University of Warsaw

Faculty of Physics

Maciej Dziubiski

Development of selected mesoscopic physical models with the aid of machine learning methods and their applications in studies of molecular systems.

Doctoral Thesis
TODO: na kierunku Fizyka
TODO: w zakresie Biofizyki

Praca wykonana pod kierunkiem
Prof. dr hab. Bogdana Lesynga
Zaklad Biofizyki, Instytut Fizyki Doswiadczalnej,
Wydzial Fizyki, Uniwersytet Warszawski

Warsaw, September 2015

Chapter 1

Introduction

The primary aim of this dissertation was the extraction of palpable observations from complex data describing (bio)molecular systems. In Chapter II we present our original method of discovering quasi-rigid parts in proteins, which relies on experimental data, or from a molecular dynamics trajectory. Chapter III presents a very different problem of identifying parts of a molecular system that propel (or hinder) a given structural transition. In both these cases we applied unsuppervised machine learning techniques (also known as clustering), which is commonly used in finding patterns in seemingly chaotic data.

In this introductory chapter we give an overview of the molecular dynamics simulation scheme and the machine learning methodology. The discussion here is broad, and was intended to motivate the application of clustering techniques described in the remainder of the dissertation. We go into more details concerning particular molecular dynamics and machine learning algorithms in Chapters II and III.

1.1   Molecular Dynamics

Molecular dynamics (MD) simulations are a wide range of numerical methods designed for the study of molecular systems. Typically, the MD approach assumes a classical potential energy function, which treats atoms as electrostatically charged points, and chemical bonds as springs and hinges. More accurate algorithms take into account quantum effects, making the simulation more reliable, but also immensely expensive in terms of computational costs. Although modelling intricate

properties of molecular systems may seem an over-simplification, MD schemes have achieved considerable success over the recent years, and are constantly improving.

Classical MD simulations are so popular mainly because of their significantly lower computational cost. Primarily, this allows for studying large, biologically relevant systems, such as enzymes, DNA, membrane channels, and many other biomolecules. Secondly, the lower computational cost enables calculations of the free energy profile of the transitions associated with the function of a molecule.

The second point is especially important, because all algorithms designed for sampling the configurational space in the classical MD set-up are naturally transferable to quantum-based methods.

Underlying the simplistic, classical description of MD simulations are complex sampling techniques used for estimating macroscopic properties of microscopic systems.

Chapter 2

ResiCon: a method for the identification of dynamic domains, hinges and interfacial regions in proteins

2.1 Abstract

Structure of most proteins is flexible. Indentification and analysis of intramolecular motions is a complex problem. Breaking a structure into relatively rigid parts, so-called *dynamic domains*, may help comprehend the complexity of protein's mobility. We propose a new approach called ResiCon (Residue Contacts analysis), which performs this task by applying a data-mining analysis of an ensemble of protein configurations and recognizes dynamic domains, hinges and interfacial regions, by considering contacts between residues. Dynamic domains found by ResiCon are more compact than those identified by two other popular methods: PiSQRD and GeoStaS. The current analysis was carried out using a known reference set of 30 NMR protein structures, as well as MD simulation data of flap opening events in HIV-1 protease. The more detailed analysis of HIV-1 protease dataset shows that ResiCon identified dynamic domains involved in structural changes of functional importance. The ResiCon server is available at URL:

http://dworkowa.imdik.pan.pl/EP/ResiCon

2.2 Introduction

Proteins are not static. Nuclear magnetic resonance (NMR) spectroscopy ([**martin1988two**]) and the spin-echo spectroscopy ([**bu2011proteins**]) experiments show that. In several cases it was proven that flexibility may be crucial to protein functionality ([**farago2010activation**], [**hamelberg2005fast**]). Although experimental methods provide only general clues about intramolecular motions, molecular dynamics (MD) simulations extend their reach by giving a higher resolution picture – both in space and time – of protein mobility. By studying an NMR ensemble or MD trajectory one may notice that it is composed of relatively rigid structural parts, often referred to as *dynamic domains* ([**hayward1997model**]).

Domains in traditional sense are regarded as parts of the protein which are: conserved (in terms of evolution), autonomous (in terms of folding), and/or compact (in terms of tertiary structure). Such "static" domains are identified through sequence homology, structural analysis of a single configuration, or both. (For conventional methods of identifying protein domains based on multiple sequences or a single structure see e.g. ([**richardson1981anatomy**], [**bork1991shuffled**]).) Conversely, dynamic domains depend on structural transitions performed by the protein.

A number of methods for identification of dynamic domains have been developed. The simplest procedures are based on normal mode analysis, which assumes a harmonic approximation of the potential energy function ([**hinsen1998analysis**]). More advanced approaches use the Gaussian Network Model and analyze correlations in motions between the residues of the protein ([**yesylevskyy2006dynamic**]). Many other approaches have also been developed ([**bahar1997direct**], [**wriggers1997protein**], [**bernhard2010optimal**], [**genoni2012identification**], [**potestio2009coarse**]), but all of them anticipate dynamic domains by analyzing a single structure of a protein. Another class of methods for dynamic domains assignment requires exactly two configurations (see [**hayward1998systematic**], [**lee2003dyndom**], [**ye2003flexible**]). However, the assumption that two "representative" structures encompass all relevant motions is rather speculative.

Experimental and *in silico* methods reach beyond single-structure representation, and are capable of producing numerous configurations of a given protein.

Rather than inferring dynamic domains from one or two structures, a more natural approach would be to determine them based on an ensemble of configurations. GeoStaS is the only method known to us that analyzes a whole ensemble of configurations and assigns each residue to a dynamic domain ([**romanowska2012determining**]). Although GeoStaS can analyze not only proteins but also nucleic acids, it fails to discover dynamic domains whenever they rotate with respect to each other. Alternative methods of analyzing ensembles of configurations assign residues to a static "core" or unstructured bundle (see [**snyder2005clustering**], [**kirchner2011objective**]).

The purpose of this study was to develop a novel methodology named ResiCon, capable of extracting dynamic domains from an ensemble of protein's configurations. ResiCon analyzes strengths of contacts between residues based exclusively on geometrical changes occuring in the provided set of structures. The set may be an NMR ensemble of configurations, or snapshots produced in the course of an MD simulation. ResiCon's main functionality is to identify dynamic domains, but it can also find hinges and interfacial (interdomain) regions.

2.3 Approach

ResiCon starts with identifying pairs of residues which are *in contact*. There are several definitions of contacts between amino-acid residues in the literature. We used the definition presented in ([**daniluk2011novel**]) and adapted it to the case when more than one structure is given (see also [**daniluk2014theoretical**]).

Next, ResiCon constructs a virtual scaffold, connecting residues which are in contact with bars. Stiffness of a given bar reflects the estimated strength of the corresponding contact.

Finally, to identify dynamic domains, ResiCon carries out a partitioning (by computing minimal cuts) of the scaffold, cutting weaker and preserving stiffer bars. This partitioning is carried out by applying a spectral clustering algorithm presented in the following section.

The fundamental underlying assumption is that stability of rigid parts results from stable interactions between its residues. However, in our approach we do not analyze physical interactions between residues – they may be hydrophobic,
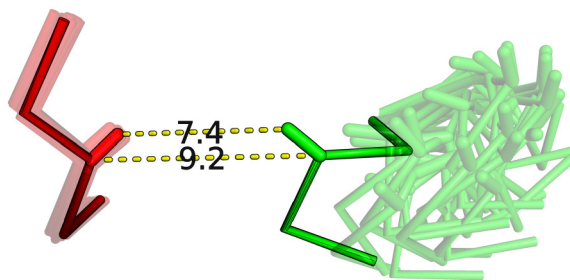
Figure 2.1: For each configuration in an ensemble, pairs of elements are constructed. In this picture residues are in contact because there exists a configuration for which the condition (2.1) is satisfied (the distance $d_C = 7.4$Å, and $d_\alpha - d_C = 1.8$Å).

electrostatic or significant in some other way. We simply assume that the measure of strength of a contact between residues is reflected by their geometrical variability across a given sample of structures.

2.4   Methods

Throughout this paper we use terms: *model*, *configuration*, *structure* and *conformation* interchangeably. We will refer to a set of structures acquired from an MD trajectory or NMR experiment as the *ensemble of configurations* or simply: *an ensemble*. Let us denote the number of structures in an ensemble by S.

Residue contact

For each pair of residues in every model we compute distances between C_α atoms (d_α) and between geometrical centers of side-chains R_C (d_C) (for glycine $R_C = C_\alpha$, and for alanine $R_C = C_\beta$). We say that two residues are *in contact*, if at least one configuration in the ensemble satisfies the condition:

$$(d_\alpha \leq 6.5\text{Å}) \quad \text{or} \quad (d_C \leq 8\text{Å} \quad \text{and} \quad d_\alpha - d_C \geq 0.75\text{Å}) \tag{2.1}$$

Threshold values are the same as in definition of contact presented in ([**daniluk2011novel**]) and relate to the range of distances in which physical interactions between residues occur. The second sub-condition favours residues whose side-chains point towards each other (see Figure 2.1). Residues that are sequential neighbors are not taken into account.

We assign a quantitative value to the strength of a contact in terms of geometrical variability of the structural part associated with that contact. Such measure is required to capture not only changes in d_α, but also rotational shifts and alterations in the backbone in the vicinity of both residues that are in contact. To do so, we constructed structural parts, comprising sequential neighbors of the two residues in contact. ResiCon assigns a numerical value to the geometric variability by using the least root mean square deviation (RMSD) ([**kabsch1976solution**]). Before elaborating on the details, we proceed with the following definitions.

Elements.

An *element* is a structural part of a protein centered around a given residue. It comprises five points, corresponding to the positions of the C_α atoms of the central residue, and its four sequential neighbors (two preceding and two following). For each model s in an ensemble, and each residue i, an element – denoted by E_i^s – is constructed. Residues for which an element cannot be built (e.g. N- and C-termini) are omitted.

Geometrical variability.

We consider pairs of elements, $E_{ij}^s := (E_i^s, E_j^s)$, and express structural deviation of a contact between two configurations r and s by RMSD of E_{ij}^r and E_{ij}^s. We use the following function to express the strength of a contact in terms of the whole ensemble:

$$G(i,j) := \max_{\text{pairs of states } (r,s)} \text{RMSD}(E_{ij}^r, E_{ij}^s).$$

The smaller the geometric variability, the stronger the contact.

We tested several statistics based on RMSDs of pairs of elements. This particular definition of G assumes that a strong contact is not "broken" in any pair of models. Conversely, a contact whose structural stability is breached at least once is assumed to be weak and not contributing to the stability of a given dynamic domain.

Note that conformational transitions may be rapid or insufficiently sampled. Therefore, defining geometrical variability in terms of some averaging statistic (e.g. mean, median) might lead to omitting significant structural changes occurring in

a protein.

Contact matrix

We now describe a matrix representation of an edge-weighted graph, in which nodes correspond to residues. Because we used a spectral clustering algorithm which required that weights in the graph were in the interval $[0, 1]$, we needed to renormalize the geometrical variability. We calculated the weight between node i and j using the following *contact function*:

$$
D(i,j) := \begin{cases} 1 & |i - j| \leq 1 \\[2mm] L_{\alpha,\beta}(G(i,j)) & \text{residues } i \text{ and } j \text{ are in contact} \\[2mm] 0 & \text{otherwise} \end{cases}
$$

where $L_{\alpha,\beta}$:

$$
L_{\alpha,\beta}(x) := \left(1 + e^{\frac{x-\alpha}{\beta}}\right)^{-1}
$$

is a logistic function. We refer to matrix $[D(i,j)]_{ij}$ as the *contact matrix*.

Parameters $\beta > 0$ and $\alpha > 0$ allow for the customization of ResiCon. The logistic function, $L_{\alpha,\beta}$, can be thought of as a rescaling transformation for the measure of geometrical variability, G. It has a simple interpretation – values of G exceeding α are smoothly cut-off, where the degree of "smoothness" is determined by β. All results presented in this paper were acquired with default values of α and β:

$$
\beta_{\text{default}} := \frac{1}{\sigma(G)} \qquad\qquad \alpha_{\text{default}} := \mu(G) \tag{2.2}
$$

where the σ and μ stand for standard deviation and mean taken over values of G for all pairs of elements.

Another feature of the logistic function becomes apparent if we consider an ensemble composed of (nearly) identical structures. This has two possible interpretations: the protein is very stable and no conformational changes exist, or that the provided ensemble does not represent such changes. Thus, ResiCon will assume that the contacts are strong – nearly as strong as the peptide bonds ($G \approx 0 \Rightarrow L_{\alpha,\beta} \approx 1$). Consequently, the contact matrix becomes a so-called *contact map*, assigning binary values to pairs of residues (i.e. 1 if a contact occurred at least once, and 0 otherwise).

Clustering

The contact matrix may be treated as a *similarity matrix*, denoted W, and be used as input in a clustering procedure. Thus, we consider residues to be similar if they are likely to belong to the same dynamic domain. The identified clusters would then correspond to quasi-rigid structural parts.

The choice of a clustering algorithm is not a trivial task and for the identification of dynamic domains two crucial requirements need to be met. Firstly, contact matrices for various proteins vary in dimension and density and the clustering algorithm needs to perform well regardless of these variabilities. Secondly, the algorithm should facilitate an automated method of choosing the optimal number of clusters.

Agglomerative hierarchical clustering algorithms are one of the most popular approaches to clustering ([**han2001data**]). They follow a greedy scheme to construct a dendrogram encoding distances between clusters. This dendrogram can be cut at a certain height, which determines the number of clusters. If the height parameter could be set so that for all similarity matrices we would obtain high-quality clusters, the hierarchical clustering would have been a good candidate for a clustering procedure. However, as we explain in the *Supplementary Materials*, estimation of this parameter is difficult, and to determine the number of clusters we would need to extend the conventional hierarchical clustering with a measure of cluster quality.

This was one of the reasons we have chosen a spectral clustering algorithm, which has an inherent indicator of a partitioning's quality.

Spectral clustering.

