220. In contrast the 58Arg variant of transthyretin decreases the mean distance deviation, while the peak values at residues 11,78,105,126,192,220 are still increased compared to the wild-type TTR. Because both variants are known to be amylogenic [58], we conclude that destabilisation is not only determined by the overall stability, but also by specific regions that cause local destabilisation of the protein that may lead to transient dissociation into monomers. The method is thus able to provide information about regions of TTR that are destabilized in disease causing variants.

GroEL-GroES chaperone complex. The existence of semi-rigid domains and their relative dynamics are essential for the functionality of large macromolecular machines. Here, we analyse the dynamics of the GroEL-GroES chaperone complex (see Figure 8), which contains 8,015 residues (72,716 atoms). The complex ensures the proper folding of many proteins [59] and avoids non-native protein aggregation. GroEL is a tetradecameric protein of 14 identical domains arranged in a *cis* and *trans* heptameric-ring. GroES is dome shaped in either un-/bound configuration and contains seven identical domains assembled as a heptamer ring.

Computational studies have provided important insights into the allosteric mechanism of the chaperonin GroEL-GroES. Protein folding within the complex involves binding, encapsula-

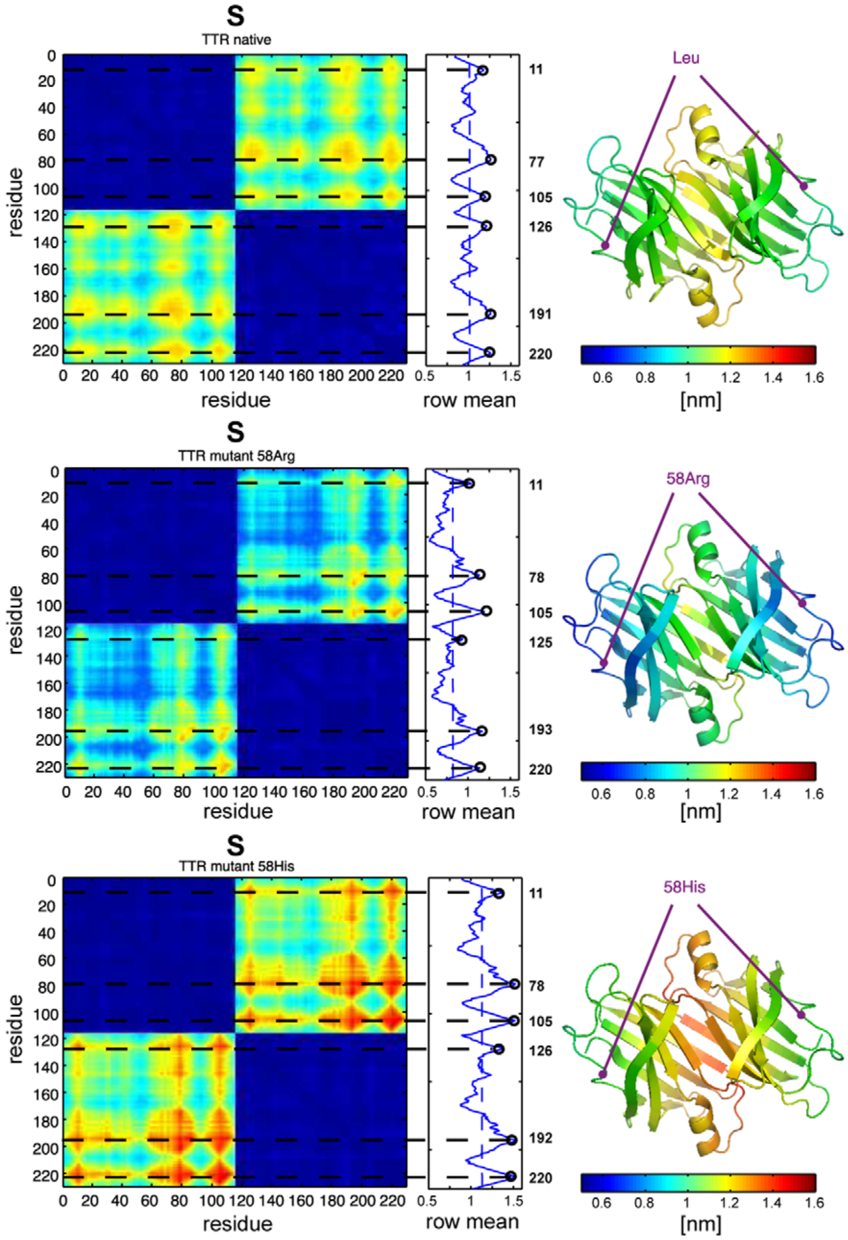

Figure 7. Distance deviation matrix for native TTR (top) compared to variants 58Arg (middle) and 58His (bottom). The mean row value of each matrix indicates flexible regions around reference residues 11, 77, 105, 191 and 220. The corresponding structures are color coded according to the average row value of S and show the location of residue 58 (purple). Large values (red) indicate flexible regions, while small values (blue) indicate rigid regions in the dimer. The data suggests that the dimer interface is destabilized for both amylogenic TTR variants of the protein.
doi:10.1371/journal.pone.0010491.g007

