Georg-August-Universität
Göttingen
Zentrum für Informatik

ISSN 1612-6793
Nummer ZFI-MSC-2009-10

Masterarbeit

im Studiengang “Angewandte Informatik”

Quantification of structure/dynamics correlation of globular proteins

René Rex

in der Abteilung für

theoretische und computergestützte Biophysik
des Max-Planck-Instituts für biophysikalische Chemie

Bachelor- und Masterarbeiten
des Zentrums für Informatik
an der Georg-August-Universität Göttingen

30. September 2009

Georg-August-Universität Göttingen
Zentrum für Informatik

Goldschmidtstraße 7
37077 Göttingen
Germany

Tel. +49 (5 51) 39-17 42010
Fax +49 (5 51) 39-1 44 15
Email office@informatik.uni-goettingen.de
WWW www.informatik.uni-goettingen.de

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Göttingen, den 30. September 2009

Masterarbeit
im Studiengang "Angewandte Informatik"

Quantification of structure/dynamics correlation of globular proteins

René Rex

Betreut durch Prof. Dr. Helmut Grubmüller

30. September 2009

Contents

1 Introduction

Protein structures, which are determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, are often presented as 3D ribbon diagrams showing α-helices, β-sheets and loops in colorful images. One can easily identify structured and less structured regions, but from the biological point of view this depiction is misleading: most proteins are no static objects and have to perform very specific movements to fulfill their task. Hence, knowledge about its dynamics is critical for an in-depth understanding of a protein's function. Or, to use the words of the authors of a recent review on this topic: the dynamic landscape is the "personality" of a protein [7].

Advanced experimental setups, e.g., those using the Förster resonance electron transfer (FRET) mechanism [4] can provide a tool for measuring the distance between two selected residues. A more suitable method is NMR spectroscopy [16]. Since new methods use isotopical labels to measure the movement of the side chains, NMR can provide nearly complete information about the dynamics of a protein up to a time scale of milliseconds. But the fastest time scales of pico- and femtoseconds are still not accessible by NMR. Often, an *in silico* experiment is the method of choice for collecting dynamic information. Advances in software and hardware efficiency during the last years allow to perform molecular dynamics (MD) simulations over a time scale of several 100 nanoseconds.

The classification of proteins based on structural similarities is widely used and accepted. But the question whether the structure of these molecules induce a comparable order among the proteins has not been addressed in full detail. This work closes this gap by an analysis directed to the detection of relations between the structure and the dynamics of a protein. For this purpose, multiple observables obtained from MD simulation trajectories of over 100 different proteins were evaluated. In this data no evidence of a natural order in the space of protein dynamics can be found. However, in comparison with a structural classification, an influence of the structure on the dynamics can be detected. These results suggest that

the structural features of a protein only partially determine the dynamics.

1.1 Related Work

Another approach dedicated to protein dynamics is the "Dynameomics" project by the Daggett group at the University of Washington. According to the website[1] the objective is "to characterize the native state dynamics and the folding / unfolding pathway of representatives from all known protein folds by molecular dynamics simulation". The first steps leading to this ambitious aim are several hundred MD simulations and the creation of a database containing all the data. To ensure comparability the simulations are carried out according to standard protocols. With respect to the multi-dimensional nature of the data and due to the fact that the database will mainly be used for the analysis with transactions being a rare event, the project uses an Online-Analytical-Processing (OLAP) approach [14].
There are some major differences between the Dynameomics project and the work presented here. First of all, the project's current main objective is the creation of a database, whereas we focus on the analysis of the MD data. Moreover, the length of the simulations used for this works is about 100 ns, and according to the statistics given on the Dynameomics website the average simulation time is 11,8 ns. Thus, only short-term dynamics can be captured by these data.

[1] `http://www.dynameomics.org`

2 Materials and Methods

2.1 MD simulations

A representative subset of 112 proteins was selected for MD simulations. All simulations were carried out using GROMACS version 3.3.1-2 [2, 15] and the OPLS-aa force field [13]. The tip4p explicit water model [12] was used and physiological sodium chloride concentrations were chosen. Each trajectory has a length of 110ns which yields data in the order of 7 terabyte in total.
The initial phase of the whole project, prior to the analysis presented here, has been conducted by Jürgen Haas[3], who carried out the simulations, and Gert Vriend[2], who helped to select a representative set of proteins.

2.2 Observables

From the simulations, numerable observables have been calculated, which can be divided into two categories. On the one hand, there are structural observables which describe the static properties of a protein (section 2.2.1). On the other hand, some observables characterize the protein dynamics. The latter can be further subdivided into those which are and those which are not based on a principal component analysis (sections 2.2.2 and 2.2.3).

2.2.1 Structural observables

SCOP class

To obtain a structural subsumption of the simulated systems the "Structural Classification of Proteins" (SCOP) database [17] was queried and the corresponding classes were extracted. All assignments in the SCOP database are purely based on

[3] http://www.mpibpc.mpg.de/home/grubmueller/ihp/aalumni/jhaas/
[2] http://swift.cmbi.ru.nl/gv/start/index.html

structural and sequential data, which is evaluated by humans. Thus, the SCOP class is not a real observable obtained from the trajectories but a starting point for a structural classification. In total, five classes are covered by the representative set. Table 2.1 shows the distribution of the systems over the classes.

SCOP class	Count
all α	12
all β	33
α / β	27
α + β	30
small	10

Table 2.1: Distribution of the simulated systems over the SCOP classes

Unlike the name implies, the all α and all β proteins are not purley made of α-helices or β-sheets. But the major fraction of them consists of these secondary structures (a remarkable exception is discussed in section 3.1). The α / β class contains proteins which are dominated by parallel β-sheets and close-by α-helices. In contrast, the proteins of the α + β class display antiparallel β-sheets and segregated α and β regions. Finally, the "small" systems are governed by metal ligand, heme, and/or disulfide bridges.

Secondary structure counts

Secondary structures play an important role concerning the 3D structure of a protein and thus a structural classification must take their presence or absence into account. Therefore, α-helices, β-sheets and turns have been counted under the condition that they are present during more than 50% of the simulation time per residue.

Radius of gyration

A value capturing the overall change of size of a molecule is the average deviation from the mean radius of gyration R_g which is defined as the root mean square

deviation from the center of mass. The formula is given by:

$$R_g = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(r_j - (\frac{1}{N}\sum_{i=1}^{N} r_i))^2}\,,$$

where r_i is the position of the ith atom and N is number of atoms.

2.2.2 Dynamic observables: PCA-based observables

Principal component analysis (PCA) is a well-known and widely used technique for dimension reduction. This method reduces the complexity of the data while it maintains as much variance as possible [11]. In MD simulations, it is commonly used to identify collective degrees of freedom which govern most of the protein dynamics. This is due to the observation that the so-called "essential subspace" consists of roughly 5-10% of the collective degrees of freedom, which have been shown to describe as much as 90% of the total atomic displacement [1]. To capture the features of this essential subspace, several PCA-based observables are included in the analysis and described in the following.