Clustering algorithms based on finding the eigensystem of the similarity matrix (or more often a matrix derived from it) are termed *spectral algorithms*. They perform a clustering by minimizing the cost of cutting a graph into subgraphs, which agrees with our intuition about finding quasi-rigid parts based on a contact matrix. Optimal clustering is achieved by discarding contacts with the lowest total weight (as few and as weak as possible) to achieve a partitioning into unconnected regions.
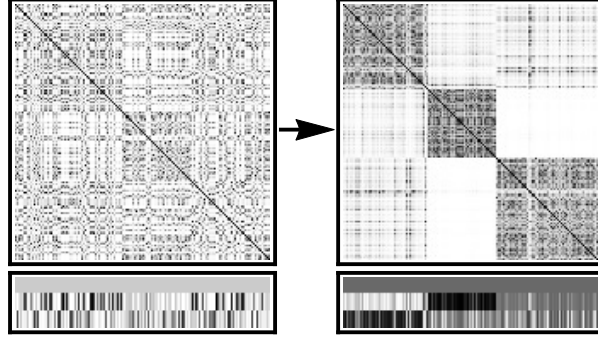
Figure 2.2: An example of a similarity matrix W and three eigenvectors of the stochastic matrix $D^{-1}W$ with two different orderings.

In the case of the clustering algorithm used in ResiCon optimal partitioning is decoded from eigenvectors corresponding to the largest eigenvalues of a stochastic matrix $D^{-1}W$, where $D := diag(d(1), \ldots, d(n))$ and $d(i) := \sum_{j=1, j\neq i}^{n} w(i, j)$ (see [**weber2004perron**]). This transformation of the similarity matrix ensures that the identified clusters tend to have similar sizes, which prevents from identifying singular nodes as clusters. Spectral algorithms make no assumptions on the shape of clusters, and, in contrast to the greedy procedures, are insensitive to the ordering of vertices.

In Figure 2.2 we present a 150×150 similarity matrix with rows and columns ordered in two different ways. The ordering on the right is dictated by the spectral clustering – elements 1–50 were assigned to the first cluster, 51–90 to the second, and 91–150 to the third. To find these three clusters we use the first three eigenvectors (corresponding to three largest eigenvalues) of the stochastic matrix $D^{-1}W$. The first eigenvector always has a constant value in all positions and corresponds to a trivial clustering into a single group. The second eigenvector encodes a partitioning into two groups: nodes 51–90 in the first, and the remaining nodes in the second group. The third eigenvector allows to discern the third cluster, composed of nodes 1–50.

Clustering algorithm.

We used a spectral clustering algorithm in which partitioning of a graph is expressed in terms of a membership matrix χ. A short description of the procedure
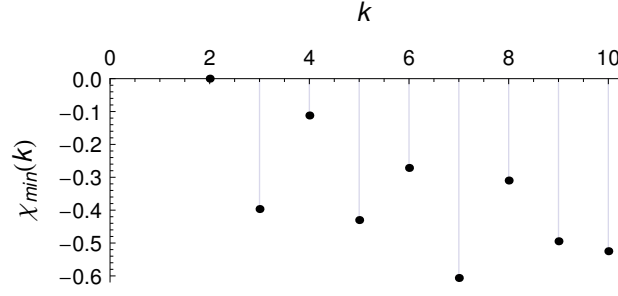
Figure 2.3: Values of the $\chi_{\min}$ indicator for $k = 2, \ldots, 10$, computed for HIV-1 protease. Because $\chi_{\min}(4)$ is closer to 0 than $\chi_{\min}(6)$, the optimal number of clusters is 4.

is presented below, for details refer to ([**weber2004perron**]).

Let Y denote the matrix containing k eigenvectors of the $D^{-1}W$ stochastic matrix, corresponding to k largest eigenvalues. The procedure computes a linear mapping $\mathcal{A}$ from the eigenvectors Y to the membership matrix:

$$\mathcal{A}Y = \chi$$

The element χ_{ij} of this matrix represents the membership of the ith node in the jth cluster. Therefore, if n is the number of nodes in the graph and k is the number of clusters, then $\chi \in \mathbb{R}^{n \times k}$.

This algorithm has two important features:

- it computes the membership matrix χ which allows for overlapping clusters,

- it offers an indicator, called $\chi_{\min}$, used to determine the optimal number of clusters.

Number of clusters.

ResiCon first checks if $k = 1$. To do so, a partitioning into two clusters is carried out. The spectral algorithm presented above finds an optimal cut ([**weber2004perron**]) leading to clusters A and B (two sets of indices, that correspond to nodes). We express the *cost* of such cut by:

$$f := \frac{\sum_{i \in A} \sum_{j \in B} w_{ij}}{\left(\sum_{k,l \in A} w_{kl}\right)\left(\sum_{k,l \in B} w_{kl}\right)}$$

where w_{ij} is the weight of the edge between nodes i and j. The validity of a clustering into clusters A and B was checked by asserting that its cost is less than a given threshold. If the criterion was met, we assumed that a clustering into two or more clusters existed. Default threshold for f used to produce all results presented in this study was 0.1. Above this value we observed that compact static proteins were partitioned into short (less than four residues long) segments, which we regarded as improper dynamic domains. On the other hand, lower values resulted in a single dynamic domain assignment in several cases where two domains were apparent.

If $k > 1$, the optimal number of clusters is determined with the use of the indicator presented in ([**weber2004perron**]). Here, we give a short overview of the properties of $\chi_{\min}$ and propose a simple procedure for computing the optimal number of clusters. The indicator $\chi_{\min}(k)$ is defined as the minimal element of the membership matrix χ found by partitioning it into k clusters.

In the case of $k = 2$ the indicator is always zero, $\chi_{\min}(2) = 0$. For $k > 2$ the indicator is less than zero, however if $\chi_{\min}(k) \approx 0$, the clustering into k clusters is the optimal one (Figure 2.3). Let us recall the notion of visualizing a clustering by a block-like similarity matrix. Roughly speaking, the value of $\chi_{\min}(k)$ resembles the deviation of the similarity matrix from the "pure" block-like form. However, it is difficult to decide which values of $\chi_{\min}$ are sufficiently close to zero to indicate the optimal number of clusters. Therefore, the problem at hand is: does the optimal number of clusters equal two, or more?

In our first approach, the optimal number of clusters was chosen based on a threshold – the optimal k was the one for which $\chi_{\min}$ was above a certain value. However, it was difficult to find the right threshold because values of $\chi_{\min}$ strongly depend on the number of nodes in the graph. Therefore, the following procedure was adapted in ResiCon:

1. Determine the cost of the optimal cut. If it exceeds the 0.1 threshold, the optimal number of clusters is $k = 1$. Otherwise, assume that $k > 1$ and continue the procedure.

2. Compute the values of $\chi_{\min}$ for partitionings into $3, 4, \ldots, M$ clusters.

3. Find $k_1 > 2$ for which $\chi_{\min}$ is closest to 0, and $k_2 > 2$ for which $\chi_{\min}$ is the second closest to 0.

4. If $\chi_{\min}(k_1) > 0.5\,\chi_{\min}(k_2)$, then the number of clusters is $k = k_1$. Otherwise, $k = 2$.

The 0.5 constant in the fourth point means that we choose k_1 as the number of clusters, if $\chi_{\min}(k_1)$ is closer to 0 then to the next best $\chi_{\min}$ (i.e. $\chi_{\min}(k_2)$). The maximal number of clusters M is set to 10 by default, but the user can specify a different number. All results presented in this paper were computed with $M = 10$.

In other words, ResiCon chooses $k > 2$ for which the indicator $\chi_{\min}$ is "relatively close" to 0. When no such value exists, a partitioning into two clusters is assumed to be optimal.

Hinges and interfacial regions

Hinges.

We define hinges as parts of the structure satisfying both of the following conditions:

1. they do not belong conclusively (in terms of membership as explained below) to any dynamic domain ,

2. they are sequentially located between dynamic domains.

The first condition is tested in terms of membership: if a residue membership in any cluster does not exceed certain threshold χ_{hinge} it may belong to a hinge. The default value of the parameter is 0.65, but the user can specify a different value.

Interfacial regions.

A residue is assumed to compose an interfacial region if two conditions are met:

1. it does not belong to any hinge,

2. it was in contact with a residue that does not belong to the same dynamic domain at least once.

Results comparison

According to our knowledge, no expert curated database of dynamic domains exists. Also, we are not aware of any quality measure for the dynamic domains assignment. Therefore, we compared different methods by analyzing agreement between their results.

We used the measure presented in ([**meilua2007comparing**]) called *Variation of Information* ($\mathcal{VI}$) to analyze the results compatibility. It has the advantage of being a metric in the space of all partitionings of a given dataset. The downside of $\mathcal{VI}$ is that its values do not lie in a fixed interval (e.g. $[0, 1]$), but instead have an upper bound that depends on the size of data. In our case data size equals the number of residues in a given protein. This means that values of $\mathcal{VI}$ for partitionings of one protein are not directly comparable to the values acquired for partitionings of another protein, with a different number of residues. Nonetheless, when considering a particular protein, the $\mathcal{VI}$ metric quantifies the agreement between different assignments of dynamic domains. Here we give an outline of the method, for details see ([**meilua2007comparing**]).

Let us denote a clustering by $\mathcal{C}$. It is composed of clusters – mutually disjoint subsets $C_1, \ldots, C_k$. That is, $\mathcal{C} = \{C_1, \ldots, C_k\}$ such that $C_i \cap C_j = \emptyset$ for all pairs i, j. Assume that the numbers of points in consecutive clusters are $n_1, \ldots, n_k$. Then, the probability that a random point from the dataset belongs to the ith cluster is

$$P(i) := \frac{n_i}{n},$$

where n is the number of all points in the set. Note that $\sum_{i=1}^{k} n_i = n$. Analogously, let another clustering of the same set of points $\mathcal{C}' = \{C'_1, \ldots, C'_{k'}\}$ be composed of clusters with $n'_1, \ldots, n'_{k'}$ points. By $n_{ij'}$ we will denote the number of points assigned to cluster i in clustering $\mathcal{C}$ and cluster j' in $\mathcal{C}'$. Then,

$$P(i, j') := \frac{n_{ij'}}{n}$$

is the probability of randomly choosing a point that belongs to both clusters.

The $\mathcal{VI}$ measure is defined in terms of entropy and joint entropy of the probability distributions defined above. That is, if the entropy of clustering $\mathcal{C}$ is expressed

by:

$$H(\mathcal{C}) := -\sum_{i=1}^{k} P(i) \log_2 P(i),$$

and the joint entropy of two clusterings is given by:

$$H(\mathcal{C}, \mathcal{C}') := -\sum_{i=1}^{k} \sum_{j'=1}^{k'} P(i, j') \log_2 P(i, j'),$$

then the variation of information of the two clusterings is defined as:

$$\mathcal{VI}(\mathcal{C}, \mathcal{C}') := 2H(\mathcal{C}, \mathcal{C}') - H(\mathcal{C}) - H(\mathcal{C}')$$

Quality of dynamic domains

We consider dynamic domains to be structural parts of the protein, which move with respect to each other, but remain internally rigid. In order to assess which method for dynamic domains identification is better, a measure was required that would quantify the quality of a given assignment. We did not find such a scoring function in the literature, and propose the following geometrical measure called *total geometrical variability*:

$$Q := \sum_{i=1}^{k} \max_{\text{pairs of states } (r,s)} \text{RMSD}(D_i^r, D_i^s),$$

where D_i^s is the set of C_α atoms comprising the domain D_i in the state s, and k is the number of domains. Smaller values of Q indicate higher quality.

Note that typically, if a domain is structurally rigid, the maximal RMSD for that domain is smaller than the sum of maximal RMSDs of its two subsets. Therefore the proposed measure favors large, compact domains. It is also worth noting that for a trivial dynamic domain (single residue) the RMSD is undefined. We set its value to 0, although this artificially reduces Q (see *Quality analysis*).

2.5 Results and discussion

We compared ResiCon with two recent methods: GeoStaS ([**romanowska2012determining**]) and PiSQRD ([**potestio2009coarse**]). The latter method represents the class of methods which identify dynamic domains by analyzing a single structure. We used

		R	G	P1	P2
R		0	1.47	3.29	0.21
G		1.47	0	3.76	1.51
P1		3.29	3.76	0	3.29
P2		0.21	1.51	3.29	0

Figure 2.4: Values of $\mathcal{VI}$ for partitionings of the `1d1d` protein molecule. A pair of results produced by PiSQRD which had the highest $\mathcal{VI}$ are denoted by P1 and P2.

a test set comprising 30 NMR-resolved protein structures exhibiting significant mobility, which was previously used in ([**snyder2005clustering**]), ([**kirchner2011objective**]), and in ([**romanowska2012determining**]). The set was initially proposed to examine the efficacy of a method of identifying structurally-stable cores in flexible proteins. These structures often contain a single rigid core with significant geometrical distortions present in peripheral regions (as indicated in [**snyder2005clustering**] and also observable in ResiCon's results).

We have also used ResiCon to analyze a canonical test case – the HIV-1 protease molecule – using an MD trajectory computed with a coarse-grained force field RedMD ([**gorecki36causality**, **gorecki2009redmd**]) as input data. This example shows that ResiCon is capable of finding dynamic domains of a protein whose functionality depends on flexibility and mobility of its rigid parts.

We also explain the so-called *zebra effect* – a peculiar result produced by ResiCon, observed in several cases. This effect is especially strong when a suboptimal numer of clusters is chosen.

Comparative analysis

ResiCon and GeoStaS are both designed to work on an ensemble of structures, and produce a single partitioning into dynamic domains. Both methods impose no assumptions about sampling and order of provided conformations. ResiCon uses maximal local distortions computed over all pairs of frames, indifferent to over- or undersampling of configurations, as long as they are present in the ensemble.