tion, and release of the substrate protein [60,61]. During the GroEL-GroES cycle the GroEL binds a mis-/unfolded protein at its apical (A) domain (see Figure 9). The binding is caused by electrostatic and hydrophobic interactions between the exposed hydrophobic residues of the substrate protein and those of the apical domain. The equatorial domain (E) plays the major role in the overall chaperonin activity. It binds and hydrolyzes ATP. The intermediate domain (I) serves as a functional bridge between the apical and equatorial domains. After ATP binding to every *cis*-subunit, GroEL is bound to the cofactor GroES. During the GroES binding large conformational changes at the apical domain of the *cis*-ring cause upwards and outwards movement of the apical GroEL domains, thereby increasing the size of the central cavity and forming a dome-shaped chamber [59,61]. By this conformational change the substrate protein is captured inside the cavity, where it will be able to undergo conformational changes toward the folded state. During ATPs hydrolysis in the *cis*-ring, ATP molecules are transferred to the *trans*-ring, which drives the release of the GroES cap and the substrate protein.

Due to the size of the system only a short molecular dynamics trajectory with duration of 2 *ns* was produced, which is certainly not converged, but can nevertheless be used for a performance test (see Section on "Computational Performance"). In Figure 8 we show the clustering error for $m \in \{2,\ldots,40\}$ and the most informative structures. We note that due to insufficient statistical information in the MD simulations (under-sampling) the optimi-

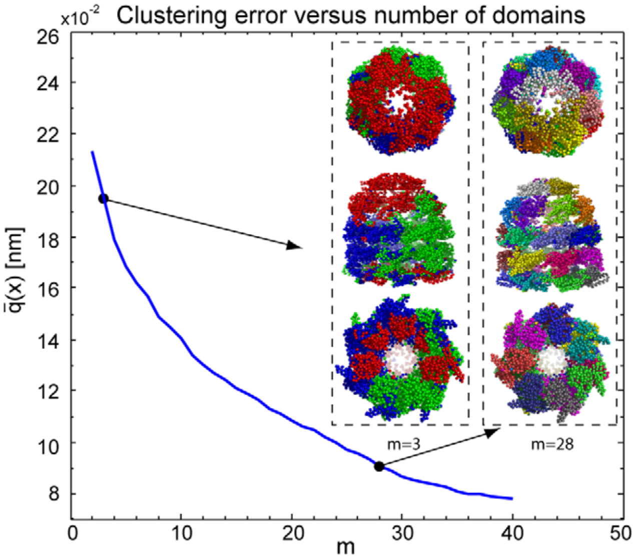

Figure 8. Clustering error for the GroEL-GroES chaperonin complex for m∈{2,...,40} and important structures. The α-carbon atoms are colored according to the coarse-graining. The poor statistics of the short MD simulation causes discontinuous domains (fragmentation) for small m. The method clearly detects the functional domains of the complex for $m=28$.
doi:10.1371/journal.pone.0010491.g008