Application of the principal component analysis

After the removal of translational and rotational motions, the covariance matrix C of the recorded C_α-atoms positions was computed. Finally, the eigenvectors a_i and their corresponding eigenvalues λ_i were obtained by diagonalizing C and thus solving the eigenvalue problem $Ca = \lambda a$.

Eigenvalues and slope of eigenvalue spectrum

The most straightforward observables which can be obtained from the PCA are the eigenvalues. They quantify the movement along the eigenvectors and thus describe the general dynamics of the protein. As we want information on the essential subspace only, no more than the first 10 eigenvalues were recorded. Another observable related to the eigenvalues is the slope of the middle third of the eigenvalue spectrum. This value indicates how fast the eigenvalues are decreasing. If the protein undergoes directed movements in few directions the slope will be low. But if the protein performs undirected movements in diverse directions the slope will be

high. The slope has been determined using a linear fit which additionally yielded a quadratic error. In the case of exponentially decreasing eigenvalues the linear slope is misleading, because the values are decreasing much faster than indicated. Therefore, the quadratic error is included in the set of observables.

Cosine content

The cosine content was used as a measure of random diffusion [9, 8]. The higher the cosine content along an eigenvector the greater is the similarity to a random diffusion. Consequently, this indicates the presence of long-term dynamics which may not have been fully explored by the simulation.

Autocorrelation function

The fluctuation along an eigenvector can be regarded as an oscillation in a harmonic potential. Assuming that the fluctuation is a random process X with mean μ and the value at time t is given by X_t the autocorrelation function

$$ACF_X(\delta) = E[(X_t - \mu) - (X_{t+\delta} - \mu)]$$

can be used to calculate correlation values depending on the given time shift δ. A least square fit to these values with the function

$$F(t) = \mathrm{e}^{-\beta\frac{t}{2}} \cdot (\cos(\omega t) + \beta\frac{\sin(\omega t)}{2\omega})$$

yields two variables (β and ω), which can be used in conjunction with the protein mass m to determine the theoretic friction f and spring constant k of the harmonic oscillation:

$$k = \sqrt{4\omega^2 m^2 + \beta^2 m^2}$$
$$f = \beta m$$

An example is shown in figure 2.1. The autocorrelation of the movement along the first eigenvector (blue graph) is fitted (red graph), and thus a damped oscillation becomes visible.

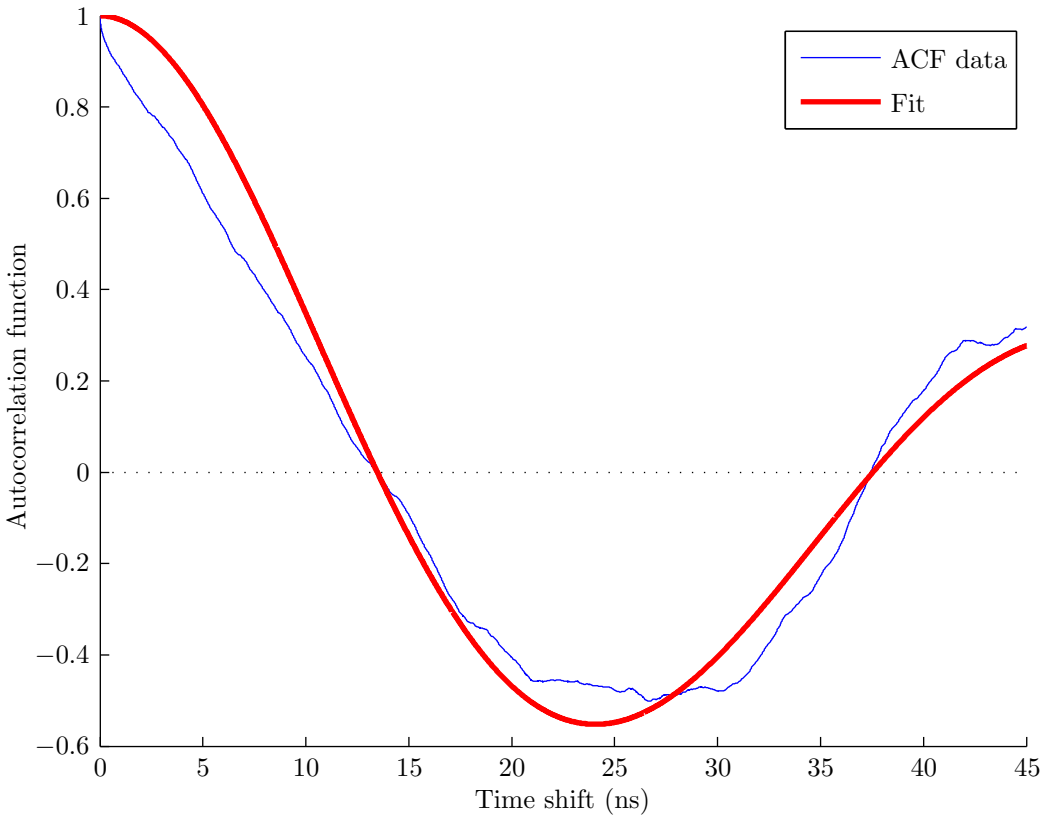

Figure 2.1: The autocorrelation along the first eigenvector (blue) and the corresponding fit (red) obtained from the simulation of an eye lens protein (PDB code: 1DSL)

Goodness of fit with a Gaussian

To shed light on the question whether the dynamics along the first eigenvectors are caused by harmonic equilibrium fluctuations or protein specific movements, a Gaussian has been fitted on each PCA mode. The quality of the fit $\hat{y}$ has been assessed by an R^2 statistic, which is given by:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \,,$$

where y_i denotes the i-th data point and $\bar{y}$ the mean. A low value indicates a poor fit and thus an anharmonic fluctuation - and vice versa.

Ruggedness of the energy landscape

The "ruggedness" is a value indicating the ruggedness of the energy landscape based on the eigenvalues of a time-dependent PCA [5]. Plotted against the eigenvalue index, the ruggedness shows a characteristic curve for each protein. A quadratic fit has been applied to these plots. The mean of the kurtosis and the skewness of these fits represent the actual variables used in the further analysis. As the ruggedness has been introduced very recently, it is not yet well-known by the community.

2.2.3 Dynamic observables: Non-PCA based observables

Solvent accessible surface (SAS)

The protein surface accessible to the solvent has been determined by a probe sphere with a radius of 0.14 nm. This observable is highly dependent on the number of residues: In the case of globular proteins, the surface increases with the sequence length. To correct this issue, the SAS has been divided by the number of residues of the corresponding protein, which yields the solvent accessible surface per residue. The standard deviation of this value is a measure for the capability of a protein to change its conformation.