However, the PiSQRD server by default analyzes a single structure and estimates the so-called *low-energy modes*, which are the eigenvectors of the structural covariance matrix (under the canonical ensemble, i.e. assuming Boltzmann distribution of configurations). These low-energy modes are assumed to carry the information relevant to dynamic domains identification. We observed that the
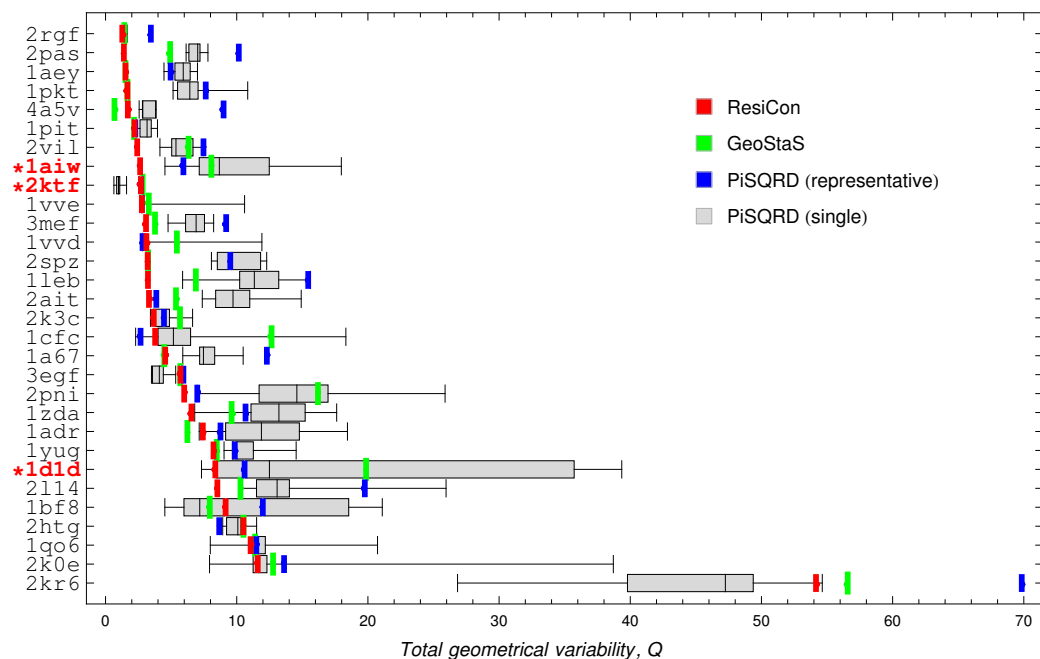
Figure 2.5: Box and whiskers plot of the dynamic domains quality score Q for ResiCon, GeoStaS, PiSQRD. In blue are Q values for PiSQRD's dynamic domains determined from structural covariance matrices (see *Supplementary Materials*).

choice of an input structure influences the results significantly, but there is no definite criterion for choosing the *right* structure for the analysis. The PiSQRD server provided with a PDB file containing NMR models by default finds dynamic domains for the first model.

This might introduce a bias unfavorable for PiSQRD's performance. We, therefore, decided to examine results produced by PiSQRD for every structure in the ensemble in order to compare ResiCon against its full capacity. The best dynamic domains were chosen and presented together with the results produced by ResiCon and GeoStaS in Table 2.1.

PiSQRD also gives a possibility of providing a set of low-energy modes extracted from a structural covariance matrix estimated from a set of structures, but this procedure is not straightforward and requires additional assumptions (see *Supplementary Materials*). We scrutinize the quality of dynamic domains found by PiSQRD from user-provided low-energy modes in section *Quality analysis*.

Because of a large number of models, it was not possible to provide a graph-

ical representation for each partitioning. We therefore focused on values of the agreement measure $\mathcal{VI}$ between results produced by the three methods.

To familiarize the Reader with values of the measure $\mathcal{VI}$, we take a look at two most distant partitionings produced by PiSQRD (denoted by P1 and P2) for an exemplary `1d1d` protein, and how they relate to the results given by ResiCon (R) and GeoStaS (G) (Figure 2.4). The clustering P2 seems to be very similar to the one produced by ResiCon, while P1 gives a sliced-and-diced picture of the protein's mobility. It seems that a partitioning produced by PiSQRD depends on physical properties embedded in a particular configuration. Although NMR models carrying information about dynamic domains may exist, others lead to chaotic partitionings. In order to produce a single clustering, PiSQRD would need a procedure for interpreting physical properties based on an ensemble of configurations.

Nearly 50% of the assignments produced by PiSQRD for `1d1d` are coherent with the clustering given by ResiCon (see *Supplementary Materials*). On the other hand, the result given by GeoStaS differs from all results produced by PiSQRD. In fact, ResiCon and PiSQRD are more coherent than GeoStaS and PiSQRD, which can be expressed by the mean of $\mathcal{VI}$. However, it would be naive to use the mean value as an indicator of self-consistency of PiSQRD. Among partitionings produced by PiSQRD, $N(N-1)/2$ agreements were calculated (where N is the number of models of `1d1d`), whereas the comparison of ResiCon or GeoStaS with PiSQRD gave N values of $\mathcal{VI}$. Consequently, the average value of $\mathcal{VI}$ between PiSQRD clusterings is not directly comparable with the average value of $\mathcal{VI}$, e.g. for ResiCon vs. PiSQRD.

Therefore, to give a better picture of the divergence of results we exploit the fact that $\mathcal{VI}$ is a metric in the space of all clusterings. For a given partitioning we computed the radius of a ball centered at that partitioning, encompassing a certain fraction of the results. In Table 2.2 we provide values of radii of balls which encompass 25, 50 and 75% of the results. In the case of PiSQRD, we constructed balls (for a given percentage of results) centered at each partitioning, and computed the mean values of their radii. The mean of radii does not carry the bias mentioned earlier. Additionally, histograms of $\mathcal{VI}$ for each protein can be found in *Supplementary Materials*.

Quality analysis

The box-and-whisker plot in Figure 2.5 provides a concise picture of quality scores Q (see *Methods*) of the dynamic domains assignments. It clearly shows that beyond a few exceptions ResiCon gives the best results. There are 7 notable exceptions: `4a5v`, `2ktf`, `3egf`, `1adr`, `1bf8`, `2htg` and `2kr6`.

The single dynamic domains found by PiSQRD (blue in Figure 2.5) were produced using low-energy modes, computed as eigenvectors of a structural covariance matrix estimated by superimposing all models in an ensemble on a representative structure. In the *Supporting Materials* we explain how this representative structure is chosen, and show qualities of dynamic domains found by PiSQRD using different methods of estimating the structural covariance matrix. It should be emphasized, however, that these dynamic domains strongly depend on the method of estimating the structural covariance matrix, which is not part of PiSQRD's functionality. Therefore, these results should only be treated as an additional insight into what the user can expect from a more complex analysis of NMR structures.

In the case of `1adr` GeoStaS gave partitionings with lower Q, by introducing a trivial domain (cutting off C- and N-termini – see *Supplementary Materials*). On the other hand, for `4a5v` GeoStaS identified a single domain, which gave a lower value of Q than any other partitioning.

For proteins `2ktf`, `3egf` and `2htg`, PiSQRD produced results with lower Q than ResiCon, by finding numerous small quasi-rigid fragments. However, for `2ktf` more than 25% of assignments found by PiSQRD contained a trivial, single-residue domain (indicated in red, with an asterisk). Although our measure does not penalize for this, we consider such behavior undesirable. Also, note that proteins `3egf` and `2htg` comprise 53 and 27 residues accordingly. It seems that in case of small proteins ResiCon often identifies a single domain, which does not necessarily result in the lowest Q.

The `2kr6` protein is an interesting example. It contains a flexible linker composed of more than 30 residues. As a consequence, partitionings with high values of Q are observed. Only PiSQRD was able to produce lower values of Q, by cutting the linker into many shorter parts. The example of the `2kr6` protein shows, that ResiCon does not consider unstructured regions to be separate dynamic do-

mains. Instead, residues constituting linkers and lacking long-distance contacts, are assigned to dynamic domains which are closest in sequence.

Comments

Although PiSQRD's capability of analyzing a single structure may be considered an advantage, results presented in Table 2.2 and Figure 2.5 show that there is a large discrepancy for different configurations of the same protein. Nevertheless, based on the histograms of $\mathcal{VI}$ presented in *Supplementary Materials* and the values of radii in Table 2.2, we conclude that in most cases results given by ResiCon and PiSQRD are mutually more coherent, than GeoStaS and PiSQRD (e.g. `1cfc`, `1qo6`, `1vvd`, `1vve`, `1yug`, `2k3c`). Notable exceptions are: `2rgf`, `2pas`, `3mef` and `1zda`. For these proteins ResiCon did not find any significant structural transitions and achieved the lowest value of Q (see Figure 2.5) by assigning a single domain.

Dynamic domains found by ResiCon are generally larger (particularly: `2k3c`, `1d1d`, `2k0e` and `2kr6`). Conversely, GeoStaS often allocates flexible N- and C-terminal parts (e.g. `1adr`, `1qo6`, `1vve`, `3egf`) as quasi-rigid parts. The size of dynamic domains is also the main difference between ResiCon and PiSQRD. In case of small, static proteins (such as `1aey`, `1pkt`, `1pit`, `2spz`, `2ait`, `3egf` and `1zda`), PiSQRD identifies numerous small and often trivial dynamic domains. ResiCon on the other hand detects no significant conformational changes (by analyzing an ensemble), and assigns a single dynamic domain. We observe that in many cases ResiCon identified a single domain which had the lowest value of Q among all partitionings (see `2rgf`, `2pas`, `1aey`, `1pkt`, `1pit`, `2vil`, `1aiw`, `3mef`, `2spz`, `1leb`, `2ait`, `2pni`, `2l14`). In these cases ResiCon correctly detected that no significant transitions were present in the protein. Therefore, unlike PiSQRD and GeoStaS, ResiCon can reliably indicate whether conformational changes occur in an ensemble of structures.

It is also noteworthy that ResiCon does not employ any post-processing procedures. This keeps the algorithm simple and clean, but results in a discontinuity of certain partitionings (see `4a5v`, `1bf8` and `1vvd`), which we refer to as the *zebra effect*. Although this seems to be an artifact of the clustering algorithm, we will take a closer look at this effect and show that it may also carry valuable information

referring to the protein's dynamics.

HIV-1 protease

An analysis of an MD trajectory of the HIV-1 protease showcases ResiCon's capabilities. This protein undergoes substantial conformational changes associated with opening/closing of its structural parts, so-called *flaps* ([**hamelberg2005fast**]). A database of X-ray-resolved structures, representing configurations which the protease can attain is available ([**vondrasek2002hivdb**]).

We examined a set of configurations of the HIV-1 protease acquired from a simulation carried out using the RedMD package ([**gorecki2009redmd**]). This coarse-grained force field was designed to simulate intramolecular motions in proteins and nucleic acids. It has correctly predicted the flap-opening motion in the HIV-1 protease, which is known biological fact ([**hamelberg2005fast**]), and was independently confirmed by an all-atom MD simulation ([**sadiq2010explicit**]). Roughly, about 1% of the trajectory seems to exhibit significant conformational transitions (flap opening events). This is a typical scenario in MD simulations, where a transition between meta-stable states is swift and fairly short. We needed a set of representative structures in order to find dynamic domains using ResiCon. From the whole trajectory, we extracted a set of 200 configurations using a generic procedure facilitating the Principal Component Analysis (PCA), implemented in the R programming language (in the `bio3d` package ([**grant2006bio3d**])) – see *Supplementary Materials*.

Values of $\chi_{\min}$ for different numbers of clusters are given in Figure 2.3. The optimal number of clusters according to our procedure is 4. Figure 2.6 depicts the dynamic domains found by ResiCon and two representative configurations of the protease, as well as results from GeoStaS and PiSQRD. Because the sample of configurations was drawn according to the Boltzmann distribution, we were able to straightforwardly estimate the structural covariance matrix and provide PiSQRD with well-founded low-energy modes, and acquire a high-quality partitioning into two domains. Results produced by GeoStaS were acquired from the whole trajectory of the protease.

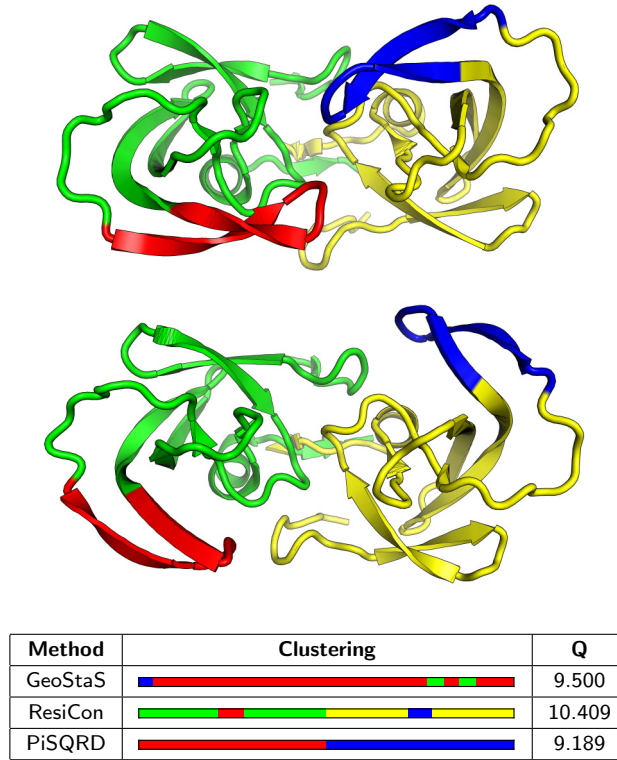| Method	Clustering	Q
GeoStaS		9.500
ResiCon		10.409
PiSQRD		9.189

Figure 2.6: Two representative conformations of HIV-1 protease with flaps being closed and open, and summary of results produced by GeoStaS, ResiCon and PiSQRD. Structures are colored according to the dynamic domains detected by ResiCon.

Dynamic domains identified by ResiCon have the highest value of the Q measure. Note that ResiCon does not try to minimize Q, but to produce a clustering with the optimal value of the χ_{min} indicator. Consequently, at the cost of a slightly higher Q (ca. $1\mathring{A}$) we arrive at a partitioning which corresponds to the biologically relevant sub-division of the protein. It is also worth mentioning that ResiCons partitioning into $k = 2$ clusters incidentally leads to an identical assignment as the one produced by PiSQRD.

Motions of the flaps between the closed and open states are crucial in the functionality of the HIV-1 protease. It is also known that throughout their motion they remain quasi-rigid ([**freedberg2002rapid**]). Therefore, using ResiCon, we successfully extracted a simplified picture of the mobility of HIV-1 protease, which agrees with experimental knowledge.

Alternative partitioning into 6 clusters (see Figure 2.7) also deserves interest.
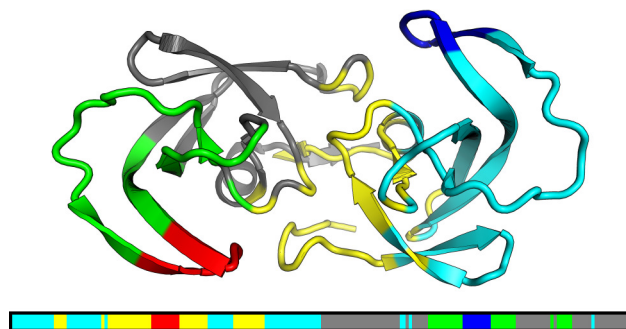
Figure 2.7: More subtle division emerges when the number of clusters is set to 6.

Apart from the flaps, additional dynamic domains resembling "arms" carrying the flaps can be observed. Note that conversely to the case of the four dynamic domains, the quasi-static regions are similar, but not ideally reflective. This indicates that motions of the two centrally symmetric sub-units of the HIV-1 protease in the provided trajectory were slightly different.