zation results in disconnected domains (fragmentation) for small M (see Figure 8). Nevertheless, the clustering with $M=3$ reveals the ring structure of the GroEL into two halves and finds the GroES as a third domain revealing the essential elements necessary to represent the conformational change caused by the complex formation of GroEL with GroES. For a large number of domains, e.g. for $M=28$, the method clearly detects functional domains that can be directly related to the heptamer-ring structure of the chaperone. All shown results ($M=3$, $M=28$) are equally "correct", but reveal different levels of detail. The fragmentation of domains for small M may be reduced by using longer molecular dynamics trajectories. Thus, the method is applicable to large molecular complexes with modest requirements of computation time, but as it is data-based the results are sensitive to the quality of the data.

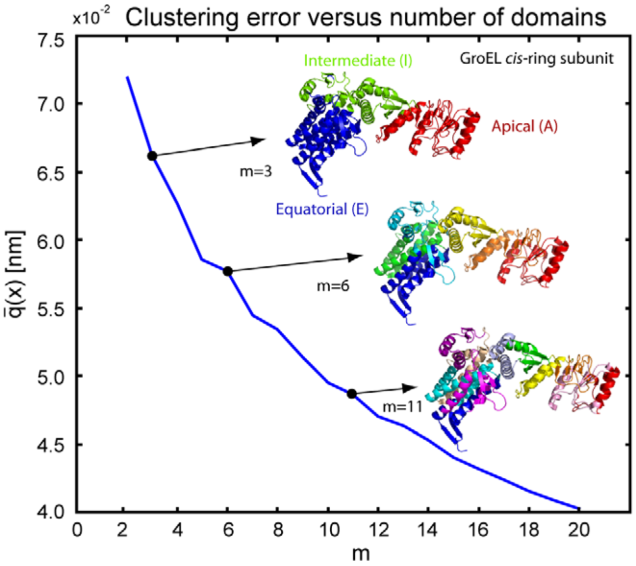

Figure 9. Clustering error for the heptameric subunit of GroEL for m∈{2,...,20}. The cartoon representation of three important structures ($M=3$,6 and 11) is colored according to the identified domains. For $M=3$ the three functional domains (apical, intermediate and equatorial) in the GroEL subunit are found.
doi:10.1371/journal.pone.0010491.g009

To study GroEL in more detail, we have further performed domain identification on the subunit of the *cis* and *trans* heptameric-rings of GroEL. The distance deviation matrix was generated by averaging over the data of seven identical subunits for the *cis* and *trans* ring respectively. This averaging enhances the statistics of the molecular dynamics data. The domains found in the optimization (see Figure 9) are in good agreement with the functional domains in the GroEL subunit [62]. For $M=3$ the major three domains (apical, intermediate and equatorial) are found in either *cis*- or *trans*-ring subunits. The distance deviation matrix for $M=3$ and the identified domain boundaries are shown in Figure 10. The average row value and the color coding for the three domains is shown on the right. These coarse-grain structures are in good agreement to those used in rigid clusters models [63] or Markov models [64]. For larger number of domains, for e.g. $M=6$, the method identifies two domains in the apical, intermediate and equatorial region respectively. For $M=11$ three domains in the apical and intermediate region respectively and four domains in the equatorial region were found.

Discussion

The coarse-graining algorithm developed in this paper is an optimal and systematic approach to decompose ensembles of molecular structures into semi-rigid domains. It consists of three steps: (i) obtaining an ensemble containing the atomic fluctuations, e.g. using molecular dynamics simulation, (ii) computation of the pair distance-deviation matrix and (iii) definition of semi-rigid domains by a quadratic optimization method, to distinguish and to quantify the rigid and flexible domains within the protein structure. The method identifies rigid regions that can vary in size and shape. The objective function minimized in the procedure is a direct measure of the clustering error and thus the within-cluster flexibility neglected by assigning the atoms into domains. We have been able to study the rigidity of proteins in systems involving 8,015 residues on a normal desktop computer.

In contrast to other methods the algorithm does not require the choice of any parameters other than the number of domains. Being able to fix the number of domains is an advantage, since it gives the user a tool to decide how much flexibility he wants to resolve and to control the magnitude of the clustering error. A straightforward automatic way to select the number of domains is by requiring the clustering error to be below a specified threshold.