Root mean square deviation

The root mean square deviation (RMSD) of the C_α-backbone can be used to compare two conformations of the same protein in terms of structural similarity. In the case of an MD simulation, a reference structure has to be chosen to which all other snapshots are compared. This observable is based on the time-averaged protein structure.

Root mean square fluctuation

The Root mean square fluctuation (RMSF) indicates the flexibility of a molecule in general. For an atom a the RMSF is given by:

$$RMSF_a = \frac{1}{T}\sum_{t=1}^{T}(x_t^a - \overline{x^a})^2 \,,$$

where T is the total number of time steps, x_t^a is the position of atom a at time t and $\overline{x^a}$ is the mean position of atom a. This value has been calculated for all C_α-atoms of the protein and the sum, the mean and the standard deviation of these values were included in the further analysis. Furthermore, the position of the minimum relative to the sequence length is recorded.

RMSF Entropy

The entropy of the RMSF is a measure indicating to what extend the fluctuation is concentrated inside a protein. In analogy to the textbook definition of the entropy, the definition of the RMSF entropy S is chosen as follows:

$$S = -\frac{1}{\sum R} \sum (R \cdot ln(R)) + ln(\sum R) \, ,$$

where R is the vector of RMSF values. Unfortunately, this term has a high dependency regarding the sequence length as can be seen in figure 2.2. For the further analysis this has been corrected by subtracting the natural logarithm of the sequence length.

2.2.4 Preprocessing and normalization

The observables have been selected to capture all possible dynamic features of a protein. Nevertheless, they may correlate with each other or with structural observables. Thus, the set was carefully reviewed prior to the further analysis to avoid the introduction of artifacts or biases into the results. Special attention was paid to possible correlations with the sequence length. This is due to the fact that we want to obtain the abstract dynamics, which does not depend on the size of the protein. Furthermore, trivial correlations have been removed, which is outlined in the following.
The matrix on figure 5.1 in the appendix shows Spearman's rank correlation coefficient [20] for all pairwise combinations of observables. Multiple strong correlations can be identified and most of them appear in blocks of related observables. Nearly all of these blocks can be removed by selecting a representative observable out of this subset. In the case of the solvent accessible surface we decided to discard the minimum and the maximum and keep the mean and the standard deviation. The same applies to the RMSD and the RMSF. From the radius of gyration and the

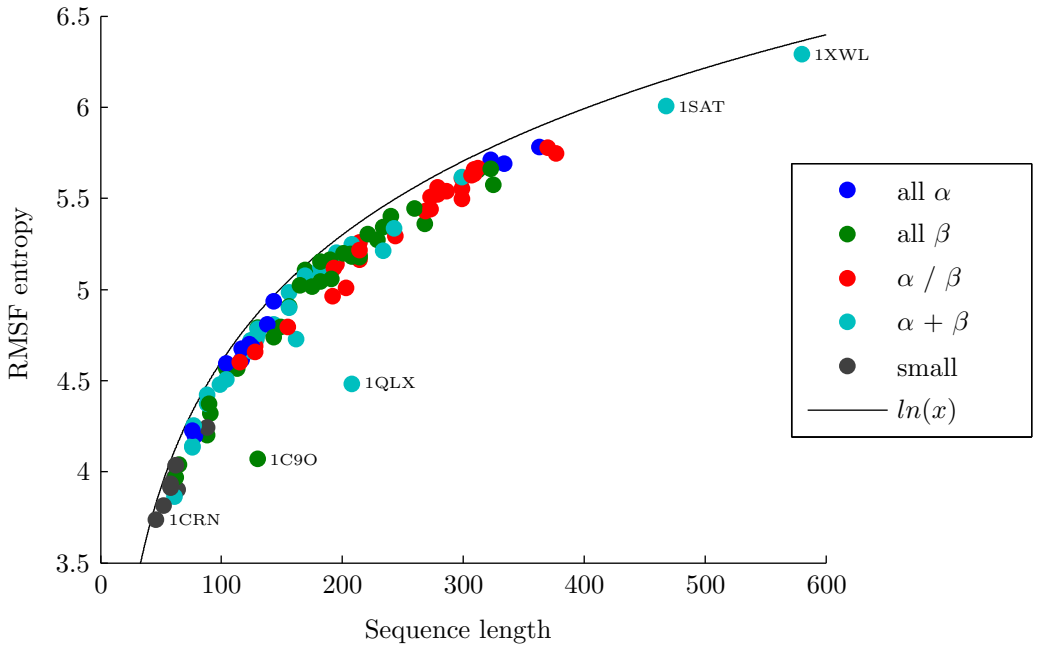

Figure 2.2: The protein sequence length plotted against the entropy of the RMSF exhibits a nearly perfect logarithmic dependency. Each dot represents a protein and is colored according to the SCOP class. PDB codes of some extreme values are shown, too: a plant seed protein (1CRN), a cold shock protein (1C9O), the human prion protein (1QLX), a hydrolase (1SAT) and a polymerase (1XWL).

structure count observables only the standard deviations remained in the set, because the means are regarded as structural observables. The friction and the spring constant of the autocorrelation function are highly correlated with each other, too. Hence, we choose to skip the spring constant. Next, all observables depending on the sequence length are divided by the number of residues. Furthermore, all standard deviations are converted to relative standard deviations by dividing them by their corresponding mean. To further decorrelate the data, the mean RMSF is divided by the mean RMSD. Moreover, the sum of the RMSF and the standard deviation of the RMSD are divided by the standard deviation of the radius of gyration. Finally, the data has been mean centered and normalized to unit variance.

2.3 Clustering and cluster validation algorithms

The analysis of the data described in the previous sections is performed using a variety of methods and algorithms which are outlined in the following.

2.3.1 k-Means

K-Means (algorithm 1) is an iterative clustering algorithm which clusters N data points into K partitions. It assumes, that the data points are elements of a vector space and that there is a distance function which calculates the distance between two data points. In this work, the well-known euclidean distance is used. The clusters are represented by the mean of the associated data points. Initially, the K means are each randomly set on one of the data points. Then each data point is associated with its nearest mean and the position of the means are updated accordingly. This procedure is repeated as long as points move from one mean to an other or until the maximum number of iterations (100 in our case) is reached. Finally, the best run, in terms of the minimum mean distance to the cluster centers, is chosen from 1000 independent runs.