It can be also seen that domains are discontinuous. Especially residues belonging to the N-terminal lobe (yellow) and to the arm (cyan) are interleaved. In the next section we will analyze this effect, and show that this partitioning also carries valuable information.

The zebra effect

Dynamic domains found by ResiCon may include residues that seem to be pulled out from another quasi-rigid fragment. This is indicated as discontinuities (stripes) in Figure 2.7. Let us take a look at an example of such an extracted residue and try to understand the source of this effect. ResiCon assigned the ASN-83 residue to the arm (gray) dynamic domain (see Figure 2.8). However, its sequential neighbors belong to the lobe domain (yellow). The reason of this discontinuity is that, although ASN-83 has peptide bonds with its yellow neighbors, they are outweighted by contacts with the gray residues (see Figure 2.8). The discontinuity suggests that throughout the trajectory ASN-83 remained docked in the cavity of the arm domain, while its peptide bonds formed an axis of a hinge.

It seems that the zebra effect is not accidental, and that it can be used to find residues which act as "pivot points". However, this effect is volatile and depends on
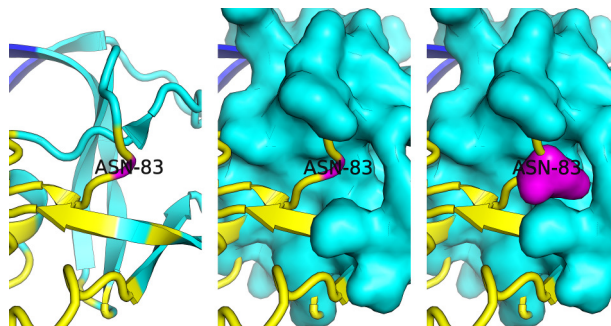
Figure 2.8: From the analysis carried out by ResiCon one can see that throughout the trajectory ASN-83 (magenta) not only remains close to residues in the N-terminal arm domain (gray) but also fills a cavity in that dynamic domain.

small fluctuations in the input. Residues are assigned to dynamic domains based on the fuzzy membership matrix. At certain positions values of membership to different domains may be almost equal, and when discretized cause emergence of stripes. Therefore, to strengthen the identification of these "pivot point" residues, a more thorough analysis of the membership matrix is required.

2.6   Conclusions

Typically, protein structures are flexible and mobile, and their conformational changes may be crucial in facilitating signaling and metabolical processes in which they participate. Breaking a structure into dynamic domains may be compared to discovering gears and levers connected by cogs and pegs analogous to those found in classical machines. Such analyses allow us to better understand molecular mechanisms responsible for biological functions (e.g. [**taylor2013classification**]), facilitate MD simulations in large time-scales with simplified forcefields (e.g. [**sinitskiy2012optimal**]), or discover potential binding sites when designing inhibitors (e.g. [**zhang2009defining**]).

We have presented a universal tool for discovering dynamic domains in proteins. ResiCon is capable of analyzing a single structure, or an NMR ensemble of structures provided in a PDB format. It can also be applied to the set of independently obtained structures (e.g. X-ray crystallographic structures obtained under different conditions). In any case, it provides a complete set of results comprising partitioning of the molecule into dynamic domains with additional highlighting of residues

composing hinges and interfacial regions. ResiCon also provides an indicator of partitioning quality, and suggests the optimal number of domains.

We tested our method using the reference set of NMR structures. It is comparable or better than the recently developed GeoStaS and PiSQRD. Apart from giving more compact dynamic domains, it is also capable of distinguishing structures composed of a single quasi-rigid region.

We have made ResiCon available online.[1] To make the analysis feasible and limit the number of uploads, queries may be rerun with changed parameters. Also, all queries and results are stored on our server.[2]

---

[1]http://dworkowa.imdik.pan.pl/EP/ResiCon
[2]These two features require registration for security reasons.

PDB code	Method	Clustering	Q	PDB code	Method	Clustering	Q
2rgf	GeoStaS		1.430	2k3c	GeoStaS		5.678
	ResiCon		1.275		ResiCon		3.676
	PiSQRD		1.421		PiSQRD		3.422
2pas	GeoStaS		4.926	1cfc	GeoStaS		12.658
	ResiCon		1.398		ResiCon		3.780
	PiSQRD		6.141		PiSQRD		2.294
1aey	GeoStaS		1.520	1a67	GeoStaS		4.478
	ResiCon		1.520		ResiCon		4.548
	PiSQRD		4.455		PiSQRD		5.887
1pkt	GeoStaS		1.670	3egf	GeoStaS		5.700
	ResiCon		1.670		ResiCon		5.713
	PiSQRD		5.147		PiSQRD		3.499
4a5v	GeoStaS		0.678	2pni	GeoStaS		16.195
	ResiCon		1.698		ResiCon		5.994
	PiSQRD		2.561		PiSQRD		6.993
1pit	GeoStaS		2.177	1zda	GeoStaS		9.610
	ResiCon		2.209		ResiCon		6.549
	PiSQRD		2.171		PiSQRD		6.798
2vil	GeoStaS		6.329	1adr	GeoStaS		6.241
	ResiCon		2.414		ResiCon		7.417
	PiSQRD		4.146		PiSQRD		7.139
1aiw	GeoStaS		8.066	1yug	GeoStaS		8.479
	ResiCon		2.636		ResiCon		8.230
	PiSQRD		4.527		PiSQRD		9.033
2ktf	GeoStaS		2.822	1d1d	GeoStaS		19.862
	ResiCon		2.693		ResiCon		8.368
	PiSQRD		0.636		PiSQRD		7.313
1vve	GeoStaS		3.299	2l14	GeoStaS		10.284
	ResiCon		2.767		ResiCon		8.525
	PiSQRD		2.767		PiSQRD		10.411
3mef	GeoStaS		3.770	1bf8	GeoStaS		7.947
	ResiCon		3.086		ResiCon		9.153
	PiSQRD		4.765		PiSQRD		4.519
1vvd	GeoStaS		5.439	2htg	GeoStaS		10.514
	ResiCon		3.110		ResiCon		10.514
	PiSQRD		2.821		PiSQRD		8.665
2spz	GeoStaS		3.244	1qo6	GeoStaS		11.403
	ResiCon		3.212		ResiCon		11.057
	PiSQRD		8.071		PiSQRD		7.985
1leb	GeoStaS		6.885	2k0e	GeoStaS		12.771
	ResiCon		3.234		ResiCon		11.604
	PiSQRD		5.874		PiSQRD		7.927
2ait	GeoStaS		5.370	2kr6	GeoStaS		56.578
	ResiCon		3.310		ResiCon		54.160
	PiSQRD		3.871		PiSQRD		26.834

Table 2.1: Summary of results produced by GeoStaS, ResiCon and PiSQRD. Dynamic domains shown for PiSQRD are those for which the lowest value of Q was achieved.

	ResiCon vs. GeoStaS	ResiCon vs. PiSQRD			GeoStaS vs. PiSQRD			PiSQRD vs. PiSQRD		
	$\mathcal{VI}$	$r_{25\%}$	$r_{50\%}$	$r_{75\%}$	$r_{25\%}$	$r_{50\%}$	$r_{75\%}$	$\langle r_{25\%}\rangle$	$\langle r_{50\%}\rangle$	$\langle r_{75\%}\rangle$
2rgf	0.62	0.45	0.48	0.57	0.26	0.26	0.32	0.01	0.14	0.20
2pas	1.94	2.43	2.45	2.47	2.00	2.08	2.11	0.61	0.96	1.23
1aey	0.00	1.99	2.33	2.37	1.99	2.33	2.37	0.77	1.10	1.52
1pkt	0.00	2.07	2.24	2.36	2.07	2.24	2.36	1.38	1.63	1.84
4a5v	1.00	1.08	1.12	1.19	1.53	1.54	1.60	0.24	0.29	0.38
1pit	1.59	1.76	2.05	2.14	2.42	2.55	2.74	1.35	1.70	1.94
2vil	1.89	1.19	1.39	1.89	2.31	2.49	2.67	1.11	1.36	1.57
1aiw	1.64	1.66	1.80	2.13	2.58	2.69	2.77	1.30	1.53	1.90
2ktf	0.80	1.69	2.01	2.06	2.15	2.36	2.48	0.93	1.17	1.74
1vve	1.05	0.00	0.00	0.00	1.05	1.05	1.05	0.62	0.62	0.62
3mef	0.32	1.30	1.43	1.50	1.21	1.32	1.40	0.64	0.86	1.09
1vvd	1.24	0.21	0.21	0.21	1.26	1.26	1.26	0.31	0.31	0.31
2spz	0.13	1.42	1.62	1.94	1.37	1.59	1.96	1.28	1.60	1.73
1leb	0.61	2.05	2.27	2.41	2.33	2.37	2.45	0.92	1.21	1.52
2ait	0.71	1.26	1.65	1.82	1.49	1.67	1.99	0.99	1.31	1.52
2k3c	1.86	0.82	0.97	0.98	1.61	1.69	1.69	0.16	0.42	0.63
1cfc	1.60	0.41	0.44	0.97	1.41	1.48	1.86	0.77	0.92	1.30
1a67	1.07	1.90	1.96	2.06	2.22	2.38	2.69	1.58	1.93	2.12
3egf	0.14	1.30	1.40	1.43	1.29	1.41	1.45	0.63	0.91	1.14
2pni	1.60	1.16	1.58	1.78	2.00	2.25	2.35	1.18	1.44	1.65
1zda	0.79	1.37	1.55	1.96	0.83	1.28	1.78	1.26	1.54	1.75
1adr	0.84	0.80	1.30	1.81	1.11	1.64	2.22	1.19	1.57	1.94
1yug	1.79	0.90	1.00	1.31	1.98	2.12	2.15	0.70	0.99	1.29
1d1d	1.47	0.00	0.21	3.15	1.47	1.51	3.80	1.08	3.08	3.17
2l14	0.55	0.49	0.57	0.75	0.91	0.98	1.18	0.61	0.71	0.95
1bf8	1.61	0.74	1.27	3.22	1.69	1.89	4.20	1.13	1.74	2.99
2htg	0.00	1.36	1.58	1.91	1.36	1.58	1.91	0.84	1.13	1.30
1qo6	0.88	0.63	0.72	0.88	0.87	0.98	1.18	0.57	0.71	0.94
2k0e	0.97	0.59	0.66	0.82	1.05	1.16	1.31	0.72	0.87	1.03
2kr6	1.11	0.79	0.81	0.84	0.85	0.87	0.93	0.61	0.68	0.82

Table 2.2: Discrepancies in assignments expressed by radii of balls encompassing 25%, 50% and 75% of results. The ordering of the results is the same as in Figure 2.5, i.e. best-scoring results are first.

Chapter 3

Towards the identification of molecular cogs

Computer simulations of molecular systems allow determination of microscopic interactions between individual atoms or groups of atoms, as well as studies of intramolecular motions. Nevertheless, description of structural transformations at the mezoscopic level and identification of causal relations associated with these transformations, is very difficult. Structural and functional properties are related to free energy changes. Therefore, in order to better understand structural and functional properties of molecular systems, it is required to deepen our knowledge of free energy contributions arising from molecular subsystems in the course of structural transformations.

The method presented in this work quantifies the energetic contribution of each pair of atoms to the total free energy change along a given collective variable. Next, with the help of a genetic clustering algorithm, the method proposes a division of the system into two groups of atoms referred to as *molecular cogs*. Atoms which cooperate to push the system forward along a collective variable are referred to as *forward cogs*, and those which work in the opposite direction as *reverse cogs*.

The procedure was tested on several small molecules for which the genetic clustering algorithm successfully found optimal partitionings into molecular cogs. The primary result of the method is a plot depicting the energetic contributions of the identified molecular cogs to the total Potential of Mean Force (PMF) change. Case-studies presented in this work should help better understand the implications of our approach, and were intended to pave the way to a future, publicly available

implementation.

INTRODUCTION

Molecular dynamics (MD) simulation methods and advanced algorithms for calculating free energy bring us closer to predicting the physical properties of biomolecules [**woo2005calculation**, **chipot2014frontiers**, **comer2014adaptive**]. However, computer simulations are not limited to interpreting experimental results. *In silico* one may also process MD data which can provide much more detailed information than that accessible in any experiment. The key to a deeper understanding of complex molecular systems is the extraction of valuable information from data produced in such simulations.

Biomolecules carry out their functions through conformational transitions between meta-stable states. Such phenomena can be simplified and described by a selected reaction coordinate which, in many cases, is a collective variable [**chipot2007free**]. A valuable result of many simulation procedures (such as Umbrella Sampling [**torrie1977nonphysi**], Thermodynamic Integration [**kirkwood1935statistical**], Adaptive Biasing Force [**comer2014ada**], and others[**gumbart2012standard**]), is a free energy profile, the so-called Potential of Mean Force (PMF) [**kirkwood1935statistical**]. In this study, the discussion is limited to one-dimensional PMFs, although it should be noted that the aforementioned methods can be formulated more generally and produce multi-dimensional free energy profiles. Of course, a one-dimensional PMF may be an oversimplification of what is going on in a complex system, but a meaningful collective variable usually leads to a free energy profile which makes the complicated transition more comprehensible [**periole2012structural**, **chipot2014frontiers**]. However, what the PMF does not provide is the information about what drives transitions between meta-stable states.

Various attempts to understand and describe the internal mechanics of molecular systems have already been reported [**arnautova2011development**, **bao2009protein**, **chiang2013molecular**, **lavery2007protein**, **sacquin2014motions**, **van1997engineering**, **seifert2013protein**]. We did not, however, find any general-purpose approach for studying a broader spectrum of cases, and in particular a methodology explaining the cause of a transition in terms of free energy contributions arising from certain parts of a molecule.

In this study we propose a new approach of analyzing the tendencies of a

molecular system to undergo a selected structural transition. The main idea is to look at a shift along a collective variable as an effect of two opposing tendencies generated by interactions within the molecule. Our method indicates two groups of atoms – referred to as *molecular cogs* – which, through cooperative interactions, are the source of these tendencies. For this purpose, we construct undirected graphs with weights between nodes expressed by energetic contributions to the free energy change arising from pair-wise interactions between atoms. These graphs are then partitioned into subsystems – corresponding to molecular cogs – using a genetic clustering algorithm. We present results for small, model systems which served as case-studies for the identification of molecular cogs and for testing if their functioning agrees with our intuitions.