The coarse-graining algorithm has been applied to a number of benchmark problems. First, the consistency and error dependence of the method was demonstrated on two short peptides, by systematically increasing the number of domains M from 2 to the number of atoms N. By using appropriate initial conditions, the clustering error was shown to be monotonically decreasing towards zero for $M=N$. The method was also used to quantify the overall stability/rigidity of several variants of the amyloidogenic protein transthyretin (58His and 58Arg) compared to its native structure. The rigidity properties could be correlated to the destabilisation and amyloid-formation properties of the protein. Compared to the wild type protein we found a decreased stability for TTR-58His variant which is in agreement with urea and thermal induced unfolding experiments of the protein variants

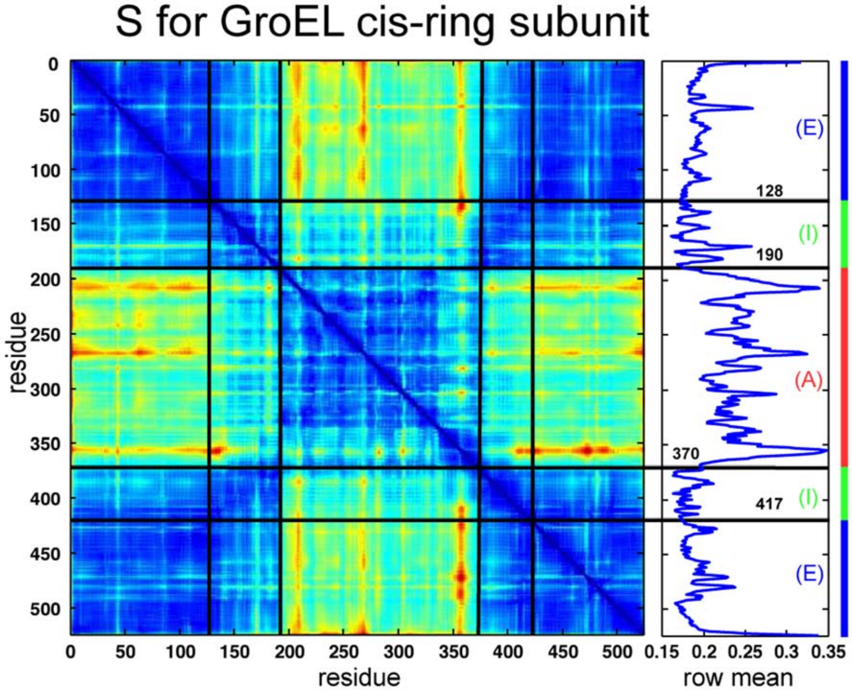

Figure 10. Distance deviation matrix S for the heptameric subunit of GroEL structured for $M=3$. The black lines indicate the identified domains boundaries between the apical (A - red), intermediate (I - green) and equatorial (E - blue) domain in the GroEL subunit. The mean value of each row and the color assignment are shown on the right.
doi:10.1371/journal.pone.0010491.g010

[51] and structural and computational studies [49]. It was further found that the TTR destabilisation is not only determined by the overall stability, but also by local destabilisation that is different in the variants. The method is able to identify the residues in the disease causing variants of the protein, that have increased flexibility compared to the wild type protein. These regions are proposed to cause local destabilisation of the protein that may lead to transient dissociation into monomers. For small number of domains the coarse-graining of x-ray structures is almost similar to the coarse-graining obtained by the clustering of molecular dynamics data. Finally, we demonstrated that the rigidity clustering of large molecular complexes like for example the 8,015-α-carbon atom system GroEL-GroES can be done within less than one CPU hour. The method clearly identifies functional and structural domains that allow to describe the conformational change of the GroEL-GroES complex formation where the ring structure is split along the long axis resulting in a deformation of the cavity. For larger number of domains the method finds the monomeric substructures in the heptameric rings of the molecular complex. The three major domains found in such a subunit are in good agreement with the apical, intermediate and equatorial domain in the GroEL monomer [62].