Input: N Data Points, Number of Cluster K, Maximum iterations m
Output: K Clusters
Random initialization of all means;
while *Movement of means* > 0 *and number of iterations* $\leq m$ **do**
    Recalculate the position of each mean;
    Assign all data points to nearest mean;

Algorithm 1: K-Means Algorithm

2.3.2 Silhouette value

Invented as a graphical display for non-hierarchical cluster methods, a silhouette plot [19] contains a bar for each point in a data set. All bars are grouped by cluster and represent the silhouette value for their corresponding data point. Assuming the existence of a distance function d (e.g. euclidean distance) and a partitioning which associates data point i with a cluster A of size N_A, the average dissimilarity of i to all other objects of A is given by

$$a(i) = \frac{1}{N_A - 1} \sum_{j \in A, j \neq i} d(i,j) \; .$$

Furthermore, the minimum average dissimilarity of i to all objects of an other cluster B is given by

$$b(i) = \min_{B \neq A}(\frac{1}{N_B} \sum_{j \in B} d(i,j)) \; .$$

Finally, the silhouette value

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

is defined as the quotient of the difference and the maximum of these two values. This expression can take values between -1 and 1. The average silhouette value $\bar{s}$ gives an estimation of the overall clustering quality. To calculate $b(i)$ the presence of at least one more cluster is required. Hence, $s(i)$ is defined for two or more clusters, only.

2.3.3 Connectivity and variance

Connectivity and variance [6] are two conflicting measures which should be minimized by an optimal clustering. The connectivity measures how many close-by data points lie within the same cluster as a data point i. This is done by evaluating the function

$$WeightNeighbor_i(j) = \begin{cases} \frac{1}{j} & if \; C^i \neq C^{NearestNeighbour_i(j)} \\ 0 & else \end{cases}$$

which is related to the nearest neighbor classifier. It returns a weight of $\frac{1}{j}$ if the j-th neighbor belongs to the same cluster as i and zero else. The connectivity for a partitioning C with N data points is given by:

$$Connectivity(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} WeightNeighbor_i(j) \; .$$

Where L is the maximum number of nearest neighbors to be evaluated. L is set to 10 for the remainder of this work.
In contrast to the connectivity, the variance within a cluster is a measure for the compactness of a clustering. For a set of clusters C containing N data points in

total and a distance function d the variance is defined as:

$$Var(C) = \sqrt{\frac{1}{N} \sum_{C_k \in C} \sum_{a \in C_k} d(a, \mu_k)} \,.$$

Here, μ_k denotes the mean of all points associated with a cluster.
To get a visual impression of the clustering quality for different numbers of clusters, the connectivity is plotted against the variance. An optimal solution is expected to minimize both measures and thus the resulting graph should display a "knee" at the corresponding number of clusters. Ideally, this point marks a pareto-optimal solution according to the two measures.

2.3.4 Jump method

Inspired from the information theoretic concept of distortion, Sugar et. al. proposed a cluster validation called "jump method" [21]. In this method the distortion d_K is estimated for different numbers of clusters K. Next, a transformation power Y has to be selected. Using this value, the distortion is transformed to "jumps"

$$J_K = d_K^{-Y} - d_{K-1}^{-Y} \qquad \text{with } d_0^{-Y} = 0 \,.$$

If an appropriate Y is chosen, the optimal number of clusters is the value which maximizes J_K. The inventors of the method state that a typical value of Y is equal to the half of the dimensionality of the data. In the case of strong correlation, which is present in most real world data sets, Y may be much smaller due to the decreased effective dimensionality. Fortunately, multiple values of Y can be probed efficiently and validated by visual inspection of the jump plots. In this work, the original implementation of the method provided by the authors is used [1].

2.3.5 Adjusted Rand index

In contrast to the validation methods presented up to now, the adjusted Rand index [18, 10] relies on external information. It compares the clustering result to a given labeling and is corrected for chance. If a completely random clustering is evaluated, the expected value is zero. In the case of a perfect agreement, it takes the value one.

[1] `http://www-rcf.usc.edu/~gareth/research/`

3 Results and Discussion

3.1 Structural classification

The SCOP classification is based on structural and sequential data and thus does not necessarily agree with the structural properties featured by a protein during an MD simulation. To verify the consistence of the SCOP classification with the structural properties, scatter plots of the relevant observables are discussed in this section.

In particular, the fraction of residues forming an α-helix or a β-sheet is an important criterion for the classification. Figure 3.1 shows the β-sheet content plotted against the α-helix content. The first three SCOP classes (all α, all β and α / β) are clearly separated from each other. However, the two remaining classes do not form clusters in this plot.

On closer inspection, one exception can be spotted: a nuclease, which is labeled with its PDB code 1SNO in the plot, is classified as all β but contains a significant fraction of α-helices. The most likely explanation is that the annotator chose the β-barrel to be the dominating feature of the protein. Another interesting aspect is highlighted in figure 3.1: The cyan dots in the yellow box on the left hand side mark proteins, which are labeled $\alpha + \beta$ by the SCOP classification, though they contain hardly any β-sheets. This seems to be a contradiction to the definition of the class. Whereas, CATH [3] (which is a protein classification database similar to SCOP) assigns the class "Mainly Alpha" to all the proteins in question. Both cases make clear that the task of structural protein classification is not solved unambiguously yet. But aside from these minor drawbacks, the classification is suitable for a comparison with a classification based on dynamic observables.