METHODOLOGY

Decomposition of the Helmholtz free energy, A, was investigated in the past, most notably by Karplus and coworkers[**boresch1995meaning**, **brady1995decomposition**]. The aim was the determination of contributions to the free energy coming from components comprising the potential energy of the system. These potential energy components might come from different interaction types, or from cooperation between subsystems of the whole molecule. The strategy of expressing the potential energy of a system as a sum of terms, $U = \sum_i U_i$, and computing the contribution of each of these terms to the free energy is ineffective because additivity in the potential energy components does not imply additive contributions to the entropy[**dill1997additivity**].

Two attempts of describing contributions to the free energy coming from parts of the system are worth mentioning. The first one was an approximate approach based on Free Energy Perturbation, in which higher order terms were neglected [**brady1995decomposition**]. These terms are not, however, negligible, which severely limits the applicability of this method. An alternative route, employing Thermodynamic Integration, was also explored [**boresch1995meaning**]. However, this approach is also limited, namely – values of the free energy contributions depend on the choice of the integration path.

We propose a different approach, in which pair-wise energetic contributions to

the free energy are readily attained. Alas, as in the aforementioned methods of free energy decomposition, our current formulation struggles with a description of the entropic contributions, and at the present is not included in our method. The current implementation is primarily applicable to small molecules for which the role of entropy in structural transitions is negligible (see an example of a PMF in the *Results* section).

Our analysis originates from the following formula [**carter1989constrained**]:

$$A'(\xi^*) = \langle m_\xi \nabla U \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*} - \langle \mathbf{v} \cdot \nabla (m_\xi \nabla \xi) \mathbf{v} \rangle_{\xi^*}, \tag{3.1}$$

where A' is the derivative of the free energy with respect to the collective variable ξ, $\mathbb{M}^{-1}$ is a diagonal matrix of inverse masses, and $\mathbf{v}$ are velocities. The term m_ξ is defined by:

$$m_\xi := \left[\sum_i m_i^{-1} \left(\frac{\partial \xi}{\partial \mathbf{x}_i} \right)^2 \right]^{-1} \tag{3.2}$$

and can be interpreted as inertia of an effective point mass moving along the ξ coordinate.

On the right-hand side of Equation (3.1) we have, respectively, energetic and entropic contributions to the free energy change, both expressed as conditional averages, with the collective variable fixed at ξ^*. In the *Supporting Information* we briefly explain how these averages are estimated *via* constrained Molecular Dynamics (cMD) simulations, and highlight an important limitation of this method.

Note, that if we consider the potential energy as a sum of interaction components:

$$U = U^{ele} + U^{vdw} + U^{bond} + \ldots = \sum_i^I U^I$$

the energetic contribution in Equation (3.1) maintains this additivity:

$$\langle m_\xi \nabla U \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*} = \sum_i^I \langle m_\xi \nabla U^I \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*}, \tag{3.3}$$

as was noted by Chipot et al. [**chipot2007free**].

In our numerical experiments we used a force field with the following set of interaction types:

- *ele* (electrostatic interactions);

- *vdw* (van der Waals interactions);

- *bond* (2-body bonded interactions);

- *angl* (3-body angle interactions);

- *tors* (4-body torsional interactions).

We shall also consider:

- *nbd* (non-bonded interactions, the sum of *ele* and *vdw* interactions);

- *conf* (conformational interactions, composed of *bond*, *angl* and *tors* interactions);

- *total* (the sum of *nbd* and *conf* interactions).

Decomposition of the energetic contribution

Our purpose was to identify the molecular cogs, i.e. sets of atoms, which cooperate and push the whole system forward/backward along a reaction coordinate. We approached this problem by converting it into the task of finding clusters in a graph. Nodes in such a graph correspond to atoms, whereas edges represent cooperation of pairs of atoms. A natural measure of such cooperation can be introduced by taking into account the energetic contribution to the free energy in Equation (3.1).

Electrostatic, van der Waals and two-atom chemical bond interactions can be readily transformed into weights in the graph. For example, electrostatic interactions between atoms α and β lead to the following contribution:

$$c_{\alpha\beta}^{ele}(\xi^*) := \frac{1}{m_\alpha} \left\langle m_\xi \frac{\partial U^{ele}}{\partial \mathbf{x}_\alpha} \cdot \frac{\partial \xi}{\partial \mathbf{x}_\alpha} \right\rangle_{\xi^*} + \frac{1}{m_\beta} \left\langle m_\xi \frac{\partial U^{ele}}{\partial \mathbf{x}_\beta} \cdot \frac{\partial \xi}{\partial \mathbf{x}_\beta} \right\rangle_{\xi^*}. \qquad (3.4)$$

Alas, n-body potential energy components cannot, in general, be decomposed into a sum of pair interactions. However, because the free energy contributions are of the form $\nabla U \cdot \mathbb{M}^{-1} \nabla \xi$, such decomposition is not required.

To clarify, let us consider a 3-body potential energy component, e.g. $U^{angl}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma)$. The weight $c_{\alpha\beta}^{angl}$ of an edge joining nodes α and β represents the contribution coming from atoms α and β which interact *via* U^{angl}, while the γ atom is kept at a fixed position. By asserting that the sum of pair-wise contributions is equal to the

total contribution arising from $U^{angl}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma)$, i.e.:

$$\sum_{i>j} c_{ij}^{angl}(\xi^*) = \sum_{i=\alpha,\beta,\gamma} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^{angl}}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*},$$

we arrive at the following definition of the cooperation term:

$$c_{\alpha\beta}^{angl}(\xi^*) := \frac{1}{2} \sum_{i=\alpha,\beta} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^{angl}}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*},$$

where the $1/2$ means that the overall *angl* contribution is evenly distributed between the cooperation terms: $c_{\alpha\beta}^{angl}$, $c_{\beta\gamma}^{angl}$ and $c_{\alpha\gamma}^{angl}$.

For the general case of an n-body energy component, U^I, we propose the following definition of the cooperation term between atoms α and β:

$$c_{\alpha\beta}^{I}(\xi^*) := \frac{1}{n-1} \sum_{i=\alpha,\beta} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^I}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*}.$$

Note that we propose to distribute the cooperation evenly among all pairs of atoms involved in the interaction I. With this definition we constructed graphs for all interaction types, I.

A matrix, $\mathcal{C}^I(\xi^*)$, such that $[\mathcal{C}^I(\xi^*)]_{\alpha\beta} := c_{\alpha\beta}^I(\xi^*)$ constructed for a particular value of the collective variable ξ^* and the interaction type I is referred to as the *transient cooperation matrix for* I (see Figure 3.1). For convenience, the transient cooperation matrix for the *total* interaction type does not contain any superscipt:

$$\mathcal{C}(\xi_i) := \sum_I \mathcal{C}^I(\xi_i).$$

Note that the sum $\sum_{\alpha>\beta}[\mathcal{C}(\xi_i)]_{\alpha\beta}$, is the overall energetic contribution to $A'(\xi_i)$ in Equation (3.1).

Comment on the entropic contribution

The second term on the right-hand side of Equation (3.1) does not explicitly depend on U. It is possible to propose pair-wise contributions of this quantity in order to construct matrices which in turn could be used as input for the genetic clustering procedure described below. However, calculating these entropic contributions would require computation of the Hessian of the collective variable (the matrix of

second-order derivatives with respect to atom coordinates) in each step of the simulation, which would make the computational cost of the procedure prohibitively high. We speculate on a possible solution of this problem in the *Conclusions* section.

Clustering

Data clustering is the procedure of finding disjoint subsets (clusters) of objects that share a high affinity, and are dissimilar to objects from other clusters. To carry out a clustering procedure we need a pair-wise affinity measure; thus, the data set is often represented by a weighted graph. Such a graph is encoded by an *affinity matrix* with elements corresponding to weights between nodes. In this study we were looking for groups of atoms whose cooperation decreased (increased) the free energy, thus pushing (pulling) the whole system forward (backward) along a reaction coordinate. Thus, we took the pair-wise energetic cooperation between atoms introduced earlier for the affinity measure.

We will refer to the clusters identified in a transient cooperation matrix as *transient molecular cogs*. The free-energy-reducing group is referred to as *forward cogs*, and the second group (which, conversely, increases the free energy gap) as *reverse cogs*. We, therefore, assume that a step taken by the molecule along a reaction coordinate is a consequence of a resultant tendency generated by the molecular cogs. But cross-cooperation between atoms from different groups can also occur, and we designate this unwanted effect 'gear grinding' (see the *Results* section).

Affinity Propagation algorithm

A considerable challenge associated with clustering is that different procedures produce different results. Therefore, it was crucial that the *right* clustering algorithm was chosen. We anticipated that the optimal algorithm should recover molecular cogs with as little gear grinding as possible, but we could not estimate how large this effect might be. To gain some insight into how molecular cogs might look like, we chose a clustering algorithm which is known to be successful in other applications.

Most clustering algorithms assume a non-negative affinity measure, whereas cooperation between atoms can be negative as well as positive. Among those al-

gorithms which accepted negative values, Affinity Propagation (AP) was the procedure found to be the most promising[**frey2007clustering**]. The AP algorithm searches for *exemplars*, i.e. nodes within the graph around which non-exemplars are grouped, thus forming a cluster. It is an iterative procedure, where, in each step, "messages" between nodes are exchanged to designate exemplars and their followers.

In our case, the sign of cooperation is arbitrary i.e. it depends on the direction of the reaction coordinate, ξ. It was crucial to ensure that molecular cogs found for a reaction coordinate $\xi' := -\xi$ were the same as those found for ξ. We resolved this problem by carrying out two clusterings with the AP method: one for $\mathcal{C}(\xi)$, and the second one for $-\mathcal{C}(\xi)$, in which the sign of all elements is changed, such as produced for ξ'. Description of the AP algorithm and our method for merging two clusterings can be found in *Supporting Information*.

One of the drawbacks of the AP method, similarly as in other clustering procedures, is that it requires several parameters, which influence the outcome. It appeared that small variations in parameters lead to markedly different results, and it was difficult to find a single set of parameters suitable for all cases (see the discussion in the *Supporting Information*). Nevertheless, the AP method gave us a first estimation of how molecular cogs look like, and it appears to be sufficiently good to initialize the genetic clustering procedure (see the *Genetic clustering* section).

Objective function

Molecular cogs identified by the AP algorithm were laden with low gear grinding, but that was not always a valuable finding. We noticed that in cases where no good partitioning was achievable, the AP method produced one-element clusters, which we considered to be artificial and incorrect.

The AP algorithm performs a clustering which maximizes the so-called *net similarity* [**frey2007clustering**]. This objective function, although helpful in many applications, does not carry any physical meaning. Note that cooperation, which we used to express affinity between atoms, translates into free energy differences, and thus into the tendency of the whole system to move along a collective variable.

Molecular cogs should not only be laden with low gear grinding, but also coopera-
tion generated by the cogs should cover the overall free energy change as much as
possible.

To clarify, given a cooperation matrix $\mathcal{C}$, the sum of all negative (positive)
elements is the overall tendency of the system to go forward (backward) along a
reaction coordinate. It is desired for forward (reverse) cogs to cover as much of this
overall tendency as possible. When there is only one atom in a cluster, there is no
coverage, even though gear grinding might be small.

Let us denote by FC the set of atom indices, which were assigned to the forward
cogs, and analogously by RC the set of atoms assigned to the reverse cogs. Given a
cooperation matrix, $\mathcal{C} = [c_{ij}]$, the following three quantities are useful in measuring
the quality of molecular cogs:

- Forward Cogs Rate:

$$\text{FCR} := \frac{\sum_{i,j \in \text{FC}} c_{ij}}{\sum_{c_{ij} < 0} c_{ij}} \tag{3.5}$$

- Reverse Cogs Rate:

$$\text{RCR} := \frac{\sum_{i,j \in \text{RC}} c_{ij}}{\sum_{c_{ij} > 0} c_{ij}} \tag{3.6}$$

- Gear Grinding Rate:

$$\text{GGR} := \frac{\sum_{\substack{i \in \text{FC} \\ j \in \text{RC} \\ c_{ij} > 0}} c_{ij}}{\sum_{c_{ij} > 0} c_{ij}} + \frac{\sum_{\substack{i \in \text{FC} \\ j \in \text{RC} \\ c_{ij} < 0}} c_{ij}}{\sum_{c_{ij} < 0} c_{ij}} \tag{3.7}$$

We assumed that both positive and negative elements exist in the cooperation
matrix, $\mathcal{C}$, i.e. that $\sum_{c_{ij} < 0} c_{ij} \neq 0$ and $\sum_{c_{ij} > 0} c_{ij} \neq 0$. In cases in which the
denominator is null, we substitute the whole fraction by 0.

FCR is the ratio of the contribution captured by the forward cogs to the total
forward propensity found in $\mathcal{C}$. RCR is analogous, and both these measures have
a maximum value of 1. Magnitude of misplaced contributions between atoms from
different clusters is indicated by GGR, which has a maximum value of 2.

We propose the following molecular cogs quality measure:

$$\text{SCORE}(\text{FC}, \text{RC}, \mathcal{C}) := 0.5(\text{FCR} + \text{RCR} - \text{GGR}), \tag{3.8}$$

and SCORE := 0 for cases when any of the cogs consists of a single atom. The above scoring function was used in our genetic clustering procedure (described in the following section). See Figure 3.2 for an example of a clustering which maximizes the SCORE.

SCORE equals 1 if the corresponding molecular cogs cover the whole propensity of the system to move along a reaction coordinate, and no gear grinding occurs. We observed that a SCORE less than 0.5 often indicates that the subdivision of the system into molecular cogs is noisy.

Note that a trivial partitioning in which all atoms are assigned to one group, e.g. FC, indicates that the system as a whole has a tendency to move forward. We denote such trivial partitionings as "all FC" and "all RC". The maximal value of SCORE is then 0.5, and although this is unfavorable, in many cases a trivial partitioning had the highest SCORE. In such cases, there was no convenient clustering into two groups, which means that the gear grinding was high.

Genetic clustering algorithm

Genetic algorithms are widely used in optimization problems when potential solutions are readily assessed and codified. They emulate the process of natural selection, promoting solutions – referred to as *specimens* – with higher scores. The codification of a solution is treated as its chromosome, which allows for *mutation* of a solution, and *crossover* with other chromosomes, thus "spawning" new specimens. At the end of each iteration of a genetic algorithm, there is a stage called *selection*, during which low-scoring solutions have a lesser chance of "survival".