Since the clustering method proposed here is a data-based method, its result will depend on the quality of that data. In principle, the result will only be globally converged, if the underlying simulations have visited all relevant conformations within the data set according to the Boltzmann probability. However the GroEL-GroES results and other studies on very large systems such as viruses [65,66] indicate that the rigidity information required to identify semi-rigid domains within one conformation converges very quickly. The advantage of data-based clustering is that it is independent of the molecular model used and can also be applied to realizations of an NMR ensemble or a series of x-ray structures of the same protein.

Besides the robustness and reliability the method is easy to implement, efficient and useful in obtaining the essential nanomechanical properties of the molecule, we expect it to become a useful tool for the analysis of large-scale molecular systems.

As an outlook the method presented here can be used as a first step to generate a simulation model for large molecules or aggregates that can for example be simulated with Brownian dynamics. In addition to the identification of mobile domains this requires also the estimation of interaction forces and diffusion constants from either simulation or experimental data. This task is a subject of ongoing work.

Acknowledgments

The authors like to thank Bert de Groot for helpful suggestions and Klaus Altland for valuable discussions about transthyretin.

Author Contributions

Conceived and designed the experiments: SB FN. Performed the experiments: SB. Analyzed the data: SB FN. Wrote the paper: SB.

References

1. Bao G (2002) Mechanics of biomolecules. J Mech Phys Solid 50: 2237–2274.
2. Bustamante C, Smith S, Liphardt J, Smith D (2000) Single-molecule studies of DNA mechanics. Curr Opin Struct Biol 10: 279–285.
3. Gardel ML, Nakamura F, Hartwig JH, Crocker JC, Stossel TP, et al. (2006) Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. Proc Natl Acad Sci Unit States Am 103: 1762–7.