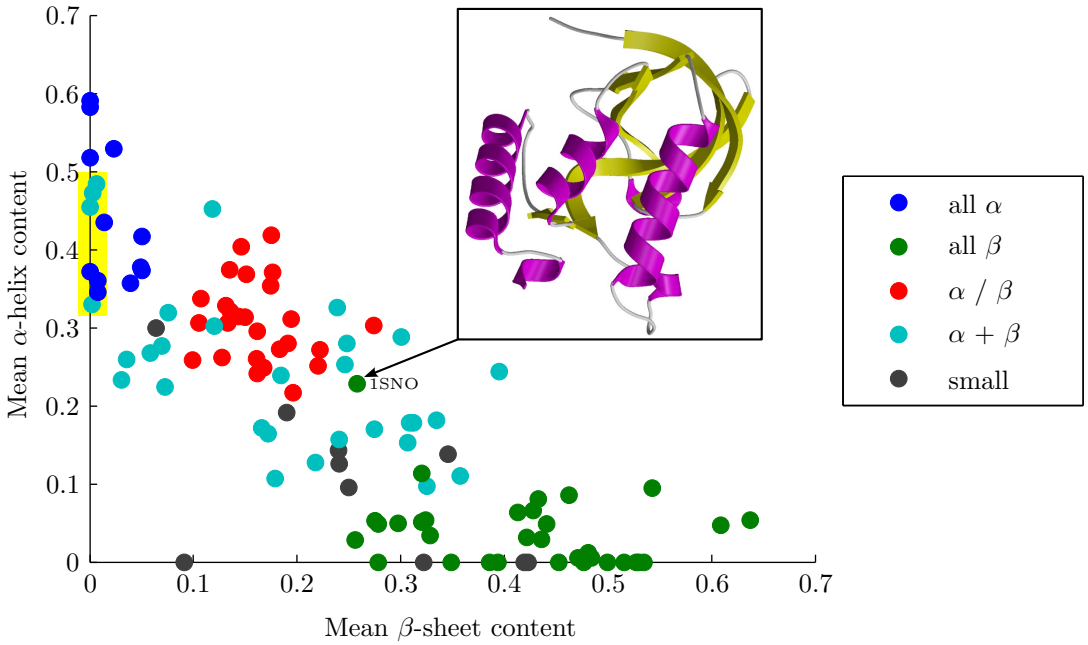

Figure 3.1: The β-sheet content plotted against the α-helix content for all simulated proteins. Each point represents a protein and is colored according to its SCOP class. The inset shows the 3D structure of the nuclease with the PDB code 1SNO which contains a β-barrel (yellow part). The yellow box on the left marks some $\alpha + \beta$ proteins which do not contain any β-sheets.

3.2 Structural influence and protein dynamics

In this section, two major questions are addressed. The first part takes a closer look at a possible partitioning of the proteins in the space of dynamics. For this purpose, an unsupervised learning algorithm is applied to the data and the results are validated by multiple cluster validation methods. In the second part, the effect of the proteins' structure on its dynamics is investigated. One the one hand, this is done by comparing the discrete assignment of the clustering with the SCOP classification. On the other hand, all simulated proteins are compared with each other, based on the structural and dynamic observables described in the previous chapter.

3.2.1 Partitioning of protein dynamics

The dynamic observables were selected to incorporate as much information on the protein dynamics as possible. But due to the high dimensional nature of the data,

an interpretation is not an easy task and requires the use of data mining methods. In particular, a k-Means algorithm was deployed to find natural clusters. An inherent problem to all clustering algorithms is the question of the correct number of clusters. In the case of the data at hand, one possibility is to choose as many clusters as SCOP classes are present in the data set. Unfortunately, this may bias the results if this number is not appropriate. Therefore, all numbers of clusters up to ten are probed and the performance is evaluated by the three cluster validation algorithms described in section 2.3. At first, the full data set with all SCOP classes is analyzed. In a second step, we study a reduced set to minimize the effect of possible misclassified proteins in terms of the SCOP class on the classification. This set consists of the three SCOP classes which form groups as shown in the previous section.

For each number of clusters from one to ten, the k-means clustering yielded a partition of the full set. These partitions are evaluated by the three cluster validation algorithms presented in section 2.3. Each algorithm produced the performance graphs shown in figure 3.2. All graphs give no reason to the assumption that a natural partitioning of the data exists. In more detail, the overall average silhouette (topmost graph) takes very low values, indicating an inadequate clustering according to the original description of Rousseeuw [19]. Note, that the silhouette value for the trivial partitioning with only one cluster is not defined. Thus, this point is missing in the graph. Furthermore, the middle graph shows the connectivity plotted against the variance. Unfortunately, all clusterings minimize only one of the measures and a pareto-optimal solution can not be identified. The decrease of the connectivity at 6 clusters can be explained by outliers which gain their own cluster at this point. Finally, the jump method has its maximum at one cluster, which suggests the absence of any natural clustering. Due to the nature of the method, it takes greater values when the maximum number of clusters is approached. The transformation power was chosen to be 3.5 by sampling a wide range of the variable and by visual inspection of the resulting graphs. Interestingly, nearly the same results were yielded by the analysis of the reduced set (figure 3.3). Due to the absence of the outliers, the connectivity anomaly does not occur anymore.

These results indicate that there are no natural groups in the data. In particular, no evidence can be found which would suggest the clustering to be influenced by the number of SCOP classes present in the data set. Under the assumption that the

chosen observables adequately describe the the protein motion, this leads to the conclusion that the effect of the structure on the dynamics is not strong enough to induce the formation of distinguishable clusters.