We used the genetic clustering algorithm to find molecular cogs with the highest SCORE, as defined in Equation (3.8) (see [**cole1998clustering**] for a good introduction to genetic clustering). The partitioning of a molecular system into molecular cogs was encoded by an array of numbers from the set $\{-1, 0, 1\}$, where the ith element of the array corresponded to the ith atom in a molecule. Values -1 and 1 translate into assignments to the forward and reverse cogs, respectively. An atom not belonging to any cluster was tagged by 0; this occurred when the atom did not cooperate with any other atom (see Figure 3.3).

To initialize the genetic clustering procedure, the AP algorithm was executed

by applying five different methods of setting the diagonal elements in the affinity matrix (see *Supporting Information*). Next, it adds the "all FC" and "all RC" trivial solutions to the starting set. Following this, solutions are iteratively chosen, mutated and added to the set, until the starting population contained 200 candidate solutions.

Once the initial set is generated, the genetic procedure repeated the following steps:

1. Compose (randomly) 50 pairs of solutions to generate offspring using the crossover procedure.

2. Select (randomly) 20 solutions to generate offspring using the mutation procedure.

3. Calculate SCORE for the offspring and add it to the population.

4. Draw (randomly) 200 solutions from the population and discard the rest.

5. Choose the best scoring solution and check if there is any improvement in the SCORE. If there was none for 10 consecutive iterations, return the best solution. Otherwise, return to 1.

All random selections are done without repeats, so that a given solution with SCORE s is chosen with a probability proportional to e^{2s}. Although the number of parameters required in the genetic algorithm is daunting, changes to the majority of these parameters only influence the speed of arriving at the optimal solution (see the *Supporting Information* for a more detailed discussion of the parameter's influence on the outcome).

Figure 3.4 shows that the genetic clustering finds the optimal solution for a range of transient cooperation matrices for *nbd* (complete results can be found in the *Results* section). The optimal solutions were found by means of a brute-force search, i.e. by producing all possible partitionings and calculations of their SCOREs. It is worth noting that in all cases there was a singular solution with the highest SCORE.

Trapezoidal rule for integrating A

The free energy difference between ξ_X and ξ_Y can be expressed as an integral:

$$\Delta A = \int_{\xi_X}^{\xi_Y} A'(\xi^*)d\xi^*. \tag{3.9}$$

The free energy derivative, $A'(\xi^*)$, can be estimated at a given ξ^* from a cMD simulations *via* Equation (3.1). The integral in Equation (3.9) can then be calculated using the trapezoidal rule [**press2007numerical**]:

$$\widehat{\Delta A} \approx \frac{\Delta\xi}{2} \left\{ \widehat{A'(\xi_1)} + 2\widehat{A'(\xi_{i+1})} + \ldots + 2\widehat{A'(\xi_{M-1})} + \widehat{A'(\xi_M)} \right\}, \tag{3.10}$$

where $\widehat{A'(\xi_i)}$ denotes an estimate of the free energy derivative at ξ_i. This requires independent cMD simulations at M values of the collective variable, $\{\xi_i\}_{i=1}^M$, with the grid size of $\Delta\xi$.

Each estimate $\widehat{A'(\xi_i)}$ has a corresponding variance, $\sigma^2[\widehat{A'(\xi_i)}]$, which in turn propagates to the variance of the integral estimate, i.e. $\sigma^2[\widehat{\Delta A}]$. In the following section we explain how we estimated the variances $\sigma^2[\widehat{A'(\xi_i)}]$. Because estimates $\widehat{A'(\xi_i)}$ are independent, the variance of the whole integral follows from Equation (3.10) straightforwardly:

$$\sigma^2[\widehat{\Delta A}] = \frac{\Delta\xi^2}{4} \left\{ \sigma^2[A'(\xi_1)] + 4\sigma^2[A'(\xi_2)] + \ldots + 4\sigma^2[A'(\xi_{M-1})] + \sigma^2[A'(\xi_M)] \right\} \tag{3.11}$$

Details concerning estimation of the variances $\sigma^2[A'(\xi_1)]$ using a bootstrapping procedure can be found in the *Supporting Information*.

The same integration rule applies to all elements of the transient cooperation matrices. A matrix in which every element is a result of the above numerical integration is referred to as the *integrated cooperation matrix* or simply: *cooperation matrix*, and denoted by $\mathcal{C}(\xi_X \to \xi_Y)$. The sum $\sum_{\alpha>\beta}[\mathcal{C}(\xi_X \to \xi_Y)]_{\alpha\beta}$ is the energetic contribution to ΔA for the $\xi_X \to \xi_Y$ path. Molecular cogs found for this matrix are called *global molecular cogs*.

RESULTS

In the first part of this section we present detailed results for a small, 11-atom molecular model, to validate the concept of our theoretical approach. We attempted

to indicate molecular cogs to verify whether the genetic algorithm finds the optimal clustering for the scoring function proposed in the *Objective function* section.

In the second part of the *Results* we show molecular cogs for the *nbd*, *ele* and *vdw* interactions for three other molecules. These case-studies allowed for testing of the transferability of the parameters used in the genetic clustering algorithm, but also uncovered certain subtleties characteristic to our approach.

The `[NH3+]CC(I)I` molecular model

We were interested in finding molecular cogs propelling a structural transition between two meta-stable states, separated by a high free energy barrier. For our first case-study we required a system with a fairly natural collective variable, in which all types of interactions are significant (*conf* as well as *nbd*). We used the 2,2-diiodoethan-1-aminium molecule, which in the SMILES format is encoded as: `[NH3+]CC(I)I` (we use the SMILES representation throughout this article because of its conciseness). The dihedral angle between atoms `N1-C5-C6-I7` was our collective variable of choice (see Figure 3.5). We used the Generalized Amber Force Field (GAFF) to model interactions between atoms, with partial charges assigned using an empirical procedure, AM1-BCC (Table 3.1).

We focused on finding molecular cogs propelling the transition between the dihedral angles of $\xi_X := -172.5°$ and $\xi_Y := -62.5°$, which correspond to two PMF minima (Figure 3.5). The free energy barrier separating these minima is high, and interactions between the `[NH3+]` group and the iodine atoms provide strong *ele* and *vdw* contributions which influence this barrier.

Constrained MD simulations were carried out for fixed values of the ξ collective variable, each simulation with 10^5 1fs timesteps, at $T = 300$K. We chose 239 points, $\{\xi_i\}_{i=1}^{239}$, equally separated by $\Delta\xi = 0.5°$, so that $\xi_1 = -179.0°$ and $\xi_{239} = -60.0°$. This was done in order to encompass the $[\xi_X, \xi_Y]$ interval; note that $\xi_{14} = \xi_X$ and $\xi_{234} = \xi_Y$.

The length of the block in the block bootstrap estimation of averages and their corresponding variances was set to 10^2, which led to 10^3 blocks for each simulation data set. We chose the length of the block with the assumption that the auto-correlation after 10^2 steps is negligible in the case of our simple system.

The cMD procedure was implemented in the Python programming language (ver. 2.7.2), using the Open Babel[**olboyle2011open**] package (ver. 2.3.2) to model the molecule (see *Supporting Information*). All simulations for the `[NH3+]CC(I)I` model took about 30h on a desktop computer.

We focused on the free energy differences and energetic contributions for the $\xi_X \to \xi_Y$ path. For these end-points we obtain the free energy difference $\Delta A \approx -16.4\frac{\text{kcal}}{\text{mol}}$, which is close to the energetic contribution, $\Delta E \approx -16.7\frac{\text{kcal}}{\text{mol}}$. The residual $0.3\frac{\text{kcal}}{\text{mol}}$ is the entropic contribution, which was negligible for our model.

Overview

Results for the *total*, *conf* and *nbd* interactions are juxtaposed in Table 3.2, and for 2-body interactions (*bond*, *ele*, *vdw*) in Table 3.3. The first row contains optimal partitionings of the integrated cooperation matrices, i.e. global molecular cogs (see Figure 3.3 for explanation), along with a picture of the model colored according to the clustering. In the second row we placed the integrated cooperation matrices rearranged in accordance with the clusterings depicted in the previous row of the table (see Figure 3.2 for explanation). This representation visualizes the "density" of cooperativity within molecular cogs, and gear grinding. The third row shows SCOREs for transient molecular cogs and compares genetic clusterings with optimal, brute-force partitionings. In the next row we see an illustration presenting transient molecular cogs, which is a concise summary of how cooperation within the molecule changes with ξ. This plot is helpful in judging whether the global molecular cogs are similar to the transient molecular cogs, and in assessing the consistency of cooperation within the molecule. Finally, the last row contains a PMF-like contribution profile which we call the *energetic contribution profile*; this is the most important result of our analysis. The green line represents the total energetic contribution of a given interaction type, orange and blue lines show contributions of the reverse and forward cogs, respectively, and the purple line – the magnitude of gear grinding. Gear grinding is quantified as the sum of absolute values of all misplaced contributions, i.e. from pairs of atoms from different clusters. Forward and reverse cogs contributions are calculated as the sum of cooperation between atoms assigned to FC and RC, respectively.

Results for the `[NH3+]CC(I)I` molecular model

In this section we comment on the results presented in Tables 3.2 and 3.3. We explain the Tables row by row, highlighting important aspects of the analysis, starting with Table 3.2 and then proceeding with Table 3.3.

In the first row of Table 3.2, global molecular cogs for *total* and *conf* interactions are trivial ("all FC"), with very low SCOREs. Thus, no interesting cooperation was detected for these interaction types – the whole system has a general propensity for preferring the state around ξ_Y. On the other hand, global molecular cogs for the *nbd* interactions have a high SCORE. The clustering suggests, in particular, that non-bonded interactions between the nitrogen atom and the iodine atoms hinder the transformation (as expected). A more surprising implication was that the hydrogen atoms from the `[NH3+]` group cooperate with the `C6` atom.

Clustered cooperation matrices are placed in the second row of Table 3.2. For the *total* and *conf* interactions, for which the partitioning was trivial, the matrices kept their original form, whereas for the *nbd* interactions rows and colums were rearranged in accordance with the partitioning. Matrices for *total* and *conf* are almost indistinguishable, which suggests that the transition is mainly governed by the *conf* interactions. It is also clear that the cooperation matrix for *nbd* is less "dense", with little gear grinding.

In the next row we see that the partitionings into transient molecular cogs for *nbd* have consistently higher SCOREs than those found for *total* and *conf*. There is a drop in SCOREs corresponding to the crossing of the free energy maximum at about $-115°$ dihedral angle. SCOREs for *nbd* molecular cogs vary significantly for different values of the collective variable, and the division into forward and reverse cogs becomes slightly more difficult after crossing the free energy maximum.

The transient molecular cogs depicted in graphs in the fourth row of the table show that the transient molecular cogs for *total* and *conf* exhibit no interesting structure and, in most cases, are of the "all FC" or "all RC" type. Transient molecular cogs for *nbd* are fairly consistent with global molecular cogs. Interestingly, the previously noted *nbd* cooperation between hydrogen atoms `H2, H3, H4` and the `C6` atom persists throughout the transition.

In the last row of the table we placed the energetic contribution profiles. Be-

cause the global molecular cogs for the *total* and *conf* interactions were of the "all FC" form, the only contributions come from the forward cogs. For the *nbd* interactions, the plot shows a beautiful separation of contributions coming from the forward ($-4.2\frac{\text{kcal}}{\text{mol}}$) and the reverse cogs ($3.7\frac{\text{kcal}}{\text{mol}}$) for non-bonded interactions, and low gear grinding for the $\xi_X \rightarrow \xi_Y$ transition.

Let us now look at the results in Table 3.3. In all three cases global molecular cogs were non-trivial, although the SCORE for *bond* is significantly lower than for *ele* and *vdw*. The partitioning for *bond* suggests that there is an impediment arising from interactions of the C6, I7 and I8 atoms, whereas the rest of the system favors the state around ξ_Y. Moving on to *ele* and *vdw* interactions we see that their global molecular cogs share a common pattern, although the SCORE for the latter is slightly lower. We can also see that the cooperation of the H2, H3, H4 and C6 atoms (indicated earlier for *nbd*) is caused by the *ele* interactions.

The next row of Table 3.3 shows graphical representations of the clustered cooperation matrices. Not surprisingly, the *bond* matrix is more sparse than any other matrix, however judging by the low SCORE for global molecular cogs, this was not sufficient to ensure a clear division into the forward and reverse cogs. Conversely, for *ele* and *vdw* there seems to be a higher degree of gear grinding, yet the SCOREs were higher. This is due to the fact that the cooperation within molecular cogs is much stronger than gear grinding between them.

SCOREs in the third row of the table show that the transient molecular cogs for the *bond* interactions were consistently low. We see the opposite for *ele*, and a completely different situation for the *vdw* transient molecular cogs. The case of the *vdw* cooperation is particularly interesting because it shows that the global molecular cogs can have a high SCORE despite the fact that most of the transient molecular cogs have SCOREs below 0.5.

In the fourth row of the table we see that the transient molecular cogs for *bond* are consistent with their global molecular cogs for dihedral angles in the $[-150°, -100°]$ interval (i.e. around the free energy maximum at $\xi \approx -115°$) For the *ele* interactions we see a stable cooperation between the H1, H2, H3 and C6 atoms, occasionally aided by atoms: H9 and I8. The reason behind the shape of the *vdw* SCORE plot becomes clearer once we see that the cooperation for

this interaction type has two stages – before and after crossing the free energy maximum. Transient molecular cogs for *vdw* on the left side of the free energy maximum are mainly trivial ("all RC") with a SCORE of about 0.5. To the right of the maximum there is a change in cooperation; we see a steady partitioning into forward cogs composed of atoms: `N1, I7, I8, H9, H10` and `H11`, and reverse cogs (atoms: `H2, H3, H4` and `C6`).

In the last row of the table we see that the *bond* interactions lead to molecular cogs with high gear grinding, which is the cause of low SCOREs. The profile shows that these interactions lower the free energy gap in the $\xi_X \to \xi_Y$ transition by $-9.9 \frac{\text{kcal}}{\text{mol}}$. However, we can also see that the contribution from atoms `C6, I7` and `I8` alone increases this gap by $5 \frac{\text{kcal}}{\text{mol}}$. For the *ele* interactions the separation into molecular cogs was clean (low gear grinding), with a $-4.3 \frac{\text{kcal}}{\text{mol}}$ contribution from the forward cogs, and a $2.8 \frac{\text{kcal}}{\text{mol}}$ contribution from the reverse cogs due to the transition. For *vdw*, the separation has also led to low gear grinding, however the overall contribution from the forward cogs is much smaller than the one from the reverse cogs. This is an important observation which occurs again in the next section, where we analyze global molecular cogs determined for other model systems.