4. Lavery R, Lebrun A, Allemand J, Bensimon D (2002) Structure and mechanics of single biomolecules: experiment and simulation. J Phys Condens Matter 14: 383–414.
5. Micheletti C, Lattanzi G, Maritan A (2007) Elastic properties of proteins: insight on the folding process and the evolutionary selection of native structures. J Mol Biol 321: 909–21.
6. Splettstoesser T, Noe F, Oda T, Smith J (2008) Nucleotide-dependence of G-actin conformation from multiple molecular dynamics simulations and observation of a putatively polymerization-competent superclosed state. Proteins: Structure, Functions, and Bioinformatics.
7. Ahmed A, Gohlke H (2006) Multiscale Modeling of Macromolecular Conformational Changes Combining Concepts From Rigidity and Elastic Network Theory. Proteins: Structure, Functions, and Bioinformatics 63: 1038–1051.
8. Ayton GS, Noid WG, Voth GA (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 17: 192–8.
9. Izvekov S, Voth G (2005) A multiscale coarse-graining method for biomolecular systems. J Phys Chem B 109: 2469–2473.
10. Atilgan A, Akan P, Baysal C (2004) Small-World Communication of Residues and Significance for Protein Dynamics. Biophys J 86: 85–91.
11. Bagci Z, Jernigan R, Bahar I (2002) Residue packing in proteins: Uniform distribution on a coarse-grained scale. J Chem Phys 116: 2269–2276.
12. Anselmi C, Bocchinfuso G, Scipioni A, Santis P (2001) Identification of protein domains on topological basis. Biopolymers 58: 218–229.
13. Nicolas WL, Rose G, Eyck L, Zimm B (1995) Rigid domains in proteins: an algorithmic approach to their identification. Proteins: Structure, Functions, and Bioinformatics 23: 38–48.
14. Liljas A, Rossman M (1974) X-ray studies of protein interactions. Annu Rev Biochem 43: 475–507.
15. Wodak SJ, Janin J (1981) Location of structural domains in proteins. Biochemistry 20: 6544–6552.
16. Lesk AM, Rose GD (1981) Folding units in globular proteins. Proc Natl Acad Sci Unit States Am 78: 4304–4308.
17. Zehfus MH (1987) Continuous compact protein domains. Proteins: Structure, Functions, and Bioinformatics 16: 90–110.
18. Xuan Z, Ling L, Chen R (2000) A new method for protein domain recognition. Eur Biophys J 29: 7–16.
19. Wriggers W, Schulten K (1997) Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins: Structure, Functions, and Bioinformatics 29: 1–14.
20. Taylor WR (1999) Protein structural domain identification. Protein Eng 12: 203–216.
21. Balsera MA, Wriggers W, Oono Y, Schulten K (1996) Principal Component Analysis and Long Time Protein Dynamics. J Phys Chem 100: 2567–2572.
22. Ma J (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 13: 373–80.
23. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 15: 586–92.
24. Hinsen K (1998) Analysis of domain motions by approximate normal mode calculations. Proteins: Structure, Functions, and Bioinformatics 33: 417–429.
25. Zhang Z, Lu L, Noid W, Krishna V (2008) A systematic methodology for defining coarse-grained sites in large biomolecules. Biophys J 95: 5073–5083.
26. Héry S, Genest D, Smith J (1998) X-ray Diffuse Scattering and Rigid-Body Motion in Crystalline Lysozyme Probed by Molecular Dynamics Simulation. J Mol Biol 279: 303–319.
27. Yesylevskyy SO, Kharkyanen VN, Demchenko AP (2006) Hierarchical clustering of the correlation patterns: new method of domain identification in proteins. Biophys Chem 119: 84–93.
28. Shibuya T (2008) Fast Hinge Detection Algorithms for Flexible Protein Structures. IEEE ACM Trans Comput Biol Bioinformatics PP: 1.
29. Potestio R, Pontiggia F, Micheletti C (2009) Coarse-Grained Description of Protein Internal Dynamics: An Optimal Strategy for Decomposing Proteins in Rigid Subunits. Biophys J 96: 4993–5002.
30. Hayward S (2004) Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. J Mol Biol 339: 1001.
31. Carlson H, McCammon J (2000) Accommodating protein flexibility in computational drug design. Mol Pharmacol 57: 213–218.
32. Mustard D, Ritchie D (2005) Docking essential dynamics eigenstructures. Proteins: Structure, Functions, and Bioinformatics 60: 269–274.
33. Navizet I, Lavery R, Jernigan R (2004) Myosin flexibility: structural domains and collective vibrations. Proteins: Structure Function and Bioinformatics 54: 384–393.
34. Menor S, de Graff A, Thorpe M (2009) Hierarchical plasticity from pair distance fluctuations. Phys Biol 6: 036017.
35. Jünger M, Reinelt G (2004) Combinatorial Optimization and Integer Programming. In: Optimization and Operations Research Encyclopedia of Life Support Systems EOLSS. pp 321–327.
36. Yesylevskyy S, Kharkyanen V (2009) Fuzzy domains: New way of describing flexibility and interdependence of the protein domains. Proteins: Structure, Functions, and Bioinformatics 74: 980–995.