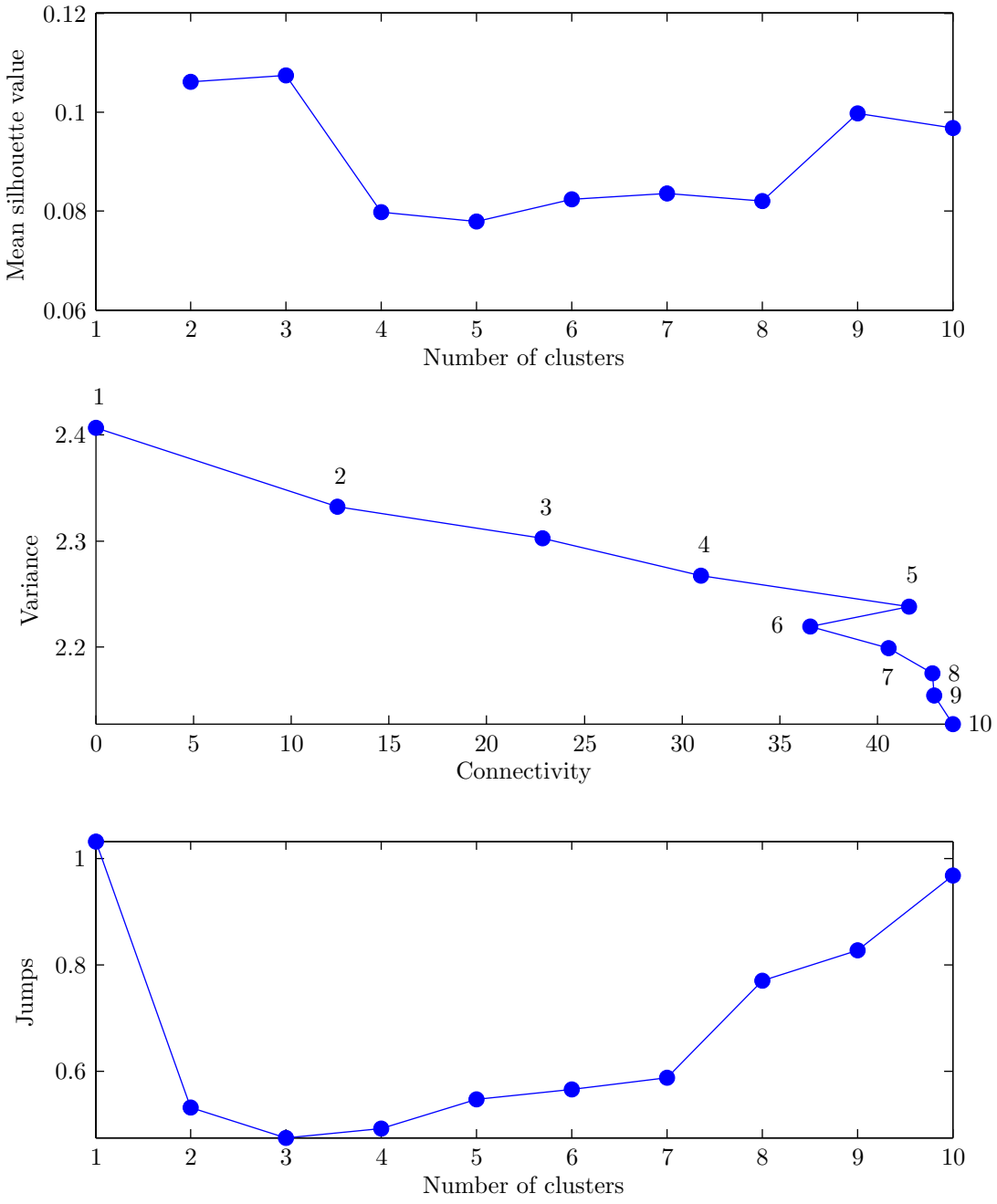

Figure 3.2: The performance graphs of three cluster validation algorithms for the full set are shown. Top: Number of clusters against silhouette value; Middle: connectivity against variance with number of clusters as label; Bottom: number of clusters against the jumps. See section 2.3 for a detailed description of the methods.

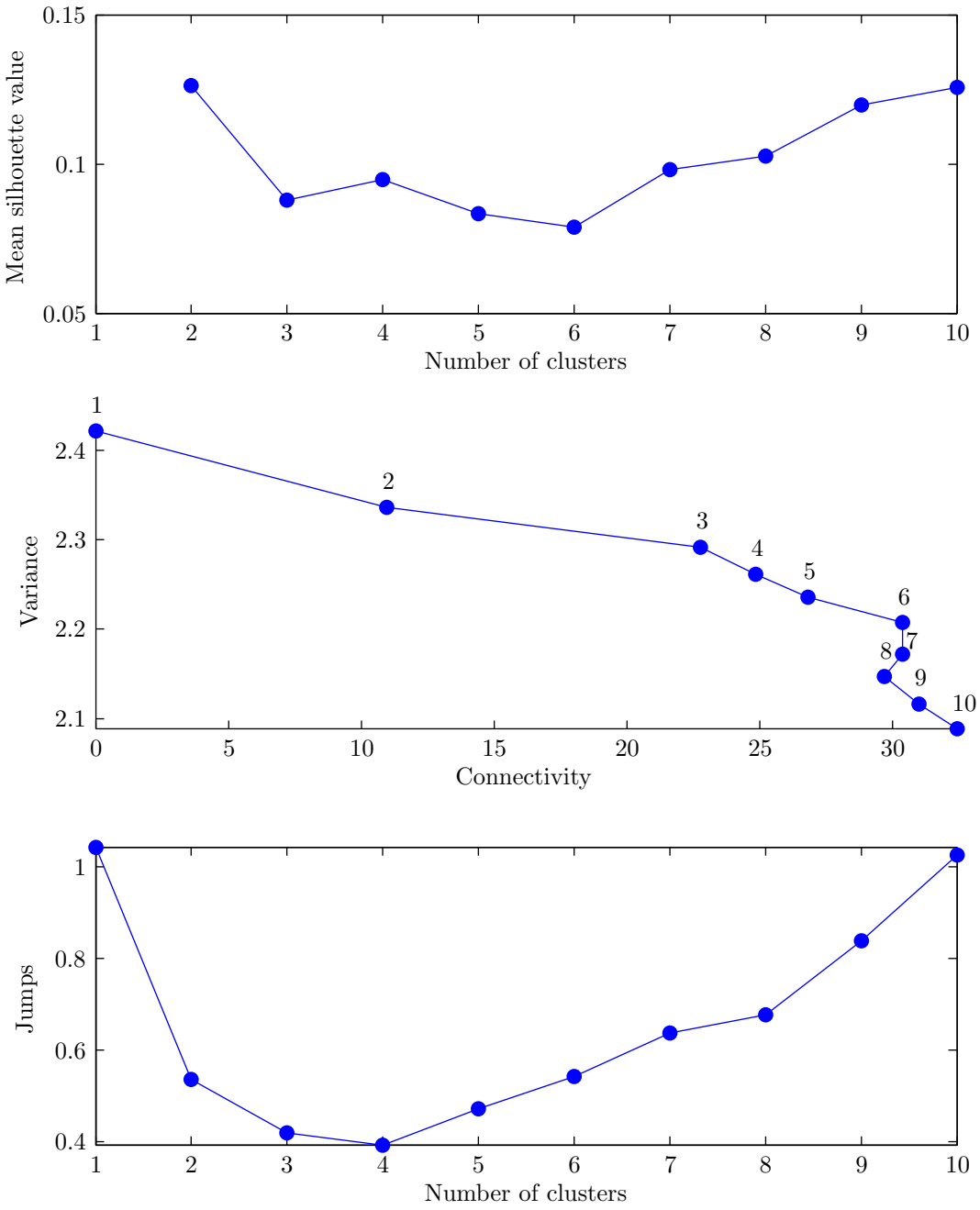

Figure 3.3: The performance graphs of three cluster validation algorithms for the reduced set are shown. Top: Number of clusters against silhouette value; Middle: connectivity against variance with number of clusters as label; Bottom: number of clusters against the jumps. See section 2.3 for a detailed description of the methods.

3.2.2 Influence of structure on the dynamics

Though the structure does not seem to dominate the dynamics of a protein it may at least have some influence. Figure 3.4, which shows the projection of the full data set on the plane spanned by the two first principal components, can give an intuition on the effect. The points are colored according to the SCOP classes of the proteins. At first glance, most SCOP classes seem to be more or less randomly distributed in the space of the dynamics. But leaving aside the two outliers, the proteins classified as small, form a cluster nearby the all α proteins.
To further investigate the distribution of the SCOP classes, the fraction of proteins belonging to a specific class in a cluster has been counted (figure 3.5). Interestingly, each class clearly dominates one cluster in case of the reduced set (chart A). On the contrary the $\alpha + \beta$ class gets more scattered when it is clustered together with the reduced set (chart B). As suggested by the PCA scatter plot, chart C shows that the small proteins share a cluster with the all α proteins. Finally, the full set displays all characteristics expressed in the other charts (chart D). In theory, all these distributions may be products of coincidence. Therefore, the SCOP labeling and the cluster results are compared using the adjusted Rand index, which is corrected for coincident similarities. Depending on the number of clusters, the index reached values between 0.07 and 0.1. To verify this outcome, the adjusted Rand index for random permutations of the clustering labels is calculated, too. This is equivalent to the clustering of random data with the same characteristics as the original data. As predicted by theory, values an order of magnitude smaller, and zero if averaged over multiple runs, were obtained for the random data.
This analysis reveals the effect of the structure on the protein dynamics, which could not bee seen in the clustering results. Furthermore, it gives an explanation for the bad performance of the k-Means algorithm. Though the structure influences the dynamics of a protein, the effect is not strong enough to create distinguishable groups in the dynamic space. One can think of multiple other parameters which may impact the protein dynamics, too. On the one hand, there are physical and chemical conditions like temperature, pressure and pH-value. For example, cold shock proteins do not start to perform their task until temperature falls below a specific value. On the other hand, biological factors like the presence or absence of other proteins influence the protein. Hence, the structure seems to be just one of many factors which influence the dynamics of a protein.