Results for related molecules

To better understand our approach, it is valuable to identify molecular cogs for other systems, related to the `[NH3+]CC(I)I` molecule scrutinized above. It was also valuable from the perspective of testing the transferability of the parameters used in the genetic clustering algorithm (as shown in the full results in the *Supporting Information*). We took into account molecules which instead of the `[NH3+]` group had the: `NH2, CH3` and `CH2Cl` groups respectively, i.e. molecules: `NCC(I)I`, `CCC(I)` and `CClCC(I)I`. Partial charges assigned by the AM1-BCC procedure for these molecules are listed in Table 3.1. In this section we report and shortly discuss global molecular cogs for the *nbd*, *ele* and *vdw* interactions (see Table 3.4).

Because we analyzed closely related molecules, it was expected that molecular cogs for the *nbd* interactions would be comparable. All molecules presented here, except for `CClCC(I)I`, have partial charges of the same sign for corresponding atoms, therefore global molecular cogs for *ele* look similar. The more interest-

ing outcome was related to discrepancies in molecular cogs identified for the *vdw* interactions.

As noted earlier, in the case of the `[NH3+]CC(I)I` molecule the overall *vdw* contribution is almost entirely explained by cooperation within the reverse cogs. The forward cogs were identified merely because their contribution was non-positive. Nevertheless, this was a valuable insight – we have learned that the *vdw* steric effects are due to interactions of particular atoms: those which comprise the reverse cogs. On the other hand, the *vdw* molecular cogs for the `NCC(I)I` molecule are trivial ("all RC") and carry no such information. No non-trivial partitioning with a higher SCORE was found because there were no legitimate forward cogs, even as ineffective as those discovered in `[NH3+]CC(I)I`. This then suggests that we gained a simplified picture of *vdw* cooperation in the `[NH3+]CC(I)I` molecule simply because we were fortunate enough to have analyzed a system in which there had been at least a minimal contribution from the forward cogs. This is a consequence of the underlying assumption that a molecule's tendency to undergo a transition is a result of two opposing cooperations. Perhaps we should approach the problem differently for instances in which the molecule as a whole has the propensity to move forward/backward (as we also saw for the *total* and *conf* interactions).

Molecular cogs for *vdw* interactions for the `CCC(I)I` molecule are similar to those of `[NH3+]CC(I)I`. This might seem natural; the `[NH3+]` and `CH3` groups share common properties. However, the energetic contribution profile for the *vdw* interactions (see *Supporting Information*) reveals that the forward cogs in `CCC(I)I` have a contribution comparable to that of the reverse cogs. Although *vdw* molecular cogs for `CCC(I)I` and `[NH3+]CC(I)I` are identical, the underlying mechanism is different.

The `CClCC(I)I` model was designed to lower the *ele* barrier. However, the resulting molecular cogs for *ele* became trivial and, as in the case of *vdw* for `NCC(I)I`, we do not know which atoms are the main source of this effect. This again suggests that perhaps an alternative method of finding molecular cogs should be considered. In cases of trivial partitionings, such method should extract information about parts of the molecule which are the main source of the free energy difference. However, we leave investigation of properties and characteristics of alternative scoring functions

for future studies.

DISCUSSION

Note that in all cases the genetic clustering algorithm gave clusterings with the best possible SCOREs. However, the efficiency of the genetic algorithm depends on the starting point, and without the help of the AP-generated initial population it took, on average, about 30 times longer, and the best result was not always achieved. To generate these initial solutions we adapted the AP clustering procedure (see *Supporting Information*).

The reason why conformational interactions lead to low-quality molecular cogs is that these are short-ranged interactions and our test molecular system is small. Partitioning of a graph into clusters is laden with a cost which depends on the weights of cross-cluster edges. But, as can be seen in Table 3.2, the integrated cooperation matrices for non-bonded interactions are more sparse than for conformational interactions. Consequently, the cost of partitioning is greater for more "dense" matrices. But this cost decreases with increasing dimensionality of a matrix. It is therefore possible that, for more complex systems, molecular cogs for conformational interactions will have a higher SCORE.

The non-bonded interactions in the GAFF force field for atoms separated by three bonds of fewer are zero. This, and the fact that our model system is small, led to a sparse *nbd* cooperation matrix. Note, however, that a larger system may yield matrices that would be more "dense" for non-bonded interactions than for short-ranged conformational interactions. Whether a partitioning of these matrices would result in higher gear grinding remains an unanswered question.

We also anticipate that the entropic contribution in Equation (3.1) should play a more notable role for larger systems. Nonetheless, we have not yet considered clustering a graph with edges weighted by the entropic contributions of pairs of atoms. The reasons are twofold: practical (we want to avoid calculating Hessians of collective variables) and conceptual (the entropic cooperation of an atom with itself is non-zero). Entropic contributions need to be considered, but this should be the subject of future studies.

In the second part of the *Results* section we identified and analyzed molecular

cogs for three additional molecules, related to the `[NH3+]CC(I)I` model. We discovered that trivial molecular cogs, which carry less information than any other partitioning, may occur whenever there is a lack of competition of cooperation within the molecule. This is a consequence of the proposed approach of dividing the system into two competing subsystems. In many cases in which global molecular cogs were trivial, we simply did not gain any insight into what is propelling the transformation. However, it seems that in such cases the analysis should be aimed at finding parts of a molecule that play the dominant role in its mobility.

Comparison with qualitative expectations

Results for the `[NH3+]CC(I)I` model, especially for the *ele* interactions, were useful in checking our intuitions against what was shown by the analysis. In this section we shortly discuss several sanity checks related to results for the *ele* interactions that were helpful in verifying whether our implementation had any critical errors, and whether the proposed approach leads to reasonable assessments.

From the plot depicting transient molecular cogs for *ele* (Table 3.3) we see that atoms `N1` and `I7` consistently hinder the transformation. These two atoms have negative partial charges (-0.85 and -0.08, respectively; see Table 3.1), and therefore repel each other throughout the transition. But the `I8` atom (partial charge of -0.08) behaves differently, i.e. for ξ lower than $-115°$ (free energy maximum) it impedes the process by repelling the `N1` atom, and aids it for ξ larger than $-115°$. This effect was correctly captured by the above analysis, because the `I8` atom finds itself in the reverse cogs in the first part of the transition, and in the forward cogs in the second part (as shown by the graph in the sixth row of Table 3.3). From the integrated contribution matrix for *ele* (second row in Table 3.3) we see, however, that the cumulative contribution from the `N1-I8` atoms is positive, which results in assigning them to one cluster.

The analysis revealed that atoms: `H2, H3, H4` and `C6` were consistently cooperating electrostatically, lowering the free energy barrier. This conclusion was more unexpected than the one concerning atoms: `N1, I7` and `I8`, but was also compatible with our intuitions, taking into account the partial charges in Table 3.1.

Our method can suggest a qualitative interpretation of the cause behind a

structural transition. However, it should be noted the most valuable information gained from this type of analysis is the quantitative description of the molecular cogs *via* the energetic contribution profile.

CONCLUSIONS

The aim of this article was to introduce a new methodology of identifying molecular cogs – parts of a molecule that propel structural transitions in forward/backward directions along a collective variable. The current framework allowed us to track energetic contributions to the free energy, leaving the problem of including entropic terms for future developments. Results show that with the use of the genetic clustering algorithm we can successfully divide small molecules and identify forward and reverse molecular cogs associated with non-bonded interactions.

In particular, we proposed the approach of defining free energy contributions originating from pairs of atoms, and a method of dividing a molecule into molecular cogs. We showed that the proposed genetic clustering algorithm efficiently finds the optimal cogs, leading to high-quality partitionings for non-bonded interactions. However, we also found that conformational interactions lead to low-scoring molecular cogs, and that the system as a whole favored one meta-stable state over the other.

Currently, our method is based on cMD simulations for computing conditional averages (Equation (3.1)). Unfortunately, this solution has a critical drawback (see *Supporting Information*), but also, in order to determine the entropic contributions, requires computing second-order derivatives of the collective variable. To resolve this problem we should facilitate the Adaptive Biasing Force (ABF) scheme for calculating the PMF [**comer2014adaptive**]. Specifically, our future work is aimed at reformulating the procedure as a plugin to the NAMD package (in which the ABF has already been implemented), to include entropic contributions and to assure scalable performance. Once this is done, numerous new opportunities will become available, some of which we mention below.

We constructed graphs in which nodes corresponded to atoms. Note, however, that they may also be assigned to amino acids (or more sizable objects) to construct graphs, which would lead to a more mosaic clustering. It could also be valuable

to represent a whole ligand as a single object in a protein-ligand complex. One could also consider the role of functional groups comprising a ligand and amino acids in a binding pocket; it might then be helpful to represent the rest of the protein as a single node in the graph. Another example is a possible treatment of solvent molecules, for example: a particular group of interesting water molecules could be transformed into separate nodes in the graph, while others reduced to a single node.

Our method may prove helpful in understanding why wild-type proteins perform better than mutants, or in explaining why certain drugs perform better than others, despite their structural similarity. It would also be interesting to consider a multi-stage induced fit docking, and to study cooperativity between amino acids and the ligand along a collective variable.

This research study demonstrates how to extract pair-wise energetic contributions to the free energy change along a reaction coordinate. In its present form[1], we could only use our method to analyze small, model molecules. However, we were able to verify the efficacy of the genetic clustering algorithm, and to learn what can be expected from the new notion of "molecular cogs". Our future aim is to analyze more complex systems, and to develop a publicly available implementation of our method for practical applications.

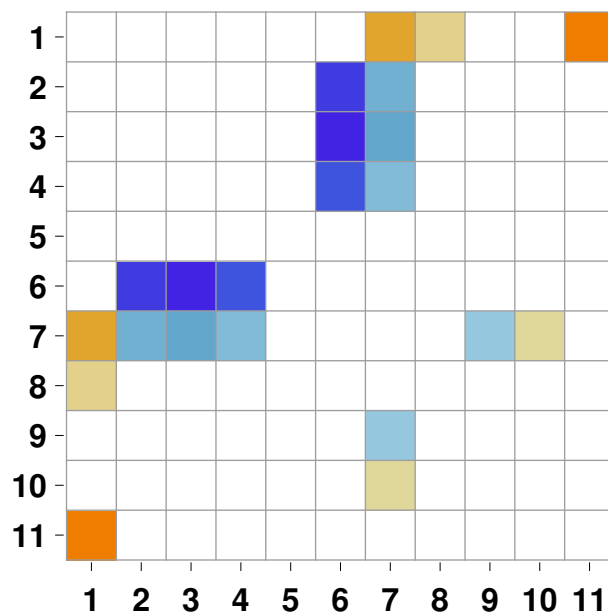[1]We made the source code available at: `http://github.com/ponadto/molecular-cogs`

Figure 3.1: **Graphical representation of a transient cooperation matrix for *nbd* interactions.** Warm colors denote positive contributions, whereas blue – negative. White squares correspond to nil values.

[NH3+]CC(I)I		NCC(I)I		CCC(I)I		CClCC(I)I	
atom	partial charge	atom	partial charge	atom	partial charge	atom	partial charge
N1	-0.85	N1	-0.92	C1	-0.10	C1	0.03
H2	0.47	H2	0.36	H2	0.04	Cl2	-0.19
H3	0.47	H3	0.36	H3	0.04	H3	0.08
H4	0.47			H4	0.04	H4	0.08
C5	0.13	C4	0.17	C5	-0.07	C5	-0.08
C6	0.08	C5	0.20	C6	0.21	C6	0.20
I7	-0.08	I6	-0.19	I7	-0.19	I7	-0.19
I8	-0.08	I7	-0.19	I8	-0.19	I8	-0.19
H9	0.13	H8	0.05	H9	0.06	H9	0.08
H10	0.13	H9	0.05	H10	0.06	H10	0.08
H11	0.13	H10	0.12	H11	0.10	H11	0.12

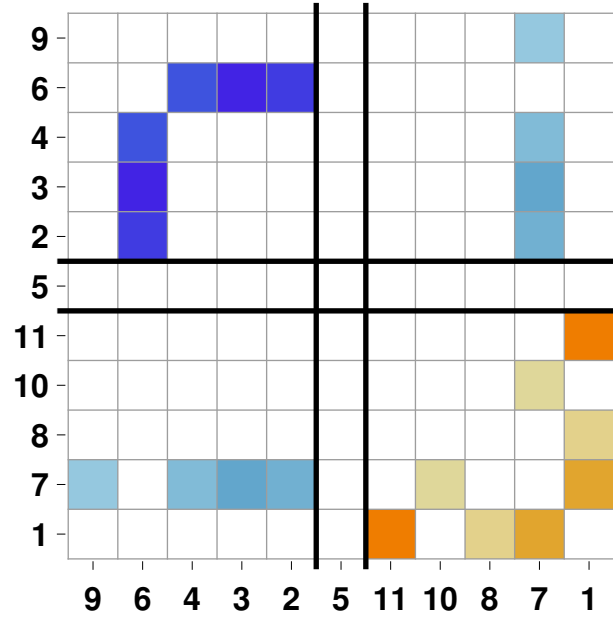Table 3.1: **Partial charges assigned by the AM1-BCC procedure.**