37. Gill P, Murray W, Saunders M (1995) User's guide for QPOPT 1.0: A Fortran package for Quadratic programming Dept of Operations Research, Stanford University.
38. Noé F, Daidone I, Smith J, di Nola A, Amadei A (2008) Solvent Electrostriction-Driven Peptide Folding Revealed by Quasi-Gaussian Entropy Theory and Molecular Dynamics Simulation. J Phys Chem B 112: 11155–11163.
39. Klabunde T, Petrassi HM, Oza VB, Raman P, Kelly JW, et al. (2000) Rational design of potent human transthyretin amyloid disease inhibitors. Nature Structural Biology 7: 312–21.
40. Sali A, Blundell T (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.
41. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7: 306–317.
42. Hess B, Bekker H, Berendsen H, Fraaije J (1997) LINCS: A Linear Constraint Solver for molecular simulations. J Comput Chem 18: 1463–1472.
43. Blake CF, Geisow M, Swan I, Rerat C, Rerat B (1974) Structure of human plasma prealbumin at 2.5 Angstrom resolution. A preliminary report on the polypeptide chain conformation, quaternary structure and thyroxine binding. J Mol Biol 88: 1–12.
44. Sunde M, Blake C (1997) The structure of amyloid fibrils by electron microscopy and X-ray diffraction. Adv Protein Chem 50: 123–159.
45. Tan SY, Pepys M (1995) Amyloidosis. Histopathology 25: 403–414.
46. Damas A, Saraiva M (2000) Review: TTR Amyloidosis—Structural Features Leading to Protein Aggregation and Their Implications on Therapeutic Strategies. J Struct Biol 130: 290–299.
47. Rochet J, Lansbury P (2000) Amyloid fibrillogenesis: themes and variations. Curr Opin Struct Biol 10: 60–68.
48. Jaroniec C, MacPhee C, Astrof N, Dobson C, Griffin R (2002) Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. Proc Natl Acad Sci Unit States Am 99: 16748–16753.
49. Cendron L, Trovato A, Seno F, Folli C, Alfieri B, et al. (2009) Amyloidogenic Potential of Transthyretin Variants. J Biol Chem 284: 25832.
50. Altland K, Winter P, Sauerborn M (1999) Electrically neutral microheterogeneity of human plasma transthyretin (prealbumin) detected by isoelectric focusing in urea gradients. Electrophoresis 20: 1349–1364.
51. Takeuchi M, Mizuguchi M, Kouno T, Shinohara Y, Aizawa T, et al. (2006) Destabilization of transthyretin by pathogenic mutations in the DE loop. Proteins: Structure, Function, and Bioinformatics 66: 716–725.
52. Jenne D, Denzel K, Blatzinger P, Winter P (1996) A new isoleucine substitution of Val-20 in transthyretin tetramers selectively impairs dimer– dimer contacts and causes systemic amyloidosis. Proc Natl Acad Sci Unit States Am 93: 6302–6307.
53. Altland K, Benson M, Costello C, Ferlini A (2007) Genetic microheterogeneity of human transthyretin detected by IEF. Electrophoresis 28: 2053–2064.
54. Foss T, Wiseman R, Kelly J (2005) The Pathway by Which the Tetrameric Protein Transthyretin Dissociates. Biochemistry 44: 15525–33.
55. Serag AA, Altenbach C, Gingery M, Hubbell W, Yates TO (2002) Arrangement of subunits and ordering of ß-strands in an amyloid sheet. Nat Struct Biol 9: 734–739.
56. Sørensen J, Hamelberg D, Schiøtt B, McCammon JA (2007) Comparative MD analysis of the stability of transthyretin providing insight into the fibrillation mechanism. Peptide Science 86: 73–82.
57. Saraiva M (2001) Transthyretin mutations in hyperthyroxinemia and amyloid diseases. Hum Mutat 17: 493–503.
58. Motozaki Y, Sugiyama Y, Ishida C, Komai K, Matsubara S, et al. (2007) Phenotypic heterogeneity in a family with FAP due to a TTR Leu58Arg mutation: A clinicopathologic study. J Neurol Sci 260: 236–239.
59. Mayhew M, Silva A, Martin J, Erdjument-Bromage H, Tempst P, et al. (1996) Protein folding in the central cavity of the GroEL-GroES chaperonin complex. Nature 379: 420–426.
60. Ma J, Sigler P, Xu Z, Karplus M (2000) A dynamic model for the allosteric mechanism of GroEL. J Mol Biol 302: 303–313.
61. Walter S (2002) Structure and function of the GroE chaperone. Cell Mol Life Sci 59: 1589–97.
62. Keskin O, Bahar I, Flatow D, Covell D, Jernigan R (2002) Molecular Mechanisms of Chaperonin GroEL- GroES Function. Biochemistry 41: 491–501.
63. Kim M, Jernigan R, Chirikjian G (2005) Rigid-Cluster Models of Conformational Transitions in Macromolecular Machines and Assemblies. Biophys J 89: 43–55.
64. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL–GroES. Mol Syst Biol 36: 1–13.
65. Arkhipov A, Larson S, McPherson A, Schulten K (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure 14: 437–449.
66. Arkhipov A, Freddolino P, Schulten K (2006) Stability and dynamics of virus capsids described by coarse-grained modeling. Structure 14: 1767–1777.