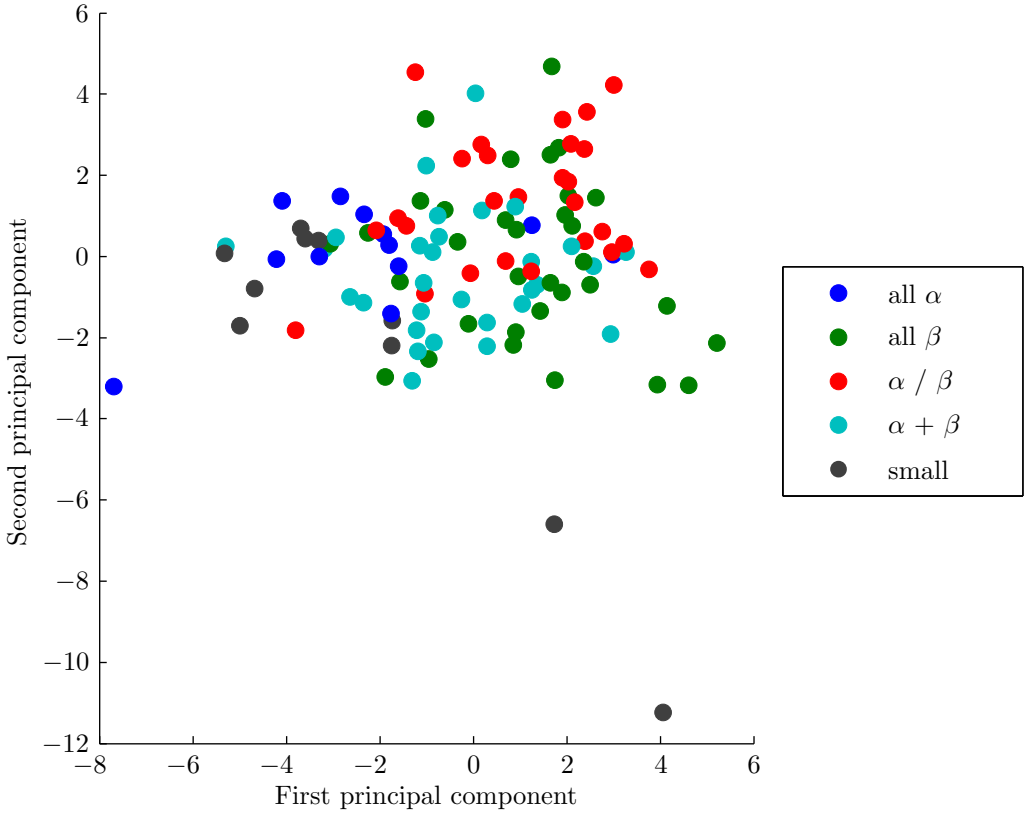

Figure 3.4: The full data set is projected on the plane spanned by the two first principal components. Again, the points are colored according to the SCOP class.