Figure 3.2: **Graphical representation of a result of a clustering procedure.** Rows and columns of the cooperation matrix in Figure 3.1 were reordered according to their assignments to: forward cogs, non-interacting atoms, and reverse cogs. In the upper left corner of this transformed matrix we have a block (square submatrix) with negative contributions, which come from the atoms $2, 3, 4$ and 6 (from the forward cogs). In the lower right corner we have a block formed by pair contributions of the atoms in the reverse cogs. The gear grinding is small and results from contributions between the atom 7 and the atoms $2, 3, 4, 9$.
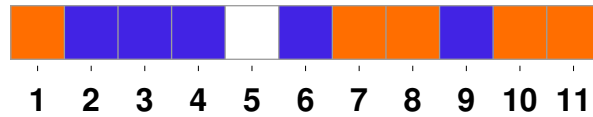
Figure 3.3: **Graphical representation of an exemplary partitioning:** $\{1, -1, -1, -1, 0, -1, 1, 1, -1, 1, 1\}$. Orange squares correspond to 1 (reverse cogs), blue to -1 (forward cogs), and white to 0 (non-interacting). The cooperation matrix in Figure 3.1 was reordered according to this assignment, yielding the transformed cooperation matrix in Figure 3.2.
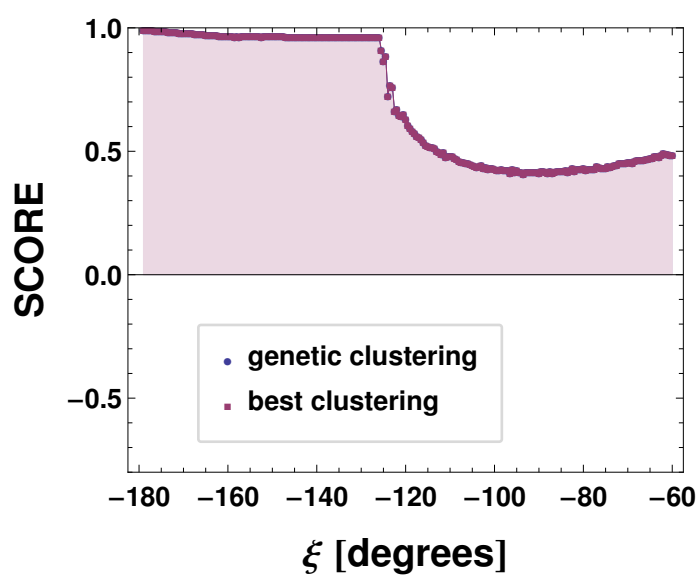
Figure 3.4: **SCORE plot for *nbd* interactions.** The plot shows that solutions found by the genetic clustering algorithm overlap perfectly with the highest scoring assignments to the transient molecular cogs.
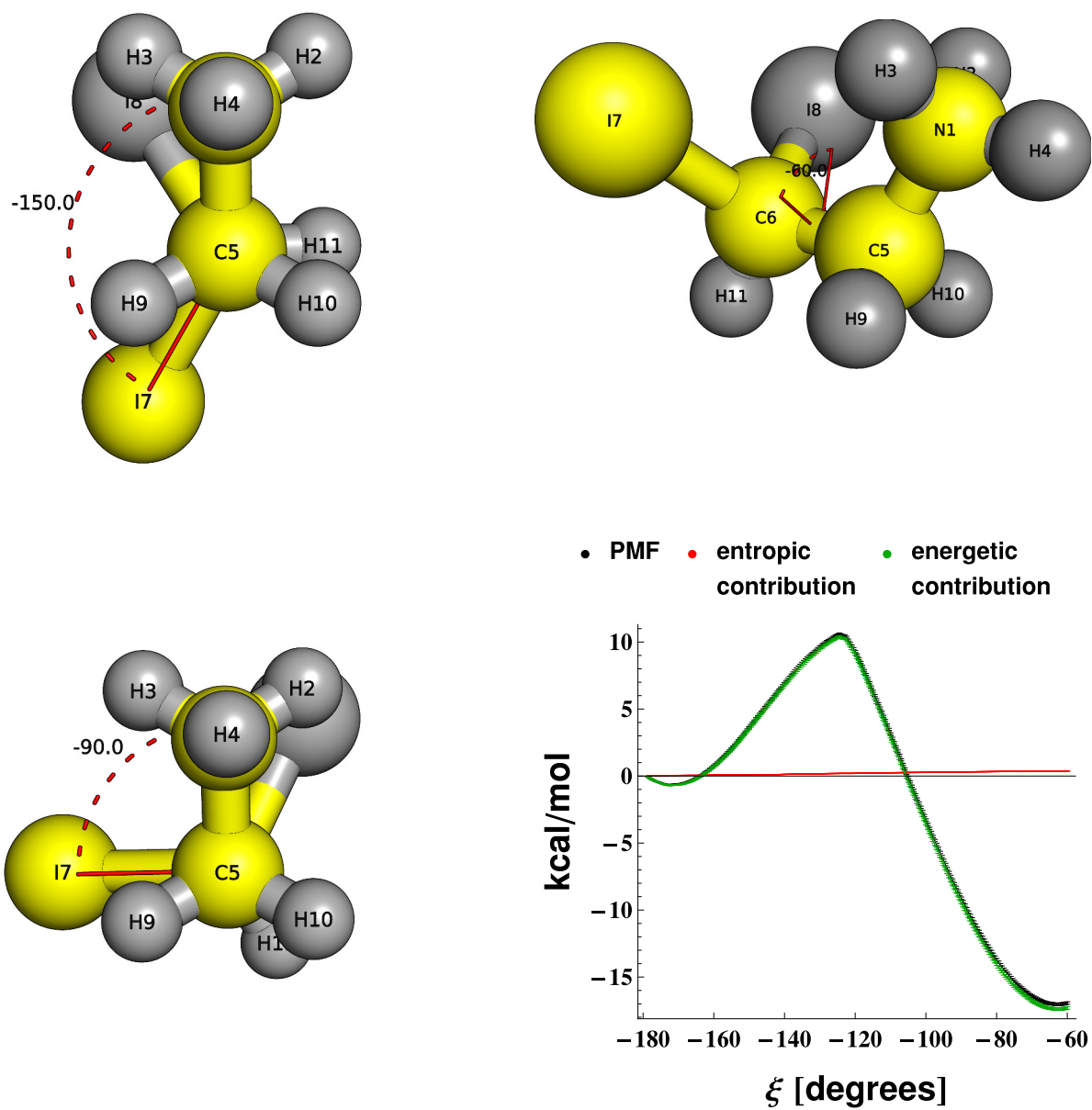
Figure 3.5: `[NH3+]CC(I)I` **molecule summary.** Three configurations of the molecule are shown, and the PMF with energetic and entropic contributions (for $T = 300$ K). Atoms `N1-C5-C6-I7` (yellow) were used to define the dihedral angle, ξ.
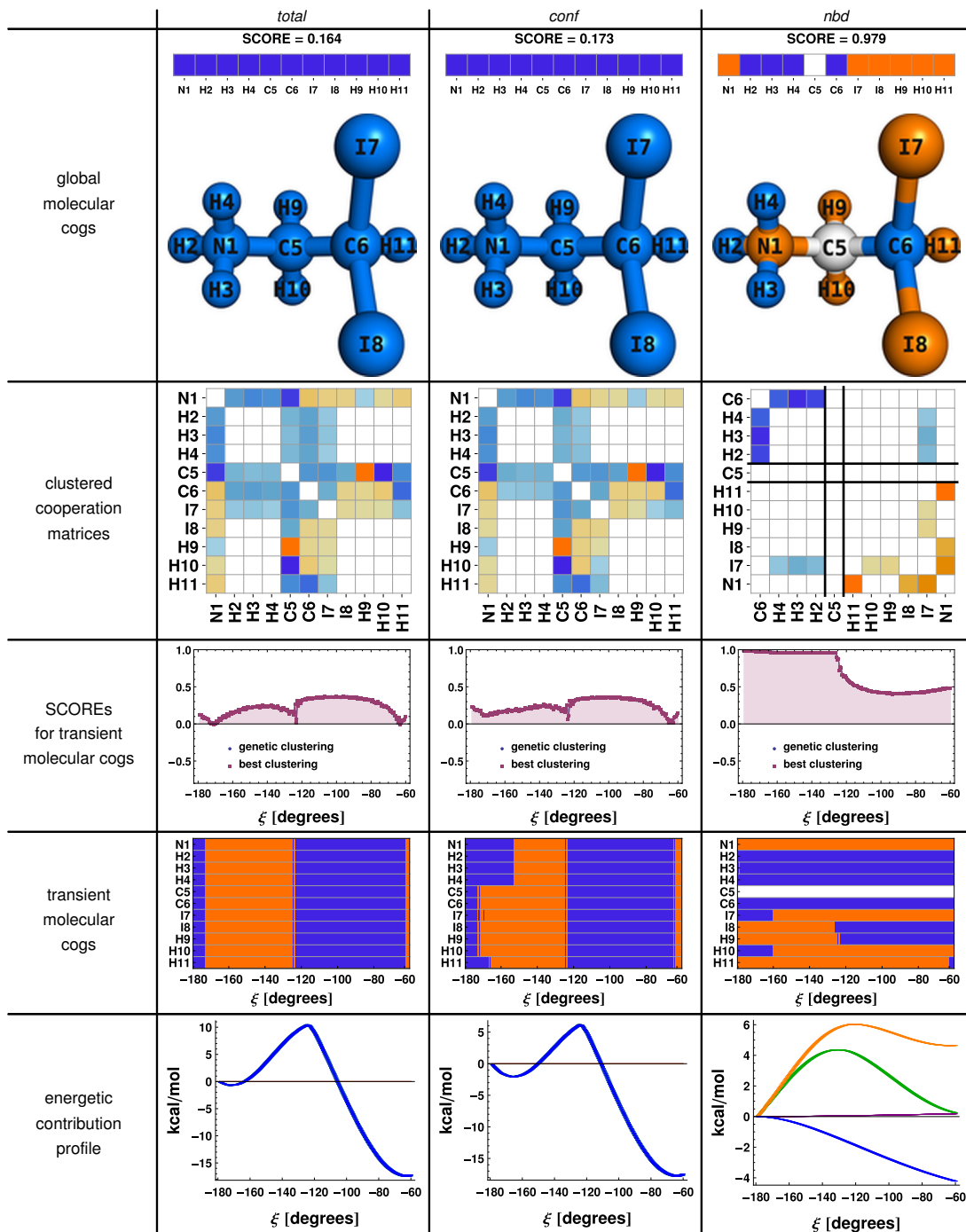
Table 3.2: **Results summary for interactions: *total, conf* and *nbd*.** Out of the triple: *total, conf* and *nbd*, only the last one leads to a high SCORE partitioning, yielding non-trivial molecular cogs. Colors: green, blue, orange and purple in the last row of the table denote contributions from the: whole system, forward cogs, reverse cogs and gear grinding, respectively (see the *Overview* section for details).
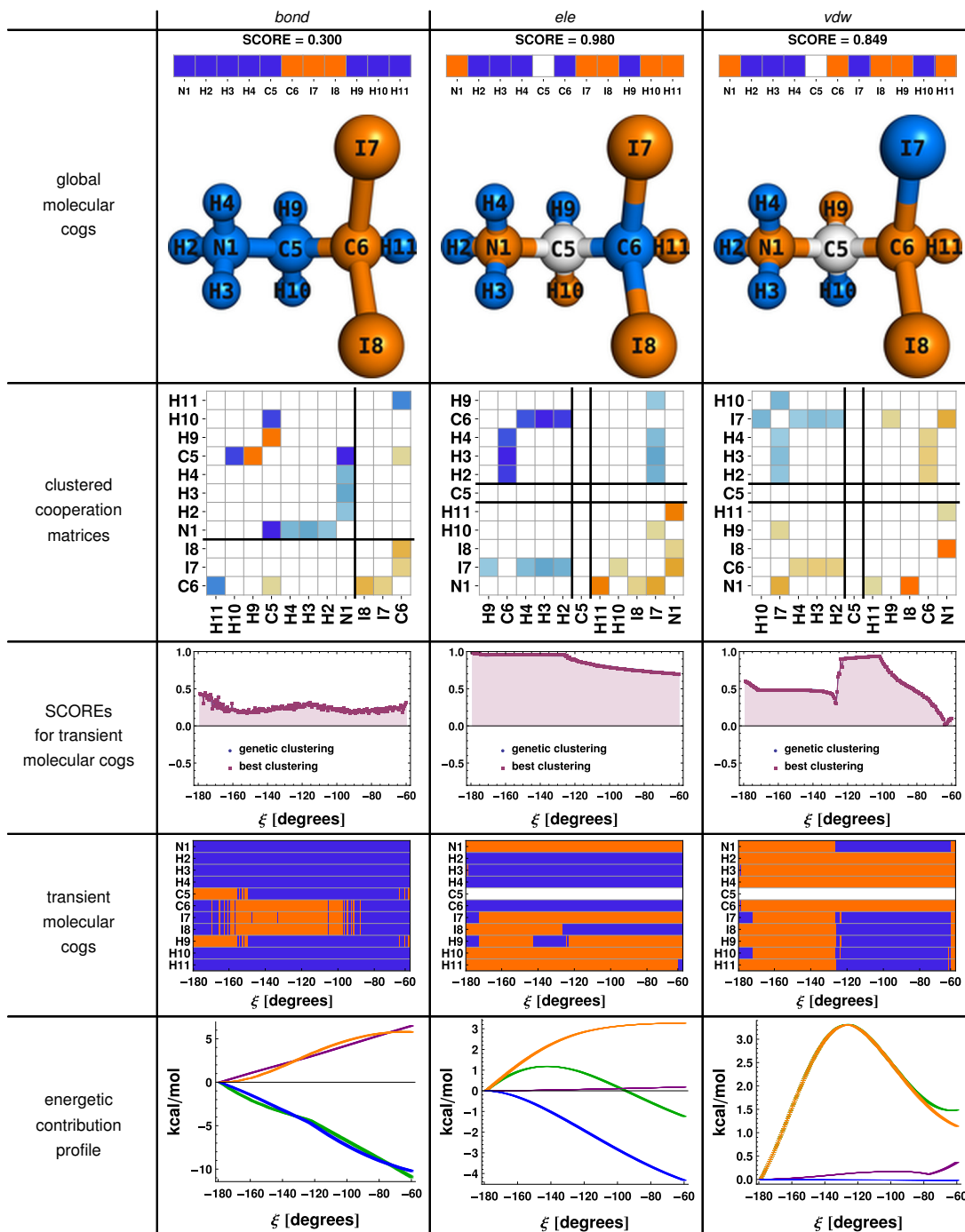
Table 3.3: **Results summary for interactions: *bond*, *ele* and *vdw*.** All three 2-atom interaction types lead to decompositions into non-trivial molecular cogs. However, the transient molecular cogs have high SCOREs only for the *ele* interactions. Colors: green, blue, orange and purple in the last row of the table denote contributions from the: whole system, forward cogs, reverse cogs and gear grinding, respectively (see the *Overview* section for details).

	`[NH3+]CC(I)I`	`NCC(I)I`	`CCC(I)I`	`CClCC(I)I`
nbd	SCORE = 0.979	SCORE = 0.980	SCORE = 0.995	SCORE = 0.748
ele	SCORE = 0.980	SCORE = 0.981	SCORE = 0.979	SCORE = 0.381
vdw	SCORE = 0.849	SCORE = 0.493	SCORE = 0.989	SCORE = 0.968

Table 3.4: **Global molecular cogs for the following molecules: `[NH3+]CC(I)I`, `NCC(I)I`, `CCC(I)I` and `CClCC(I)I`.** We focused our discussion on partitionings into forward and reverse cogs for the *nbd, ele* and *vdw* interactions. Full results are detailed in the *Supporting Information.*

Chapter 4

Summary

4.1   X—-X