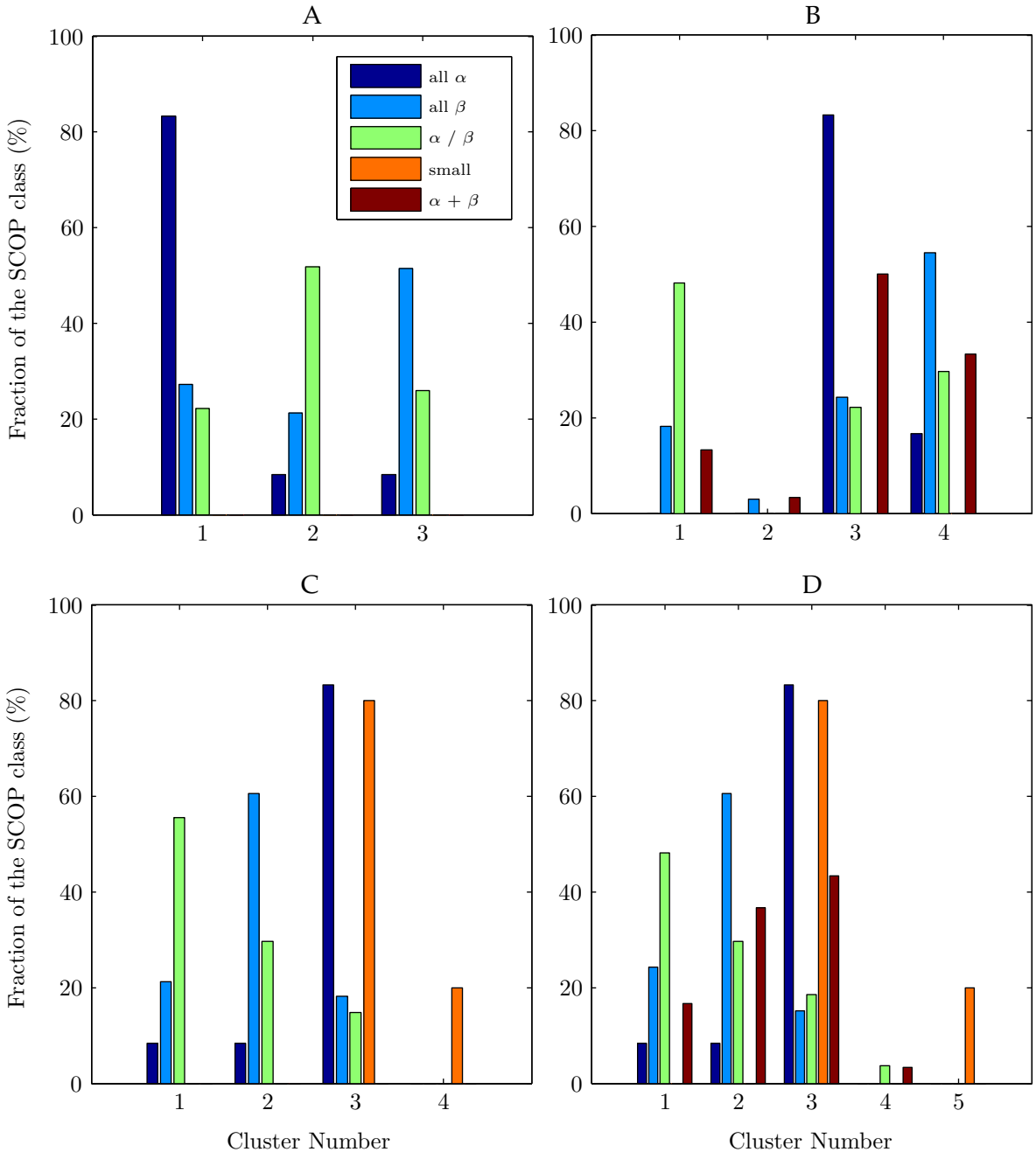

Figure 3.5: Frequency distribution of the SCOP classes under different clustering conditions: A) reduced set with 3 clusters, B) reduced set and $\alpha + \beta$ proteins with 4 clusters, C) reduced set and smallproteins with 4 clusters, D) full set with five clusters

4 Conclusion

Based on MD simulations of a wide variety of proteins, numerable observables have been created and prepared to reflect all dynamic features. On the one hand, these observables were analyzed using an unsupervised learning algorithm and different validation methods. On the other hand, the data has been used in conjunction with a structural classification to uncover and quantify the influence of a proteins' structure on its dynamics. As the cluster analysis did not reveal any natural groups, but the effect of structural properties is still noticeable, we draw the conclusion that the dynamics must be significantly influenced by other parameters like temperature or the presence of further proteins.

4.1 Outlook

The work presented here can be extended in multiple directions. Only one factor influencing the protein dynamics has been addressed. The elucidation of the other parameters is still an open issue. Moreover, the structural classification and comparison can be improved. One possibility is to use structural alignment tools to compare and group proteins based on structural features. Such a method can capture even more complex similarities, e.g. beta barrels.

Bibliography

[1] A. Amadei, A. Linssen, and H. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, 17(4), 1993. 2.2.2

[2] H. Berendsen, D. Van der Spoel, and R. Van Drunen. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995. 2.1

[3] A. Cuff, I. Sillitoe, T. Lewis, O. Redfern, R. Garratt, J. Thornton, and C. Orengo. The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(Database issue):D310, 2009. 3.1

[4] T. Forster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik*, 437, 1948. 1

[5] J. Haas and H. Grubmueller. Probing the energy landscape governing protein motions. *Computer Simulation and Theory of Macromolecules*, 2006. 2.2.2

[6] J. Handl, J. Knowles, and D. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005. 2.3.3

[7] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964, 2007. 1

[8] B. Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62(6):8438–8448, Dec 2000. 2.2.2

[9] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65(3):031910, Mar 2002. 2.2.2

[10] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 2.3.5

[11] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988. 2.2.2

[12] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983. 2.1

[13] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996. 2.1

[14] C. Kehl, A. Simms, R. Toofanny, V. Daggett, and A. Fersht. Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data. *Protein Engineering Design and Selection*, 2008. 1.1

[15] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*, 7(8):306–317, 2001. 2.1

[16] A. Mittermaier and L. Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228, 2006. 1

[17] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995. 2.2.1

[18] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850, 1971. 2.3.5

[19] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20(1):53–65, 1987. 2.3.2, 3.2.1

[20] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, pages 441–471, 1987. 2.2.4

[21] C. Sugar and G. James. Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association*, 98(463):750–763, 2003. 2.3.4

5 Appendix

5.1 Table of abbreviations

ACF	Autocorrelation function
FRET	Förster Resonance Electron Transfer
MD	Molecular dynamics
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SAS	Solvent accessible surface
SCOP	Structural Classification of Proteins

5.2 Additional figures

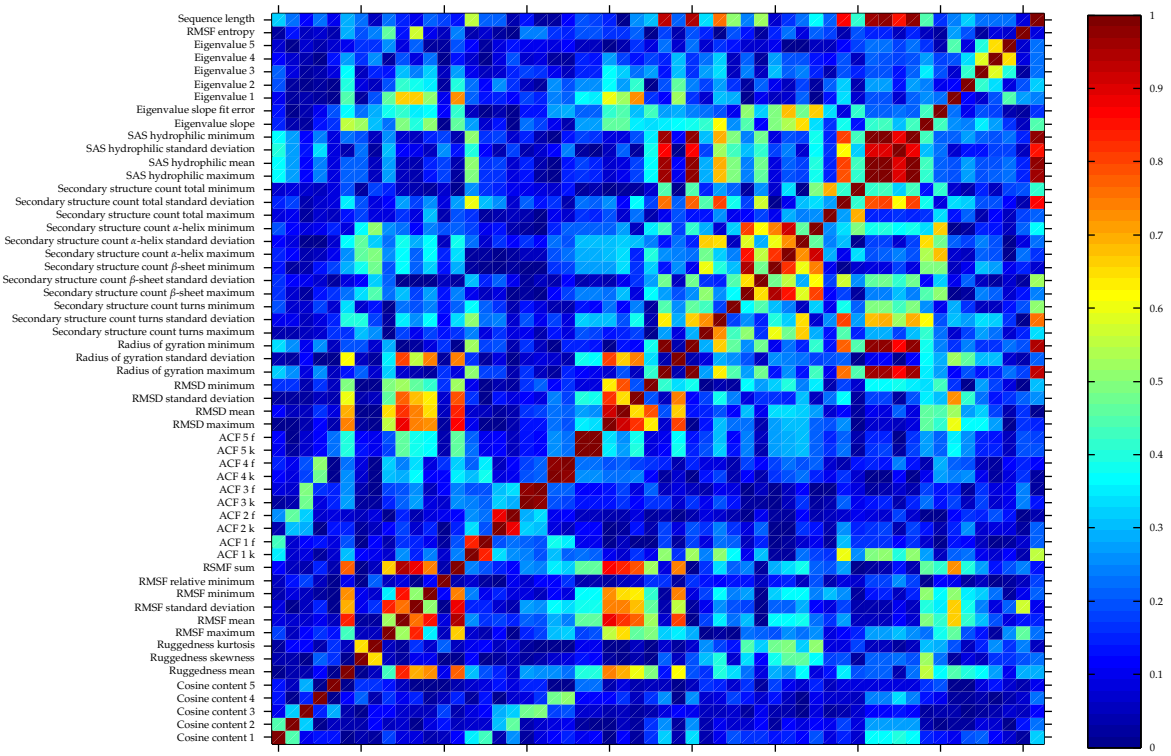

Figure 5.1: Correlation matrix of all dynamic observables and the sequence length before decorrlation and normalization. See chapter 2 for a detailed description.